ISIPTA '13

Proceedings of the Eighth International Symposium on

Imprecise Probability: Theories and Applications

July 2–5 2013, Compiègne, France

www.sipta.org/isipta13

Université de Technologie de Compiègne

Edited by

Fabio Cozman
Thierry Denœux
Sébastien Destercke
Teddy Seidenfeld

Published by SIPTA
Society for Imprecise Probability: Theories and Applications
www.sipta.org

Printed 2013 by
Université de Technologie de Compiègne
Rue Roger Couttolenc
60200 Compiègne, France

ISBN 978-2-913923-35-5

The book was typset using LaTeX.

Contents

CONFERENCE PAPERS

CONFERENCE POSTER ABSTRACTS

Preface

The *Eighth International Symposium on Imprecise Probability: Theories and Applications* is held in Compiègne, France, 2–5 July 2013.

The ISIPTA meetings are a primary forum for presenting and discussing advances in imprecise probability and are organized once every two years. The first meeting was held in Gent in 1999, followed by meetings in Ithaca (Cornell University), Lugano (IDSIA), Pittsburgh (Carnegie Mellon University), Prague, Durham (UK) and Innsbruck. In the decade since the first meeting, imprecise probability has come a long way, which is reflected by the wide range of topics presented at the 2013 meeting, but particularly also in the increased presence of imprecise probability in journals and at other conferences.

As with previous ISIPTA meetings, the program only contains plenary sessions. In total, 38 papers are presented by a short talk and a poster, which guarantees ample time for discussion. The papers are included in these proceedings and are also available on the SIPTA webpage (`www.sipta.org`). Submitted papers have undergone a high quality reviewing process by members of the Program Committee, ensuring the quality of the presented research results.

To provide a platform for preliminary ideas and challenging applications for which the research is not yet completed, poster-only presentations have been introduced at ISIPTA'09 and the initiative pursued in ISIPTA'11. We continue with this tradition; short abstracts of these poster-only presentations are included in the proceedings and are available on the SIPTA webpage.

As with previous ISIPTA meetings, a wide variety of theories and applications of imprecise probability are presented. New application areas and novel ways for dealing with limited information prove the increasing success of imprecise probability.

Most participants having a good knowledge of the basics of imprecise probabilities, the two introductory tutorials introducing have been scheduled the day before the start of the conference. We thank Thierry Denœux and Matthias Troffaes for preparing and presenting tutorials on Belief functions and imprecise probabilities, respectively.

Invited talks are intended to present both recent developments in selected fields and topics that are related but not directly linked to the main topics of ISIPTA. We thank Alessio Benavoli (Switzerland) for preparing and presenting the talk *Pushing Dynamic Estimation to the Extremes: from the Moon to Imprecise Probability*, Linda van der Gaag (Netherlands) for preparing and presenting the talk *Recent Advances in Sensitivity Analysis of Bayesian Networks*, Christophe Labreuche (France) for preparing and presenting the talk *Robustness in Multi-Criteria Decision Making and its relation with Imprecise Probabilities* and Jean-Marc Tallon (France) for preparing and presenting the talk *Ambiguity and ambiguity attitudes in economics.*

During the conference two prizes are awarded: the *Best Poster Award*, sponsored by Springer-Verlag, and the *IJAR Young Researcher Award*, granted by the International Journal of Approximate Reasoning.

We believe that, in the fourteen years since ISIPTA'99, imprecise probability has found a solid place in research on uncertainty quantification and related fields. Because applications are increasing, both in number and success, we are optimistic about the future impact of imprecise probability. We think that the current format of ISIPTA is successful, and we hope that all participants will find the meeting pleasant, informative, and beneficial. We hope that ISIPTA'13 provides a good platform to present and discuss work, and also leads to new ideas and collaborations.

Finally, we wish to thank several people for their support. We thank Thomas Fetz and Matthias Troffaes for their precious advices, inherited from the organization of past ISIPTAs. We thank Serafín Moral for his extensive and expert help in managing the system supporting the conference website.

We thank the members of the Program Committee for their excellent reviewing activities. Special thanks also to the Local Organizing Committee, in particular, to Nathalie Alexandre and Cécile Poncin for their help.

We thank all our sponsors for their support and help in organizing this conference.

Finally, we thank all who have contributed to the success of ISIPTA'13, be it by submitting their research results, presenting them at the conference, or by attending sessions and participating in discussions. We hope that these proceedings will convey the state of the art of imprecise probability, raise interest and contribute to the further dissemination of the fascinating ideas of this active and highly relevant research field.

Fabio Cozman
Thierry Denœux
Sébastien Destercke
Teddy Seidenfeld

Compiègne, France

Organization, Supporters and Sponsors

Steering Committee

Fabio Cozman, Brazil
Thierry Denœux, France
Sébastien Destercke, France
Thomas Fetz, Austria
Serafín Moral, Spain
Teddy Seidenfeld, USA
Matthias Troffaes, UK

Program Committee Board

Fabio Cozman, Brazil
Thierry Denœux, France
Sébastien Destercke, France
Teddy Seidenfeld, USA

Program Committee Members

Joaquín Abellán, Spain
Alessandro Antonucci, Switzerland
Thomas Augustin, Germany
Alessio Benavoli, Switzerland
Salem Benferhat, France
Dan Berleant, USA
Alberto Bernardini, Italy
Cassio Campos, Switzerland
Andrea Capotorti, Italy
Marco Cattaneo, Germany
Giorgio Corani, Switzerland
Inés Couso, Spain
Richard Crossman, UK
Fabio Cuzzolin, UK
Gert de Cooman, Belgium
Serena Doria, Italy
Didier Dubois, France
Love Ekenberg, Sweden
Malcolm Farrow, UK
Terrence Fine, USA
Angelo Gilio, Italy
Michel Grabisch, France
Robert Hable, Germany
Joseph Halpern, USA
Nathan Huntley, UK
Manfred Jaeger, Denmark
Radim Jiroušek, Czech Republic
Cliff Joslyn, USA
Erich Peter Klement, Austria

Igor Kozine, Denmark
Vladik Kreinovich, USA
Tomaš Kroupa, Czech Republic
Rudolf Kruse, Germany
Isaac Levi, USA
Weiru Liu, Ireland
Andres Masegosa, Spain
Enrique Miranda, Spain
Ilya Molchanov, Switzerland
Serafín Moral, Spain
Robert Nau, USA
Renato Pelessoni, Italy
Erik Quaeghebeur, Belgium
David Rï¿½os Insï¿½a, Spain
Fabrizio Ruggeri, Italy
Prakash SHenoy, USA
Damjan Škulj, Slovenia
Michael Smithson, Australia
Joerg Stoye, USA
Jean-Marc Tallon, France
Choh M. Teng, USA
Fulvio Tonon, USA
Matthias Troffaes, UK
Lev Utkin, Russia
Linda van der Gaag, Netherlands
Barbara Vantaggi, Italy
Jiřina Vejnarova, Czech Republic
Paolo Vicig, Italy
Nic Wilson, UK
Marco Zaffalon, Switzerland

Additional reviewers

Christian Braune, Germany
Pascal Held, Germany
Christian Moewes, Germany

Davide Petturiti, Italy
Arthur Van Camp, Belgium

Local Organizing Committee

Cedric Baudrit
Veronique Cherfaoui
Thierry Denœux
Sébastien Destercke
Mylène Masson
Benjamin Quost
Mohamed Sallak

Supporters and Sponsors

Université de Technologie de Compiègne

Centre National de la Recherche Scientifique

Springer

Elsevier

City of Compiègne

AgroParisTech

Picardie Region

Heudiasyc Laboratory

Labex MS2T

Conference Papers

Inclusion/exclusion principle for belief functions

Felipe Aguirre[1] felipe.aguirre@utc.fr
Christelle Jacob[2] jacob@isae.fr
Sébastien Destercke[1] sebastien.destercke@utc.fr
Didier Dubois[3] dubois@irit.fr
Mohamed Sallak[1] mohamed.sallak@utc.fr

Abstract

The inclusion-exclusion principle is a well-known property of set cardinality and probability measures, that is instrumental to solve some problems such as the evaluation of systems reliability or of uncertainty over Boolean formulas. However, when using sets and probabilities conjointly, this principle no longer holds in general. It is therefore useful to know in which cases it is still valid. This paper investigates this question when uncertainty is modelled by belief functions. After exhibiting necessary and sufficient conditions for the principle to hold, we illustrate its use on some applications, i.e. reliability analysis and uncertainty over Boolean formulas. [1]

1 Introduction

Probability theory is the most well-known approach to model uncertainty. However, even when the existence of a single probability is assumed, it often happens that the distribution is partially known, in which case one is forced to use a selection principle (e.g., maximum entropy [13]) to work within probability theory. This is particularly the case in the presence of severe uncertainty (few samples, imprecise or unreliable data, ...) or when subjective beliefs are elicited. Many authors claim that in situations involving imprecision or incompleteness, uncertainty cannot be modelled faithfully by a single probability, and they have proposed frameworks to properly model such uncertainty: possibility theory [11], belief functions [16], imprecise probabilities [17], info-gap theory [3], ...

A known practical drawback of belief functions and of other imprecise probabilistic theories is that their manipulation can be computationally more demanding than probabilities. Indeed, the fact that belief functions are more general than classical probabilities prevents the use of some properties that hold for the latter but not for the former. This is the case, for instance, of the well known and useful inclusion-exclusion principle (also known as the sieve formula or Sylvester-Poincaré equality).

Given a space $\mathscr{X}$, a probability measure P over this space and a collection $\mathscr{A}_N = \{A_1,\dots,A_N | A_i \subseteq \mathscr{X}\}$ of measurable subsets of $\mathscr{X}$, the inclusion-exclusion principle states that

$$P(\cup_{i=1}^{n} A_i) = \sum_{\mathscr{I} \subseteq \mathscr{A}_n} (-1)^{|\mathscr{I}|+1} P(\cap_{A \in \mathscr{I}} A) \qquad (1)$$

where $|\mathscr{I}|$ is the cardinality of $\mathscr{I}$. This equality allows to easily compute the probability of $\cup_{i=1}^{n} A_i$. This principle is used in numerous problems, including the evaluation of the reliability of complex systems when using minimal paths.

In this paper, we investigate in Section 2 necessary and sufficient conditions under which a similar equality holds for belief functions. Section 3 then studies how the results apply to the practically interesting case where events A_i and focal sets are Cartesian products. Section 4 then shows that such conditions are met for specific events of monotone functions, and applies this result to the reliability analysis of multi-state systems. Finally, Section 5 computes the belief and plausibility of Boolean formulas expressed in normal forms.

2 General Additivity Conditions for Belief Functions

After introducing notations, Section 2.2 provides general conditions for families of subsets for which the inclusion-exclusion principle holds for belief functions. We then interest ourselves to the specific case where focal sets of belief functions are Cartesian products of subsets.

2.1 Setting

A mass distribution [16] defined on a (finite) space $\mathscr{X}$ is a mapping $m : 2^{\mathscr{X}} \to [0,1]$ from the power set of $\mathscr{X}$ to the

[1] HEUDIASYC, UMR 7253. Université de Technologie de Compiègne. Centre de Recherches de Royallieu. 60205 COMPIEGNE, France
[2] Institut supérieur de l'aéronautique et de l'espace (ISAE) 10, avenue E. Belin - Toulouse
[3] Institut de Recherche en Informatique de Toulouse (IRIT), ADRIA

unit interval such that $m(\emptyset) = 0$ and $\sum_{E \subseteq \mathscr{X}} m(E) = 1$. A set E that receives a strictly positive mass is called *focal set*, and the set of focal sets of m is denoted by $\mathscr{F}_m$. From the mapping m are usually defined two set-functions, the plausibility and the belief functions, respectively defined for any $A \subseteq \mathscr{X}$ as

$$Pl(A) = \sum_{E \cap A \neq \emptyset} m(E), \tag{2}$$

$$Bel(A) = \sum_{E \subseteq A} m(E) = 1 - Pl(A^c). \tag{3}$$

They are such that $Bel(A) \leq Pl(A)$. The plausibility function measures how much event A is possible, while the belief function measures how much event A is certain. In the theory of evidence [16], belief and plausibility functions are interpreted as confidence degrees not necessarily related to probabilities. However, the mass distribution m can also be interpreted as the random set corresponding to an imprecisely observed random variable [8], in which case Bel, Pl can be interpreted as probability bounds inducing a convex set $\mathscr{P}(Bel)$ such that

$$\mathscr{P}(Bel) = \{P | \forall A, Bel(A) \leq P(A) \leq Pl(A)\}$$

is the set of all probabilities bounded by Bel and Pl. Note that, since Bel and Pl are dual ($Bel(A) = 1 - Pl(A^c)$), we can concentrate on one of them. A distribution m can be seen as a probability distribution over sets, and in this sense it captures both probabilistic and set-based modelling: any probability p can be modelled by a mass m such that $m(\{x\}) = p(x)$ and any set E can be modelled by the mass $m(E) = 1$.

Consider now a collection of events $\mathscr{A}_n = \{A_1, \ldots, A_n | A_i \subseteq \mathscr{X}\}$ of subsets of $\mathscr{X}$ and a mass distribution m from which can be computed a belief function Bel. Usually, we have the inequality [16]

$$Bel(\cup_{i=1}^n A_i) \geq \sum_{\mathscr{I} \subseteq \mathscr{A}_n} (-1)^{|\mathscr{I}|+1} Bel(\cap_{A \in \mathscr{I}} A) \tag{4}$$

that is to be compared to Eq. (1). Belief functions are said to be *n-monotonic* for any $n > 0$. Note that we can assume without loss of generality that for any i, j, $A_i \not\subseteq A_j$ (otherwise A_i can be suppressed from Equation 4), that is there is no inclusion between the sets of $\mathscr{A}_n$. If Equation 4 becomes an equality, we will say that the belief is *additive* for collection $\mathscr{A}_n$, or $\mathscr{A}_n$-additive for short.

2.2 General necessary and sufficient conditions

In the case of two events A_1 and A_2, none of which is included in the other one, the basic condition for the inclusion-exclusion law to hold is that focal sets included in $A_1 \cup A_2$ should only lie (be included) in A_1 or A_2. Indeed, otherwise, if $\exists E \not\subseteq A_1$ and $E \not\subseteq A_2$ with $m(E) > 0$, then

$$\begin{aligned} Bel(A_1 \cup A_2) &\geq m(E) + Bel(A_1) + Bel(A_2) - Bel(A_1 \cap A_2) \\ &> Bel(A_1) + Bel(A_2) - Bel(A_1 \cap A_2). \end{aligned}$$

So, one must check that $\mathscr{F}_m$ satisfies:

$$\mathscr{F}_m \cap 2^{A_1 \cup A_2} = \mathscr{F}_m \cap \left(2^{A_1} \cup 2^{A_2}\right)$$

where 2^C denote the set of subsets of C. So, one must check that $\forall E \in \mathscr{F}_m$ such that $E \subseteq A_1 \cup A_2$, either $E \subseteq A_1$ or $E \subseteq A_2$, or equivalently

Lemma 1. *A belief function is additive for $\{A_1, A_2\}$ if and only if $\forall E \subseteq A_1 \cup A_2$ such that $(A_1 \setminus A_2) \cap E \neq \emptyset$ and $(A_2 \setminus A_1) \cap E \neq \emptyset$ then $m(E) = 0$.*

Proof. Immediate, as E overlaps A_1 and A_2 without being included in one of them if and only if $(A_1 \setminus A_2) \cap E \neq \emptyset$ and $(A_2 \setminus A_1) \cap E \neq \emptyset$. □

This result can be extended to larger collections of sets $\mathscr{A}_n, n > 2$ in quite a straightforward way

Proposition 1. *$\mathscr{F}_m$ satisfies the property $\mathscr{F}_m \cap 2^{A_1 \cup \ldots \cup A_n} = \mathscr{F}_m \cap \left(2^{A_1} \cup \ldots \cup 2^{A_n}\right)$ if and only if $\forall E \subseteq (A_1 \cup \ldots \cup A_n)$, if $E \in \mathscr{F}_m$ then $\nexists A_i, A_j$ such that $(A_i \setminus A_j) \cap E \neq \emptyset$ and $(A_j \setminus A_i) \cap E \neq \emptyset$.*

Proof. $\mathscr{F}_m \cap 2^{A_1 \cup \ldots \cup A_n} = \mathscr{F}_m \cap \left(2^{A_1} \cup \ldots \cup 2^{A_n}\right)$
if and only if $\nexists E \in \mathscr{F}_m \cap \left(2^{A_1 \cup \ldots \cup A_n} \setminus \left(2^{A_1} \cup \ldots \cup 2^{A_n}\right)\right)$
if and only if $\nexists E \subseteq (A_1 \cup \ldots \cup A_n), E \in \mathscr{F}_m$ such that $\forall i = 1, \ldots, n, E \not\subseteq A_i$
if and only if $\nexists i \neq j, E \in \mathscr{F}_m, E \not\subseteq A_i, E \not\subseteq A_j, E \cap A_i \neq \emptyset, E \cap A_j \neq \emptyset$
if and only if $\nexists i \neq j, E \in \mathscr{F}_m$, with $(A_i \setminus A_j) \cap E \neq \emptyset$ and $(A_j \setminus A_i) \cap E \neq \emptyset$ □

So, based on Proposition 1, we have:

Theorem 2. *The equality*

$$Bel(\cup_{i=1}^n A_i) = \sum_{\mathscr{I} \subseteq \mathscr{A}_n} (-1)^{|\mathscr{I}|+1} Bel(\cap_{A \in \mathscr{I}} A) \tag{5}$$

holds if and only if $\forall E \subseteq (A_1 \cup \ldots \cup A_n)$, if $m(E) > 0$, then $\nexists A_i, A_j$ such that $(A_i \setminus A_j) \cap E \neq \emptyset$ and $(A_j \setminus A_i) \cap E \neq \emptyset$.

Theorem 2 shows that going from $\mathscr{A}_2$-additivity for 2 given sets to $\mathscr{A}_n$-additivity is straightforward, as ensuring $\mathscr{A}_n$-additivity comes down to checking the additivity conditions for every pair of subsets in $\mathscr{A}$.

Note that by duality one also can write a form of inclusion-exclusion property for plausibility functions:

$$Pl(\cap_{i=1}^n B_i) = \sum_{\mathscr{I} \subseteq \mathscr{B}_n} (-1)^{|\mathscr{I}|+1} Pl(\cup_{B \in \mathscr{I}} B) \tag{6}$$

for a family of sets $\mathscr{B}_n = \{\overline{A}_i : A_i \in \mathscr{A}_n\}$ where $\mathscr{A}_n$ satisfies the condition of Proposition 1.

3 When focal sets are Cartesian products

In this section, we investigate a practically important subcase where focal sets and events $A_i, i = 1,\dots,n$ are Cartesian products. That is, we assume that $\mathscr{X} = \mathscr{X}^1 \times \dots \times \mathscr{X}^D := \mathscr{X}^{1:D}$ is the product space of *finite* spaces $\mathscr{X}^i$, $i = 1,\dots,D$. We will call the spaces $\mathscr{X}^i$ *dimensions*. We will denote by X_i the value of a variable (e.g., the state of a component, the value of a propositional variable) on $\mathscr{X}^i$.

Given $A \subseteq \mathscr{X}$, we will denote by A^i the projection of A on $\mathscr{X}^i$. Let us call *rectangular* a subset $A \subseteq \mathscr{X}$ that can be expressed as the Cartesian product $A = A^1 \times \dots \times A^D$ of its projections (in general, we only have $A \subseteq A^1 \times \dots \times A^D$ for any subset A). Note that a rectangular subset A is completely characterized by its projections.

In the following we study the additivity property for families $\mathscr{A}_n$ containing rectangular sets only, when the focal sets of mass functions defined on $\mathscr{X}$ are also rectangular (to simplify the proofs, we will also assume that all rectangular sets are focal sets). Note that, in practice, assuming sets of $\mathscr{A}$ to be rectangular is not very restrictive, as in the finite case, any set $A \subseteq \mathscr{X}$ can be decomposed into a union of rectangular subsets.

3.1 Two sets, two dimensions

Let us first explore the case $n = 2$ and $D = 2$, that is $\mathscr{A}_2 = \{A_1, A_2\}$ with $A_i = A_i^1 \times A_i^2$ for $i = 1,2$. The main idea in this case is that if $A_1 \setminus A_2$ and $A_2 \setminus A_1$ are rectangular with disjoint projections, then $\mathscr{A}_2$-additivity holds for belief functions and this is characteristic.

Lemma 2. *If A and B are rectangular and have disjoint projections, then there is no rectangular subset of $A \cup B$ overlapping both A and B*

Proof. Consider $C = C^1 \times C^2$ overlapping both A and B. So there is $a^1 \times a^2 \in A \cap C$ and $b^1 \times b^2 \in B \cap C$. Since C is rectangular, $a^1 \times b^2$ and $b^1 \times a^2 \in C$. However if $C \subseteq A \cup B$ then $a^1 \times b^2 \in A \cup B$ and either $b^2 \in A^2$ or $a^1 \in B^1$. Since $a^1 \in A^1$ and $b^2 \in B^2$ by assumption, we reach a contradiction since projections are not disjoint. □

We can now study characteristic conditions for additivity for belief functions on two sets:

Theorem 3. *Additivity applied to $\mathscr{A}_2 = \{A_1, A_2\}$ holds for belief functions if and only if one of the following condition holds*

1. $A_1^1 \cap A_2^1 = A_1^2 \cap A_2^2 = \emptyset$
2. $A_1^1 \subseteq A_2^1$ *and* $A_2^2 \subseteq A_1^2$ *(or changing both inclusion directions)*

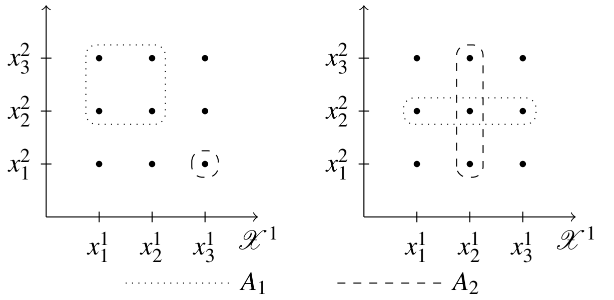


Figure 1: Situations satisfying Theorem 3

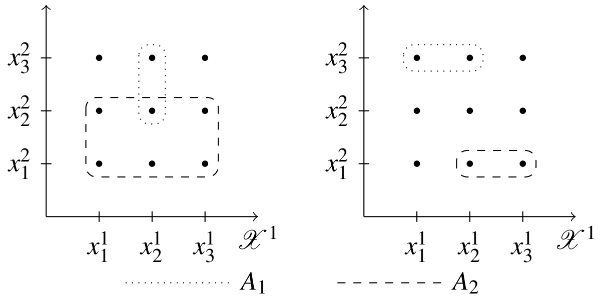


Figure 2: Situations not satisfying Theorem 3

Proof. First note that inclusions of Condition 2 can be considered as strict, as we have assumed A_1, A_2 to not be included in each other (otherwise the result is trivial).
$\Leftarrow$ 1.: If $A_1^1 \cap A_2^1 = A_1^2 \cap A_2^2 = \emptyset$, A_1 and A_2 are disjoint, as well as their projections. Then by Lemma 2 additivity holds for belief functions on any two sets.
$\Leftarrow$ 2.: $A_1^1 \subset A_2^1$ and $A_2^2 \subset A_1^2$ implies that $A_1 \setminus A_2 = A_1^1 \times (A_1^2 \setminus A_2^2)$ and $A_2 \setminus A_1 = (A_2^1 \setminus A_1^1) \times A_2^2$. As they are rectangular and have disjoint projections, Lemma 2 applies.
$\Rightarrow$ 1.: Suppose $A_1 \cap A_2 = \emptyset$ with $A_1^1 \cap A_2^1 \neq \emptyset$. Then $(A_1^1 \cap A_2^1) \times (A_1^2 \cup A_2^2)$ is rectangular, not contained in A_1 nor A_2 but contained in $A_1 \cup A_2$, so additivity does not hold.
$\Rightarrow$ 2.:Suppose $A_1^1 \subset A_2^1$ but $A_2^2 \not\subset A_1^2$. Again, $(A_1^1 \cap A_2^1) \times (A_1^2 \cup A_2^2) = A_1^1 \times (A_1^2 \cup A_2^2)$ is rectangular, neither contained in A_1 nor A_2 but contained in $A_1 \cup A_2$. □

Figure 1 and 2 show various situations where conditions of Theorem 3 are satisfied and not satisfied, respectively.

3.2 The multidimensional case

We can now proceed to extend Theorem 3 to the case of any number D of dimensions. However, this extension will not be as straightforward as going from Lemma 1 to Proposition 1, and we need first to characterize when the union of two disjoint singletons is rectangular. We will call such rectangular unions *minimal rectangles*. A singleton is a degenerated example of minimal rectangle.

Lemma 3. *Let $a = \{a^1\} \times \dots \times \{a^D\}$ and $b = \{b^1\} \times \dots \times \{b^D\}$ be two distinct singletons in $\mathscr{X}$. Then, $a \cup b$ forms a* *minimal rectangle if and only if there is only one $i \in [1,D]$ such that $a^i \neq b^i$*

Proof. $\Rightarrow$: If $a^i \neq b^i$ for only one i, then $a \cup b = \{a^1\} \times \ldots \times \{a^i, b^i\} \times \ldots \{a^D\}$ is rectangular.

$\Leftarrow$: Consider the case where singletons differ on two components, say $a^1 \neq b^1$ and $a^2 \neq b^2$, without loss of generality. In this case,

$$a \cup b = \{\{a^1\} \times \{a^2\} \times \{a^3\} \times \ldots \times \{a^D\},\\ \{b^1\} \times \{b^2\} \times \{a^3\} \times \ldots \times \{a^D\}\}.$$

The projections of $a \cup b$ on dimensions 1 and 2 of $\mathscr{X}$ are $\{a^1, b^1\}$ and $\{a^2, b^2\}$ respectively, $\{a^i\}$ for $i > 2$. Hence, the Cartesian product of the projections of $a \cup b$ is the set $\{a^1, b^1\} \times \{a^2, b^2\} \times \{a^3\} \times \ldots \times \{a^D\}$. It contains elements not in $a \cup b$ (e.g. $\{a^1\} \times \times \{b^2\} \times \{a^3\} \times \ldots \times \{a^D\}$). Since $a \cup b$ is not characterised by its projections on dimensions $\mathscr{X}_i$, it is not rectangular, and this finishes the proof. □

As mentioned before, any set can be decomposed into rectangular sets, and in particular any rectangular set can be decomposed into minimal rectangles. Let us now show how Theorem 3 can be extended to D dimensions.

Theorem 4. *Additivity holds on $\mathscr{A}_2 = \{A_1, A_2\}$ for belief functions if and only if one of the following condition holds*

1. *$\exists$ distinct $p, q \in \{1, \ldots, D\}$ such that $A_1^p \cap A_2^p = A_1^q \cap A_2^q = \emptyset$*
2. *$\forall i \in \{1, \ldots, D\}$ either $A_1^i \subseteq A_2^i$ or $A_2^i \subseteq A_1^i$*

Proof. Again, we can consider that there are at least two distinct $p, q \in \{1, \ldots, D\}$ such that inclusions $A_1^p \subset A_2^p$ and $A_2^q \subset A_1^q$ of Condition 2 are strict, as we have assumed A_1, A_2 to not be included in each other (otherwise the result is trivial).

$\Leftarrow$ 1.: Any two singletons $a_1 \in A_1$ and $a_2 \in A_2$ will be such that $a_1^i \in A_1^i$ and $a_2^i \in A_2^i$ must be distinct for $i = p, q$ since $A_1^p \cap A_2^p = A_1^q \cap A_2^q = \emptyset$. Thus it will be impossible to create minimal rectangles included in $A_1 \cup A_2$, and therefore any rectangular set in it.

$\Leftarrow$ 2.: Let us denote by P the set of indices p such that $A_1^p \subset A_2^p$ and by Q the set of indices q such that $A_2^q \subset A_1^q$. Now, let us consider two singletons $a_1 \in A_1 \setminus A_2$ and $a_2 \in A_2 \setminus A_1$. Then

- $\exists p \in P$ such that $a_1^p \in A_1^p \setminus A_2^p$, otherwise a_1 is included in $A_1 \cap A_2$
- $\exists q \in Q$ such that $a_2^q \in A_2^q \setminus A_1^q$, otherwise a_2 is included in $A_1 \cap A_2$

but since $a_1^q \in A_1^q$ and $a_2^p \in A_2^p$ by definition, a_1 and a_2 must differ at least on two dimensions, hence one cannot form a minimal rectangle not in $A_1 \cap A_2$.

$\Rightarrow$ 1: Suppose $A_1 \cap A_2 = \emptyset$ with $A_1^q \cap A_2^q \neq \emptyset$ only for q. Then the following rectangular set contained in $A_1 \cup A_2$

$$(A_1^1 \cap A_2^1) \times \cdots \times (A_1^{q-1} \cap A_2^{q-1}) \times (A_1^q \cup A_2^q)\\ \times (A_1^{q+1} \cap A_2^{q+1}) \ldots \times (A_1^D \cap A_2^D) \tag{7}$$

is neither contained in A_1 nor A_2, so additivity will not hold.

$\Rightarrow$ 2.: suppose $A_1 \cap A_2 \neq \emptyset$ and $A_1^q \not\subseteq A_2^q$, $A_1^q \not\supseteq A_2^q$ for some q. Again, the set (7) is rectangular, neither contained in A_1 nor A_2 but contained in $A_1 \cup A_2$. □

Using Proposition 1, the extension of $\mathscr{A}_n$-additivity to D-dimensional sets is straightforward:

Theorem 5. *Additivity holds on $\mathscr{A}_N = \{A_1, \ldots, A_N\}$ for belief functions if and only if, for each pair A_i, A_j, one of the following condition holds*

1. *$\exists$ distinct $p, q \in \{1, \ldots, D\}$ such that $A_i^p \cap A_j^p = A_i^q \cap A_j^q = \emptyset$*
2. *$\forall \ell \in \{1, \ldots, D\}$ either $A_i^\ell \subseteq A_j^\ell$ or $A_j^\ell \subseteq A_i^\ell$*

3.3 On the practical importance of rectangular focal sets

While limiting ourselves to rectangular subsets in $\mathscr{A}$ is not especially restrictive, the assumption that focal sets have to be restricted to rectangular sets may seem restrictive (as it is not allowed to cut any focal set into smaller rectangular subsets without redistributing the mass). However, such mass assignments on rectangular sets are found in many practical situations:

- such masses can be obtained by defining marginal masses m^i on each space $\mathscr{X}^i$, $i = 1, \ldots, D$ and then combining them under an assumption of (random set) independence [7]. In this case, the joint mass m assigns to each rectangular set E the mass

  $$m(E) = \prod_{i=1}^{D} m^i(E^i). \tag{8}$$

  Additionally, computing belief and plausibility functions of any rectangular set A becomes easier in this case, as

  $$Bel(A) = \prod_{i=1}^{D} Bel^i(A^i), \quad Pl(A) = \prod_{i=1}^{D} Pl^i(A^i), \tag{9}$$

  where Bel^i, Pl^i are the measures induced by m^i;

- as all we need is to restrict masses to product events, we can also consider cases of unknown independence or of partially known dependence, as long as this knowledge can be expressed by linear constraints on the marginal masses [1];
- using more generic models than belief functions is possible [9], since the mass positivity assumption can be dropped without modifying our results.

4 Inclusion-exclusion for monotone functions

In this section, we show that the inclusion-exclusion principle can be applied to evaluate some events of interest for monotone functions, and we provide an illustration from Multi-State Systems (MSS) reliability.

4.1 Checking the conditions

Let $\phi : \mathscr{X}^{1:D} \to \mathscr{Y}$ be a D-placed function, where $\mathscr{X}^j = \{x^j_1, \ldots, x^j_{k_j}\}$ is a finite ordered set, for every $j = 1, \ldots, D$. We note $\leq_j$ the order relation on $\mathscr{X}_j$ and assume (without loss of generality) that elements are indexed such that $x^j_i <_j x^j_k$ iff $i < k$. We also assume that the output space $\mathscr{Y}$ is ordered and we note $\leq_{\mathscr{Y}}$ the order on $\mathscr{Y}$, assuming an indexing such that $y_i <_{\mathscr{Y}} y_k$ iff $i < k$. Given two elements $x, y \in \mathscr{X}^{1:D}$, we simply write $x \geq y$ if $x^j \geq_j y^j$ for $j = 1, \ldots, n$, and $x < y$ if moreover $x \neq y$ (i.e., $x^j <_j y^j$ for at least one j).

We assume that the function is non-decreasing in each of its arguments X^j, that is

$$\phi(x^1_{i_1}, \ldots, x^\ell_{i_\ell}, \ldots, x^D_{i_D}) \leq_{\mathscr{Y}} \phi(x^1_{i_1}, \ldots, x^\ell_{i'_\ell}, \ldots, x^D_{i_D}) \quad (10)$$

iff $i_\ell \leq i'_\ell$. Note that a function monotone in each variable X^j can always be transformed into a non-decreasing one, simply by reversing$\leq_j$ for those variables X^j in which ϕ is non-increasing.

We now consider the problem of estimating the uncertainty of some event $\{\phi(\cdot) \geq d\}$ (or $\{\phi(\cdot) < d\}$, obtained by duality). Evaluating the uncertainty over such events is instrumental in a number of applications, such as risk analysis [2]. Given a value $d \in \mathscr{Y}$, let us define the concept of minimal path and minimal cut vectors.

Definition 1. A minimal path (MP) vector x of function ϕ for value d is an element $x \in \mathscr{X}^{1:D}$ such that $\phi(x) \geq d$ and $\phi(y) < d$ for any $x > y$ (x is a minimal element in $\{x : \phi(x) \geq d\}$).

Definition 2. A minimal cut (MC) vector x of function ϕ for value d is an element $x \in \mathscr{X}^{1:D}$ such that $\phi(x) < d$ and $\phi(y) \geq d$ for any $x < y$ (x is a maximal element in $\{x : \phi(x) < d\}$).

Let $p_1, \ldots, p_P$ be the set of all minimal path vectors of some function for a given performance level d (means to obtain minimal paths are provided by Xue [18]). We note $A_{p_i} = \{x \in \mathscr{X}^{1:D} | x \geq p_i\}$ the set of configurations dominating the minimal path vector p_i and $\mathscr{A}_{\mathscr{P}} = \{A_{p_1}, \ldots, A_{p_P}\}$ the set of events induced by minimal path vectors. Note that

$$A_{p_i} = \times_{j=1}^{D} \{x^j | x^j \geq_j p^j_i\} \quad (11)$$

is rectangular, hence we can use results from Section 3.

Lemma 4. *The rectangular sets $\mathscr{A}_{\mathscr{P}}$ induced by minimal path vectors satisfy Theorem 5*

Proof. Consider two minimal path vectors A_{p_i}, A_{p_j} and a dimension ℓ, then either $\{x^\ell \geq_\ell p^\ell_i\} \subseteq \{x^\ell \geq_\ell p^\ell_j\}$ or $\{x^\ell \geq_\ell p^\ell_i\} \supseteq \{x^\ell \geq_\ell p^\ell_j\}$. □

It can be checked that $\{x \in \mathscr{X}^{1:D} | \phi(x) \geq d\} = \cup_{i=1}^{P} A_{p_i}$. We can therefore write the inclusion/exclusion formula for belief functions:

$$\begin{aligned} Bel(\phi(x) \geq d) &= Bel(A_{p_1} \cup \ldots \cup A_{p_P}) \\ &= \sum_{\mathscr{I} \subseteq \mathscr{A}_{\mathscr{P}}} (-1)^{|\mathscr{I}|+1} Bel(\cap_{A \in \mathscr{I}} A), \\ &= 1 - Pl(\phi(x) < d) \end{aligned} \quad (12)$$

Under the hypothesis of random set independence, computing each term simplifies into

$$Bel(A_{p_j} \cap \ldots \cap A_{p_k}) = \prod_{i=1}^{D} Bel(\{x^i \geq \max\{p^i_j, \ldots, p^i_k\})$$

The computation of $Bel(\phi(x) < d)$ can be done similarly by using minimal cut vectors. Let $\mathscr{C}_1, \ldots, \mathscr{C}_C$ be the set of all minimal cut vectors of ϕ. Then $A_{\mathscr{C}_i} = \{x \in \mathscr{X}^{1:D} | x \leq \mathscr{C}_i\} = \times_{j=1}^{D} \{x^j | x^j \leq_j \mathscr{C}^j_i\}$ is rectangular and we have the following result, whose proof is similar to the one of Lemma 4.

Lemma 5. *The rectangular sets $\mathscr{A}_{\mathscr{C}}$ induced by minimal cut vectors satisfy Theorem 5*

Denoting by $\mathscr{A}_{\mathscr{C}} = \{A_{\mathscr{C}_1}, \ldots, A_{\mathscr{C}_C}\}$ the set of events induced by minimal cut vectors, we have that $\{x \in \mathscr{X}^{1:D} | \phi(x) < d\} = \cup_{i=1}^{C} A_{\mathscr{C}_i}$, hence applying the inclusion/exclusion formula for belief functions gives

$$\begin{aligned} Bel(\phi(x) < d) &= Bel(A_{\mathscr{C}_1} \cup \ldots \cup A_{\mathscr{C}_C}) \\ &= \sum_{\mathscr{I} \subseteq \mathscr{A}_{\mathscr{C}}} (-1)^{|\mathscr{I}|+1} Bel(\cap_{A \in \mathscr{I}} A), \\ &= 1 - Pl(\phi(x) \geq d). \end{aligned} \quad (13)$$

Let us now illustrate how this result can be applied to reliability problems.

4.2 Application to Multi-State Systems (MSS) reliability

Using the inclusion/excusion formula is a classical way of estimating system reliability. In this section we show that, thanks to our results, we can extend it to the case where system components can be in multiple states and where the uncertainty about these states is given by belief functions. We refer to Lisnianski and Levitin [15] for a detailed review of the problem.

MSS analysed in this section are such that

- their components are s-independent, meaning that the state of one component has no influence over the state of other components;
- the states of each component are mutually exclusive;
- the MSS is coherent (if one state component efficiency increases, the overall efficiency increases).

Let us now show that for such systems, we can define minimal path sets and minimal cut sets that satisfy the exclusion/inclusion principle.

In reliability analysis, variables X^j, $j = 1,\dots,D$ correspond to the D components of the system and the value x_i^j is the ith state of component j. Usually, states are ordered according to their performance rates, hence we can assume the spaces $\mathscr{X}^j$ to be ordered. $\mathscr{X}^{1:D}$ corresponds to the system states and $\mathscr{Y} = \{y_1,\dots,y_Y\}$ is the ordered set of global performance rates of the system.

The structure function $\phi : \mathscr{X}^{1:D} \to \mathscr{Y}$ links the system states to its global performance. As the system is coherent, function ϕ is non-decreasing, in the sense of Eq. (10).

As a typical task in multi-state reliability analysis is to estimate with which certainty a system will guarantee a level d of performance, results from Section 4.1 directly apply.

Example 1. Let us now illustrate our approach on a complete example, inspired from Ding and Lisnianski [10].

In this example, we aim to evaluate the availability of a flow transmission system design presented in Fig. 3 and made of three pipes. The flow is transmitted from left to right and the performance levels of the pipes are measured by their transmission capacity (tons of per minute). It is supposed that elements 1 and 2 have three states: a state of total failure corresponding to a capacity of 0, a state of full capacity and a state of partial failure. Element 3 only has two states: a state of total failure and a state of full capacity. All performance levels are precise.

The state performance levels and the state probabilities of the flow transmitter system are given in Table 2. These probabilities could have been obtained the imprecise Dirichlet model [4] considered in Li *et al.* [14]. We aim to estimate the availability of the system when $d = 1.5$. The minimal paths are

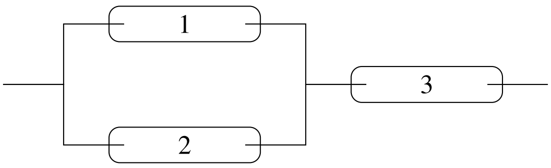


Figure 3: Flow transmission system

$$p_1 = (x_1^1, x_2^2, x_3^3) = (0, 1.5, 4),\ p_2 = (x_3^1, x_1^2, x_3^3) = (1.5, 0, 4).$$

The set A_{p_1} and A_{p_2} of vectors a such that $a \geq p_1$, $b \geq p_2$ are

$$A_{p_1} = \{0,1,1.5\} \times \{1.5,2\} \times \{4\} \text{ and}$$
$$A_{p_2} = \{1.5\} \times \{0,1.5,2\} \times \{4\},$$

and their intersection $A_{p_1} \cap A_{p_2}$ consists of vectors c such that $c \geq p_1 \vee p_2$ (with $\vee = \max$), that is:

$$A_{p_1} \cap A_{p_2} = \{1.5\} \times \{1.5,2\} \times \{4\}.$$

Applying the inclusion/exclusion formula for a requested level $d = 1.5$, we obtain

$$Bel(\phi \geq 1.5) \quad = Bel(A_{p_1}) + Bel(A_{p_2}) - Bel(A_{p_1} \cap A_{p_2})$$

For example, we have

$$\begin{aligned} Bel(A_{p_1}) &= Bel(\{0,1,1.5\} \times \{1.5,2\} \times \{4\}) \\ &= Bel(\{0,1,1.5\}).Bel(\{1.5,2\}).Bel(\{4\}) \\ &= 1 * 0.895 * 0.958 \\ &= 0.8574 \end{aligned}$$

and $Bel(A_{p_2})$, $Bel(A_{p_1} \cap A_{p_2})$ can be computed similarly. Finally we get

$$Bel(\phi \geq 1.5) \quad = \quad 0.8574 + 0.7654 - 0.6851 = 0.9377$$

and by duality with $Bel(\phi < 1.5)$, we get

$$Pl(\phi \geq 1.5) \quad = \quad 1 - Bel(\phi < 1.5) = 0.9523.$$

The availability A_s of the flow transmission system for a requested performance level $d = 1.5$ is given by $[Bel(A), Pl(A)] = [0.9377, 0.9523]$.

5 The case of Boolean formulas

In this section, we consider binary spaces $\mathscr{X}^i$, and lay bare conditions for applying the inclusion/exclusion property to Boolean formulas expressed in Disjunctive Normal Form (DNF).

x^1	0	0	0	0	0	0	1	1	1	1	1	1	1.5	1.5	1.5	1.5	1.5	1.5
x^2	0	0	1.5	1.5	2	2	0	0	1.5	1.5	2	2	0	0	1.5	1.5	2	2
x^3	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4	0	4
$y = \Phi(x^1,x^2,x^3)$	0	0	0	1.5	0	2	0	1	0	2.5	0	3	0	1.5	0	3	0	3.5

Table 1: Performance rates of the oil transmission system

X^j	1	2	3
p_1^j	[0.096,0.106]	[0.095,0.105]	-
p_2^j	[0.095,0.105]	[0.195,0.205]	[0.032,0.042]
p_3^j	[0.799,0.809]	[0.7,0.71]	[0.958,0.968]
g_1^j	0	0	-
g_2^j	1	1.5	0
g_3^j	1.5	2	4

Table 2: Parameters of the flow transmission system

In propositional logic, each $\mathscr{X}^i = \{x^i, \overline{x}^i\}$ can be associated to a variable also denoted by x_i, and $\mathscr{X}^{1:D}$ is the set of interpretations of the propositional language generated by the set $\mathscr{V}$ of variables x_i. In this case, x^i is understood as an atomic proposition, while $\overline{x}^i$ denotes its negation. Any rectangular set $A \subseteq \mathscr{X}^{1:D}$ can then be interpreted as a conjunction of literals (often called a partial model), and given a collection of n such partial models $\mathscr{A}_n = \{A_1, \dots, A_n\}$, the event $A_1 \cup \dots \cup A_n$ is a Boolean formula expressed in Disjunctive Normal Form (DNF - a disjunction of conjunctions). All Boolean formulas can be written in such a form.

A convenient representation of a partial model A is in the form of an orthopair [6] (P,N) of disjoint subsets of indices of variables $P,N \subset [1,D]$ such that $A_{(P,N)} = \bigwedge_{k\in P} x^k \wedge \bigwedge_{k \in N} \overline{x}^k$. Then a singleton in $\mathscr{X}^{1:D}$ is of the form $\bigwedge_{k\in P} x^k \wedge \bigwedge_{k\in \overline{P}} \overline{x}^k$, i.e. corresponds to an orthopair $(P,\overline{P})$.

We consider that the uncertainty over each Boolean variable x^i is described by a belief function Bel^i. For simplicity, we shall use x^i as short for $\{x^i\}$ in the argument of set-functions. As $\mathscr{X}^i$ is binary, its mass function m^i only needs two numbers to be defined, e.g., $l^i = Bel^i(x^i)$ and $u^i = Pl^i(x^i)$. Indeed, we have $Bel^i(x^i) = l^i = m^i(x^i)$, $Pl^i(x^i) = 1 - Bel^i(\overline{x}^i) = 1 - m^i(\overline{x}^i)$ and $m^i(\mathscr{X}^i) = u^i - l^i$. For D marginal masses m^i on $\mathscr{X}^i$, $i = 1, \dots, D$, the joint mass m on $\mathscr{X}^{1:D}$ can be computed as follows for any partial model $A_{(P,N)}$, applying Equation(8):

$$m(A_{(P,N)}) = \prod_{i\in P} l^i \prod_{i \in N} (1-u^i) \prod_{i \notin P\cup N} (u^i - l^i) \qquad (14)$$

We can particularize Theorem 5 to the case of Boolean formulas, and identify conditions under which the belief or the plausibility of a DNF can be easily estimated using Equality (1), changing probability into belief. Let us see how the conditions exhibited in this theorem can be expressed in the Boolean case.

Consider the first condition of Theorem 5

$$\exists p \neq q \in \{1, \dots, D\} \text{ such that } A_i^p \cap A_j^p = A_i^q \cap A_j^q = \emptyset.$$

Note that when spaces are binary, $A_i^p = x^p$ (if $p \in P_i$), or $A_i^p = \overline{x}^p$ (if $p \in N_i$), or yet $A_i^p = \mathscr{X}^i$ (if $p \notin P_i \cup N_i$). $A_i \cap A_j = \emptyset$ therefore means that for some index p, $p \in (P_i \cap N_j) \cup (P_j \cap N_i)$ (there are two opposite literals in the conjunction).

The condition can thus be rewritten as follows, using orthopairs (P_i, N_i) and (P_j, N_j):

$$\exists p \neq q \in \{1, \dots, D\} \text{ such that } p,q \in (P_i \cap N_j) \cup (P_j \cap N_i).$$

For instance, consider the equivalence connective $x^1 \iff x^2 = (x^1 \wedge x^2) \vee (\overline{x}^1 \wedge \overline{x}^2)$ so that $A_1 = x^1 \wedge x^2$ and $A_2 = \overline{x}^1 \wedge \overline{x}^2$. We have $p = 1 \in P_1 \cap N_2, q = 2 \in P_1 \cap N_2$, hence the condition is satisfied and $Bel(x^1 \iff x^2) = Bel(x^1 \wedge x^2) + Bel(\overline{x}^1 \wedge \overline{x}^2)$ (the remaining term is $Bel(\emptyset)$.

The second condition of Theorem 5 reads

$$\forall \ell \in \{1, \dots, D\} \text{ either } A_i^\ell \subseteq A_j^\ell \text{ or } A_j^\ell \subseteq A_i^\ell$$

and the condition $A_i^\ell \subseteq A_j^\ell$ can be expressed in the Boolean case as:

$$\ell \in (P_i \cap \overline{N}_j) \cup (N_i \cap \overline{P}_j) \cup (\overline{P}_i \cap \overline{N}_i \cap \overline{P}_j \cap \overline{N}_j).$$

The condition can thus be rewritten as follows, using orthopairs (P_i, N_i) and (P_j, N_j):

$$P_i \cap N_j = \emptyset \text{ and } P_j \cap N_i = \emptyset$$

For instance consider the disjunction $x^1 \vee x^2$, where $A_1 = x^1$ and $A_2 = x^2$, so that $P_1 = \{1\}, P_2 = \{2\}, N_1 = N_2 = \emptyset$. So $Bel(x^1 \vee x^2) = Bel(x^1) + Bel(x^2) - Bel(x^1 \wedge x^2)$.

We can summarize the above results as

Proposition 6. *The set of partial models $\mathscr{A}_n = \{A_1, \dots, A_n\}$ satisfies the inclusion/exclusion principle if and only if, for any pair A_i, A_j one of the two following conditions is satisfied:*

- $\exists p \neq q \in \{1, \dots, D\}$ *s.t.* $p,q \in (P_i \cap N_j) \cup (P_j \cap N_i)$.
- $P_i \cap N_j = \emptyset$ *and* $P_j \cap N_i = \emptyset$

This condition tells us that for any pair of partial models, :

- either conjunctions A_i, A_j contain at least two opposite literals,
- or events A_i, A_j have a non-empty intersection and have a common model.

These conditions allow us to check, once a formula has been put in DNF, whether or not the inclusion/exclusion principle applies. Important particular cases where it applies are disjunctions of partial models having only positive (negative) literals, of the form $A_1 \cup \ldots \cup A_n$, where $N_1 = \ldots = N_n = \emptyset$ ($P_1 = \ldots = P_n = \emptyset$). This is the typical Boolean formula one obtains in fault tree analysis, where the system failure is due to the failures of some subsets of components, the latter failures being modelled by positive literals. More generally, the inclusion/exclusion principle applies to disjunctions of partial models which can, via a renaming, be rewritten as a disjunction of conjunctions of positive literals: namely, whenever a single variable never appears in a positive and negative form in two of the conjunctions.

As an example where the inclusion/exclusion principle cannot be applied, consider the formula $x^1 \vee (\overline{x}^1 \wedge x^2)$ (which is just the disjunction $x^1 \vee x^2$ we already considered above). It does not hold that $Bel(x^1 \vee (\overline{x}^1 \wedge x^2) = Bel(x^1) + Bel(\overline{x}^1 \wedge x^2)$, since the latter sum neglects $m(x^2)$, where x^2 is a focal set that implies neither x^1 nor $\overline{x}^1 \wedge x^2$. Note that this remark suggests that normal forms that are very useful to compute the probability of a Boolean formula efficiently may not be useful to speed up computations of belief and plausibility degrees. For instance, $x^1 \vee (\overline{x}^1 \wedge x^2)$ is a binary decision diagram (BDD) [5] for the disjunction, and this form prevents $Bel(x^1 \vee x^2)$ from being computed using the inclusion/exclusion principle.

We can give explicit expressions for the belief and plausibility of conjunctions or disjunctions of literals in terms of marginal mass functions:

Proposition 7. *The belief of a conjunction $C_{(P,N)} = \bigwedge_{k\in P} x^k \wedge \bigwedge_{k\in N} \overline{x}^k$, and that of a disjunction $D_{(P,N)} = \bigvee_{k\in P} x^k \vee \bigvee_{k\in N} \overline{x}^k$ of literals forming an orthopair (P,N) are respectively given by:*

$$Bel(C_{(P,N)}) = \prod_{i\in P} l^i \prod_{i\in N} (1-u^i), \tag{15}$$

$$Bel(D_{(P,N)}) = 1 - \prod_{i\in P} (1-l^i) \prod_{i\in N} u^i. \tag{16}$$

Proof. $Bel(C_{(P,N)})$ can be obtained by applying Equation (9) to $C_{(P,N)}$.

For $Bel(D_{(P,N)})$, we have

$$\begin{aligned} Pl(C_{(N,P)}) &= Pl(\wedge_{i\in N} x^i \wedge \wedge_{i\in P} \overline{x}^i) \\ &= \prod_{i\in N}(1-l^i) \prod_{i\in P} u^i \\ &= 1-(1-\prod_{i\in N}(1-l_i)\prod_{i\in P} u_i) \\ &= 1-Bel(\vee_{i\in N} \overline{x}^i \vee \vee_{i\in P} x^i) \\ &= 1-Bel(D_{(P,N)}) \end{aligned}$$

where the second equality following from Equation (9). □

Using the fact that $Bel(C_{(N,P)}) = 1 - Pl(D_{(P,N)})$, we can deduce

$$Pl(D_{(P,N)}) = 1 - \prod_{i\in P} l^i \prod_{i\in N} (1-u^i). \tag{17}$$

$$Pl(C_{(P,N)}) = \prod_{i\in P} u^i \prod_{i\in N} (1-l^i). \tag{18}$$

To compute the plausibility of a formula ϕ, we can put it in conjunctive normal form, that is as a conjunction of clauses $\wedge_{i=1}^k \kappa_i$ where the κ_i's are disjunctions of literals. Then we can write:

$$Pl(\phi) = 1 - Bel(\neg(\wedge_{i=1}^k \kappa_i)) = 1 - Bel(\vee_{i=1}^k \neg\kappa_i) \tag{19}$$

Noticing that the terms $\neg\kappa_i$ are rectangular (partial models), we can apply Proposition 6 again (this trick can be viewed as an application of results of Subsection 4.1 to ordered scale $\mathscr{X} = \{0 < 1\}$). As a consequence we can compute the belief and the plausibility of any logical formula that obeys the conditions of Proposition 6 in terms of the belief and plausibilities of atoms x^i.

Example 2. For instance consider the formula $\phi = (x^1 \wedge \overline{x}^2) \vee (\overline{x}^1 \wedge x^2) \vee x^3$, with $A_1 = x^1 \wedge \overline{x}^2, A_2 = \overline{x}^1 \wedge x^2, A_3 = x^3$. It satisfies Proposition 6, and

$$\begin{aligned} &Bel(\phi) = Bel(x^1 \wedge \overline{x}^2) + Bel(\overline{x}^1 \wedge x^2) \\ &+ Bel(x^3) - Bel(x^1 \wedge \overline{x}^2 \wedge x^3) - Bel(\overline{x}^1 \wedge x^2 \wedge x^3) \\ &= l_1(1-u_2) + (1-u_1)l_2 + l_3(1-l_1(1-u_2) - (1-u_1)l_2) \end{aligned}$$

In CNF, this formula reads : $(x^1 \vee x^2) \wedge (\overline{x}^1 \vee \overline{x}^2) \wedge x^3$. Then:

$$\begin{aligned} &Pl(\phi) = 1 - Bel((x^1 \wedge x^2) \vee (\overline{x}^1 \wedge \overline{x}^2) \vee \overline{x}^3); \\ &= 1 - Bel(x^1 \wedge x^2) - Bel(\overline{x}^1 \wedge \overline{x}^2) - Bel(\overline{x}^3) \\ &+ Bel(x^1 \wedge x^2 \wedge \overline{x}^3) + Bel(\overline{x}^1 \wedge \overline{x}^2 \wedge \overline{x}^3) \\ &= 1 - l^1 l^2 - (1-u^1)(1-u^2) - 1 + u^3 + l_1 l_2 (1-u^3) \\ &+ (1-u^1)(1-u^2)(1-u^3) \end{aligned}$$

6 Conclusion

We provided necessary and sufficient conditions for the inclusion/exclusion principle to hold with belief functions.

To demonstrate the usefulness of those results, we discussed their application to system reliability and to uncertainty evaluation over DNF and CNF Boolean formulas.

We can mention several lines of research that would complement the present results: (1) find necessary and sufficient conditions for the inclusion/exclusion principle to hold for plausibilities in the general case (a counterpart to Proposition 5); (2) investigate the relation between the assumption of random set independence (made in this paper) and other types of independence [12]; (3) investigate how to decompose an event / a formula into a set of event satisfying the inclusion/exclusion principle (e.g., classical BDDs do not always provide adequate solutions).

Acknowledgements

This work was carried out in the framework of the Labex MS2T, which was funded by the French Government, through the program 'Investments for the future" managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02) and is also partially supported by @MOST Prototype, a joint project of Airbus, LAAS, ONERA and ISAE.

References

[1] C. Baudrit and D. Dubois. Comparing methods for joint objective and subjective uncertainty propagation with an example in a risk assessment. In *Proc. Fourth International Symposium on Imprecise Probabilities and Their Application (ISIPTA'05)*, pages 31–40, Pittsburg (USA, Pennsylvanie), 2005.

[2] C. Baudrit, D. Guyonnet, and D. Dubois. Joint propagation and exploitation of probabilistic and possibilistic information in risk assessment. *IEEE Trans. Fuzzy Systems*, 14:593–608, 2006.

[3] Y. Ben-Haim. *Info-gap decision theory: decisions under severe uncertainty*. Academic Press, 2006.

[4] J.M. Bernard. An introduction to the imprecise dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(2):123–150, 2005.

[5] R.E. Bryant. Symbolic boolean manipulation with ordered binary-decision diagrams. *ACM Computing Surveys (CSUR)*, 24(3):293–318, 1992.

[6] D. Ciucci. Orthopairs: A simple and widely usedway to model uncertainty. *Fundam. Inform.*, 108(3-4):287–304, 2011.

[7] I. Couso, S. Moral, and P. Walley. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5:165–181, 2000.

[8] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[9] S. Destercke. Independence and 2-monotonicity: Nice to have, hard to keep. *International Journal of Approximate Reasoning (In press)*, 2012.

[10] Y. Ding, M. J. Zuo, A. Lisnianski, and Z. G. Tian. Fuzzy multi-state system: General definition and performance assessment. *IEEE Transactions on Reliability*, 57:589 – 594, 2008.

[11] D. Dubois and H. Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York, 1988.

[12] C. Jacob, D. Dubois, and J. Cardoso. Evaluating the uncertainty of a boolean formula with belief functions. *Advances in Computational Intelligence*, pages 521–531, 2012.

[13] E.T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[14] C-Y. Li, X. Chen, X-S. Yi, and J y. Tao. Interval-valued reliability analysis of multi-state systems. *IEEE Transactions on Reliability*, 60:323 – 330, 2011.

[15] A. Lisnianski and G. Levitin. M*ulti-*S*tate* S*ystem Reliability: Assessment, Optimization and Applications*. World Scientific Publishing Co Pte Ltd, 2003.

[16] G. Shafer. *A mathematical Theory of Evidence*. Princeton University Press, New Jersey, 1976.

[17] P. Walley. *Statistical reasoning with imprecise Probabilities*. Chapman and Hall, New York, 1991.

[18] J. Xue. On multistate system analysis. *IEEE Transactions on Reliability*, pages 329–337, 1985.

Temporal Data Classification by Imprecise Dynamical Models

Alessandro Antonucci
IDSIA (Switzerland)
alessandro@idsia.ch

Rocco de Rosa
Università di Milano (Italy)
rocco.derosa@unimi.it

Alessandro Giusti
IDSIA (Switzerland)
alessandrog@idsia.ch

Fabio Cuzzolin
Oxford Brookes (UK)
fabio.cuzzolin@brookes.ac.uk

Abstract

We propose a new methodology to classify temporal data with imprecise hidden Markov models. For each sequence we learn a different model by coupling the EM algorithm with the imprecise Dirichlet model. As a model descriptor, we consider the expected value of the observable variable in the limit of stationarity of the Markov chain. In the imprecise case, only the bounds of this descriptor can be evaluated. In practice the sequence, which can be regarded as a trajectory in the feature space, is summarized by a hyperbox in the same space. We classify these static but interval-valued data by a credal generalization of the k-nearest neighbors algorithm. Experiments on benchmark datasets for computer vision show that the method achieves the required robustness whilst outperforming other precise and imprecise methods.

Keywords. Time-series classification, credal sets, Markov chains, credal classification.

1 Introduction

The theory of imprecise probability (IP, [17]) extends Bayesian theory of subjective probability to cope with sets of distributions, this providing more general and robust models of uncertainty. These ideas have been applied to classification and a number of IP-based, so-called *credal*, classifiers for static data have been proposed (e.g., [19]). A key feature of these approaches is the ability of discriminating between hard-to-classify instances (e.g., for Bayesian-like approaches, those prior-dependent) for which multiple class labels are returned in output, and the others "easy" instances to which single labels are assigned. On the other side, dynamical models such as Markov chains and hidden Markov models (HMMs) have been also extended to IP in order to model the non-stationarity of a process (see e.g., [5, 6]). It seems therefore natural to merge these two lines of research and develop a credal classifier for temporal data based on imprecise HMMs, thus generalizing methods already developed for precise HMMs (e.g., [13]).

This is achieved as follows. First, from each sequence, we learn an imprecise HMM by means of a technique, already tested in [3] and [16], which combines the EM algorithm, commonly used to learn precise HMMs, with the *imprecise Dirichlet model* (IDM, [18]), a popular approach to learn IPs from (complete) data. After this step, each sequence is associated with an imprecise HMM. As a descriptor of this model (and hence of the sequence), we evaluate the lower and upper bounds of the expected values of the features in the limit of stationarity. This is based on a characterization of the limit behaviour of imprecise Markov chains provided in [6]. As a result, the sequence is associated with a hyperbox in the feature space. This represents a static, but interval-valued, datum which can be processed by a classifier. To achieve that, a generalization of the k-nearest neighbors algorithm to support interval data is proposed. Overall this corresponds to a *credal classifier* (i.e., a classifier which might return more than a single class) for temporal data. Its performances are tested on some of the most important computer vision benchmarks. The results are promising: the methods we propose achieve the required robustness in the evaluation of the class labels to be assigned to a sequence and outperform the competing imprecise method proposed in [3] with respect to state-of-the-art metrics [20] for performance evaluation. The performance is also good when comparing the algorithm with the *dynamic time warping*, a state-of-the-art approach to the classification of temporal sequences, whose performance degrades when coping with multidimensional data [14].

2 Temporal data

Let us introduce the key features of our approach and the necessary formalism for the precise case. Variables $O_1, O_2, \dots, O_T$ denote the observations of a particular phenomenon at T different (discrete) times. These are assumed to be observable, i.e., their actual (real) values are available and denoted by $o_1, o_2, \ldots, o_T$.

If the observations are all sampled from the same distribution, say $P(O)$, the empirical mean converges to its theoretical value (strong law of large numbers):

$$\lim_{T\to+\infty} \frac{\sum_{i=1}^T o_i}{T} = \int_{-\infty}^{+\infty} o \cdot P(o) \cdot \mathrm{d}o. \tag{1}$$

Under the stationarity assumption, the empirical mean is therefore a sensible descriptor of the sequence. More generally, observations at different times can be sampled from different distributions (i.e., the process can be non-stationary). Such a situation can be modeled by pairing O_t with an auxiliary discrete variable X_t, for each $t = 1, \ldots, T$. Variables $\{X_t\}_{t=1}^T$ are in correspondence with the generating distributions: they all take values from the same set, say $\mathcal{X}$, whose M elements are in one-to-one correspondence with the different distributions. In other words, for each $t = 1, \ldots, T$, O_t is sampled from $P(O_t|X_t = x_t)$, and $P(O|x_{t'}) = P(O|x_{t''})$ if and only if $x_{t'} = x_{t''}$.

Variables $\{X_t\}_{t=1}^T$ are, generally speaking, *hidden* (i.e., their values are not directly observable). The modeling of the generative process requires therefore the assessment of the joint mass function $P(X_1, \ldots, X_T)$. This becomes particularly simple under the *Markovian assumption*: given X_{t-1}, all previous values of X are irrelevant to X_t, i.e., $P(X_t|x_{t-1}, x_{t-2}, \ldots, x_1) = P(X_t|x_{t-1})$. Together with chain rule, this implies the factorization:[1]

$$P(x_1, \ldots, x_T) := P(x_1) \cdot \prod_{t=2}^{T} P(x_t|x_{t-1}), \tag{2}$$

for each $(x_1, \ldots, x_T) \in \mathcal{X}^T$. If the transition probabilities among the hidden variables are time-homogeneous, the specification of the joint model reduces to the assessment of $P(X_1)$ and $P(X_t|X_{t-1})$, i.e., $M^2 + M$ parameters. A model of this kind is called a *Markov chain* and, in the time-homogeneous case, it is known to assume a stationary behaviour on long sequences, i.e., the following limit exists:

$$\tilde{P}(x) := \lim_{T\to\infty} P(X_T = x), \tag{3}$$

where the probability on the right-hand side is obtained by marginalizing out all the variables in the joint in Eq. (2) apart from X_T. The marginal probability mass function $\tilde{P}$ over $\mathcal{X}$ is called the *stationary mass function* of the chain and it can be computed by standard algorithms.

In this limit, also the generation of the observations becomes stationary, i.e.,

$$\tilde{P}(O) = \sum_{x\in\mathcal{X}} P(O|x) \cdot \tilde{P}(x). \tag{4}$$

Again, as in Eq. (1), the empirical mean converges to the theoretical value, which is now:

$$\lim_{T\to+\infty} \frac{\sum_{i=1}^T o_i}{T} = \sum_{x\in\mathcal{X}} \tilde{P}(x) \cdot \int_{-\infty}^{+\infty} o \cdot P(o|x) \cdot \mathrm{d}o. \tag{5}$$

The two key points of this paper are the following: (i) emphasize the fact that, although coincident in the limit of infinite sequences, the weighted average of the means on the right-hand side of Eq. (5) provides a better descriptor than the empirical mean on the left-hand side; (ii) extend Eq. (5) to the imprecise-probabilistic framework and then use the new descriptor for robust classification of temporal data.

Concerning (i), the important remark is that the arithmetic mean does not take into account the temporal correlation of the data, while the learning of the transition probabilities $P(X_t|X_{t-1})$ and hence the corresponding value of the stationary mass function takes that into account. An empirical validation of this point is reported in Section 5. A discussion of point (ii) is in the next two sections.

3 Imprecise hidden Markov models

By merging the Markov chain defined in the previous section together with the time-homogeneous emission terms $P(O_t|X_t)$, we define a probabilistic model over the whole set of variables $X_1, O_1, \ldots, X_T, O_T$ which is called a *hidden Markov model* (HMM). An imprecise HMM is obtained by simply replacing with *credal sets*, i.e., convex sets of probability mass functions over the same variables, the precise local models $P(X_1)$, $\{P(X_{t+1}|x_t)\}_{x_t\in\mathcal{X}}$ and $\{P(O_t|x_t)\}_{x_t\in\mathcal{X}}$. While a precise HMM defines a single distribution over its whole set of variables, an imprecise HMM defines a joint credal set, which is the convex closure of the whole set of joint distributions obtained when each local model takes its values in the corresponding credal set. In the following we explain, respectively: (i) how to learn an imprecise HMM from a sequence; (ii) how to extend Eq. (5) to the case of imprecise HMMs; (iii) how to perform classification with these models.

3.1 Learning

The hidden variables $X_1, \ldots, X_T$ of a HMM, no matter whether precise or imprecise, are by definition directly unobservable. Algorithms to learn model parameters from incomplete data in HMMs are therefore

[1] We use P for both probability mass functions and densities.

needed. A typical choice in the precise case is the EM algorithm, which finds a local optimum of the likelihood by an iterative procedure. Extending EM to IP is not trivial: credal sets can be described by a variable number of parameters (e.g., its extreme points), which cannot be easily tracked during the iteration.[2]

Despite the lack of a sound version of EM for IP, a simple heuristic approach based on the IDM has been shown to provide reasonable estimates [3]. In practice the counts required by the IDM to learn IPs, which are not available for incomplete data, are just replaced by the expectations provided by the standard EM. For the first variable in the chain, this corresponds to the following constraints:

$$\frac{E[n(x_1)]}{\sum_{x_1} E[n(x_1)] + s} \leq P(x_1) \leq \frac{E[n(x_1)] + s}{\sum_{x_1} E[n(x_1)] + s}, \tag{6}$$

for each $x_1 \in \mathcal{X}$, where $E[n(x_1)]$ is the EM expectation, after convergence, for $X_1 = x_1$, the sum is over all the elements of $\mathcal{X}$ and s is a nonnegative real parameter which describes the level of cautiousness in the learning process. Intervals in Eq. (6) are used to compute the credal set $K(X_1)$ made of the probability mass functions consistent with these (linear) constraints. We similarly proceed for the transition credal sets $\{K(X_t|x_{t-1})\}_{x_{t-1}\in\mathcal{X}}$. Considering the freedom in the choice of the number of hidden states M, it is worth noticing that the above IDM-based probability intervals are invariant with respect to that number.

Regarding the *emission* part of the model (i.e., the relation between hidden and observable variables), note that the discussion was introduced in the case of a scalar observable O just for sake of simplicity. In real-world problems, we often need to cope with sequences of arrays of $F > 1$ features, say $\mathbf{o}_1, \ldots, \mathbf{o}_T$, with $\mathbf{o}_t \in \mathbb{R}^F$ for each $t = 1, \ldots, T$. To define a joint model over the features we assume their conditional independence given the corresponding hidden variable. A Gaussian distribution is indeed used, for each feature, to model the relation between hidden and observable variables:

$$P(\mathbf{o}_t|x_t) \cdot \mathrm{d}\mathbf{o}_t = \prod_{f=1}^{F} \mathcal{N}_{\sigma_f(x_t)}^{\mu_f(x_t)}(o_t^f) \cdot \mathrm{d}o_t^f, \tag{7}$$

where o_t^f is the f-th component of the array $\mathbf{o}_t$, $\mathcal{N}_\sigma^\mu$ is a Gaussian density with mean μ and standard deviation σ, and $\mu_f(x_t)$ and $\sigma_f(x_t)$ are the EM estimates for the mean and standard deviation of the Gaussian over O_t^f given that $X_t = x_t$.[3]

Regarding the choice of the number of hidden states $M := |\mathcal{X}|$, with Gaussian emission terms the clustering method in [12] provides an optimal criterion to assess this value. The cluster information (means and standard deviations) also defines a possible initialization of for the emission terms in the EM, while uniform choices are adopted for the transition and the prior. Overall, after this learning step, the sequence of observations in the F-dimensional space is associated with a time-homogeneous imprecise HMM, with imprecise specification of the transition and prior probabilities and precise specification of the (Gaussian) emission terms.

3.2 An interval-valued descriptor for imprecise HMMs

In this section we show how the descriptor proposed in Eq. (5) for precise HMMs can be generalized to the case of the imprecise HMM we learn from a sequence of feature vectors. In the imprecise case the stationary mass function of a Markov chain is replaced by a *stationary credal set*, say $\tilde{K}(X)$. Its computation, which is briefly summarized in Appendix A, can be obtained by Choquet integration [6]. Thus, in this generalized setup, distribution $\tilde{P}(X)$ in Eq. (5) is only required to belong to $\tilde{K}(X)$. Note that $\tilde{K}$ is a finitely generated credal set which can be equivalently characterized by (a finite number of) linear constraints. Regarding the emission terms, nothing changes as they are assumed to be precise. Thus, for each feature o_f, with $f = 1, \ldots, F$, we evaluate the bounds of the expectation as

$$\underline{o}^f := \min_{\tilde{P}(X)\in\tilde{K}(X)} \sum_{x\in\mathcal{X}} \tilde{P}(x) \cdot \mu_f(x), \tag{8}$$

$$\overline{o}^f := \max_{\tilde{P}(X)\in\tilde{K}(X)} \sum_{x\in\mathcal{X}} \tilde{P}(x) \cdot \mu_f(x). \tag{9}$$

Both $\underline{o}^f$ and $\overline{o}^f$ are solutions of linear programs with $|\mathcal{X}|$ optimization variables and an equal number of linear constraints (see Appendix A). The interval $[\underline{o}^f, \overline{o}^f]$ represents therefore the range of the descriptor in Eq. (5) in the case of imprecise HMMs.

The lower and upper vectors $\underline{\mathbf{o}}, \overline{\mathbf{o}} \in \mathbb{R}^F$ are indeed obtained by applying the optimization is Eqs. (8) and (9) to each feature. They define a hyperbox in the feature space, which can be regarded as the range of the F-dimensional version of the descriptor in Eq. (5) when

[2] An exception is the EM for belief functions proposed in [7]. Yet, belief functions correspond to a special class of credal sets parametrized by a fixed number of elements.

[3] The choice of using a single Gaussian, separately for each feature, is just for the sake of simplicity. An extension of the methods proposed in this paper to a single multivariate Gaussian with non-diagonal covariance matrix would be straightforward, even with mixtures.

IPs are introduced in the model. Overall, a static interval-valued summary of the information contained in the temporal sequence has been obtained: the sequence, which is a trajectory in the feature space is described by a hyperbox in the same space (Fig. 1). In the next section, a standard approach to the classification of static data is extended to the case of interval data like the ones produced by this method.

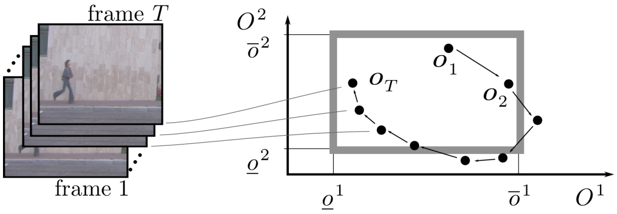


Figure 1: From trajectories to hyperboxes in the feature space. The example refers to footage data from which two features are extracted at the frame level.

4 K-nearest neighbors for interval data

4.1 Distances between hyperboxes

Consider the F-dimensional real space $\mathbb{R}^F$. Let us make it a metric space by considering, for instance, the *Manhattan* distance which, given $\mathbf{x}, \mathbf{y} \in \mathbb{R}^F$, defines their distance δ as

$$\delta(\mathbf{x}, \mathbf{y}) := \sum_{f=1}^{F} |x_f - y_f|. \tag{10}$$

Given two points $\underline{\mathbf{x}}, \overline{\mathbf{x}} \in \mathbb{R}^F$ such that, for each $f = 1, \ldots, F$, $\underline{x}_f \leq \overline{x}_f$, the *hyperbox* associated with these two points is denoted by $[\underline{\mathbf{x}}, \overline{\mathbf{x}}]$ and defined as

$$[\underline{\mathbf{x}}, \overline{\mathbf{x}}] := \left\{ \mathbf{x} \in \mathbb{R}^F \,\middle|\, \underline{x}_f \leq x_f \leq \overline{x}_f \right\}. \tag{11}$$

The problem of extending a distance defined over points to hyperboxes can be solved by considering the general ideas proposed in [1].

Given two hyperboxes, their distance can be characterized by means of a real interval whose bounds are, respectively, the minimum and the maximum distance (according to the distance defined for points) between every possible pair of elements in the two hyperboxes. Accordingly, the lower distance between two boxes is:

$$\underline{\delta}([\underline{\mathbf{x}}, \overline{\mathbf{x}}], [\underline{\mathbf{y}}, \overline{\mathbf{y}}]) := \min_{\mathbf{x} \in [\underline{\mathbf{x}}, \overline{\mathbf{x}}], \mathbf{y} \in [\underline{\mathbf{y}}, \overline{\mathbf{y}}]} \delta(\mathbf{x}, \mathbf{y}), \tag{12}$$

and similarly, with the maximum instead of the minimum for the upper distance $\overline{\delta}([\underline{\mathbf{x}}, \overline{\mathbf{x}}], [\underline{\mathbf{y}}, \overline{\mathbf{y}}])$. With the Manhattan distance in Eq. (10), the evaluation of the lower (and similarly for the upper) distance as in Eq. (12) takes a particularly simple form:

$$\underline{\delta}([\underline{\mathbf{x}}, \overline{\mathbf{x}}], [\underline{\mathbf{y}}, \overline{\mathbf{y}}]) = \sum_{f=1}^{F} \min_{\substack{\underline{x}_f \leq x_f \leq \overline{x}_f, \\ \underline{y}_f \leq y_f \leq \overline{y}_f}} |x_f - y_f|. \tag{13}$$

The optimization in the F-dimensional space is in fact reduced to F, independent, optimizations on the one-dimensional real space. Each task can be reduced to linear program whose optimum is in a combination of the extremes, unless intervals overlap. In other words:

$$\min_{\substack{\underline{x}_f \leq x_f \leq \overline{x}_f \\ \underline{y}_f \leq y_f \leq \overline{y}_f}} |x_f - y_f| = \min \left\{ \begin{array}{c} |\underline{x}_f - \underline{y}_f|, |\overline{x}_f - \underline{y}_f|, \\ |\underline{x}_f - \overline{y}_f|, |\overline{x}_f - \overline{y}_f| \end{array} \right\}, \tag{14}$$

unless $\overline{x}_f \geq \underline{y}_f$ or $\overline{y}_f \geq \underline{x}_f$, a case where the lower distance is clearly zero. A dual relation holds for the upper distance case with no special discussion in case of overlapping.

Replacing the Manhattan with the Euclidean distance makes little difference if we consider only the sum of the squared differences of the coordinates without the square root.[4] In this case the lower distance is the sum, for $f = 1, \ldots, F$ of the following terms:

$$\min_{\substack{\underline{x}_f \leq x_f \leq \overline{x}_f, \\ \underline{y}_f \leq y_f \leq \overline{y}_f}} (x_f - y_f)^2. \tag{15}$$

This is the minimum of a convex function, which is attained on the border of its (rectangular) domain. It is straightforward to check that the minimum should lie on one of the four extreme points of the domain. Thus, the minimum in Eq. (15) is the minimum of the squares of the four quantities in Eq. (14). Again, the only exception is when the two intervals overlap (the global minimum is in $x_f = y_f$), and the lower distance becomes zero. Similar considerations hold for the upper distance.

4.2 Hyperboxes classification

The above defined interval-valued distance for hyperboxes is the key to extend the *k-nearest neighbors* (k-NN) algorithm to the case of interval-valued data. First, let us review the algorithm for pointwise data.

Let C denote a *class* variable taking its values in a finite set $\mathcal{C}$. Given a collection of supervised data $\{c^d, \mathbf{x}^d\}_{d=1}^D$ classification is intended as the problem of assigning a class label $\tilde{c} \in \mathcal{C}$ to a new instance $\tilde{\mathbf{x}}$ on the basis of the data. The k-NN algorithm for $k = 1$

[4] The square root is a monotone function, which has no effect on the ranking-based classification method we define here.

assigns to $\tilde{\mathbf{x}}$ the label associated with the instance nearest to $\tilde{\mathbf{x}}$, i.e., the solution is $\tilde{c} := c^{d^*}$ with

$$d^* = \operatorname{argmin}_{d=1,\dots,D} \delta(\mathbf{x}, \mathbf{x}^d). \tag{16}$$

For $k > 1$, the k nearest instances need to be considered instead: a voting procedure among the relative classes decides the label of the test instance.

To extend this approach to interval data just replace the sharp distance among points used in Eq. (16) with the interval-valued distance for hyperboxes proposed in Section 4.1. Yet, to compare intervals instead of points a decision criterion is required.

To see that, consider for instance three hyperboxes and the two intervals describing the distance between the first hyperbox and, respectively, the second and the third. If the two intervals do not overlap, we can trivially identify which is the hyperbox nearer to the first one. Yet, in case of overlapping, this decision might be controversial. The most cautious approach is *interval dominance*, which simply suspends any decision in this case.

When applied to classification, interval dominance produces therefore a *credal* classifier, which might return more than a class in output. If the set of optimal classes according to this criterion is defined as $\mathcal{C}^*$, we have that $c \in \mathcal{C}^*$ if and only if there exists a datum $(c^i, \mathbf{x}^i)$ such that $c = c^i$ and

$$\overline{\delta}([\underline{\mathbf{x}}^i, \overline{\mathbf{x}}^i], [\underline{\mathbf{x}}, \overline{\mathbf{x}}]) < \underline{\delta}([\underline{\mathbf{x}}^d, \overline{\mathbf{x}}^d], [\underline{\mathbf{x}}, \overline{\mathbf{x}}]) \tag{17}$$

for each $d = 1, \dots, D$ such that $c^d \neq c^i$. Classes in the above defined set are said to be *undominated* because they correspond to instances in the dataset whose interval-valued distance from the test instance is not clearly bigger that the interval distance associated to any other instance. A demonstrative example is in Fig. 2. Note also that the case $k > 1$ simply requires the iteration of the evaluation in Eq. (17).

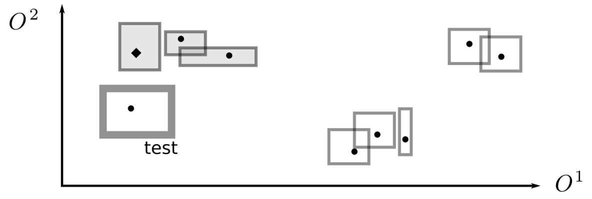


Figure 2: Rectangular data processed by the 1-NN classifier. Gray background denotes data whose interval distance from the test instance is undominated. Points inside the rectangles describe consistent precise data and the diamond is the nearest instance.

4.3 Summary and related work

By merging the discussions in Sections 3 and 4 we have a classifier, to be called iHMM-kNN, for temporal data based on imprecise HMMs. In summary, for each sequence we: (i) learn an imprecise HMM (Section 3.1); (ii) compute its stationary credal set (Appendix A); (iii) solve the LP tasks required to compute the hyperbox associated with the sequence (Section 3.2). These supervised hyperboxes are finally used to learn a credal classifier (Section 4).

Another credal classifier for temporal data based on imprecise HMMs, called here iHMM-Lik, has been proposed in [3]. Each imprecise HMM learned from a supervised sequence is used to "explain" the test instance, i.e., the lower and upper bounds of the probability of the sequence are evaluated. These (probability) intervals are compared and the optimal classes according to interval dominance returned.

Regarding traditional (i.e., not based on IP) classifiers, *dynamic time warping* (DTW) is a popular state-of-the-art approach. Yet, its performance degrades in the multi-feature (i.e., $F > 1$) case [14]. Both these methods will be compared with our classifier in the next section.

Other approaches to the specific problem of classifying interval data have been also proposed. E.g., remaining in the field of IP, the approach proposed in [15] can be used to define a SVM for interval data. Yet, time complexity increases exponentially with the number of features, thus preventing an application of the method to data with high feature dimensionality. This is not the case for iHMM-kNN, whose complexity is analyzed below.

4.4 Complexity analysis

Our approach to the learning of imprecise HMMs has the same time complexity of the precise case, namely $O(M^2TF)$. The computation of the stationary credal set is $O(T)$, while to evaluate the hyperboxes a LP task should be solved for each feature, i.e., roughly, $O(M^3F)$. Also the distance between two hyperboxes can be computed efficiently: the number of operations required is roughly four times the number of operations required to compute the distance between two points, both for Manhattan and Euclidean metrics. To classify a single instance as in Eq. (17), lower and upper distances should be evaluated for all the sequences, i.e., $O(DF)$. Overall, the complexity is linear in the number of features and in the length of the sequence and polynomial in the number of hidden states. Similar results can be found also for space.

4.5 Metrics for credal classifiers

Credal classifiers might return multiple classes in output. Evaluating their performance requires therefore specific metrics, which are reviewed here. First, a characterization of the level of indeterminacy is achieved by: the *determinacy* (*det*), i.e., percentage of instances classified with a single label; the *average output size* (*out*), i.e., average number of classes on instances for which multiple labels are returned. For accuracy we distinguish between: *single-accuracy* (*sing-acc*), i.e., accuracy over instances classified as a single label; *set-accuracy* (*set-acc*), i.e., the accuracy over the instances classified with multiple labels[5].

A utility-based measure has been recently proposed in [20] to compare credal and precise classifiers with a single indicator. In our view, this is the most principled approach to compare the 0-1 loss of a traditional classifier with a utility score defined for credal classifiers. The starting point is the *discounted accuracy*, which rewards a prediction containing q classes with $1/q$ if it contains the true class, and with 0 otherwise. This indicator can be already compared to the accuracy achieved by a determinate classifier.

Yet, risk aversion demands higher utilities for indeterminate-but-correct outputs when compared with wrong-but-determinate ones (see [20] for details). Discounted accuracy is therefore modified by a (monotone) transformation u_w with $w \in [.65, .80]$. A conservative approach consists in evaluating the whole interval $[u_{.65}, u_{.80}]$ for each credal classifier and compare it with the (single-valued) accuracy of traditional classifiers. Interval dominance can be used indeed to rank performances.

The precise counterpart of a credal classifier is a classifier always returning a single class included in the output of the credal classifier. E.g., a counterpart of iHMM-kNN is obtained by setting $s = 0$ in the IDM. If a precise counterpart is defined, it is also possible to evaluate: the *precise single accuracy* (*p-sing-acc*), i.e., the accuracy of the precise classifier when the credal returns a single label; the *precise set-accuracy* (*p-set-acc*), i.e., the accuracy of the precise classifier when the credal returns multiple labels.

5 Experiments

5.1 Benchmark datasets

To validate the performance of the iHMM-kNN algorithm we use two of the most important computer vision benchmarks: the Weizmann [8] and KTH [11] datasets for *action recognition.* For this problem, the class is the action depicted in the sequence (Fig. 3).

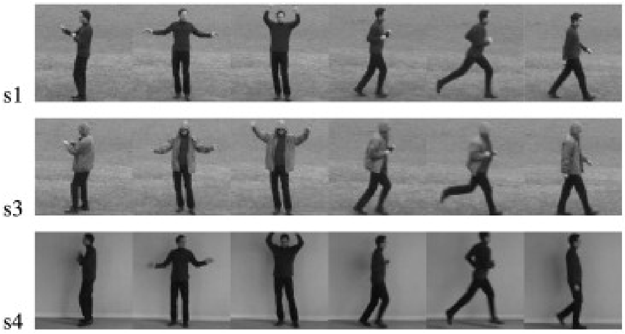


Figure 3: Frames extracted from the KTH dataset.

These data are footage material which requires a *features extraction* procedure at the frame level. Our approach is based on histograms of oriented optical flows [4], a simple technique which describes the flows distribution in the whole frame as an histogram with 32 bins representing directions (Fig. 4).

For a through validation also the AUSLAN dataset [9] based on gestures in the Australian sign language and the JAPVOW dataset [10] with speech about Japanese vowels are considered. Table 1 reports relevant information about these benchmark datasets.

Dataset	$\|\mathcal{C}\|$	F	D	T
KTH$_1$	6	32	150	51
KTH$_2$	6	32	150	51
KTH$_3$	6	32	149	51
KTH$_4$	6	32	150	51
KTH	6	32	599	51
Weizmann	9	32	72	105-378
AUSLAN	95	22	1865/600	45-136
JAPVOW	9	12	370/270	7-29

Table 1: Datasets used for benchmarking. The columns denotes, respectively, name, number of classes, number of features, size (test/training datasets sizes if no cross validation has been done) and the number of frames of each sequence (or their range if this number is not fixed). As usually done, the KTH dataset is also split in four subgroups.

To avoid features with small ranges being penalized by the k-NN with respect to others spanning larger domains a feature normalization step has been performed. This is a just a linear transformation in the feature space which makes the empirical mean of the sample equal to zero and the variance equal to one.

[5] In this case, classification is considered correct if the set of labels includes the true class.

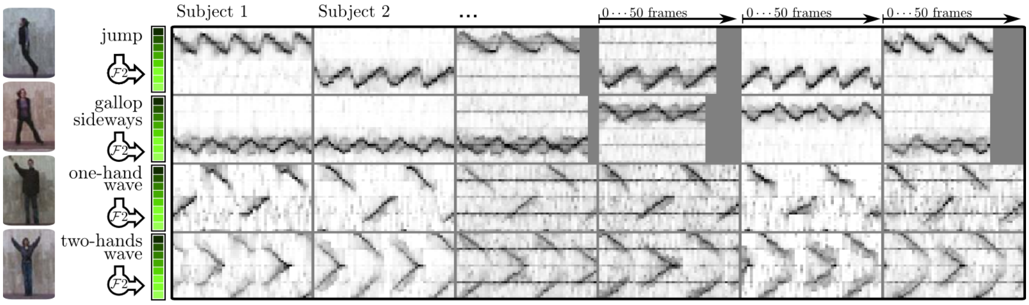


Figure 4: Low-level feature extraction. Rows correspond to different actions (i.e., class labels), columns to subjects. In each cell, feature values are shown as gray levels, with the different feature variables on the y axis, and frames on the x axis. Characteristic time-varying patterns are visible for each action.

5.2 Results

Our iHMM-kNN algorithm is empirically tested against the iHMM-Lik algorithm and the DTW on the seven datasets described in the previous section. Five runs of ten-fold cross validation are considered for KTH and Weizmann. A single run with fixed test and training set is considered instead for AUSLAN and JAPVOW. We implemented in Matlab both iHMM-kNN and iHMM-Lik.[6] Regarding DTW, the Mathworks implementation for Matlab has been used.

Our classification algorithm has only two parameters to be specified: the integer value of k in the k-NN and the real parameter s of the IDM as in Eq. (6).[7] We choose $k = 1$ because higher values could make the classifier too indeterminate. As reported in the second column of Table 2, small values are used also for s. The remaining columns of that table report the determinacies and average output size of both our algorithm and iHMM-Lik (with the same value of s). As a comment, with the selected values of s, either the determinacy is high or the average output size is consistently lower than the number of class labels. For AUSLAN, in particular, despite the very high number of classes the classifier is mostly determinate and, if not, much fewer than the original 95 classes are returned. When compared to iHMM-Lik, iHMM-kNN is less determinate and its average output size smaller. This can be explained by the high dimensionality of the feature space.

Tables 3 and 4 report information about accuracy. Results in Table 3 about single and set accuracy clearly report a higher performance of iHMM-kNN when compared to iHMM-Lik.

As noted in Section 4.5, the interval $[u_{.65}, u_{.80}]$ pro-

Dataset	s	iHMM-kNN		iHMM-Lik	
		det	*out*	*det*	*out*
KTH$_1$	.5	.311	2.85	.700	2.28
KTH$_2$	.5	.055	3.96	.565	2.13
KTH$_3$	.5	.135	2.91	.820	2.00
KTH$_4$	.5	.040	3.31	.600	2.42
KTH	.5	.111	3.51	.601	2.28
Weizmann	.5	.053	4.00	.766	2.00
AUSLAN	.01	.749	6.77	.935	2.37
JAPVOW	.01	.968	2.00	.965	2.15

Table 2: Determinacies and average output sizes for the benchmark datasets.

Dataset	iHMM-kNN		iHMM-Lik	
	sing-acc	*set-acc*	*sing-acc*	*set-acc*
KTH$_1$	.989	.990	.301	.017
KTH$_2$	.534	.981	.180	.384
KTH$_3$	.901	.972	.070	.083
KTH$_4$	.680	1.000	.269	.524
KTH	.883	.986	.299	.448
Weizmann	1.000	1.000	.275	.143
AUSLAN	.782	.675	.021	.062
JAPVOW	.958	.917	.283	.462

Table 3: Single and set accuracies on the benchmark.

[6] Both these tools are available as a free software at `http://ipg.idsia.ch/software`.

[7] Remember that the method described in [12] is used to fix the number M of states of the hidden variables. In our experiments this number ranges between 2 and 30.

vides a better summary of the credal classifiers performance by also allowing for a comparison with a traditional classifier like DTW. The results are in Table 4. Also this descriptor shows that iHMM-kNN clearly outperforms iHMM-Lik. This basically means that our interval-valued descriptor provides a better summary of a sequence rather than the interval-valued likelihood. Impressively, iHMM-kNN also competes with the DTW, showing both the quality of our approach and the (known) degradation of the DTW performance in the multiple-features case.

Dataset	iHMM-kNN		iHMM-Lik		DTW
	$u_{.65}$	$u_{.80}$	$u_{.65}$	$u_{.80}$	*acc*
KTH$_1$	**.659**	**.752**	.211	.212	.613
KTH$_2$	**.409**	**.517**	.201	.225	.369
KTH$_3$	**.550**	**.662**	.073	.076	.529
KTH$_4$	**.474**	**.597**	.281	.310	**.480**
KTH	**.495**	**.604**	.283	.309	**.525**
Weizmann	**.463**	**.575**	.236	.242	**.540**
AUSLAN	.680	.702	.021	.022	**.838**
JAPVOW	**.946**	**.951**	.283	.285	.697

Table 4: Accuracies for the benchmark datasets. Best performances are boldfaced.

Moreover, we already noted that iHMM-kNN has a precise counterpart obtained by setting $s = 0$ in the IDM constraints as in Eq. (6) and corresponding to the precise approach described in Section 2. This allows to check whether the classifier discriminates between "easy" instances (on which a single class is returned) and "difficult" ones. Results in Table 5 show that the precise single accuracy is larger than the precise set accuracy. KTH$_4$ is the only exception which can be explained by its low determinacy.

Dataset	*p-sing-acc*	*p-set-acc*	*acc*
KTH$_1$	.989	.787	.849
KTH$_2$	.534	.447	.451
KTH$_3$	.901	.671	.703
KTH$_4$	.680	.782	.779
KTH	.883	.674	.698
Weizmann	1.000	.842	.853
AUSLAN	.782	.351	.674
JAPVOW	.958	.333	.938

Table 5: Precise single and set accuracy of iHMM-kNN. The same classifier with $s = 0$ is used as a precise counterpart and its accuracy is in the last column. The values of *p-sing-acc* in this table coincide therefore with the *sing-acc* in Table 3.

As already discussed in Section 3.1, the adopted IDM-EM approach to the learning is the most critical part of the whole methodology. An alternative method, again heuristic and very naive, is therefore tested: LIN-VAC adopts a credal set corresponding to a *linear-vacuous mixture* [17] of the probability mass functions estimated by the EM.[8] The results of a comparison with this method for the Weizmann dataset are in Table 6. To determine the value of ϵ, we choose that leading to a determinacy comparable with that of IDM-EM. The $[u_{.65}, u_{.80}]$ intervals obtained in this way are overlapping, this suggesting the need of new, more sophisticated, models for this learning step.

Method	IDM-EM	LIN-VAC
parameter	$s = .5$	$\epsilon = .03$
det	.053	.054
out	4.00	4.38
$[u_{.65}, u_{.80}]$	$[.463, .575]$	$[.400, .504]$

Table 6: An alternative to the IDM-EM learning approach tested on the Weizmann dataset.

Finally, to validate our argument about the descriptor on the right-hand side of Eq. (5) being better than the sample mean, we compare the two descriptors in the precise case over datasets with different time lengths. When coping with short sequences the difference is in favor of our method (+2% on JAPVOW, +5% KTH$_2$) while the gap disappear with longer sequences (e.g., $-.4$% on Weizmann). This remark makes our method especially suited for the classification of short sequences.

6 Conclusions and outlooks

A new credal classifier for temporal data has been presented. Imprecise HMMs are learned from each sequence, and described as hyperbox in the feature space. These data are finally classified by a generalization of the k-NN approach. The results are promising: the algorithm outperforms another credal classifier proposed for this task and competes with the state-of-the-art method DTW. As a future work, we want to investigate novel, more reliable, learning techniques like for instance the likelihood-based approach already considered for complete data in [2]. Also more complex topologies should be considered.

[8] Given a mass function $P_0(X)$, its linear-vacuous mixture is a credal set $K(X)$ defined by the constraints $(1-\epsilon)P_0(x) \leq P(x) \leq (1-\epsilon)P_0(x)+\epsilon$. This corresponds to the vacuous credal set for $\epsilon = 1$ and to the original mass function for $\epsilon = 0$.

A Computation of the stationary credal set

Given an imprecise Markov chain as in Section 2, for each $\mathcal{X}' \subseteq \mathcal{X}$, define $Q_{\mathcal{X}'} : \mathcal{X} \to \mathbb{R}$, such that, $\forall x \in \mathcal{X}$:

$$\overline{Q}_{\mathcal{X}'}(x) := \min \left\{ \sum_{x \in \mathcal{X}'} \overline{P}(x'|x), 1 - \sum_{x \in \mathcal{X} \setminus \mathcal{X}'} \underline{P}(x'|x) \right\}. \tag{18}$$

Given this function, $\forall g : \mathcal{X} \to \mathbb{R}$, define $\overline{R}_g : \mathcal{X} \to \mathbb{R}$, such that:

$$\overline{R}_g(x) := \underline{g} + \int_{\underline{g}}^{\overline{g}} \overline{Q}_{\{x' \in \mathcal{X} : g(x') \geq t\}}(x) \mathrm{d}t, \tag{19}$$

for each $x \in \mathcal{X}$, with $\underline{g} := \min_{x \in \mathcal{X}} g(x)$ and $\overline{g} := \max_{x \in \mathcal{X}} g(x)$. Proceed similarly for the unconditional probability of the first hidden variable. In this way the following numbers (instead of functions) are defined:

$$\overline{Q}^0_{\mathcal{X}'} := \min \left\{ \sum_{x \in \mathcal{X}'} \overline{P}(x'), 1 - \sum_{x \in \mathcal{X}'} \underline{P}(x') \right\}. \tag{20}$$

$$\overline{R}^0_g := \underline{g} + \int_{\underline{g}}^{\overline{g}} \overline{Q}^0_{\{x' \in \mathcal{X} : g(x') \geq t\}} \mathrm{d}t. \tag{21}$$

A "lower" version of these functions and numbers can be obtained by simply replacing the lower probabilities with the uppers, maxima with the minima, and vice versa. For each $i = 1, \ldots, n$ let $h_i : \mathcal{X} \to \mathbb{R}$. To characterize the stationary credal set $\tilde{K}(X)$, consider $\overline{P}^*(x') := \max_{P(X) \in \tilde{K}(X)} P(x')$. Given the recursion:

$$h_{j+1}(x) := \overline{R}_{h_j}(x), \tag{22}$$

with initialization $h_1 := I_{x'}$[9], we obtain:

$$\overline{P}^*(x') := \lim_{n \to \infty} \overline{R}^0_{h_n}, \tag{23}$$

and similarly for the upper.

References

[1] A. Antonucci. An interval-valued dissimilarity measure for belief functions based on credal semantics. In T. Denoeux and Masson M.H., editors, *Belief Functions: Theory and Applications - Proceedings of the 2nd International Conference on Belief Functions*, volume 164 of *Advances in Soft Computing*, pages 37–44. Springer, 2012.

[2] A. Antonucci, M. Cattaneo, and G. Corani. Likelihood-based naive credal classifier. In *ISIPTA '11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, pages 21–30. SIPTA, 2011.

[3] A. Antonucci, R. de Rosa, and A. Giusti. Action recognition by imprecise hidden markov models. In *Proceedings of the 2011 International Conference on Image Processing, Computer Vision and Pattern Recognition, IPCV 2011*, pages 474–478. CSREA Press, 2011.

[4] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[5] G. de Cooman, F. Hermans, A. Antonucci, and M. Zaffalon. Epistemic irrelevance in credal networks: the case of imprecise Markov trees. *International Journal of Approximate Reasoning*, 51(9):1029–1052, 2010.

[6] G. de Cooman, F. Hermans, and E. Quaeghebeur. Sensitivity analysis for finite markov chains in discrete time. In *Uncertainty in Artificial Intelligence: Proceedings of the Twenty-Fourth Conference*, pages 129–136, 2008.

[7] T. Denoeux. Maximum likelihood from evidential data: an extension of the EM algorithm. In C. et al. Borgelt, editor, *Combining soft computing and statistical methods in data analysis (Proceedings of SMPS 2010)*, Advances in Intelligent and Soft Computing, pages 181–188. Springer, 2010.

[8] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.

[9] J.K. Kies. *Empirical Methods for Evaluating Video-Mediated Collaborative Work.* PhD thesis, Virginia Tech, March 1997.

[10] J. Toyama M. Kudo and M. Shimbo. Multidimensional curve classification using passing-through regions. *Pattern Recognition Letters*, 20:1103–1111, 1999.

[11] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Proc. of International Conference on Pattern Recognition*, 2004.

[9] For each $x' \in \mathcal{X}$, $I_{x'}$ is the *indicator function* of x', i.e., a function $\mathcal{X} \to \mathbb{R}$ such that $I_{x'}(x)$ is equal to one if $x = x'$ and zero otherwise.

[12] T. Shi, M. Belkin, and B. Yu. Data spectroscopy: Eigenspaces of convolution operators and clustering. *The Annals of Statistics*, 37(6B):3960–3984, 2009.

[13] P. Smyth. Clustering sequences with hidden Markov models. In *Advances in Neural Information Processing Systems*, pages 648–654. MIT Press, 1997.

[14] G. A. Ten Holt, M. J. T. Reinders, and E.A. Hendriks. Multi-dimensional dynamic time warping for gesture recognition. *Time*, 5249:23–32, 2007.

[15] L.V. Utkin and F.P.A. Coolen. Interval-valued regression and classification models in the framework of machine learning. In *ISIPTA '11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, pages 371–380. SIPTA, 2011.

[16] A. Van Camp and G. de Cooman. A new method for learning imprecise hidden markov model. In S. Greco, B. Bouchon-Meunier, G. Coletti, B. Matarazzo, and R.R. Yager, editors, *Communications in Computer and Information Science*, volume 299, pages 460–469. Springer, 2012.

[17] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, 1991.

[18] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58(1):3–34, 1996.

[19] M. Zaffalon. The naive credal classifier. *J. Stat. Plann. Inference*, 105(1):5–21, 2002.

[20] M. Zaffalon, G. Corani, and D.D. Mauá. Utility-based accuracy measures to empirically evaluate credal classifiers. In *ISIPTA '11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, pages 401–410. SIPTA, 2011.

The description of extreme 2-monotone measures

Andrey G. Bronevich
National Research University
"Higher School of Economics"
Moscow, Russia
brone@mail.ru

Igor N. Rozenberg
JSC "Research, Development and Planning
Institute for Railway Information Technology,
Automation and Telecommunication"
Moscow, Russia
I.Rozenberg.gismps.ru

Abstract

The paper is devoted to the description of extreme points in the set of 2-monotone measures. We describe them using lattices on which an extreme 2-monotone measure is additive. We also propose the way of generation extreme monotone measures based on the aggregation of extreme measures with the help of multilinear extension. We describe also the class of extreme 2-monotone measures that are additive on the filter on which a 2-monotone measure has positive values.

Keywords. 2-monotone measures, extreme points, additivity on lattices, filters, partially ordered sets, multilinear extension.

1 Introduction

2-monotone measures play an important role in the theory of imprecise probabilities [17], because for imprecise probabilities represented by 2-monotone measures it is possible to find analytical solutions for many problems and, therefore, such models are more attractive in a computational point of view. Meanwhile, some unsolved problems concerning 2-monotone measures can be solved [5] if we know the structure of extreme points of the set of all 2-monotone measures. It is worth to mention that finding description of extreme points of a convex set is usually a hard problem. This problem is solved for the set of all monotone measures [13,15], *p*-symmetrical measures [7,8], but for some convex families, e.g. k-additive measures [7], is far from the final solution.

The aim of this paper is to make one step forward in this direction, providing some general necessary and sufficient conditions that a 2-monotone measure is an extreme point and giving descriptions of some special families of them.

The paper has the following structure. We remind first some results concerning monotone measures and criteria of 2-monotonicity. After that we provide general necessary and sufficient conditions that a 2-monotone measure is an extreme point through lattices on which it is additive. After that we remind the multilinear extension of monotone measures and using it we define the composition of monotone measures. We show that the composition of extreme 2-monotone measures is an extreme 2-monotone measure again. The paper is ended by describing a special class of 2-monotone measures which are additive on the filter of sets on which a 2-monotone measure has positive values.

2 Monotone measures

Let X be a finite set and let $\mu : 2^X \to [0,1]$ be a set function on the powerset 2^X. Then μ is called a *monotone measure* [9] if the following conditions hold:

1) $\mu(\varnothing) = 0$ and $\mu(X) = 1$;

2) $A \subseteq B$ for $A, B \in 2^X$ implies $\mu(A) \leq \mu(B)$.

Let us denote the set of all monotone measures on 2^X by $M_{mon}(X)$ or briefly M_{mon} if the set X is clearly defined from the context. For monotone measures $\mu_1, \mu_2 \in M_{mon}$ we define their convex sum as $\mu(A) = a\mu_1(A) + (1-a)\mu_2(A)$, where $a \in [0,1]$ and $A \in 2^X$. Clearly, $\mu \in M_{mon}$, i.e. the set M_{mon} is convex and it is possible to show [13,15] that extreme points of M_{mon} are $\{0,1\}$-valued monotone measures, i.e. monotone measures with values in $\{0,1\}$. Let the algebra 2^X be considered as a partially ordered set w.r.t. inclusion of sets. By definition, a filter $\mathbf{f}$ in 2^X is a nonempty subset of 2^X such that $A \in \mathbf{f}$, $A \subseteq B$ implies $B \in \mathbf{f}$. Any filter can be uniquely defined by the set of its minimal elements $\{A_1, ..., A_m\}$. This fact is denoted by $\mathbf{f} = \langle A_1, ..., A_m \rangle$. The connection between filters of algebra 2^X and $\{0,1\}$-valued monotone measures is shown in the following lemma [13].

Lemma 1. *Any* $\{0,1\}$*-valued monotone measure* η *defines a filter* $\mathbf{f} = \{A \in 2^X \mid \eta(A) > 0\}$ *such that* $\varnothing \notin \mathbf{f}$. *Conversely, any filter* $\mathbf{f}$ *with* $\varnothing \notin \mathbf{f}$ *defines a* $\{0,1\}$*-valued monotone measure* η *by*

$$\eta(A) = \begin{cases} 1, & A \in \mathbf{f}, \\ 0, & A \notin \mathbf{f}. \end{cases} \quad (1)$$

In the sequel we denote a $\{0,1\}$-valued measure as $\eta_{\mathbf{f}}$ if it corresponds to a filter $\mathbf{f}$.

Remark 1. Clearly, a set $\mathbf{f}(t) = \{A \in 2^X \mid \mu(A) > t\}$ for any given $\mu \in M_{mon}$ and $t \in [0,1)$ is a filter in algebra

2^X, moreover, $\mu(A)=\int\limits_0^1 \eta_{\mathbf{f}(t)}(A)dt$ and if $\left\{t_1,t_2,...,t_k\right\}$ is the set of all values of μ and $0=t_1<t_2<...<t_k=1$, then $\mu=\sum\limits_{i=1}^{k-1}(t_{i+1}-t_i)\eta_{\mathbf{f}(t_i)}$.

3 2-monotone measures

A monotone measure μ is called *2-monotone* [9] if the following inequality

$$\mu(A)+\mu(B)\le\mu(A\cap B)+\mu(A\cup B) \qquad (2)$$

is fulfilled for any $A,B\in 2^X$. We denote the set of all 2-monotone measures on the algebra 2^X by $M_{2-mon}(X)$. The condition (2) can be simplified [3]. It is sufficient to check inequalities of the following type:

$$\mu(A\cup\{x_i\})+\mu(A\cup\{x_j\})\le\mu(A)+\mu(A\cup\{x_i,x_j\}) \text{ , (3)}$$

for all $A\in 2^X$ and $x_i,x_j\in 2^X$ such that $\left|X\setminus A\right|\ge 2$, $x_i,x_j\in X\setminus A$ and $x_i\ne x_j$.

In the next we can also consider nonnegative set functions μ on 2^X with $\mu(\varnothing)=0$. Such set functions are called 2-monotone if they are monotone and inequalities (2) or equivalently inequalities (3) are fulfilled. The next proposition shows that the monotonicity of μ is not necessary to check.

Proposition 1. Let μ *be a nonnegative set function on* 2^X *with* $\mu(\varnothing)=0$. *Then it is 2-monotone iff inequalities* (3) *are fulfilled* for all $A\in 2^X$ and $x_i,x_j\in 2^X$ such that $\left|X\setminus A\right|\ge 2$, $x_i,x_j\in X\setminus A$ and $x_i\ne x_j$.

Let us consider how Proposition 1 can be strengthened if we know that the sets on which a nonnegative set function is positive, form a filter. In this case we say that μ is *2-monotone on the filter* $\mathbf{f}=\left\{A\in 2^X \mid \mu(A)>0\right\}$, if inequalities (3) are fulfilled, when $A\cup\{x_i\}\in\mathbf{f}$ and $A\cup\{x_j\}\in\mathbf{f}$. In addition, we say that a set function μ, which is 2-monotone on the filter $\mathbf{f}$, is also *2-monotone on its borders* if inequalities (3) are fulfilled if at least $A\cup\{x_i\}\in\mathbf{f}$ or $A\cup\{x_j\}\in\mathbf{f}$. Next proposition is the direct consequence of Proposition 1.

Proposition 2. *Given a nonnegative set function* μ *such that* $\mathbf{f}=\left\{A\in 2^X \mid \mu(A)>0\right\}$ *is a filter. Then* μ *is 2-monotone iff it is 2-monotone on the filter* $\mathbf{f}$ *and its borders.*

In some cases the 2-monotonicity on the filter can imply the 2-monotonicity on its borders. The description of such a case is given in the following proposition.

Proposition 3. *Let a nonnegative set function be 2-monotone on the filter* $\mathbf{f}\supseteq\left\{A\in 2^X \mid \mu(A)>0\right\}$ *and let* $\mathfrak{C}=\{C_1,...,C_m\}$ *be the set of its minimal elements. Then* μ *is 2-monotone on borders of* $\mathbf{f}$, *if for every* $C_k\in\mathfrak{C}$ *and every* $x_i\notin C_k$ *there exists a* $C_l\in\mathfrak{C}$, *such that* $\{x_i\}=C_l\setminus C_k$.

4 Additivity properties of 2-monotone measures on lattices

We denote by $M_{pr}(X)$ the set of all probability measures on the algebra 2^X . Let $\mu\in M_{2-mon}$, then the *core* of μ is the set of probability measures defined by $core(\mu)=\left\{P\in M_{pr} \mid P\ge\mu\right\}$. It is well known [16] that $core(\mu)$ is a nonempty convex set for any $\mu\in M_{2-mon}$ and its extreme points are probability measures P_γ , where $\gamma:\{1,2,...,n\}\to\{1,2,...,n\}$ is a permutation of the set $\{1,2,...,n\}$ and any P_γ is constructed with the help of the chain of sets $B_1=\{x_{\gamma(1)}\}$, $B_2=\{x_{\gamma(1)},x_{\gamma(2)}\}$, …, $B_n=\{x_{\gamma(1)},...,x_{\gamma(n)}\}$ by the rule: $P_\gamma(B_i)=\mu(B_i)$, $i=1,...,n$. Let us remind the result from [2,4], that can be also found in [10].

Proposition 4. *Let* $\mu\in M_{2-mon}$, *then the system of sets* $\mathcal{L}_\gamma(\mu)=\{A\in 2^X \mid \mu(A)=P_\gamma(A)\}$ *is a lattice w.r.t. operations* $\cap$ *and* $\cup$, *i.e.* $A,B\in\mathcal{L}_\gamma(\mu)$ *implies* $A\cap B,A\cup B\in\mathcal{L}_\gamma(\mu)$ *and it is a maximal lattice, on which* μ *is additive.*

Remark 2. Additivity of μ on $\mathcal{L}_\gamma(\mu)$ means that if $A,B\in\mathcal{L}_\gamma(\mu)$, then

$$\mu(A)+\mu(B)=\mu(A\cap B)+\mu(A\cup B) .$$

As we will see in the next such maximal lattices play an important role for the extreme points description of 2-monotone measures. Therefore, we also present here some results showing the connections between such lattices and partially ordered sets.

Let us assume that a maximal lattice $\mathfrak{L}$, on which a 2-monotone measure μ is additive, contains maximal chains described by a set of permutations $\Gamma=\left\{\gamma_i\right\}$. We put into correspondence the linear order ρ_γ on X to each permutation $\gamma\in\Gamma$ in a way that $x_i\rho_\gamma x_j$ if $\gamma(i)\le\gamma(j)$. Then the following theorem is valid.

Theorem 1. *Let* $\mathfrak{L}$ *be a maximal lattice, on which a 2-monotone measure* μ *is additive, and let the maximal chains in* $\mathfrak{L}$ *be described by a a set of permutations* $\Gamma=\left\{\gamma_i\right\}$. *Consider the partial order* $\rho_\Gamma=\bigcap\limits_{\gamma\in\Gamma}\rho_\gamma$ [1], *where linear orders* ρ_γ *are defined as above. Then* $\left\{\rho_\gamma\right\}_{\gamma\in\Gamma}$ *is the set of all linear orders satisfying* $\rho_\gamma\supseteq\rho_\Gamma$.

[1] Here is used the usual intersection of relations, i.e. if $\rho_1,\rho_2\in\{1,...,n\}\times\{1,...,n\}$, then $\rho_1\cap\rho_2$ is the usual intersection defined for sets.

Next result can be considered as a corollary of more general result that can be found in [10].

Theorem 2. *Let ρ be a partial order on X and let $\{\rho_\gamma\}_{\gamma\in\Gamma}$ be the set of all linear extensions ρ, i.e. each ρ_γ is a linear order and $\rho_\gamma \supseteq \rho$. Then any ρ_γ, $\gamma\in\Gamma$, induces a chain of sets $B_0=\varnothing$, $B_1=\{x_{\gamma(1)}\}$, $B_2=\{x_{\gamma(1)},x_{\gamma(2)}\}$, ..., $B_n=\{x_{\gamma(1)},x_{\gamma(2)},...,x_{\gamma(n)}\}$ and the union of all such chains is a lattice of sets w.r.t. union and intersection.*

5 The description of extreme 2-monotone measures through lattices

Proposition 5. *Let us consider the set of all maximal lattices $\mathfrak{L}_\gamma(\mu)$, on which a 2-monotone measure μ is additive, and let $\mathbf{f}_\mu=\{A\in 2^X \mid \mu(A)>0\}$. Then μ is not an extreme point in M_{2-mon} iff there exists a 2-monotone measure ν ($\nu\neq\mu$) such that*

1) $\mathfrak{L}_\gamma(\mu)\subseteq\mathfrak{L}_\gamma(\nu)$ for any permutation γ;

2) $\mathbf{f}_\nu\subseteq\mathbf{f}_\mu$.

Corollary 1. *Let μ be an extreme point in M_{2-mon}. Then the filter $\mathbf{f}_\mu$ and the system of lattices $L_\gamma(\mu)$ define μ uniquely.*

6 Multilinear extension and composition of monotone measures

In this section we will use the notion of pseudo-Boolean functions [12]. Any pseudo-Boolean function is a mapping $\varphi:\{0,1\}^n\to\mathbb{R}$. For our purpose, it is sufficient to consider pseudo-Boolean functions taking their values in $[0,1]$, i.e. we assume that $\varphi:\{0,1\}^n\to[0,1]$. It is easy to see that there is a one-to-one correspondence between pseudo-Boolean functions and set functions. For this purpose, we consider set functions defined on the algebra 2^Z, where $Z=\{1,...,n\}$, and consider vectors $\mathbf{1}_A=(x_1,...,x_n)$, where $A\in 2^Z$ and $x_i=1$ if $i\in A$ and $x_i=0$ otherwise. Then obviously $\mu(A)=\varphi(\mathbf{1}_A)$, where $A\in 2^Z$ is a set function on 2^Z. If we consider the class of monotone pseudo-Boolean functions $\varphi:\{0,1\}^n\to[0,1]$ with $\varphi(\mathbf{0})=0$ and $\varphi(\mathbf{1})=1$, where $\mathbf{0}=(0,...,0)$ and $\mathbf{1}=(1,...,1)$, then it corresponds to the class of monotone measures on 2^Z.

Any pseudo-Boolean function can be uniquely represented as a multilinear polynomial [14] as

$$\varphi(\mathbf{x})=\sum_{A\in 2^Z} m(A)\prod_{i\in A} x_i\,, \qquad (4)$$

where m is the Möbius transform m of the set function $\mu(A)=\varphi(\mathbf{1}_A)$, defined by

$$m(A)=\sum_{B\subseteq A}(-1)^{|A\backslash B|}\mu(B)\,.$$

We see that there is a one-to-one correspondence between multilinear polynomials and pseudo-Boolean functions. In addition, we can assume that the vector $\mathbf{x}$ in formula (4) can take values in $[0,1]^n$. In this case, the function $\tilde{\varphi}:[0,1]^n\to[0,1]$ is called [14] the *multilinear extension* of φ.

The next proposition [3] shows how to check monotonicity and 2-monotonicity of a set function using its multilinear extension.

Proposition 6. *Let $\mu:2^Z\to[0,1]$ and let φ be its corresponding pseudo-Boolean function. Then μ is a monotone measure iff the multilinear extension $\tilde{\varphi}$ of φ has the following properties:*

1) $\tilde{\varphi}(\mathbf{0})=0$ and $\tilde{\varphi}(\mathbf{1})=1$;

2) $\dfrac{\partial\tilde{\varphi}(\mathbf{x})}{\partial x_i}\geq 0$ for any x_i and at any point $\mathbf{x}\in[0,1]^n$.

In addition, μ is 2-monotone iff

3) $\dfrac{\partial^2\tilde{\varphi}(\mathbf{x})}{\partial x_i\partial x_j}\geq 0$ for any x_i,x_j and at any point $\mathbf{x}\in[0,1]^n$.

Proposition 6 shows that the multilinear extension of a monotone measure is an aggregation function. Let us remind that, by definition [11], an aggregation function $\tilde{\varphi}$ is a mapping $\tilde{\varphi}:[0,1]^n\to[0,1]$ such that

1) $\tilde{\varphi}(\mathbf{0})=0$ and $\tilde{\varphi}(\mathbf{1})=1$;

3) $\tilde{\varphi}(\mathbf{x})\leq\tilde{\varphi}(\mathbf{y})$ for $\mathbf{x},\mathbf{y}\in[0,1]^n$ if $\mathbf{x}\leq\mathbf{y}$ ($\mathbf{x}\leq\mathbf{y}$ means for $\mathbf{x}=(x_1,...,x_n)$ and $\mathbf{y}=(y_1,...,y_n)$ that $x_i\leq y_i$, $i=1,...,n$).

We can generate monotone measures using aggregation functions as follows. Let $\tilde{\varphi}:[0,1]^n\to[0,1]$ be an aggregation function and $X_1,...,X_n$ be mutually disjoint finite nonempty sets and let μ_i, $i=1,..,n$, be monotone measures on 2^{X_i}. Then a set function μ on 2^X, where $X=X_1\cup...\cup X_n$, defined by

$$\mu(A)=\tilde{\varphi}\left(\mu_1(A\cap X_1),...,\mu_n(A\cap X_n)\right),\ A\in 2^X, \qquad (5)$$

is also a monotone measure. For the measure μ, defined by formula (5), we will use the notation $\mu=\tilde{\varphi}\circ\boldsymbol{\mu}$, where $\boldsymbol{\mu}=(\mu_1,...,\mu_n)$.

In this section, we will use multilinear polynomials as aggregation functions. It can be shown [3] that if φ is a multilinear extension of a 2-monotone measure and μ_i, $i=1,...,n$, be 2-monotone measures on 2^{X_i}, then $\mu=\varphi\circ\boldsymbol{\mu}$ is also a 2-monotone measure.

Obviously, we can introduce the same representation like (5) for pseudo-Boolean functions. Let $\tilde{\varphi}:[0,1]^n\to[0,1]$ be an aggregation function and let $\mu_i(\mathbf{x}^{(i)})$, $i=1,...,n$, be pseudo-Boolean functions. Then the aggregation of these functions is defined as

$$\mu(\mathbf{x})=\tilde{\varphi}\left(\mu_1(\mathbf{x}^{(1)}),...,\mu_n(\mathbf{x}^{(n)})\right), \text{ where } \mathbf{x}=\left(\mathbf{x}^{(1)},...,\mathbf{x}^{(n)}\right).$$

Proposition 7. *Let $\tilde{\varphi}:[0,1]^n\to\mathbb{R}$ be the multilinear extension of a pseudo-Boolean function φ and let*

$\mu_i(\mathbf{x}^{(i)})$, $i=1,...,n$, be pseudo-Boolean functions. Let us consider the pseudo-Boolean function $\mu(\mathbf{x})=\tilde{\varphi}\left(\mu_1(\mathbf{x}^{(1)}),...,\mu_n(\mathbf{x}^{(n)})\right)$, where $\mathbf{x}=\left(\mathbf{x}^{(1)},...,\mathbf{x}^{(n)}\right)$. Then the multilinear extension of μ can be computed as

$$\tilde{\mu}(\mathbf{x})=\tilde{\varphi}\left(\tilde{\mu}_1(\mathbf{x}^{(1)}),...,\tilde{\mu}_n(\mathbf{x}^{(n)})\right).$$

Remark 3. Proposition 7 allows us to represent the aggregation (5) using more simple aggregations as follows. Let $\tilde{\varphi}\left(t_1,...,t_n\right)$ be a multilinear extension of a pseudo-Boolean function $\varphi:\{0,1\}^n\to[0,1]$. Then we can consider the following sequence of pseudo-Boolean functions

$$\varphi_0=\tilde{\varphi}\left(t_1,...,t_n\right),\ \varphi_1=\tilde{\varphi}\left(\mu_1(\mathbf{x}^{(1)}),t_2,...,t_n\right),$$
$$\varphi_2=\tilde{\varphi}\left(\mu_1(\mathbf{x}^{(1)}),\mu_2(\mathbf{x}^{(2)}),t_3,...,t_n\right),$$
$$\varphi_n(\mathbf{x})=\tilde{\varphi}\left(\mu_1(\mathbf{x}^{(1)}),...,\mu_n(\mathbf{x}^{(n)})\right),$$

where $t_i\in\{0,1\}$, $i=1,...,n$, and corresponding aggregation functions:

$$\tilde{\varphi}_0=\tilde{\varphi}\left(t_1,...,t_n\right),\ \tilde{\varphi}_1=\tilde{\varphi}\left(\tilde{\mu}_1(\mathbf{x}^{(1)}),t_2,...,t_n\right),$$
$$\tilde{\varphi}_2=\tilde{\varphi}\left(\tilde{\mu}_1(\mathbf{x}^{(1)}),\tilde{\mu}_2(\mathbf{x}^{(2)}),t_3,...,t_n\right),$$
$$\tilde{\varphi}_n(\mathbf{x})=\tilde{\varphi}\left(\tilde{\mu}_1(\mathbf{x}^{(1)}),...,\tilde{\mu}_n(\mathbf{x}^{(n)})\right),$$

that have to be obviously multilinear extensions of corresponding pseudo-Boolean functions. Each φ_i is generated from φ_{i-1} by replacing variable t_i with the pseudo-Boolean function $\mu_i(\mathbf{x}^{(i)})$.

The interpretation of simple aggregations, considered in Remark 3, through set functions is given in the following lemma.

Lemma 2. *Let $\varphi_1:\{0,1\}^n\to[0,1]$ and $\varphi_2:\{0,1\}^m\to[0,1]$ be pseudo-Boolean functions and let $\tilde{\varphi}_i$, $i=1,2$, be their multilinear extensions. Consider their aggregation of the following type:*

$$\varphi(x_1,...,x_{n-1},x_{n+1},....,x_{n+m})=\tilde{\varphi}_1(x_1,...,x_{n-1},\varphi_2(x_{n+1},....,x_{n+m})),$$

and corresponding set functions on 2^Z, where $Z=\{1,...,n+m\}$:

$$\mu_1(A)=\varphi_1(\mathbf{1}_A),\ \textit{where}\ A\subseteq\{1,2,...,n\};$$
$$\mu_2(B)=\varphi_2(\mathbf{1}_B),\ \textit{where}\ B\subseteq\{n+1,...,n+m\};$$
$$\mu(C)=\varphi(\mathbf{1}_C),\ \textit{where}\ C\subseteq\{1,...,n-1,n+1,...,n+m\}.$$

Then

$$\mu(A\cup B)=\mu_1(A)+(\mu_1(A\cup\{n\})-\mu_1(A))\mu_2(B),$$

where $A\subseteq\{1,...,n-1\}$ and $B\subseteq\{n+1,...,n+m\}$.

Like in the theory of Boolean functions, let us introduce the notion of essential variable for pseudo-Boolean functions. Let $\varphi:\{0,1\}^n\to[0,1]$ be a pseudo-Boolean function. The variable x_i is called *essential* for φ if there are vectors $\mathbf{x}_1=\left(x_1,...,x_{i-1},0,x_{i+1},...,x_n\right)$ and $\mathbf{x}_2=\left(x_1,...,x_{i-1},1,x_{i+1},...,x_n\right)$ in $\{0,1\}^n$ such that $\varphi(\mathbf{x}_1)\neq\varphi(\mathbf{x}_2)$. It is easy to express such a property using set functions. Let $\mu(A)=\varphi(\mathbf{1}_A)$, where $A\in 2^Z$. Then the variable x_i is essential if the set function $\nu(A)=\mu(A\cup\{i\})-\mu(A)$, where $A\in 2^Z$, is not identical to zero.

Proposition 8. *Let $\varphi:\{0,1\}^n\to[0,1]$ be a pseudo-Boolean function and let $\tilde{\varphi}:[0,1]^n\to[0,1]$ be its multilinear extension. Then the variable x_i is essential for φ iff there is a $\mathbf{x}\in[0,1]^n$ such that $\dfrac{\partial\tilde{\varphi}(\mathbf{x})}{\partial x_i}\neq 0$.*

Proposition 9. Let *$\mu=\tilde{\varphi}\circ\boldsymbol{\mu}$ be the aggregation defined by formula (5), and let $\tilde{\varphi}:[0,1]^n\to[0,1]$ be a multilinear extension of a monotone measure φ. Then representation $\mu=\tilde{\varphi}\circ\boldsymbol{\mu}$ for fixed sets $X_1,...,X_n$ is defined uniquely iff each variable in φ is essential.*

In this section we will prove the following result.

Theorem 3. *Let $\tilde{\varphi}:[0,1]^n\to[0,1]$ be a multilinear extension of a 2-monotone measure φ on 2^Z and let all variables of $\tilde{\varphi}$ be essential. Let us assume that μ_i are 2-monotone measures on 2^{X_i}, where $X_1,...,X_n$ are mutually disjoint finite nonempty sets. Then $\mu=\tilde{\varphi}\circ\boldsymbol{\mu}$ is an extreme point iff 2-monotone measures φ, $\mu_1,...,\mu_n$ are extreme points too.*

7 Examples of extreme 2-monotone measures

Let μ be an extreme 2-monotone measure. Then we call it *perfect* if it is uniquely defined by a filter $\mathbf{f}=\left\{A\in 2^X\,|\,\mu(A)>0\right\}$, in other words, an extreme measure is not perfect if there is another extreme 2-monotone measure with the same filter $\mathbf{f}$, on which it has positive values. We will describe next the class of such extreme 2-monotone measures.

Let μ be a set function on 2^X. We say that μ is *additive on a filter* $\mathbf{f}$ if

a) $\mu(A)=0$ for any $A\notin\mathbf{f}$;

b) $\mu(A\cup\{x_i\})+\mu(A\cup\{x_j\})=\mu(A)+\mu(A\cup\{x_i,x_j\})$ for any sets $A\cup\{x_i\},A\cup\{x_j\}\in\mathbf{f}$ such that $x_i,x_j\notin A$ and $x_i\neq x_j$.

Lemma 3. *Let a set function μ be additive on a filter $\mathbf{f}$. Consider any $A\in\mathbf{f}$ and $x_i\notin A$. Then $\mu(A\cup\{x_i\})-\mu(A)=\mu(C\cup\{x_i\})-\mu(C)$ for any $C\in\mathbf{f}$ with $C\subseteq A$.*

Corollary 2. *If the set function μ is additive on a filter $\mathbf{f}$, then $\mu(A\cup\{x_i\})-\mu(A)=\mu(C\cup\{x_i\})-\mu(C)$ for any $A,C\in\mathbf{f}$ such that $A,C\subseteq X\setminus\{x_i\}$.*

The results formulated in Lemma 3 and Corollary 2 can be better described by the function

$$\nu(x_i)=\mu(A\cup\{x_i\})-\mu(A),$$

where $A \in \mathbf{f}$ and $x_i \notin A$. Let us notice that $\nu(x_i)$ does not depend on the choice of A. The value $\nu(x_i)$ is called *the weight* of x_i on filter $\mathbf{f}$ for a set function μ.

Proposition 10. *Let a nonnegative set function μ be additive on a filter $\mathbf{f}$, and $\nu(x_i) \geq 0$ for all $x_i \in X$. Then μ is 2-monotone.*

Proposition 11. *Let a nonnegative set function μ be additive on a filter $\mathbf{f}$. Let us consider the system of sets $2^X \setminus \mathbf{f}$ and the set of its maximal elements $\{C_1,...,C_k\}$. Then μ is 2-monotone if $\{\bar{C}_1,...,\bar{C}_k\}$ is a covering of X.*

Let us consider how to construct 2-monotone measures that are additive on a filter. We prove first the following auxiliary lemma.

Lemma 4. *Let $\mathbf{f}$ be a filter of the algebra 2^X. Then the system of sets*

$$\mathbf{f}_0 = \{A \mid A \cup \{x_i\}, A \cup \{x_j\} \in \mathbf{f}, x_i, x_j \notin A\}$$

is also a filter and $\mathbf{f}_0 \supseteq \mathbf{f}$.

Proposition 12. *Let we use the notations from Lemma 4, $A \in \mathbf{f}_0$, $x_j \notin A$, and let a set function μ be additive on the filter $\mathbf{f}$. Then the value $\nu(x_j) = \mu(A \cup \{x_j\}) - \mu(A)$ does not depend on the choice of $A \in \mathbf{f}_0$.*

Corollary 3. *Let $A \in \mathbf{f}_0$, $B \supseteq A$, and let μ be additive on the filter $\mathbf{f}$. Then*

$$\mu(B) = \mu(A) + \sum_{x_i \in B \setminus A} \nu(x_i).$$

Corollary 4. *Let $A \cap B \in \mathbf{f}_0$ for sets A and B, and let μ be additive on the filter $\mathbf{f}$. Then*

$$\mu(A) + \mu(B) = \mu(A \cap B) + \mu(A \cup B).$$

Proposition 13. *Let a set function μ be additive on the filter $\mathbf{f}$. Then values of ν obeys the following system of equations*:

$$\sum_{x_i \notin B} \nu(x_i) = \mu(X) \text{ for all } B \in \mathbf{f}_0 \setminus \mathbf{f}, \qquad (6)$$

in addition

$$\mu(B) = \begin{cases} 0, & B \notin \mathbf{f}, \\ \mu(X) - \sum\limits_{x_i \notin B} \nu(x_i), & B \in \mathbf{f}, \end{cases} \qquad (7)$$

Conversely, each set function μ obeying equalities (6) *and* (7) *is additive on the filter $\mathbf{f}$.*

Remark 4. Solving equations (6) and (7) w.r.t. $\nu(x_i)$ we can find all set functions that are additive on the filter $\mathbf{f}$, i.e. it is guaranteed that any such function satisfies

1) $\mu(A \cup \{x_i\}) + \mu(A \cup \{x_j\}) = \mu(A) + \mu(A \cup \{x_i, x_j\})$ for $A \cup \{x_i\}, A \cup \{x_j\} \in \mathbf{f}$ and $A \cap \{x_i, x_j\} = \varnothing$;

2) $\mu(A) = 0$ for $A \notin \mathbf{f}$.

However, we can not guarantee that μ is 2-monotone, because 2-monotonicity of μ in this case is equivalent to $\nu(x_i) \geq 0$ for all $x_i \in X$ by Proposition 10.

Proposition 14. *Let μ be a 2-monotone measure that is additive on the filter $\mathbf{f} = \{A \in 2^X \mid \mu(A) > 0\}$. Then μ is an extreme 2-monotone measure if it is defined uniquely.*

Proposition 15. *Let the filter $\mathbf{f}$ obey the conditions formulated in Proposition 3. Then if an extreme 2-monotone measure μ, which is additive on $\mathbf{f} = \{A \in 2^X \mid \mu(A) > 0\}$, exists, then it is perfect.*

Proposition 16. *Let the filter $\mathbf{f}$ obey the conditions formulated in Proposition 11. Then if an extreme 2-monotone measure μ, which is additive on $\mathbf{f} = \{A \in 2^X \mid \mu(A) > 0\}$, exists, then it is defined uniquely.*

Let us consider examples of perfect 2-monotone measures that are additive on filter. A monotone measure is called *symmetrical* if its values depend only on the cardinality of the corresponding set. The next proposition gives the description of extreme symmetrical 2-monotone measures.

Proposition 17. *Let $X = \{x_1, x_2, ..., x_n\}$. Then any symmetrical monotone measure, defined by*

$$\mu_k(A) = \begin{cases} 0, & |A| < k-1, \\ (m-k+1)/(n-k+1), & |A| = m \geq k-1, \end{cases}$$

where $k = 2,...,n$, is a perfect extreme 2-monotone measure.

Remark 5. It is easy to show that the set of all symmetrical 2-monotone measures on 2^X, where $X = \{x_1, x_2, ..., x_n\}$, is convex and the extreme points of it are measures $\mu_k, k = 1,...,n$. Obviously, μ_1 is not an extreme point of M_{2-mon} if $n = 1$, because it is represented as $\mu_1 = (1/n)\sum_{k=1}^{n} \eta_{\langle\{x_k\}\rangle}$.

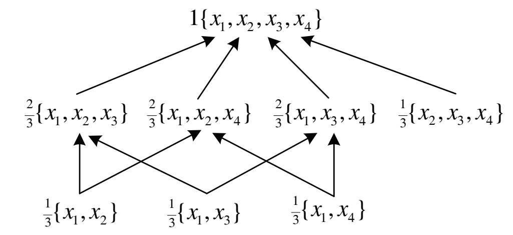

$1\{x_1, x_2, x_3, x_4\}$
$\frac{2}{3}\{x_1, x_2, x_3\}$
$\frac{2}{3}\{x_1, x_2, x_4\}$
$\frac{2}{3}\{x_1, x_3, x_4\}$
$\frac{1}{3}\{x_2, x_3, x_4\}$
$\frac{1}{3}\{x_1, x_2\}$
$\frac{1}{3}\{x_1, x_3\}$
$\frac{1}{3}\{x_1, x_4\}$

Figure. 1: A perfect extreme 2-monotone measure that is additive on the filter.

Remark 6. Let $X = \{x_1, x_2, x_3\}$, then the extreme points of $M_{2-mon}(X)$ are perfect and they are additive on the filter of their positive values. These measures are $\eta_{\langle A \rangle}$, where $|A| > 0$, and the symmetrical measure μ_2 for $n = 3$. If $X = \{x_1, x_2, x_3, x_4\}$, then extreme points of $M_{2-mon}(X)$ are not necessarily measures described in Proposition 17. For example, let us consider the filter $\mathbf{f} = \langle\{x_1, x_2\}, \{x_1, x_3\}, \{x_1, x_4\}, \{x_2, x_3, x_4\}\rangle$. Let us try to find a 2-monotone measure μ that is additive on

$\mathbf{f}=\{A\in 2^X \mid \mu(A)>0\}$. In this case, $\mathbf{f}_0=\langle\{x_1\},\{x_2,x_3\},\{x_2,x_4\},\{x_3,x_4\}\rangle$, and the function $\nu(x_i)$, $x_i\in X$, should obey the following linear system of equations:

$$\begin{cases}\nu(x_2)+\nu(x_3)+\nu(x_4)=1,\\ \nu(x_1)+\nu(x_2)=1,\\ \nu(x_1)+\nu(x_3)=1,\\ \nu(x_1)+\nu(x_4)=1,\end{cases}$$

that has the following unique solution $\nu(x_1)=2/3$, $\nu(x_2)=\nu(x_3)=\nu(x_4)=1/3$. After that we can calculate the values of 2-monotone measure μ by the formula (7). This measure is depicted on Figure 1. Using Proposition 15 it easy to check that μ is a perfect extreme 2-monotone measure.

Let us consider an example of extreme 2-monotone measure μ, depicted on Figure 2, that is not additive on the filter $\mathbf{f}=\{A\in 2^X \mid \mu(A)>0\}$. Using Corollary 1, it is easy to show that it is an extreme point of M_{2-mon}. In addition, it is possible to show that μ is a perfect extreme 2-monotone measure.

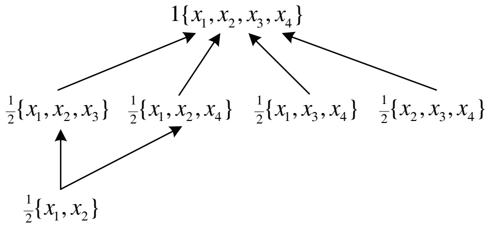

$1\{x_1,x_2,x_3,x_4\}$
$\frac{1}{2}\{x_1,x_2,x_3\}$
$\frac{1}{2}\{x_1,x_2,x_4\}$
$\frac{1}{2}\{x_1,x_3,x_4\}$
$\frac{1}{2}\{x_2,x_3,x_4\}$
$\frac{1}{2}\{x_1,x_2\}$

Figure. 2: A perfect extreme 2-monotone measure that is not additive on the filter.

It is easy to find extreme 2-monotone measures that are not perfect. Such measures are depicted on Figure 3, with parameters α,β,γ, and λ, given in Table 1.

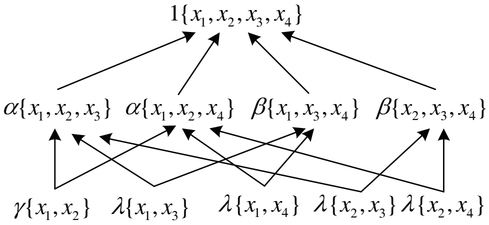


Figure 3: Extreme 2-monotone measures that are not perfect.

No.	α	β	γ	λ
1.	2/3	1/2	1/2	1/6
2.	2/3	1/2	1/3	1/6
3.	1/3	1/2	1/6	1/6
4.	1/3	1/3	1/6	1/6
5.	5/6	1/3	2/3	1/6

Table 1: Values of parameters $\alpha,\beta,\gamma,\lambda$.

8 Conclusion

In this paper we give general necessary and sufficient conditions under which 2-monotone measures are extreme points of M_{2-mon}, describe some important classes of them, and give ways of their generation. As shown by examples, the introduced class of extreme 2-monotone measures, that are additive on filters do not cover all possible extreme 2-monotone measures, and we cannot generate all possible extreme 2-monotone measures based on aggregations with the help of multilinear extension. However, this paper can be considered as the first step to the desirable solution. As one can see from the examples, general extreme 2-monotone measures have structures that are similar to a structure of extreme 2-monotone measures that are additive on filter and there is a possibility to generalize it. This can be the topic for the future research.

Appendix

Proof of Proposition 1. We should prove monotonicity, i.e. $\mu(A\cup\{x_k\})-\ \mu(A)\ge 0$ for $A\in 2^X$ and $x_k\notin A$. Let us consider a chain of sets $B_0=\varnothing$, $B_1=\{y_1\}$, $B_1=\{y_1,y_2\}$,…, $B_m=\{y_1,...,y_m\}$. Then inequalities (2) imply

$$0\le\mu(B_0\cup\{x_k\})-\mu(B_0)\le\mu(B_1\cup\{x_k\})-\mu(B_1)\le...\le$$

$\mu(B_m\cup\{x_k\})-\mu(B_m)$, i.e. $\mu(A\cup\{x_k\})-\mu(A)\ge 0$. ■

Proof of Proposition 3. It is necessary to show that (3) is valid for $A\cup\{x_i\}\in\mathbf{f}$ and $A\cup\{x_j\}\notin\mathbf{f}$. Since $\mu(A\cup\{x_j\})=0$ and $\mu(A)=0$, this inequality is transformed to

$$\mu(A\cup\{x_i\})\le\mu(A\cup\{x_i\}\cup\{x_j\}).$$

Let us prove that $\mu(B\cup\{x_j\})-\mu(B)\ge 0$ for any $B\in\mathbf{f}$ and $x_j\notin B$. Since $B\in\mathbf{f}$, then there exists a minimal element $C_k\in\mathcal{C}$ such that $C_k\subseteq B$. Let us show first that

$$\mu(C_k\cup\{x_j\})-\mu(C_k)\ge 0 \qquad \text{(A1).}$$

According to the statement of the proposition for $C_k\in\mathcal{C}$ and $x_j\notin C_k$ there exists $C_l\in\mathcal{C}$ such that $\{x_j\}=C_l\setminus C_k$. Since $C_k\setminus C_l\neq\varnothing$, there is some $x_i\in C_k\setminus C_l$, and obviously $C_l\subseteq(C_k\setminus\{x_i\})\cup\{x_j\}$, i.e. $(C_k\setminus\{x_i\})\cup\{x_j\}\in\mathbf{f}$. Because μ is 2-monotone on $\mathbf{f}$, we have $\mu((C_k\setminus\{x_j\})\cup\{x_i\})+\mu(C_k)\le\mu(C_k\setminus\{x_j\})+\mu(C_k\cup\{x_i\})$. Let us notice that in the last inequality $\mu((C_k\setminus\{x_j\})\cup\{x_i\})>0$ and $\mu(C_k\setminus\{x_j\})=0$, therefore, the inequality (A1) is valid.

Let us show next that this inequality is fulfilled for B if $C_k\subseteq B$. For this purpose, consider the following chain of sets

$$B_0=C_k,\ B_1=C_k\cup\{x_{i_1}\},...,B_r=C_k\cup\{x_{i_1},...,x_{i_r}\}=B.$$

Since μ is 2-monotone on the filter $\mathbf{f}$, the following inequalities are valid:

$$\mu(B_0\cup\{x_j\})-\mu(B_0)\le\mu(B_1\cup\{x_j\})-\mu(B_1)\le...$$
$$\le\mu(B_r\cup\{x_j\})-\mu(B_r),$$

i.e. $\mu(C_k\cup\{x_i\})-\mu(C)\le\mu(B\cup\{x_i\})-\mu(B)$.■

Proof of Theorem 1. Obviously, $\rho_\gamma\supseteq\rho_\Gamma$ for any order ρ_γ with $\gamma\in\Gamma$. Let us show next that if $\rho_{\gamma'}\supseteq\rho_\Gamma$, then $\gamma'\in\Gamma$. For this purpose, it is necessary to show that sets $B_1=\{y_1\}$, $B_2=\{y_1,y_2\}$, ..., $B_n=\{y_1,y_2,...,y_n\}$, where $y_i=x_{\gamma'(i)}$, $i=1,2,...,n$, are in $\mathfrak{L}$. Let us show first that $B_1\in\mathfrak{L}$. Let us put into correspondence to each permutation $\gamma\in\Gamma$ the set

$B_\gamma(y_1)=\{x_{\gamma(1)},x_{\gamma(2)},...,x_{\gamma(m)}\}$ such that $y_1=x_{\gamma(m)}$. It is easy to see that conditions $\rho_\Gamma=\bigcap_{\gamma\in\Gamma}\rho_\gamma$ and $\rho_{\gamma'}\supseteq\rho_\Gamma$ imply $\bigcap_{\gamma\in\Gamma}B_\gamma(y_1)=\{y_1\}$, i.e. $B_1\in\mathfrak{L}$. We then prove $B_k\in\mathfrak{L}$, $k=2,...,n$, by induction. Let us assume that $B_1,...,B_{k-1}\in\mathfrak{L}$ and show that $B_k\in\mathfrak{L}$. In this case the conditions $\rho_\Gamma=\bigcap_{\gamma\in\Gamma}\rho_\gamma$ and $\rho_{\gamma'}\supseteq\rho_\Gamma$ imply $\bigcap_{\gamma\in\Gamma}B_\gamma(y_k)\subseteq B_{k-1}\cup\{y_k\}$. Therefore, $B_{k-1}\cup\bigcap_{\gamma\in\Gamma}B_\gamma(y_k)=B_k$, i.e. $B_k\in\mathfrak{L}$.■

Proof of Proposition 5. *Necessity***.** Let us assume that μ is not an extreme point in M_{2-mon}. Then it can be represented in the form $\mu=a\mu_1+(1-a)\mu_2$, where $a\in(0,1)$, $\mu_1,\mu_2\in M_{2-mon}$ and $\mu_1\neq\mu$. Clearly, μ_1 obeys conditions on ν 1) and 2) in this proposition.

Sufficiency. Let us assume to the contrary that there exists $\nu\in M_{2-mon}$ with properties 1) and 2). We will show that in this case μ is not an extreme point in M_{2-mon}. For this purpose, let us consider a set function $\theta_a(A)=\mu(A)-a\nu(A)$, parametrically depending on $a\in[0,1]$ and also

$$\varepsilon_1=\max\{a\in[0,1]\,|\,\theta_a(A)\geq 0\ \mathit{for\ all}\ A\in 2^X\},$$

$$\varepsilon_2=\max\{a\in[0,1]\,|\,\theta_a(A)+\theta_a(B)\leq$$
$$\theta_a(A\cap B)+\theta_a(A\cup B)\ \mathit{for\ all}\ A,B\in 2^X\}.$$

It is easy to see that conditions 1) and 2) imply that $\varepsilon_1>0$ and $\varepsilon_2>0$. Therefore, a set function θ_b, where $b=\min\{\varepsilon_1,\varepsilon_2\}$, is nonnegative and 2-monotone. Thus, μ is represented as

$$\mu=b\nu(A)+(1-b)\mu_2,$$

where $\mu_2=\theta_b/(1-b)$ and, obviously, $\nu,\mu_2\in M_{2-mon}$, i.e. μ is not an extreme point in M_{2-mon}.■

Proof of Proposition 7. Clearly, $\tilde{\mu}(\mathbf{x})=\tilde{\varphi}\left(\tilde{\mu}_1(\mathbf{x}^{(1)}),...,\tilde{\mu}_n(\mathbf{x}^{(n)})\right)$ for every binary vector $\mathbf{x}$ and $\tilde{\varphi}\left(\tilde{\mu}_1(\mathbf{x}^{(1)}),...,\tilde{\mu}_n(\mathbf{x}^{(n)})\right)$ is a multilinear polynomial. Therefore, the proposition follows from the uniqueness of such a polynomial for the pseudo-Boolean function μ. ■

Proof of Lemma 2. Using the Taylor decomposition at the point $\mathbf{x}=(x_1,...,x_{n-1},0)$, we get

$$\varphi(x_1,...,x_{n-1},x_{n+1},....,x_{n+m})=\varphi_1(x_1,...,x_{n-1},0)+$$
$$\frac{\partial\tilde{\varphi}_1(x_1,...,x_{n-1},0)}{\partial x_n}\varphi_2(x_{n+1},...,x_{n+m}).$$

Then we find that if $(x_1,...,x_{n-1},x_{n+1},....,x_{n+m})=\mathbf{1}_{A\cup B}$, then

$$\varphi_1(x_1,...,x_{n-1},0)=\mu_1(A),$$

$$\frac{\partial\tilde{\varphi}_1(x_1,...,x_{n-1},0)}{\partial x_n}=\mu_1(A\cup\{n\})-\mu_1(A),$$

$$\varphi_2(x_{n+1},...,x_{n+m})=\mu_2(B).\blacksquare$$

Proof of Proposition 8. Let $\mu(A)=\varphi(\mathbf{1}_A)$, where $A\in 2^Z$. Then the multilinear extension of φ can be represented as

$$\tilde{\varphi}(\mathbf{x})=\sum_{A\subseteq Z}\mu(A)\prod_{k\in A}x_k\prod_{k\notin A}(1-x_k).$$

Taking partial derivative, we get

$$\frac{\partial\tilde{\varphi}(\mathbf{x})}{\partial x_i}=\sum_{A\subseteq Z\setminus\{i\}}\left(\mu(A\cup\{i\})-\mu(A)\right)\prod_{k\in A}x_k\prod_{k\notin A}(1-x_k).$$

The proposition follows from the last formula. ■

Proof of Proposition 9. Let us show that φ is defined uniquely. Let $\mathbf{x}\in\{0,1\}^n$ and $A=\bigcup_{i=1}^n A_i$, where $A_i\in 2^{X_i}$, is chosen such that $A_i=X_i$ if $x_i=1$ and $A_i=\varnothing$ if $x_i=0$. Then $\left(\mu_1(A\cap X_1),...,\mu_n(A\cap X_n)\right)=\mathbf{x}$, i.e. $\mu(A)=\varphi(\mathbf{x})$. This means that φ is defined uniquely by μ.

Let us show that vector $\boldsymbol{\mu}$ is defined uniquely if each variable x_i is essential for φ. Let us assume that the variable x_i is essential for φ. Then by definition there are vectors $\mathbf{x}_1=\left(x_1,...,x_{i-1},0,x_{i+1},...,x_n\right)$ and $\mathbf{x}_2=\left(x_1,...,x_{i-1},1,x_{i+1},...,x_n\right)$ in $\{0,1\}^n$ such that $\varphi(\mathbf{x}_1)\neq\varphi(\mathbf{x}_2)$. Let $A=\bigcup_{k=1}^n A_k$, where $A_k\in 2^{X_k}$, such that $A_k=X_k$ if $x_k=1$ and $k\neq i$; $A_k=\varnothing$ if $x_k=0$ and $k\neq i$; and A_i is chosen arbitrary in 2^{X_i}. Then using the Taylor decomposition, we get

$$\mu(A)=\tilde{\varphi}\left(x_1,...,x_{i-1},\mu_i(A_i),x_{i+1},...,x_n\right)=$$
$$\tilde{\varphi}(\mathbf{x}_1)+\mu_i(A_i)\frac{\partial\tilde{\varphi}(\mathbf{x}_1)}{\partial x_i}.$$

Therefore, we can calculate

$$\mu_i(A_i)=\left(\mu(A)-\tilde{\varphi}(\mathbf{x}_1)\right)\Big/\frac{\partial\tilde{\varphi}(\mathbf{x}_1)}{\partial x_i},$$

because $\frac{\partial\tilde{\varphi}(\mathbf{x}_1)}{\partial x_i}\neq 0$ according to Proposition 8. Thus, each set function μ_i is defined uniquely if every variable x_i is essential. Let us notice that if φ contains a nonessential variable x_i, then $\tilde{\varphi}$ does not depend on x_i. This implies that the representation $\mu=\tilde{\varphi}\circ\boldsymbol{\mu}$ is not defined uniquely, since any μ_i has no influence on the result of aggregation.■

Proof of Theorem 3. *Necessity.* Consider 2 possible cases.

1) Let us assume to the contrary that φ is not an extreme 2-monotone measure, however, μ is an extreme 2-monotone measure. Then $\varphi=a\varphi_1+(1-a)\varphi_2$, where $a\in(0,1)$ and φ_1,φ_2 are different 2-monotone measures on 2^Z. Therefore, $\tilde{\varphi}=a\tilde{\varphi}_1+(1-a)\tilde{\varphi}_2$ and $\mu=\tilde{\varphi}\circ\boldsymbol{\mu}=$ $a\tilde{\varphi}_1\circ\boldsymbol{\mu}+(1-a)\tilde{\varphi}_2\circ\boldsymbol{\mu}$, where $\tilde{\varphi}_1\circ\boldsymbol{\mu},\tilde{\varphi}_2\circ\boldsymbol{\mu}$ are different 2-monotone measures by Proposition 9. But this contradicts our assumption that μ is an extreme 2-monotone measure.

2) Let us assume to the contrary that μ_i is not an extreme 2-monotone measure for some $i\in\{1,...,n\}$, however, μ is an extreme 2-monotone measure. Then μ_i can be represented as a convex sum of two different 2-monotone measures: $\mu_i=a\mu_i^{(1)}+(1-a)\mu_i^{(2)}$, where $a\in(0,1)$, therefore,

$$\mu=\tilde{\varphi}\circ\left(\mu_1,...,\mu_{i-1},a\mu_i^{(1)}+(1-a)\mu_i^{(2)},\mu_{i+1}...,\mu_n\right)=$$
$$a\tilde{\varphi}\circ\left(\mu_1,...,\mu_{i-1},\mu_i^{(1)},\mu_{i+1}...,\mu_n\right)+$$
$$(1-a)\tilde{\varphi}\circ\left(\mu_1,...,\mu_{i-1},\mu_i^{(2)},\mu_{i+1}...,\mu_n\right),$$

where

$$\tilde{\varphi}\circ\left(\mu_1,...,\mu_{i-1},\mu_i^{(1)},\mu_{i+1}...,\mu_n\right),\tilde{\varphi}\circ\left(\mu_1,...,\mu_{i-1},\mu_i^{(2)},\mu_{i+1}...,\mu_n\right),$$

are different 2-monotone measures. But this contradicts the assumption that μ is an extreme 2-monotone measure.

Sufficiency. We will prove sufficiency by induction. According to Remark 3 any aggregation (5) can be represented as a composition of simple aggregations described in Lemma 2. Therefore, if we prove that any simple aggregation of a type $\varphi(x_1,...,x_{n-1},x_{n+1},....,x_{n+m})=\tilde{\varphi}_1(x_1,...,x_{n-1},\varphi_2(x_{n+1},....,x_{n+m}))$, whe re the corresponding set functions μ_1 and μ_2 are extreme 2-monotone measures (see notations from Lemma 2), generates the extreme 2-monotone measure μ. Then we can also say that the general aggregation produces the extreme 2-monotone measure if the conditions of the theorem are fulfilled. Let us assume to the contrary that μ is not an extreme 2-monotone measure. Then there are 2 different 2-monotone measures $\mu^{(0)}$ and $\mu^{(1)}$ such that

$$\mu = a\mu^{(0)} + (1-a)\mu^{(1)}\text{, where } a \in (0,1)\text{ .}$$

Let us consider 2-monotone measures, generated by a mapping

$$\psi(i)=\begin{cases} i, & i\in\{1,...,n-1\}, \\ n, & i\in\{n+1,...,m\}, \end{cases}$$

Obviously, $\mu^{\psi}=a(\mu^{(0)})^{\psi}+(1-a)(\mu^{(1)})^{\psi}$ [2], $\mu^{\psi}=\mu_1$, and $\mu^{(0)}$, $\mu^{(1)}$ are 2-monotone measures. But according to our assumption μ_1 is an extreme 2-monotone measure. Therefore, this implies that $(\mu^{(0)})^{\psi}=\mu_1$.

Our next step is to show that if μ_2 is also an extreme 2-monotone measure, then $\mu^{(0)}=\mu^{(1)}=\mu$.

By Lemma 2, μ can be represented as

$$\mu(A\cup B)=\mu_1(A)+(\mu_1(A\cup\{n\})-\mu_1(A))\mu_2(B)\text{ ,} \tag{A2}$$

where $A\subseteq\{1,...,n-1\}$ and $B\subseteq\{n+1,...,n+m\}$. Let us denote $Y=\{n+1,...,n+m\}$. Then, taking in account the correspondence between pseudo-Boolean and set functions, the formula (A2) can be rewritten as

$$\mu(A\cup B)=\mu(A)+(\mu(A\cup Y)-\mu(A))\mu_2(B)\text{ ,}$$

and we can calculate

$$\mu_2(B)=\frac{\mu(A\cup B)-\mu(A)}{\mu(A\cup Y)-\mu(A)}\text{,}$$

for any $A\subseteq\{1,...,n-1\}$ such that $\mu(A\cup Y)-\mu(A)>0$. Let us consider set functions:

$$\mu_2^{(i)}(B)=\frac{\mu^{(i)}(A\cup B)-\mu^{(i)}(A)}{\mu^{(i)}(A\cup Y)-\mu^{(i)}(A)}\text{, } i=1,2\text{ ,}$$

of $B\subseteq Y$ for any $A\subseteq\{1,...,n-1\}$ with $\mu(A\cup Y)-\mu(A)>0$ and $\mu^{(i)}(A\cup Y)=\mu(A\cup Y)$. It is easy to show that these set functions are 2-monotone. Let us notice that we have proved that $\mu^{(i)}(A\cup Y)=\mu(A\cup Y)$ and $\mu^{(i)}(A)=\mu(A)$. After that we easily derive that

$$a\mu_2^{(0)}(B)+(1-a)\mu_2^{(1)}(B)=\mu_2(B)\text{ .}$$

By our assumption, μ_2 is an extreme 2-monotone measure. This implies that $\mu_2^{(0)}=\mu_2^{(1)}=\mu_2$. Thus, we can write

$$\mu^{(i)}(A\cup B)=\mu(A)+(\mu(A\cup Y)-\mu(A))\mu_2(B)=\mu(A\cup B)$$

for any $A\subseteq\{1,...,n-1\}$ and $B\subseteq\{n+1,...,n+m\}$, i.e. $\mu^{(0)}=\mu^{(1)}=\mu$, but this contradicts our assumption that measures $\mu_2^{(0)}$ and $\mu_2^{(1)}$ are different.■

Proof of Lemma 3. Let us consider the sequence of sets

$$B_0=C\text{ , } B_1=C\cup\{x_{i_1}\},...,B_m=C\cup\{x_{i_1},...,x_{i_m}\}=A\text{ .}$$

Since μ is additive on the filter $\mathbf{f}$, we can write

$$\mu(C\cup\{x_j\})-\mu(C)=\mu(B_1\cup\{x_i\})-\mu(B_1)=...$$
$$=\mu(B_m\cup\{x_i\})-\mu(B_m)\text{ ,}$$

i.e. $\mu(C\cup\{x_i\})-\mu(C)=\mu(A\cup\{x_i\})-\mu(A)$. Thus, the required equality is valid.■

Proof of Corollary 2. By Lemma 3

$$\mu(X)-\mu(X\setminus\{x_i\})=\mu(C\cup\{x_i\})-\mu(C)$$

for any $C\in\mathbf{f}$ with $C\subseteq X\setminus\{x_i\}$. This implies the result.■

Proof of Proposition 10. Let us check inequality (3), considering the following possible cases:

a) if $A\cup\{x_i\}\in\mathbf{f}$ and $A\cup\{x_j\}\in\mathbf{f}$, then the inequality (3) follows from the additivity of μ on $\mathbf{f}$;

b) if $A\cup\{x_i\}\in\mathbf{f}$ and $A\cup\{x_j\}\notin\mathbf{f}$, then (3) is transformed to

$$\mu(A\cup\{x_i\})\le\mu(A\cup\{x_i,x_j\})\text{ .}$$

The last inequality is valid, because according to our assumption $\nu(x_j)\ge 0$;

c) if $A\cup\{x_i\}\notin\mathbf{f}$ and $A\cup\{x_j\}\notin\mathbf{f}$, then inequality (3) is obviously true. ■

Proof of Proposition 11. It is sufficient to show that

$$\mu(A\cup\{x_i\})-\mu(A)\ge 0 \tag{A3}$$

for all $A\in\mathbf{f}$ and any $x_i\in X$. By the assumption $\left\{\bar{C}_1,...,\bar{C}_k\right\}$ is a covering of X, therefore, there is a set $\bar{C}_l$ such that $x_i\in\bar{C}_l$. Let us consider 2 possible cases.

If $\left|\bar{C}_l\right|=1$, i.e. $\bar{C}_l=\{x_i\}$, then $x_i\in A$ for all $A\in\mathbf{f}$. Obviously, in this case the inequality (A3) is valid.

If $\left|\bar{C}_l\right|\ge 2$, then there is $x_j\in\bar{C}_l$ such that $x_j\ne x_i$. Since C_l is a maximal element in $2^X\setminus\mathbf{f}$, then $x_i\cup C_l\in\mathbf{f}$, $x_j\cup C_l\in\mathbf{f}$, and additivity of μ on $\mathbf{f}$ implies

$$\nu(x_i)=\mu(C_l\cup\{x_i,x_j\})-\mu(C_l\cup\{x_j\})=$$
$$\mu(C_l\cup\{x_i\})-\mu(C_l)=\mu(C_l\cup\{x_i\})\ge 0\text{ ,}$$

i.e. the inequality (A3) is valid for all $A\in\mathbf{f}$.■

Proof of Lemma 4. Clearly $\mathbf{f}_0\supseteq\mathbf{f}$. Let us show that $B\in\mathbf{f}_0$ and $B\subseteq C$ implies $C\in\mathbf{f}_0$. It is sufficient to consider the case, when $B\notin\mathbf{f}$ and $C\notin\mathbf{f}$. Then there exist $x_i,x_j\notin B$ such that

$$B=(B\cup\{x_i\})\cap(B\cup\{x_j\})\text{ .}$$

By our assumption $C\notin\mathbf{f}$, therefore $x_i,x_j\notin C$. This implies that $C=(C\cup\{x_i\})\cap(C\cup\{x_j\})$, i.e. $C\in\mathbf{f}_0$. ■

Proof of Proposition 12. It is necessary to show that $\mu(A\cup\{x_i\})-\mu(A)=\nu(x_i)$ for any $A\in\mathbf{f}_0$ and $x_j\notin A$. Let us show first that if $A\in\mathbf{f}_0$, then

$$\mu(A)+\sum_{x_i\notin A}\nu(x_i)=\mu(X)\text{ .}$$

Let us consider two possible cases. Let $A\in\mathbf{f}$ and $X\setminus A=\left\{y_1,y_2,...,y_m\right\}$. Then

[2] μ^{ψ} denotes a measure on $2^{\{1,...,n\}}$ such that $\mu^{\psi}(A)=\mu(\psi^{-1}(A))$, where $\psi^{-1}(A)=\left\{i\in\{1,...,n+m\}\,|\,\psi(i)\in A\right\}$.

$$\mu(A\cup\{y_1\})=\mu(A)+\nu(y_1),$$
$$\mu(A\cup\{y_1,y_2\})=\mu(A\cup\{y_1\})+\nu(y_2)$$
$$=\mu(A)+\nu(y_1)+\nu(y_2),$$
$$\vdots$$
$$\mu(X)=\mu(A)+\sum_{i=1}^{m}\nu(y_i),$$

i.e. the required equality is valid for $A\in\mathbf{f}$. Let us consider the case, when $A\in\mathbf{f}_0\backslash\mathbf{f}$. Then there exist $x_i,x_j\notin A$ such that $A\cup\{x_i\},A\cup\{x_j\}\in\mathbf{f}$, and

$$\mu(A\cup\{x_i\})+\mu(A\cup\{x_j\})=\mu(A\cup\{x_i,x_j\})+\mu(A),$$

i.e.

$$\mu(A\cup\{x_j\})-\mu(A)=\mu(A\cup\{x_i,x_j\})-$$
$$\mu(A\cup\{x_i\})=\nu(x_j).$$

After that we see that $\mu(A\cup\{x_j\})=\mu(A)+\nu(x_j)$ and

$$\mu(X)=\mu(A\cup\{x_j\})+\sum_{x_i\notin A\cup\{x_j\}}\nu(x_i)=\mu(A)+\sum_{x_i\notin A}\nu(x_i).$$

Thus, we can write

$$\mu(A\cup\{x_j\})-\mu(A)=\mu(X)-\sum_{x_i\notin A\cup\{x_j\}}\nu(x_i)-$$
$$\left(\mu(X)-\sum_{x_i\notin A}\nu(x_i)\right)=\nu(x_j).\blacksquare$$

Proof of Corollary 4.

$$\mu(A)+\mu(B)=\mu(A\cap B)+\sum_{x_i\in A\backslash B}\nu(x_i)+\mu(A\cap B)+$$
$$\sum_{x_i\in B\backslash A}\nu(x_i)=\mu(A\cap B)+\mu(A\cap B)+$$
$$\sum_{x_i\in(A\cup B)\backslash(A\cap B)}\nu(x_i)=\mu(A\cap B)+\mu(A\cup B).\blacksquare$$

Proof of Proposition 13. The first part of the proposition follows from the results considered above. Let us prove the second part. For this purpose, let us show that any set function μ, obeying (6) and (7) is additive on $\mathbf{f}$, i.e.

$$\mu(A\cup\{x_i\})+\mu(A\cup\{x_j\})=\mu(A)+\mu(A\cup\{x_i,x_j\}),$$

for $A\cup\{x_i\}$, $A\cup\{x_j\}\in\mathbf{f}$ and $A\cap\{x_i,x_j\}=\varnothing$.

Let us consider 2 possible cases. Let $A\in\mathbf{f}$, then

$$\mu(A\cup\{x_i\})+\mu(A\cup\{x_j\})=\mu(X)-$$
$$\sum_{x_k\notin A\cup\{x_i\}}\nu(x_k)+\mu(X)-\sum_{x_k\notin A\cup\{x_i\}}\nu(x_k)=$$
$$\mu(X)-\sum_{x_k\notin A}\nu(x_k)+\mu(X)-\sum_{x_k\notin A\cup\{x_i,x_j\}}\nu(x_k)=$$
$$\mu(A)+\mu(A\cup\{x_i,x_j\}).$$

Let $A\in\mathbf{f}_0\backslash\mathbf{f}$, then

$$\mu(A\cup\{x_i\})+\mu(A\cup\{x_j\})=\mu(X)-$$
$$\sum_{x_k\notin A\cup\{x_i\}}\nu(x_k)+\mu(X)-\sum_{x_k\notin A\cup\{x_i\}}\nu(x_k)=$$
$$\mu(X)-\sum_{x_k\notin A}\nu(x_k)+\mu(X)-\sum_{x_k\notin A\cup\{x_i,x_j\}}\nu(x_k).$$

In the last expression $\sum_{x_k\notin A}\nu(x_k)=\mu(X)$, in addition, $\mu(A)=0$. This implies that

$$\mu(A\cup\{x_i\})+\mu(A\cup\{x_j\})=\mu(A)+\mu(A\cup\{x_i,x_j\}).\blacksquare$$

Proof of Proposition 14. Let us assume to the contrary that μ is uniquely defined by the filter, but it is not extreme. Then there are 2 different 2-monotone measures μ_1 and μ_2 such that $\mu=a\mu_1+(1-a)\mu_2$, where $a\in(0,1)$. It easy to check that both measures μ_1 and μ_2 are additive on filter $\mathbf{f}$ but this contradicts our assumption.■

Proof Proposition 15. Let us assume to the contrary that μ obeys conditions of the proposition, however, there is another extreme 2-monotone measure ν with $\mathbf{f}=\left\{A\in 2^X\,|\,\nu(A)>0\right\}$. Let us consider the set function $\theta_a(A)=\nu(A)-a\mu(A)$, parametrically depending on $a\in[0,1]$ and

$$b=\max\left\{a\in[0,1]\,|\,\theta_a(A)\geq 0\ for\ all\ A\in 2^X\right\}.$$

Clearly, $b>0$ and the set function θ_b is 2-monotone on the filter $\mathbf{f}$. According to Proposition 3 θ_b is 2-monotone on 2^X. Therefore, we can represent ν as

$$\nu=b\mu+(1-b)\mu_2,$$

where $\mu_2=\theta_b/(1-b)$ is a 2-monotone measure, but this contradicts our assumption.■

Proof of Proposition 16. Let us assume to the contrary that μ obeys conditions of the proposition, however, there is another extreme 2-monotone measure ν, which is additive on $\mathbf{f}=\left\{A\in 2^X\,|\,\nu(A)>0\right\}$. Let us consider the set function $\theta_a(A)=\mu(A)-a\nu(A)$, parametrically depending on $a\in[0,1]$ and

$$b=\max\left\{a\in[0,1]\,|\,\theta_a(A)\geq 0\ for\ all\ A\in 2^X\right\}.$$

Clearly, $b>0$ and the set function θ_b is additive on the filter $\mathbf{f}$. According to Proposition 11, θ_b is 2-monotone on 2^X. Therefore, we can represent μ as

$$\mu=b\nu+(1-b)\mu_2,$$

where $\mu_2=\theta_b/(1-b)$ is a 2-monotone measure, but this contradicts our assumption.■

Proof of Proposition 17. Let us notice that $\mu_n=\eta_{\langle X\rangle}$ and for this case the proposition is obviously true. Let us check that μ_k, where $k\in\{2,...,n-1\}$, is additive on the filter $\mathbf{f}=\{A\in 2^X\,\|A|\geq k\}$, i.e. the following equality holds

$$\mu(A\cup\{x_i\})+\mu(A\cup\{x_j\})=\mu(A)+\mu(A\cup\{x_i,x_j\}), \text{(A4)}$$

for $A\cup\{x_i\}$, $A\cup\{x_j\}\in\mathbf{f}$ and $A\cap\{x_i,x_j\}=\varnothing$. Let $|A|=m\geq k-1$, then the equality (A4) is transformed to

$$\frac{m-k+2}{n-k+1}+\frac{m-k+2}{n-k+1}=\frac{m-k+1}{n-k+1}+\frac{m-k+3}{n-k+1},$$

i.e. (A4) is valid for this case.

Let us show that μ_k is an extreme 2-monotone measure. In this case $\mathbf{f}_0=\{A\in 2^X\,\|A|\geq k-1\}$ and by Proposition 13 all possible 2-monotone measures that are additive on $\mathbf{f}$ can be found by solving the following linear system of equations:

$$\sum_{x_i\notin A}\nu(x_i)=1 \text{ for all } A\in 2^X \text{ with } |A|=k-1.$$

It is easy to check that the solution is uniquely defined by $\nu(x_i)=1/(n-k+1)$, $i=1,...,n$. This implies that μ_k is an extreme 2-monotone measure. It is easy to check that μ_k is a perfect extreme 2-monotone measure by Proposition 15. ■

Acknowledgements

Andrey G. Bronevich sincerely thanks the German Academic Exchange Service (DAAD), Ludwig Maximilian University, and Prof. Thomas Augustin for research opportunity provided. Authors thank anonymous reviewers for very useful remarks allowing to make considerable improvements of the paper.

References

[1] G. Birkhoff. *Lattice Theory*. American Mathematical Society Colloquium Publications, v. 25 (3rd ed.), Providence, R.I, 1979.

[2] Bronevich A.G. Canonical sequences of fuzzy measures. In Proc. of International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU-2004), Perugia-Italy, 2004, 8 pp.

[3] Bronevich A.G. On the closure of families of fuzzy measures under eventwise aggregations. *Fuzzy sets and systems,* v. 153, 2005, 45-70.

[4] Bronevich A.G., Augustin T. Approximation of coherent lower probabilities by 2-monotone measures. Proc. of the 6th International Symposium on Imprecise Probability: Theories and Their Applications, Durham, United Kingdom, 2009, 9 pp.

[5] Bronevich A.G., Klir G.J. Measures of uncertainty for imprecise probabilities: An axiomatic approach. International *Journal of Approximate Reasoning*, v. 51, 2010, 365-390.

[6] Chateauneuf A., Jaffray J.Y. Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. *Mathematical Social Sciences*, v. 17, 1989, 263-283.

[7] Combarro E., Miranda P. Extreme points of some families of non-additive measures. *European Journal of Operational Research*, v. 174, 2006, 1865-1884.

[8] Combarro E., Miranda P. On the polytope of non-additive fuzzy measures. *Fuzzy Sets and Systems*, v. 159, 2008, 2145-2162.

[9] Denneberg D. *Non-additive Measure and Integral*. Kluwer, Dordrecht, 1997.

[10] Fujishige S. *Submodular functions and optimization.* Series: Annals of Discrete Mathematics, v. 58, second edition, Elsevier, Amsterdam, 2005.

[11] Grabisch M., Marichal J.-L., Mesiar R., Pap E. *Aggregation Functions*, Cambridge University Press, Cambridge. UK, 2009.

[12] Hammer P.L., Rudeanu S. *Boolean Methods in Operational Research and Related Arreas*. Springer, Berlin, Germany, 1968.

[13] Karkishchenko A.N. Invariant fuzzy measures on a finite algebra. Proc. of the North American Fuzzy Information society, NAFIPS'96, USA, Berkeley, June 20-22, 1996, v. 1, pp. 588- 592.

[14] Owen G. Multilinear extensions of games. In A.E. Roth (ed.) *The Shapley Value. Essays in Honor of Lloyd S. Shapley*. Cambridge University Press. Cambridge. UK, 1988, 139-151.

[15] Radojevic D. The logical representation of the discrete Choquet integral. *Belgian Journal of Operations Research, Statistics and Computer Science*, v. 38, 1998, 67-89.

[16] Shapley L.S. Cores of convex games. *International Journal of Game Theory*, v. 1, 1971, 11-26.

[17] Walley P. *Statistical reasoning with imprecise probabilities.* Chapman & Hall, London, 1991.

On the Robustness of Imprecise Probability Methods

Marco E. G. V. Cattaneo
Department of Statistics, LMU Munich
cattaneo@stat.uni-muenchen.de

Abstract

Imprecise probability methods are often claimed to be robust, or more robust than conventional methods. In particular, the higher robustness of the resulting methods seems to be the principal argument supporting the imprecise probability approach to statistics over the Bayesian one. The goal of the present paper is to investigate the robustness of imprecise probability methods, and in particular to clarify the terminology used to describe this fundamental issue of the imprecise probability approach.

Keywords. Robustness, imprecise probabilities, Bayesian analysis, credibility, decision making, indecision, sensitivity analysis, imprecise Dirichlet model.

1 Introduction

The theories of imprecise probability replace probability measures by more general mathematical objects, which can often be identified with particular sets of probability measures. Such sets appear naturally also in Bayesian sensitivity analysis (also called robust Bayesian analysis) [6, 27] and robust statistics [4, 20]. Hence, there is a strong connection between imprecise probability and robustness. In fact, methods resulting from the imprecise probability approaches to inference and decision making are often claimed to be "robust" (or "more robust" than alternative methods) [1, 14, 36], usually without specifying the meaning of "robust". The goal of the present paper is to investigate the robustness of imprecise probability methods. We will focus in particular on the most developed theory of imprecise probability: the theory of lower and upper previsions [33, 35].

The question of the robustness of imprecise probability methods is particularly important in statistics, where the imprecise probability approach can be seen as an alternative to the Bayesian approach. In fact, when comparing these two approaches to statistics, the latter has clear advantages in terms of technical and conceptual simplicity [12, 13], also thanks to important invariances [3, 18, 21]. On the other hand, the (higher) robustness of the resulting methods seems to be one of the few general advantages claimed by the proponents of the imprecise probability approach. That is, the alleged (higher) robustness of the imprecise probability methods seems to be the principal argument for preferring the imprecise probability approach to statistics over the Bayesian one.

The present paper examines various aspects of the question of the robustness of imprecise probability methods, and in particular tries to clarify the terminology used to describe this fundamental issue of the imprecise probability approach. The paper is organized as follows. In the next section the concept of robustness is introduced. The robustness of imprecise probability methods is then investigated in Section 3, which is the core of the paper. In particular, in Subsection 3.1 the higher credibility of imprecise probability analyses over Bayesian analyses is discussed. These two kinds of analyses are then compared with regard to decision making: Subsection 3.2 considers the case when a decision has to be made, while the case when indecision is allowed is studied in Subsection 3.3. The final section summarizes the results.

2 Robustness

Robustness means "insensitivity to small deviations from the assumptions" [19, p. 2]. In the Bayesian approach to inference and decision making it mainly refers to "possible misspecification of the prior distribution" [7, p. 195]. Hence, the conclusions of a Bayesian analysis are not robust if there are several reasonable choices for the prior distribution and the conclusions depend on which prior is actually chosen, as in the following example.

Example 1 *In the Bayesian framework, given an exchangeable sequence of Bernoulli random variables*

$X_1, X_2, \ldots$, de Finetti's theorem [16, § 11.4] implies that they are independent and $Ber(\theta)$-distributed conditional on the success probability $\theta \in [0,1]$. That is, to complete the Bayesian model we must choose a (prior) probability distribution for θ. Suppose that we have (almost) no prior information about θ: several prior probability distributions have been suggested in this situation. In particular, Bayes [5] and Jeffreys [24] proposed the prior uniform distribution of θ on $[0,1]$ and of $\arcsin\sqrt{\theta}$ on $[0, \pi/2]$, respectively. Using Walley's (s,t)-parametrization of the beta distribution [33, 34], these two proposals correspond to the priors $\theta \sim Beta(2, 1/2)$ and $\theta \sim Beta(1, 1/2)$, respectively.

Assume now that we observe $X_1 + \cdots + X_7 = 6$. That is, of the first seven Bernoulli trials, six were successes and one was a failure. In general, on the basis of these data, the conjugate prior distribution $Beta(s,t)$ is updated to the posterior distribution

$$Beta\left(s+7, \frac{s\,t+6}{s+7}\right). \tag{1}$$

In particular, Bayes' and Jeffreys' priors are updated to the posteriors $\theta \sim Beta(9, 7/9)$ and $\theta \sim Beta(8, 13/16)$, respectively.

Finally, suppose that we must choose between two courses of action with uncertain payoffs $A = 5\,X_8 - 4$ and $B = 4 - 5\,X_8$, respectively, expressed in a linear utility scale. This can be interpreted as choosing the side of a bet with odds of 4 to 1 on a success in the next Bernoulli trial, where the total stake is a fixed small amount of money. In general, the conjugate prior distribution $Beta(s,t)$ leads to the posterior expected utilities

$$E(A) = \frac{s}{s+7}\,(5\,t-4) + \frac{7}{s+7}\left(5\,\frac{6}{7} - 4\right) \tag{2}$$

and $E(B) = -E(A)$. These are plotted in Figure 1 as functions of $s \in (0,3]$, in the case $t = 1/2$ and in the limit cases $t \to 1$ and $t \to 0$. In particular, Jeffreys' prior would lead to the choice of the first course of action (that is, betting on success), since $E(A) > E(B)$ when $(s,t) = (1, 1/2)$, while Bayes' prior would lead to the choice of the second course of action (that is, betting on failure), since $E(B) > E(A)$ when $(s,t) = (2, 1/2)$.

Therefore, in this situation the decision resulting from the Bayesian approach is not robust, if both Bayes' and Jeffreys' priors are considered as reasonable choices in the case of (almost) no prior information about θ. The Bayesian answer to this non-robustness issue would be to give more careful consideration to the prior information about θ, in order to be able to identify more precisely the prior probability distribution for θ.

Exactly as for the Bayesian approach, the conclusions resulting from the imprecise probability approach to inference and decision making are robust if they are not too sensitive to small deviations from the assumptions in general, and to possible misspecification of the prior (imprecise) probability distribution in particular. More precise definitions of robustness would be possible, but would have a high degree of arbitrariness, while the above informal definition is sufficient for the scope of the present paper.

3 Imprecise Probability Methods

The robustness of some kinds of conclusions resulting from an imprecise probability analysis has been studied in [32], with comforting results. However, this study did not consider the robustness of the conclusions when the imprecise probabilities have been updated in the light of new data. In this situation, which is obviously very important for the imprecise probability approach to statistics, the conclusions resulting from an imprecise probability analysis are in general not robust (and not more robust that the ones resulting from a Bayesian analysis), as shown in the following example.

Example 2 *Let X be a random variable taking value in the set $\{1,2,3\}$. Assume that our prior imprecise probabilities are determined by the unique assessment $\underline{P}(X) = x$, where $x \in [1,3]$ is a real number. Suppose now that we learn that the value of X is not 2. That is, we observe the event $X \in \{1,3\}$. If we update our prior imprecise probabilities by regular extension [33, Appx. J], then the posterior lower prevision of X is*

$$\underline{P}(X) = \begin{cases} 1 & \textit{if } x < 2, \\ x & \textit{if } x \geq 2, \end{cases} \tag{3}$$

while if we update them by natural extension, then it is

$$\underline{P}(X) = \begin{cases} 1 & \textit{if } x \leq 2, \\ x & \textit{if } x > 2, \end{cases} \tag{4}$$

since the prior lower probability of the observed event is 0 if and only if $x \leq 2$. In both cases (3) and (4), the posterior lower prevision of X, as a function of $x \in [1,3]$, has a discontinuity at $x = 2$.

Therefore, the posterior lower prevision of X is not robust, if for example both values $x = 1.99$ and $x = 2.01$ are considered as reasonable choices for the prior lower prevision. By contrast, in a Bayesian analysis of this situation, the posterior expectation of X would be a continuous function of the prior probability values, although it would be very sensitive to these values if the prior probability of the observed event were very small. Anyway, in this situation the posterior

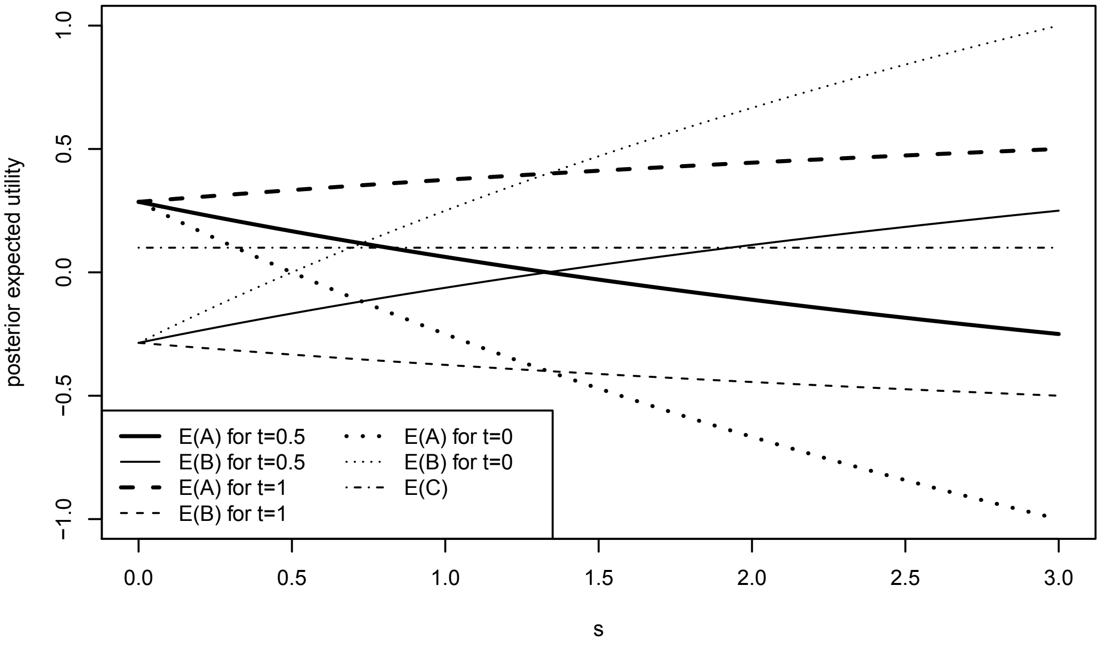


Figure 1: Expected utilities according to the posterior distribution (1), as functions of $s \in (0, 3]$, in the case $t = 1/2$ and in the limit cases $t \to 1$ and $t \to 0$.

distribution of the imprecise probability analysis is in general not more robust than the one of a Bayesian analysis.

However, the situation analyzed in Example 2 is artificial, and consequently its importance for the imprecise probability methods suggested in the literature is not clear. For this reason, in the remainder of the present section we shall consider further the situation of Example 1, focusing on the imprecise probability model that seems to be by far the most studied and used: the imprecise Dirichlet model [8, 34], in the special case of Bernoulli random variables [33, § 5.3].

The imprecise Dirichlet model satisfies some important invariance properties, and in particular the representation invariance principle [34]. This principle describes a particular kind of robustness with respect to assumptions about the statistical model, and it cannot be satisfied by objective Bayesian analyses. However, it can be satisfied by subjective Bayesian analyses, and its appropriateness is questionable anyway [34, p. 52]. On the other hand, the imprecise Dirichlet model is highly non-robust with respect to other aspects of the statistical model [28, 29]. Therefore, to keep things simple, in the remainder of this section we shall consider only the robustness with respect to the choice of the prior distribution.

3.1 Credibility

From the standpoint of the theory of lower and upper previsions, a Bayesian analysis corresponds to the special case of an imprecise probability analysis in which we have so much prior information that the previsions are linear. Hence, from this standpoint, a lower prevision can be interpreted as being based on less information (or assumptions) than a linear prevision dominating it. In this case, the Bayesian analysis can thus be considered as less credible than the imprecise probability analysis, according to a "law of decreasing credibility" [26, p. 1], stating that the credibility of the conclusions decreases when additional assumptions are made.

Such a law seems reasonable when inferences such as confidence or credible regions are considered as conclusions, but it does not necessarily seem reasonable when decisions or point estimates are considered. Anyway, for the sake of argument, let's agree that imprecise probability analyses are more credible than Bayesian analyses (when the linear previsions dominate the lower previsions). Does this imply that they are also more robust?

Example 3 *In the imprecise probability framework, given an exchangeable sequence of Bernoulli random variables $X_1, X_2, \ldots$, a generalization of de Finetti's theorem [15] implies that they are independent and*

$Ber(\theta)$-distributed conditional on the success probability $\theta \in [0,1]$. That is, to complete the imprecise probability model we must choose a (prior) imprecise probability distribution for θ. The usual choice of the prior imprecise probability distribution in the case of (almost) no prior information about θ is the imprecise Dirichlet model, which corresponds to the set of all $Beta(s,t)$ distributions with $t \in (0,1)$. That is, the parameter s must still be chosen: the most popular choices appear to be $s = 2$ and $s = 1$ [8, 34, 36]. In this context, it is important to note that the imprecise previsions resulting from different choices of s are nested, the more imprecise corresponding to the larger values of s.

When observing $X_1 + \cdots + X_7 = 6$, the imprecise Dirichlet model is updated by regular extension to the posterior imprecise probability distribution corresponding to the set of all distributions (1) with $t \in (0,1)$. The posterior lower and upper previsions, $\underline{P}(A)$ and $\overline{P}(A)$, of the utility of the first course of action described in Example 1 are the limits of (2) as $t \to 0$ and as $t \to 1$, respectively. By contrast, the posterior lower and upper previsions, $\underline{P}(B) = -\overline{P}(A)$ and $\overline{P}(B) = -\underline{P}(A)$, of the utility of the second course of action are the limits of $E(B) = -E(A)$ as $t \to 1$ and as $t \to 0$, respectively. These two pairs of posterior lower and upper previsions are plotted in Figure 1 as functions of $s \in (0,3]$.

The posterior imprecise previsions with $s = 1$ are thus more credible (in the sense considered above) than the posterior expectations resulting from Jeffreys' prior, and the posterior imprecise previsions with $s = 2$ are more credible than the posterior expectations resulting from both Bayes' and Jeffreys' priors. However, it is not clear why these posterior imprecise previsions should be more robust than the posterior expectations of Example 1, since they too depend strongly on the choice of s.

The question of the alleged higher robustness of imprecise probability analyses compared to Bayesian analyses can perhaps be better clarified by considering the choice of a probability distribution as consisting of two steps. First we choose a lower prevision $\underline{P}$, and then we select a linear prevision P dominating it. The second step can be seen as an additional assumption, and therefore the imprecise probability analysis based on $\underline{P}$ is more credible than the Bayesian analysis based on P. Moreover, since there is certainly some arbitrariness in the second step, the imprecise probability analysis can appear to be more robust than the Bayesian analysis. However, once P has been selected, it does not depend on the choice of $\underline{P}$ anymore. That is, the robustness of the imprecise probability analysis is relative to the arbitrariness in the choice of $\underline{P}$, while the robustness of the Bayesian analysis is relative to the arbitrariness in the choice of P (and not in both choices of $\underline{P}$ and P). So it is not clear that in general the Bayesian analysis is less robust that the imprecise probability analysis, even when the latter is more credible (in the above sense).

Of course, the imprecise probability analysis would be more robust than the Bayesian analysis, if there were no arbitrariness in the choice of the lower prevision. In this case, "conclusions drawn from the imprecise model are automatically robust, because they do not rely on arbitrary or doubtful assumptions" [33, p. 5]. Unfortunately, this is never the case, because there is always some arbitrariness in the choice of a model, even when we choose the vacuous model. In fact, if the vacuous prevision is a reasonable choice, then probably also a slightly more determined imprecise prevision would be reasonable.

In particular, the choice of the prior distribution in the imprecise probability analysis of Example 3 does not seem to be less arbitrary than the choice of the prior distribution in the Bayesian analysis of Example 1. In fact, thanks to symmetry arguments, in the Bayesian analysis the choice of $t = 1/2$ is less problematic than the choice of s, which must be chosen also in the imprecise probability analysis. In analogy to the discussion above, we could see the choice of the prior probability distribution in Example 1 as consisting of two steps. First we choose to restrict attention to the beta distributions and we select the value of s, while in a second step we also choose the value of t. With this description, it appears that the imprecise Dirichlet model (corresponding to the choices in the first step) has one assumption less than the Bayesian beta prior (the assumption of a particular value for t). However, this appearance is misleading, because in the imprecise Dirichlet model we also make a choice about t: we choose to let it vary in the whole interval $(0,1)$. In fact, replacing this interval for instance with the interval $[\varepsilon, 1-\varepsilon]$, for some small positive ε, could also be a reasonable choice [11].

An important difference between the choices of s in Examples 1 and 3 is that in the latter case the imprecise previsions resulting from different values of s are nested, and this could make the choice "less crucial" than in the former case [34, p. 12]. The importance of this property of the imprecise Dirichlet model for the question of the robustness of the imprecise probability analysis of Example 3 depends on how the imprecise previsions are used. Therefore, in the following subsections we shall consider the decision problem of Example 1 in the imprecise probability framework of Example 3.

3.2 Decision

Several decision criteria have been suggested in the literature on imprecise probabilities [2, 17, 31]. Some of these criteria, like Γ-maximin, induce a total preorder on the possible decisions, and usually identify a single optimal decision. When such criteria are used in an imprecise probability analysis, the resulting conclusions are in general not more robust than those resulting from a Bayesian analysis, as shown in the following example.

Example 4 *Consider the decision problem of Example 1 in the imprecise probability framework of Example 3. In particular, Figure 1 shows that* $\underline{P}(A) > \underline{P}(B)$ *when* $s = 1$, *while* $\underline{P}(B) > \underline{P}(A)$ *when* $s = 2$. *Hence, the* Γ*-maximin decision would correspond to the first course of action (that is, betting on success) when* $s = 1$, *and to the second course of action (that is, betting on failure) when* $s = 2$. *We would obtain the same decisions if we used the* Γ*-maximax, Hurwicz [2, 22], or interval bound dominance [17] criteria instead of* Γ*-maximin.*

Therefore, in this situation the decision resulting from the imprecise probability approach is not robust, if one of these criteria is used and both $s = 1$ *and* $s = 2$ *are considered as reasonable choices for the parameter* s *of the imprecise Dirichlet model in the case of (almost) no prior information about* θ. *In complete analogy with the Bayesian analysis of Example 1, an answer to this non-robustness issue would be to give more careful consideration to the prior information about* θ, *in order to be able to identify more precisely the prior imprecise probability distribution for* θ.

Other decision criteria, like maximality, E-admissibility, or interval dominance, often do not identify a unique optimal decision, and are perhaps more in keeping with the spirit of imprecise probabilities. When such criteria are used, imprecise probability analyses can be seen as descriptions of the robustness or non-robustness of Bayesian analyses. In fact, if one of these criteria identifies a single optimal decision in an imprecise probability analysis based on a lower prevision $\underline{P}$, then this decision is the unique optimal one in each Bayesian analysis based on a linear prevision P dominating $\underline{P}$ (assuming that in these Bayesian analyses there are optimal decisions). By contrast, the two approaches diverge when the Bayesian analysis is not robust, in the sense that different linear previsions P dominating $\underline{P}$ lead to different optimal decisions. In this case, all these decisions are optimal in the imprecise probability analysis based on $\underline{P}$, when one of the above criteria is used. However, this situation has very different meanings for the two approaches to decision making. In the Bayesian approach the non-robustness issue can be tackled by identifying more precisely the linear prevision P, while in the imprecise probability approach there is not necessarily a more precise lower prevision $\underline{P}$ that would still be a reasonable choice.

Therefore, since the goal of decision making is to select one of the possible decisions, in the imprecise probability approach we often still have to choose one of the optimal decisions, when one of the above criteria is used. This choice can be based on a second decision criterion selected among the ones usually identifying a single optimal decision, like Γ-maximin [25]. However, when such two-stage decision procedures are used in an imprecise probability analysis, the resulting conclusions are in general not more robust than those resulting from a Bayesian analysis, as shown in the following example.

Example 5 *Figure 1 shows that in the decision problem of Example 1, when* $s = 1$ *we have* $E(A) > E(B)$ *if* $t \in (0, 1)$ *is sufficiently large, and* $E(B) > E(A)$ *if* $t \in (0, 1)$ *is sufficiently small. That is, the decision resulting from the Bayesian approach is not robust, if all* $Beta(1, t)$ *distributions with* $t \in (0, 1)$ *are considered as reasonable choices for the prior probability distribution. Therefore, in the imprecise probability framework of Example 3, when* $s = 1$ *both courses of action would correspond to optimal decisions according to the criteria of maximality, E-admissibility, or interval dominance. Exactly the same holds in the case with* $s = 2$. *By contrast, when* $s = 1/3$ *these criteria would lead to a single optimal decision, corresponding to the first course of action (that is, betting on success), since in this case* $\underline{P}(A) > \overline{P}(B)$, *as can be seen in Figure 1. That is, the decision resulting from the Bayesian approach is robust, if only the* $Beta(1/3, t)$ *distributions with* $t \in (0, 1)$ *are considered as reasonable choices for the prior probability distribution.*

However, if the goal of the imprecise probability analysis is decision making (and not the study of the robustness or non-robustness of Bayesian analyses), then when $s = 1$ *or* $s = 2$ *we still have to select one of the two possible decisions. If we choose one of the four criteria considered in Example 4 as the second decision criterion in a two-stage decision procedure, then we obviously obtain the same conclusions as in Example 4.*

Another possibility (besides a second criterion in a two-stage procedure) for choosing a decision when there are multiple optimal decisions, is to select it arbitrarily. Of course, there is no real hope that the resulting decisions can be robust, since arbitrariness is antithetical to robustness. However, one could maintain that such an arbitrary choice cannot be nonrobust, because from the point of view of the decision criterion all optimal decisions are in a certain sense "equivalent". But even from this point of view the decisions resulting from the imprecise probability approach are not robust in general, as shown in the following example.

Example 6 *Consider again the decision problem of Example 1 in the imprecise probability framework of Example 3, with as decision criterion maximality, E-admissibility, or interval dominance. In Example 5 we have seen that in this case the first course of action (that is, betting on success) would correspond to the unique optimal decision when $s = 1/3$, while both courses of action would correspond to optimal decisions when $s = 1$. Hence, if we would choose one of the two optimal decisions arbitrarily when $s = 1$, then we could choose the second course of action (that is, betting on failure), which does not correspond to the single optimal decision when $s = 1/3$.*

Therefore, in this situation the decision resulting from the imprecise probability approach is not robust, if both $s = 1/3$ and $s = 1$ are considered as reasonable choices for the parameter s of the imprecise Dirichlet model. Of course, $s = 1/3$ is not a usual choice for this parameter, but it would suffice to slightly modify the decision problem, in order to obtain that the difference in the decisions is between the cases $s = 1$ and $s = 2$ (instead of $s = 1/3$ and $s = 1$). For instance, it would suffice to consider the decision problem corresponding to choosing the side of a bet with odds of 5 to 2 (instead of 4 to 1) on a success in the next Bernoulli trial, where the total stake is a fixed small amount of money (in this situation, the decision resulting from the Bayesian approach would be the same for both Bayes' and Jeffreys' priors: betting on success).

Hence, in this subsection we have seen that when a decision has to be made, the imprecise probability approach is in general not more robust than the Bayesian one. In particular, the choice of s in the imprecise probability analyses of Examples 4, 5, and 6 does not appear to be "less crucial" than in the Bayesian analysis of Example 1. In this context, it is important to note that the results would remain substantially unchanged if randomized decisions were allowed in these examples. In this case, we would have infinitely many possible decisions, but the (sets of) randomization probabilities of the optimal decisions would still change in a discontinuous way at either $s = 4/3$ or $s = 1/2$ (depending on the example being considered).

3.3 Indecision

As discussed in Subsection 3.2, decision criteria like maximality, E-admissibility, or interval dominance often do not identify a unique optimal decision, when used in an imprecise probability analysis. Instead of choosing a decision from the set of all optimal decisions, the set itself is sometimes considered as the conclusion resulting from the imprecise probability approach [1, 14, 36]. That is, (partial) indecision is sometimes allowed.

In this case, the set of all possible decisions of the original decision problem is practically replaced by its power set (without the empty set). The resulting new decision problem is in a certain sense smoother than the original one, because the indecision about two (originally) possible decisions can be seen as a middle course between them. Therefore, non-robustness issues regarding the new decision problem can be less serious than those regarding the original one. However, the Bayesian approach too can be applied to the new decision problem, as shown in the following example.

Example 7 *In Example 5 we have considered the decision criteria of maximality, E-admissibility, and interval dominance for the decision problem of Example 1, in the imprecise probability framework of Example 3. We have seen that the first course of action (that is, betting on success) would correspond to the unique optimal decision when $s = 1/3$, while both courses of action would correspond to optimal decisions when $s = 1$ or $s = 2$. Hence, if indecision is allowed, then we would stick to the first course of action when $s = 1/3$, but we would have indecision between the two courses of action when $s = 1$ or $s = 2$.*

In order to apply the Bayesian approach when indecision is allowed, we can define the utility C of the indecision between the two courses of action. Assuming risk aversion, this utility must be larger than the utility of choosing one of the two courses of action at random (by tossing a fair coin) [37]: that is, $C > 0$. The choice $C = 1/10$ is plotted in Figure 1: we can see that in this case the decision resulting from the Bayesian approach would still be the first course of action (that is, betting on success) when $(s,t) = (1/3, 1/2)$, and the second course of action (that is, betting on failure) when $(s,t) = (2, 1/2)$, but it would be the indecision between the two courses of action when $(s,t) = (1, 1/2)$.

The new decision problem in Example 7 can be considered as smoother than the original one in Example 1, because in a certain sense there is a new possible choice (the indecision) somewhere in between the two courses of action. In particular, with the new decision problem the choice of s is perhaps “less crucial” than with the original one, but this holds for the Bayesian analysis as well as for the imprecise probability analysis.

Apparently, the imprecise probability approach has the advantage of not needing to define the utilities of the cases of (partial) indecision. However, this appearance can be misleading. First, the definition of these utilities can be avoided in the Bayesian approach too, for instance by replacing the posterior expectations of the utilities of the original decisions with their highest posterior density intervals (for a given probability level), and using interval dominance as a decision criterion. Second, and most important, the definition of the utilities for the cases of (partial) indecision is necessary anyway to evaluate and compare the resulting imprecise probability methods: much work has recently been done in this direction [37]. The trouble is that the imprecise probability methods are obtained on the basis of one decision problem (without utilities for the cases of indecision), and are then evaluated on the basis of another (with utilities for the cases of indecision).

The difficulty in evaluating and comparing imprecise probability methods is strictly related to a fundamental issue in the imprecise probability approach to inference and decision making: the difficulty in comparing models with different degrees of imprecision [30]. The discussion of this issue goes far beyond the scope of the present paper, but it is important to note the connection with the difficulty in the choice of the parameter s of the imprecise Dirichlet model of Example 3, since the degree of imprecision of this model increases with s.

4 Conclusion

Imprecise probability methods are often claimed to be robust, or more robust than Bayesian methods. Sometimes the expression “more robust” is simply used as a synonym for “more imprecise” or “less determinate” [23]. However, this use is misleading, if not wrong. In fact, “more robust” has a positive connotation, which “more imprecise” or “less determinate” do not have, and which derives from its usual interpretation in science and engineering as meaning something like “less sensitive to small changes in the conditions or in the assumptions”.

In particular, in the Bayesian approach to inference and decision making, robustness mainly refers to changes in the choice of prior probability distribution. A Bayesian sensitivity analysis (also called robust Bayesian analysis) is the study of the robustness of the conclusions of a Bayesian analysis. The fact that Bayesian sensitivity analyses are often performed by letting the prior vary in a set of probability distributions can suggest the idea that imprecise probability analyses are robust (since imprecise probability measures can be identified with particular sets of probability measures). In fact, as discussed in Subsection 3.1, imprecise probability analyses can perhaps be considered as more credible than Bayesian ones, and as noted in Subsection 3.2, they can be seen as descriptions of the robustness or nonrobustness of Bayesian analyses, when decision criteria like maximality, E-admissibility, or interval dominance are used. However, the robustness of imprecise probability analyses does not refer to the variability of a (precise) prior in a set of probability distributions, but rather to the variability of the (imprecise) prior in a set of imprecise probability distributions.

Another source of confusion about the robustness of imprecise probability methods (besides the meaning of “robust” in the expression “robust Bayesian analysis”) seems to be the idea that they are allowed to be inconclusive, while Bayesian methods are not. In fact, the Bayesian approach to a particular decision problem is sometimes compared to the imprecise probability approach to a modified version of the decision problem, in which (partial) indecision is allowed. As discussed in Subsection 3.3, the new decision problem is in a certain sense smoother than the original one, and so robustness can be less of an issue. However, both approaches can be applied to both decision problems, and a fair comparison is possible only if they are applied to the same one.

In conclusion, imprecise probability methods are in general not robust, and not more robust than Bayesian methods. The robustness of the imprecise probability approach to inference and decision making can be increased by introducing a second-order possibility distribution, allowing a smoother and more efficient updating rule [9, 10], but this goes beyond the scope of the present paper, and will be the subject of future work.

Acknowledgments

The author wishes to thank Alessio Benavoli and other participants in the session “Imprecise Probability Theories for Statistical Problems: Why? How? When?” at CFE-ERCIM 2012 for stimulating discussions, and the anonymous referees for their valuable comments and suggestions.

References

[1] Antonucci, A., Cattaneo, M., and Corani, G. (2012). Likelihood-based robust classification with Bayesian networks. In *Advances in Computational Intelligence, Part 3*. Springer, 491–500.

[2] Augustin, T. (2002). Expected utility within a generalized concept of probability – a comprehensive framework for decision making under ambiguity. *Stat. Pap.* 43, 5–22.

[3] Augustin, T. (2003). On the suboptimality of the generalized Bayes rule and robust Bayesian procedures from the decision theoretic point of view: A cautionary note on updating imprecise priors. In *ISIPTA '03*. Carleton Scientific, 31–45.

[4] Augustin, T., and Hable, R. (2010). On the impact of robust statistics on imprecise probability models: A review. *Struct. Safety* 32, 358–365.

[5] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lond.* 53, 370–418.

[6] Berger, J. O. (1984). The robust Bayesian viewpoint. In *Robustness of Bayesian Analyses*. North-Holland, 63–124.

[7] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. 2nd edn. Springer.

[8] Bernard, J.-M. (2005). An introduction to the imprecise Dirichlet model for multinomial data. *Int. J. Approx. Reasoning* 39, 123–150.

[9] Cattaneo, M. (2008). Fuzzy probabilities based on the likelihood function. In *Soft Methods for Handling Variability and Imprecision*. Springer, 43–50.

[10] Cattaneo, M. (2009). A generalization of credal networks. In *ISIPTA '09*. SIPTA, 79–88.

[11] Corani, G., and Benavoli, A. (2010). Restricting the IDM for classification. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods*. Springer, 328–337.

[12] Couso, I., Moral, S., and Walley, P. (2000). A survey of concepts of independence for imprecise probabilities. *Risk Decis. Policy* 5, 165–181.

[13] Cozman, F. G. (2005). Graphical models for imprecise probabilities. *Int. J. Approx. Reasoning* 39, 167–184.

[14] De Bock, J., and De Cooman, G. (2011). State sequence prediction in imprecise hidden Markov models. In *ISIPTA '11*. SIPTA, 159–168.

[15] De Cooman, G., Quaeghebeur, E., and Miranda, E. (2009). Exchangeable lower previsions. *Bernoulli* 15, 721–735.

[16] De Finetti, B. (1975). *Theory of Probability.* Vol. 2. Wiley.

[17] Destercke, S. (2010). A decision rule for imprecise probabilities based on pair-wise comparison of expectation bounds. In *Combining Soft Computing and Statistical Methods in Data Analysis*. Springer, 189–197.

[18] Elga, A. (2010). Subjective probabilities should be sharp. *Philos. Imprint* 10.

[19] Huber, P. J., and Ronchetti, E. M. (2009). *Robust Statistics*. 2nd edn. Wiley.

[20] Huber, P. J., and Strassen, V. (1973). Minimax tests and the Neyman–Pearson lemma for capacities. *Ann. Stat.* 1, 251–263.

[21] Huntley, N., and Troffaes, M. C. M. (2009). Characterizing factuality in normal form sequential decision making. In *ISIPTA '09*. SIPTA, 239–248.

[22] Hurwicz, L. (1951). Some specification problems and applications to econometric models. *Econometrica* 19, 343–344.

[23] Hutter, M. (2003). Robust estimators under the imprecise Dirichlet model. In *ISIPTA '03*. Carleton Scientific, 274–289.

[24] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond., Ser. A* 186, 453–461.

[25] Levi, I. (1974). On indeterminate probabilities. *J. Philos.* 71, 391–418.

[26] Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer.

[27] Pericchi, L. R., and Walley, P. (1991). Robust Bayesian credible intervals and prior ignorance. *Int. Stat. Rev.* 58, 1–23.

[28] Piatti, A., Zaffalon, M., and Trojani, F. (2005). Limits of learning from imperfect observations under prior ignorance: the case of the imprecise Dirichlet model. In *ISIPTA '05*. SIPTA, 276–286.

[29] Piatti, A., Zaffalon, M., Trojani, F., and Hutter, M. (2009). Limits of learning about a categorical latent variable under prior near-ignorance. *Int. J. Approx. Reasoning* 50, 597–611.

[30] Seidenfeld, T., Schervish, M. J., and Kadane, J. B. (2011). Forecasting with imprecise probabilities. In *ISIPTA '11*. SIPTA, 317–326.

[31] Troffaes, M. C. M. (2007). Decision making under uncertainty using imprecise probabilities. *Int. J. Approx. Reasoning* 45, 17–29.

[32] Troffaes, M. C. M., and Hable, R. (2011). Robustness of natural extension. In *ISIPTA '11*. SIPTA, 361–370.

[33] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall.

[34] Walley, P. (1996). Inferences from multinomial data: Learning about a bag of marbles. *J. R. Stat. Soc., Ser. B* 58, 3–57.

[35] Williams, P. (2007). Notes on conditional previsions. *Int. J. Approx. Reasoning* 44, 366–383.

[36] Zaffalon, M. (2001). Statistical inference of the naive credal classifier. In *ISIPTA '01*. Shaker, 384–393.

[37] Zaffalon, M., Corani, G., and Mauá, D. (2012). Evaluating credal classifiers by utility-discounted predictive accuracy. *Int. J. Approx. Reasoning* 53, 1282–1301.

An approach to uncertainty in probabilistic assignments with application to vibro-acoustic problems

Alice Cicirello
Department of Engineering, University of Cambridge
ac685@cam.ac.uk

Robin S. Langley
Department of Engineering, University of Cambridge
rsl21@cam.ac.uk

Abstract

In this paper a novel imprecise probability description is applied to vibro-acoustic problems in engineering. Frequently little data is available concerning the variability of the key input parameters required for a predictive analysis. This has led to widespread use of several uncertainty descriptions. The hybrid Finite Element/Statistical Energy Analysis (FE/SEA) approach to the analysis of vibro-acoustic systems is based on subdividing a system into: (i) SEA components which incorporate a non-parametric model of uncertainty and (ii) FE components with parametric uncertainty. This approach, combined with the Laplace asymptotic method, allows the evaluation of the failure probability. A novel strategy for establishing bounds on the failure probability when an imprecise probability model (based on expressing the probability density function of a random variable in the form of a maximum entropy distribution with bounded parameters) is employed is presented. The approach is illustrated by application to a built-up plate system.

Keywords. Uncertainties in probabilistic assignments, hybrid FE/SEA method, reliability analysis, parametric and non-parametric uncertainty models, maximum entropy distribution, vibro-acoustic analysis.

1 Introduction

In engineering problems it is frequently the case that little data is available concerning the variability of the key input parameters (geometry, material properties, and boundary conditions) required for a predictive analysis, and yet an engineering assessment of a design must nonetheless be performed. This topic has been the subject of much recent research, and various analytical and computational approaches have been proposed (for example, [1-9]). Such methods require some description of the underlying uncertainties (for example, uncertainties in material properties, loading conditions, and fabrication details) which could be probabilistic (parametric [1-4], non-parametric [5-7] or a combination of both [8,9]) or non-probabilistic [1,4].

Reliability methods aim to estimate the probability that design targets will be met [10,11]; this probability is referred to as the reliability of the system. These methods are often based on a parametric probabilistic description of the uncertain parameters of the system and rely on the assumption that the statistical distributions (i.e. probability density function (pdf)) of these parameters are precisely known [12]. The parametric probabilistic description requires a large amount of empirical data if the pdf is constructed using a frequentist view. Alternatively the pdf may be interpreted as a statement of belief based on expert opinion, as in the subjective approach to probability theory [13]. The more common frequentist approach is concerned with the outcome of experiments performed (hypothetically or in reality) on large ensembles of systems; these ensembles may either be real (for example cars from a production line), or virtual but realizable in principle (such as an ensemble of manufactured satellites, when only one satellite may actually be built). In contrast, with the subjective approach, no ensemble is necessarily involved. The frequentist and subjective views can be roughly aligned to the notions of aleatory and epistemic uncertainty; aleatory uncertainty is an irreducible uncertainty associated with an inherent variability of the properties of the system, while epistemic uncertainty is reducible, being associated with a lack of knowledge of the analyst with respect to the system's properties which are fixed [4]. Clearly, the interpretation employed for defining the pdf of the uncertain parameters will affect the interpretation of the results obtained with a predictive analysis.

In practice, only a limited amount of data may be available and therefore it is often difficult to identify the form of the distribution of the random variable and/or the parameters of the distribution. Moreover, the analyst may have uncertainties in belief, meaning that the specified pdf is itself subject to doubt. Using a pdf which differs from the actual one can significantly affect the prediction of the system performance with respect to safety, quality, design or cost constraints [12,14]. One way around this difficulty is to employ imprecise probability descriptions in the reliability assessment in order to establish bounds on the failure probability (that is the probability that the response exceeds a critical level). These bounds allow: (i) the evaluation of the sensitivity of the system response to the uncertainty of the system parameters; (ii) the identification of the worst case scenario (the highest failure probability expected). Many reliability approaches which includes imprecise probability descriptions have been developed in the past years, among which there are: (i) First Order Reliability Method (FORM) [10] approaches which employ pdfs with one [15] or two [16] bounded parameters (mean, variance or another distribution parameter), [15,16]; (ii) Dempster-Shafer theory (DST) [17,18] and P-box models of imprecise probabilities [19-21] applied to reliability analysis [22-25]; (iii) reliability analysis with random sets [26,27]; (iv) reliability assessment by means of Fuzzy Probabilities [4,28]; (v) Reliability models which account for the lack of information about the independence of the stress and strength, and about the parameters of each pdf [29]; (vi) reliability models based on imprecise Bayesian inference models [30]; (vii) Interval importance sampling methods combined with specified pdf with bounded parameters [31]. However, the application of these approaches is often limited to simple models, mainly because of the computational burden associated to the propagation of the imprecise probability description.

In automotive and aerospace industries there are design requirements to ensure vibro-acoustic performance is met. Vibro-acoustic problems usually involve a very broad frequency range due to the broadband nature of the loadings acting on the system. Broadly speaking three frequency ranges can be identified: low-, mid- and high-frequency ranges. In the low-frequency range the length scale of deformation of the system components is relatively long with respect to their overall dimension so that: (i) few degrees of freedom are required to model their dynamic behavior; (ii) the system response is insensitive to small changes in the system properties. The Finite Element method (FE) [32] is a well-established deterministic technique for acoustics and vibration analysis in the low-frequency range. In the high-frequency range, instead, the length scale of deformation is comparable to small manufacturing imperfections producing high sensitivity to uncertainty and requiring a large number of degrees of freedom for capturing the components' dynamic behavior. An alternative to FE is to employ Statistical Energy Analysis (SEA) [6,33], a probabilistic technique which was developed specifically to deal with high frequency vibration. In SEA the system is modeled as an assembly of subsystems, whose response is described by their vibrational energy (defined as twice the time-averaged kinetic energy). The number of degrees of freedom employed is drastically reduced compared to the FE approach, since a single degree of freedom SEA subsystem might replace thousand of finite element nodes. The interaction between the SEA subsystems is described using the principle of conservation of energy flow, and this leads to a set of equations that can be solved to yield the subsystem energies. This method can predict both the ensemble average vibrational energy levels [33] (averaged across an ensemble of nominally identical structures) and the ensemble variance of the energy levels [6]. The application of this approach is limited to high frequency because of its underlying assumptions (i.e. each structural component is sufficiently random and that the coupling between subsystems is sufficiently weak [6]). Between the respective ranges of validity of FE and SEA there is a *mid-frequency* region and much research effort has been directed at the development of efficient analytical methods that can be applied in this range. One such method is the hybrid FE/SEA method [7,34]. This approach is based on subdividing a system into SEA components (which incorporate a non-parametric probabilistic model of uncertainty), and deterministic FE components. This partition leads to a large reduction of the number of degrees of freedom employed in the model and a large gain in numerical efficiency. Moreover the method enables the prediction of the mean and variance of the response (such as the energy response of a SEA subsystem or the mean squared amplitude of the finite element degrees of freedom) over a collection of systems with random SEA subsystems properties [7,34] without employing Monte Carlo simulations. The hybrid FE/SEA method has been recently generalized by introducing parametric uncertainty into the FE components [8] in order to provide an enhanced description of those components which may contain a degree of randomness, but cannot be appropriately modeled as SEA subsystems. The vibro-acoustic performance of a complex system in a broad frequency range can be established by applying the hybrid FE/SEA method in combination with the Laplace's method (hybrid FE/SEA + Laplace) [35]. With this approach both parametric and non-parametric probabilistic uncertainty models are employed and the failure probability over the combined ensembles of uncertainty can be assessed. This approach is enhanced in this paper by considering a system with uncertain properties modeled with non-parametric, parametric and also imprecise parametric probabilistic descriptions in order to account for those input parameters of the FE components which are imprecisely known. In particular, the hybrid FE/SEA + Laplace is extended in this paper by employing a recently developed model of imprecise probability [36] in order to establish bounds on the failure probability. The imprecise model employed is based on expressing the probability density function of a random variable in the form of a maximum entropy distribution with bounded parameters [36]. This parametric probabilistic uncertainty model will be described in more details in Section 2. The hybrid FE/SEA + Laplace approach will be summarized in Section 3. In Section 4 an efficient approach for establishing bounds on the failure probability is presented. The method is illustrated by application to a built-up plate system in Section 5.

2 Probability Density Function with Bounded Parameters

In this Section a recently developed parametric model of uncertainty which admits uncertainty in the probabilistic assignments is described [36]. This uncertainty model requires as input bounded statistical expectations of specified functions of the random variable and it can be used to describe both aleatory and epistemic uncertainties. The uncertainty model is briefly described in Subsection 2.1. In Subsection 2.2, a procedure for treating the bounded statistical expectations is summarised.

2.1 Basic Concepts

The model of uncertainty is based on considering that the pdf of a random variable x itself is subject to doubt. The pdf is expressed as the exponential of a series expansion, but the parameters within this model, the so-called basic variables, are allowed to have bounded description [36]:

$$p(x|\mathbf{a} \in \mathrm{S}) = \exp\left[\sum_{j=1}^{n} a_j f_j(x)\right]. \tag{1}$$

Eq. (1) represents a family of distributions defined over the set of basic variables $\mathbf{a}$ (which has entries a_j with $j = 2,3...,n$) that lie within an admissible region S. A "basic variable" is defined here as one which can have any possible pdf within certain bounds, including the extreme case of a delta function at any point between the bounds. If a parameter is not "basic", then its pdf can be expressed in terms of the basic parameters, and thus only this type of parameter is considered in what follows. The admissible region S can be an interval, a convex region, etc. The term $f_j(x)$ is a specified function of the uncertain variable, such that $f_1(x) = 1$. The coefficient a_1 is dependent on the bounded basic variables a_j and it is chosen to satisfy the normalisation condition.
Eq. (1) describes a single distribution when the basic variables have fixed values, and accounts for a more general description (a set of pdfs) when these parameters are bounded. In particular, for fixed basic variables, the pdf expression corresponds to the maximum entropy distribution [13] that arises from specifying the expected values $\mathrm{E}\left[f_j(x)\right]$, where the basic variables are replaced by the Lagrange multipliers (which are constant values). It can be therefore argued that Eq. (1) represents a family of *maximum entropy continuous distributions*. When the constraints are expressed in terms of statistical expectation inequality constraints, such as:

$$v_{j,\min} \le v_j = \mathrm{E}\left[f_j(x)\right] = \int f_j(x)\, p(x|\mathbf{a})\,\mathrm{d}x \le v_{j,\max}, \quad j = 2,3,...,n \tag{2}$$

where $v_{j,\min}$ and $v_{j,\max}$ are the lower and upper bound on the jth statistical expectation v_j, within a class of distribution (for example, polynomial distributions, maximum entropy distribution, etc.), there are many distributions which are consistent with the statistical expectation inequality constraints. The Principle of Maximum Entropy (MAXENT) selects, among the class of maximum entropy distributions, the distribution with the largest entropy [37]. The proposed approach, instead, constructs a family of maximum entropy distributions consistent with the statistical expectation inequality constraints and selects, among this family of pdfs, the pdf which maximises (or equivalently minimises) a specified engineering metric (for example, the probability of exceeding a specified limit value, the probability of being within a certain region). This pdf is potentially different from the pdf which maximises the entropy (which can be recovered as well); therefore the proposed approach is more useful from an engineering point of view. This aspect of the approach will be illustrated by a numerical application in Section 5 of this paper.
The inequality constraints on the statistical expectation of the uncertain variable may arise by analysing a small data set or can be provided by an expert who may prefer to assign bounds rather than specifying a single value. If $f_j(x) = x$ then the inequality constraints are specified on the mean value, alternatively if $f_j(x) = x^2$ they are specified on the second moment. $f_j(x)$ can be also defined as an interval of possible values that the uncertain variable may take, i.e. $f_j(x) = [b,c]$; in this case the constraints corresponds to the probability of finding the random variable within those bounds.
The family of pdfs defined in Eq. (1) is constructed as follows:

1. The form of the pdf which maximises the entropy is computed, as for the maximum entropy approach, by using the Lagrange multipliers method.
2. The Lagrange multipliers are substituted by the basics variables $\mathbf{a}$.
3. The bounds on the statistical expectations of the uncertain variable are used to establish bounds on the basic variables.

A procedure for obtaining an approximate mapping of the basic variables domain (a-domain) starting from a bounded description of the statistical expectations (m-domain) of the uncertain variable [36,38] is summarized in the next Subsection.

2.2 Bounds Conversion

Consider the case for which two statistical expectations of the uncertain variable x lie within a rectangle, as described in Figure 1.

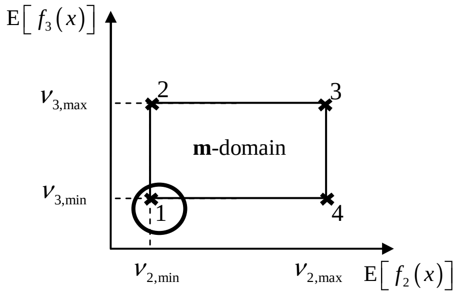


Figure 1: Moment domain (m-domain).

The first step of the approach requires the evaluation of the maximum entropy distribution, which for the present case take the form

$$p(x|\mathbf{a}) = \exp[a_1 + a_2 f_2(x) + a_3 f_3(x)]. \tag{3}$$

In principle, each point of the basic variables domain (a-domain), which is depicted in Figure 2, can be evaluated by solving a set of two non-linear equations in terms of the statistical expectations of the random variable.

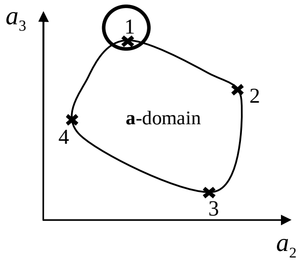


Figure 2: Basic variables domain (a-domain).

For example, point 1 of the m-domain can be mapped in the corresponding point 1 of the a-domain by solving:

$$\begin{cases} \int f_2(x)\exp\left[a_1 + a_2 f_2(x) + a_3 f_3(x)\right]\mathrm{d}x = v_{2,\min} \\ \int f_3(x)\exp\left[a_1 + a_2 f_2(x) + a_3 f_3(x)\right]\mathrm{d}x = v_{3,\min} \end{cases} \tag{4}$$

where a_2 and a_3 are the unknown coefficients, and a_1 is chosen to satisfy the normalisation condition.

In practice, considering enough points along the edges of the m-domain would allow a good approximation of the shape of the a-domain to be obtained, reducing the number of sets of equations to be solved. The problem is that, even for a simple problem (like the 2D case depicted in Figure 1), the solution of each set of non-linear equations can be time consuming and convergence problems may occur.

An approximate mapping of the a-domain can be obtained by [36,38]:

I. Evaluating the mid-points of the surfaces of the hypercube defining the m-domain ($v_j^* = \mathrm{E}\left[f_j^*(x)\right] = \left(v_{j,\max} - v_{j,\min}\right)/2$).
II. Estimate the corresponding point a^* solving a set of non-linear equations for the mid-point of the m-domain.
III. Each point of the a-domain is then calculated by using an approximate expression of the sth moment:

$$\begin{aligned} v_s = v_s^* &+ \sum_{j=2}^{n} c_j^{s*}\left(a_j - a_j^*\right) \\ &+ \frac{1}{2}\sum_{j=2}^{n}\sum_{k=2}^{n} c_{jk}^{s*}\left(a_j - a_j^*\right)\left(a_k - a_k^*\right), \end{aligned} \tag{5}$$

where:

$$c_j^{s*} = \mathrm{E}\left\{\left(f_s(x) - \mathrm{E}\left[f_s^*(x)\right]\right)\left(f_j(x) - \mathrm{E}\left[f_j^*(x)\right]\right)\right\}, \tag{6}$$

$$c_{jk}^{s*} = \mathrm{E}\begin{Bmatrix} \left(f_s(x) - \mathrm{E}\left[f_s^*(x)\right]\right)\left(f_j(x) - \mathrm{E}\left[f_j^*(x)\right]\right) \\ \times\left(f_k(x) - \mathrm{E}\left[f_k^*(x)\right]\right) \end{Bmatrix}. \tag{7}$$

This approach is expected to yield less accurate results when the variation of the sth moment value with respect to the mid-point moment domain value becomes large.

3 The Hybrid FE/SEA Method Combined with the Laplace Asymptotic Method

In this Section the hybrid FE/SEA approach and its combination with the Laplace asymptotic method are briefly reviewed.

3.1 Basic Concepts

The hybrid Finite Element/Statistical Energy Analysis (FE/SEA) method [7,34] is a vibro-acoustic analysis technique which combines the strength of a well established low-frequency deterministic technique, the Finite Element method (FE) [32], with a high-frequency probabilistic method, the Statistical Energy Analysis method (SEA) [6,33], by means of the diffuse field reciprocity relation [39,40]. With this approach, within the frequency range of interest of the problem on hand, a complex system is considered as an assembly of (i) components with very few local modes, collectively called the "master" system and modelled by using FE; and (ii) components with many local modes, called "subsystems", which are modelled with SEA, and it is assumed that all the SEA subsystems are coupled exclusively through the master system. For example, a generic class of engineering systems characterised by thin panels coupled through stiff structural components is often encountered in aerospace structures, where a frame is coupled with a skin panel, or in automotive structures, where the frame of the car is coupled to the roof panel and window panel. Within the hybrid FE/SEA modelling strategy, the panels would be modelled as SEA subsystems, and the stiff components would be modelled using FE. The response of the master system is described by a set of nodal degrees of freedom $\mathbf{q}$, and the response of the SEA subsystems is described by a set of vibrational energies $\mathbf{E}$ (defined as twice the time-averaged kinetic energy).

The properties of the hybrid FE/SEA model components (such as density, Young's modulus, geometry, etc.) are represented by two groups of parameters to distinguish different models of uncertainty [8]: the master system properties are represented by a set of parameters $\mathbf{b}$, while the properties of the SEA subsystem are represented by a set of parameters $\mathbf{s}$. The effect of the uncertain parameters $\mathbf{s}$ is accounted for via a non-parametric statistical approach based on the fact that at high frequency the statistics of the natural frequencies and mode shapes of the subsystems can approach certain universal distributions, regardless of the detailed nature of the underlying uncertainty [7,8,40]. The effect of the uncertain parameters $\mathbf{b}$ is accounted for by a probabilistic parametric uncertainty model [8]. The system is therefore varying over two ensembles: a non-parametric ensemble (a collection of systems with random subsystem properties) and a parametric ensemble (a collection of systems with random master system properties).

For fixed master system properties, the hybrid FE/SEA method enables the calculation of the conditional non-parametric ensemble average $\mu_j(\mathbf{b})$ and ensemble variance $\sigma_j^2(\mathbf{b})$ of a response variable w (which can be the vibrational energy of the SEA subsystem j, or the cross spectrum of the finite element degrees of freedom) [7,34]. The ensemble is non-parametric in the sense that the details of the parameters $\mathbf{s}$ are never considered in the model, but rather the Gaussian Orthogonal Ensemble (GOE) is used to described the statistics of the subsystem natural frequencies and mode shapes [7,40]. This approach obviates the need for any detailed knowledge of the variability or uncertainty of the parameters $\mathbf{s}$ and does not require Monte Carlo Simulations to be performed to propagate the uncertainty. The equations necessary for the evaluation of $\mu_j(\mathbf{b})$ are reviewed in the following Subsection.

3.2 The Hybrid FE/SEA Equations for Fixed FE Properties

The hybrid FE/SEA equations for evaluating the ensemble average response ($\mu_j(\mathbf{b})$) at the excitation frequency ω are [34]:

a) Subsystem energy balance equations

$$\omega(\eta_j+\eta_{d,j})E_j+\sum_k \omega\eta_{jk}n_j(E_j/n_j-E_k/n_k)=P_{in,j}^{ext}+P_{in,j}, \quad (8)$$

where η_j is the damping loss factor of the subsystem j, $\eta_{d,j}$ is an additional loss factor on the subsystem j due to the energy dissipated in the FE components, η_{jk} is the coupling loss factor between subsystem j and subsystem k, n_j is the modal density of subsystem j (which is defined as the average number the average number of natural frequencies within a unit frequency band), E_j is the ensemble average vibrational energy of subsystem j, $P_{in,j}^{ext}$ is the external power input to the subsystem arising from the loads acting on the master system and $P_{in,j}$ is the power input arising from external loads directly applied to the subsystem j.

Eq. (8) states that the power dissipated through damping ($\omega(\eta_j+\eta_{d,j})E_j$) plus the net power transmitted to other subsystems ($\sum_k \omega\eta_{jk}n_j(E_j/n_j-E_k/n_k)$) is balanced by the power input to the subsystem ($P_{in,j}^{ext}+P_{in,j}$), and it is based on the assumption that the power transmitted is proportional to the difference of the average modal energies (defined as E_j/n_j) of the coupled subsystems. Eq. (8) has the same form as the standard SEA equations [33], but also contains two additional terms relating to: (i) the contribution of the master system to the power input $P_{in,j}^{ext}$, and (ii) the power dissipated in the master system, $\omega\eta_{d,j}E_j$. These two terms can be expressed in terms of: (i) the total dynamic stiffness matrix $\mathbf{D}_{tot}=\sum_k \mathbf{D}_{dir}^{(k)}+\mathbf{D}_d$, where $\mathbf{D}_d$ is the dynamic stiffness matrix associated with the FE model ($\mathbf{D}_d=-\omega^2\mathbf{M}+i\omega\mathbf{C}+\mathbf{K}$, where $\mathbf{M}$, $\mathbf{C}$ and $\mathbf{K}$ are respectively the FE component mass, damping and stiffness matrices), and $\mathbf{D}_{dir}^{(k)}$ is the so-called direct field dynamic stiffness matrix for subsystem k which can be computed using various techniques [34]; (ii) the cross-spectral matrix of the loading applied directly to the master system $\mathbf{S}_{ff}=\left[\mathbf{ff}^{*\mathrm{T}}\right]$, so that

$$\omega\eta_{d,j}=\left(\frac{2\,\alpha_k}{\pi n_j}\right)\sum_{rs}\mathrm{Im}\left\{D_{d,rs}\right\}\left(\mathbf{D}_{tot}^{-1}\,\mathrm{Im}\left\{\mathbf{D}_{dir}^{(j)}\right\}\mathbf{D}_{tot}^{-1*\mathrm{T}}\right)_{rs}, \quad (9)$$

$$P_{in,j}^{ext}=(\omega/2)\sum_{rs}\mathrm{Im}\left\{D_{dir,rs}^{(j)}\right\}\left(\mathbf{D}_{tot}^{-1}\mathbf{S}_{ff}\mathbf{D}_{tot}^{-1*\mathrm{T}}\right)_{rs}, \quad (10)$$

where the superscript $*$ indicates the complex conjugate, the superscript T denotes the transpose, Im represents the imaginary part of the matrix, and α_k is a factor which takes into account the fact that the subsystem wave field may not be perfectly diffuse [7]. Generally α_k is equal to 1 when the subsystem wave field is diffuse, and close to 2 when the subsystem is excited predominantly by motion of the master system [7].

Three of the remaining terms in Eq. (8), specifically η_j, n_j, and $P_{in,j}$, are evaluated by using standard SEA procedures [33], while the coupling loss factors are expressed analytically as a function of the total dynamic stiffness matrix in the form [34]

$$\omega\eta_{jk}n_j = \left(\frac{2\ \alpha_k}{\pi}\right)\sum_{rs}\mathrm{Im}\left\{D_{dir,rs}^{(j)}\right\}\left(\mathbf{D}_{tot}^{-1}\,\mathrm{Im}\left\{\mathbf{D}_{dir}^{(k)}\right\}\mathbf{D}_{tot}^{-1*\mathrm{T}}\right)_{rs}. \quad (11)$$

Writing Eq. (8) for each subsystem leads to a set of equations that can be solved to yield the ensemble average vibrational energy E_j of each subsystem. This set of E_j is then used to calculate the average response of the master system.

b) Master system response equation

$$\mathbf{S}_{qq} = \mathbf{D}_{tot}^{-1}\left[\mathbf{S}_{ff} + \sum_k\left(\frac{4\alpha_k E_k}{\omega\pi n_k}\right)\mathrm{Im}\left\{\mathbf{D}_{dir}^{(k)}\right\}\right]\mathbf{D}_{tot}^{-1*\mathrm{T}}, \quad (12)$$

here $\mathbf{S}_{qq}$ is the cross-spectrum of the response of the master system (averaged over the non-parametric ensemble), and the two terms on the right-hand side correspond to the forcing arising from external excitation (expressed in terms of the cross spectrum of the forces, $\mathbf{S}_{ff}$) and the forcing arising from the subsystems, as yielded by the diffuse field reciprocity relation [39,40].

By using the hybrid FE/SEA variance theory [7] it is also possible to estimate the covariance of the subsystem energies ($\mathrm{Cov}\left[\bar{E}_j, \bar{E}_k\right]$, where $\bar{E}_j = E_j / n_j$) and the variance of the cross-spectral matrix of the response of the master system ($\mathrm{Var}\left[S_{qq}\right]$) over the non-parametric ensemble, which are indicated in what follows as $\sigma_j^2(\mathbf{b})$. These equations are required in the following developments of the theory for estimating the probability density of the general response variable, but for brevity they will not be included in this paper. The reader is referred to the paper by Langley and Cotoni [7] where their full derivations can also be found.

3.3 Hybrid FE/SEA + Laplace

The hybrid FE/SEA method has been recently combined with the Laplace's method [35] (a technique used to approximate integrals expressed in the Laplace form [41]) in order to establish the failure probability of a complex built-up system with input parameters described by a combination of parametric and non-parametric probabilistic uncertainty models.

The failure probability is defined as the probability that a deterministic limit value w_0 is reached and/or exceeded by the general response variable $w = w(\mathbf{b},\mathbf{s})$ (which can be the vibrational energy of subsystem j, or the cross spectrum response of the master system). This condition can be expressed as:

$$P_f = \mathrm{P}\left[w \geq w_0\right] = \int_{w_0}^{\infty} p(w)\,\mathrm{d}\,w. \quad (13)$$

The application of the hybrid FE/SEA method for fixed $\mathbf{b}$ yields the conditional non-parametric ensemble mean and variance of the response ($\mu_j(\mathbf{b})$ and $\sigma_j^2(\mathbf{b})$, respectively), which can then be used to evaluate the probability density function of the general response variable conditional on $\mathbf{b}$, $p(w|\mathbf{b})$; for example, the pdf of the non-parametric ensemble vibrational energy is usually log-normal, and therefore the mean and variance yield the complete pdf [6,8,42]. Eq. (13) can be conveniently rewritten in terms of $p(w|\mathbf{b})$:

$$P_f = \int_{w_0}^{\infty}\int_{\mathbf{b}} p(w|\mathbf{b})\,p(\mathbf{b})\,\mathrm{d}\,\mathbf{b}\,\mathrm{d}\,w. \quad (14)$$

The failure probability conditional on $\mathbf{b}$ can be now defined as:

$$P_f(\mathbf{b}) = \int_{u_0}^{\infty} p(w|\mathbf{b})\,\mathrm{d}\,w; \quad (15)$$

and therefore Eq. (13) can be written as an unbounded integral:

$$P_f = \int_{\mathbf{b}} P_f(\mathbf{b})\,p(\mathbf{b})\,\mathrm{d}\mathbf{b}. \quad (16)$$

The integral in Eq. (16) can be evaluated numerically by considering a grid of integration points (direct integration), although this approach is unpractical when a large number of uncertain input parameters is considered [10]. Alternatively, an approximate evaluation of this integral can be obtained by applying the Laplace's method to the integral expressed in the form $\int_{\mathbf{b}}\exp\left[\ln\left[P_f(\mathbf{b})\,p(\mathbf{b})\right]\right]\mathrm{d}\mathbf{b}$. In particular, the failure probability can be approximated as [35]:

$$P_f \approx \sum_{j=1}^{\psi} P_f\left(\mathbf{b}_j^*\right)p\left(\mathbf{b}_j^*\right)(2\pi)^{d/2}\det\left[\mathbf{H}\left(\mathbf{b}_j^*\right)\right]^{-1/2}, \quad (17)$$

where ψ stands for the number of local maxima of $\ln\left[P_f(\mathbf{b})\,p(\mathbf{b})\right]$ at locations $\mathbf{b}_j^*$, d is the dimension of the set of random variables $\mathbf{b}$ involved in the problem, $\det[\ ]$ is the matrix determinant operator and $\mathbf{H}\left(\mathbf{b}_j^*\right)$ is the Hessian matrix whose elements are given by

$$H_{ij}(\mathbf{b}) = -\frac{\partial^2}{\partial b_i \partial b_j}\left[\ln\left(P_f(\mathbf{b})\,p(\mathbf{b})\right)\right]. \quad (18)$$

This approximation (Eq. (17)) corresponds to replacing the integrand function with an n-dimensional Gaussian distribution with mean equal to $\mathbf{b}_j^*$ and covariance matrix equal to the inverse of $\mathbf{H}\left(\mathbf{b}_j^*\right)$. Conditions for the accuracy of Eq. (17) are discussed in references [41,43].

4 Bounds on the Failure Probability

4.1 Hybrid FE/SEA + Laplace Using Imprecise Probabilities

The hybrid FE/SEA + Laplace approach can be generalised considering the case in which the uncertain input parameters $\mathbf{b}$ of the FE components can be subdivided into two groups: (i) a set of parameters $\hat{\mathbf{b}}$ described by a specified probability density function $p\left(\hat{\mathbf{b}}\right)$; and (ii) a set of parameters $\mathbf{b}_{imp}$ imprecisely known described in terms of bounded statistical expectations (derived from small data set or specified by an analyst). The second set of parameters $\mathbf{b}_{imp}$ can be modelled by using the imprecise probability uncertainty model presented in Section 2. With this approach, the joint pdf of the random variables $p\left(\mathbf{b}_{imp}\middle|\mathbf{a}\right)$ is expressed in the form of a maximum entropy distribution (Eq. (1)), and the bounds on the statistical expectations are converted into bounds on the so-called basic variables $\mathbf{a}$ (as described in Section 2). If these basic variables are taken to have fixed values $\mathbf{a}$, then a single pdf $p\left(\mathbf{b}_{imp}\middle|\mathbf{a}\right)$ is identified.
According to Eq. (17), the failure probability conditional on the basic variables is then given by

$$\begin{aligned} P_f\left(\mathbf{a}\right) &= \mathrm{P}\left[w\left(\hat{\mathbf{b}},\mathbf{b}_{imp}\middle|\mathbf{a},\mathbf{s}\right)\ge w_0\right] \\ &= \int_{\mathbf{b}} P_f\left(\hat{\mathbf{b}},\mathbf{b}_{imp}\middle|\mathbf{a}\right) p\left(\hat{\mathbf{b}}\right) p\left(\mathbf{b}_{imp}\middle|\mathbf{a}\right) \mathrm{d}\mathbf{b}. \end{aligned} \tag{19}$$

where $P_f\left(\hat{\mathbf{b}},\mathbf{b}_{imp}\middle|\mathbf{a}\right)$ is the failure probability conditional on $\left(\hat{\mathbf{b}},\mathbf{b}_{imp}\middle|\mathbf{a}\right)$.

The hybrid FE/SEA + Laplace approach [35] can be then employed to estimate the failure probability as:

$$\begin{aligned} P_f\left(\mathbf{a}\right) \approx & \sum_{j=1}^{\psi} P_f\left(\hat{\mathbf{b}}_j^*,\mathbf{b}_{imp,j}^*\middle|\mathbf{a}\right) p\left(\hat{\mathbf{b}}_j^*\right) p\left(\mathbf{b}_{imp,j}^*\middle|\mathbf{a}\right) \\ & \times\left(2\pi\right)^{d/2} \det\left[\mathbf{H}\left(\hat{\mathbf{b}}_j^*,\mathbf{b}_{imp,j}^*\middle|\mathbf{a}\right)\right]^{-1/2} \end{aligned} \tag{20}$$

The evaluation of the failure probability requires:

I. Evaluation of $p\left(w\middle|\left(\hat{\mathbf{b}}_j,\mathbf{b}_{imp,j}\middle|\mathbf{a}\right)\right)$ by using the results yielded by the hybrid FE/SEA method.
II. Calculation of $P_f\left(\hat{\mathbf{b}}_j,\mathbf{b}_{imp,j}\middle|\mathbf{a}\right)$ by using Eq. (15).
III. Evaluation of $\left(\hat{\mathbf{b}}_j^*,\mathbf{b}_{imp,j}^*\middle|\mathbf{a}\right)$ by applying a standard unconstrained minimization algorithm to $-\ln\left[P_f\left(\hat{\mathbf{b}}_j^*,\mathbf{b}_{imp,j}^*\middle|\mathbf{a}\right)p\left(\hat{\mathbf{b}}_j^*\right)p\left(\mathbf{b}_{imp,j}^*\middle|\mathbf{a}\right)\right]$.
IV. Evaluation of the Hessian matrix.

If the basic variables are allowed to vary, a family of response pdfs is obtained and the bounds on the failure probability can be established as

$$\min_{\mathrm{S}}\left(P_f\left(\mathbf{a}\right)\right)\le P_f\le\max_{\mathrm{S}}\left(P_f\left(\mathbf{a}\right)\right). \tag{21}$$

These bounds give an indication of the sensitivity of the system reliability with respect to the uncertainty on the pdf of the input parameters. If the bounds are wide, the uncertainty in the input parameter description is significantly affecting the system reliability. On the other hand, if the bounds are narrow then the system reliability is little affected by the uncertainty in the pdf of the uncertain parameters.

4.2 Steps for Implementing the Proposed Approach

The reliability analysis can be summarised as follows:

I. The system is subdivided into: (i) FE components with uncertain properties $\mathbf{b}$; and (ii) SEA components with uncertain properties $\mathbf{s}$.
II. The effect of the uncertain parameters $\mathbf{s}$ of the SEA components is accounted for by using non-parametric statistical methods.
III. The uncertain parameters of the FE components $\mathbf{b}$ are partitioned into two sets of parameters: (i) $\hat{\mathbf{b}}$ modelled by using a specified pdf $p\left(\hat{\mathbf{b}}\right)$; and (ii) $\mathbf{b}_{imp}$ modelled via the imprecise probability model $p\left(\mathbf{b}_{imp}\middle|\mathbf{a}\right)$ where $\mathbf{a}$ are the basic variables which define the family of pdfs (Eq. (1)).
IV. The admissible region of the basic variable a-domain) associated to the random variables $\mathbf{b}_{imp}$ (obtained as described in Section 2 from the knowledge of the bounds on statistical expectations) is overlaid with a grid of points. This grid is chosen in order to capture enough sampled points within and along the a-domain.
V. For each sampled point of this grid, the corresponding a_1 is calculated via normalization. The set of basic variables associated to each point of the domain identifies a single $p\left(\mathbf{b}_{imp}\middle|\mathbf{a}\right)$.
VI. For fixed basic variable $\mathbf{a}$, $P_f\left(\mathbf{a}\right)$ is calculated using Eq. (20).
VII. The bounds on the failure probability are then calculated by using Eq. (21).

5 Numerical Application

The example system is composed by two simply supported plates coupled via a spring/mass system in order to represent with the simplest possible dynamic model a generic class of systems in which thin panels are coupled to stiff structural components (such as the frame of a car coupled to the roof and the window panels). The coupling is realised using three springs attached in the interior of each plate (point connections) linked to the second mass of the spring/mass system (Figure 3). The system is excited with a unit force applied to the first mass of the spring/mass system. The two plates are made of aluminium (Young's modulus $Y = 71\times10^9\ N/m^2$, density $2700\ Kg/m^3$ and Poisson's ratio $\nu = 0.3$) and their properties are summarised in Table 1.

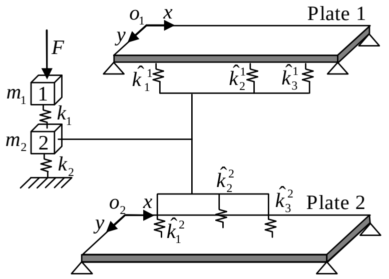


Figure 3: Built-up plate system under investigation.

Elements	Thickness (mm)	Size $L_x\times L_y$ ($m\times m$)	Loss factor η (%)	Modal density n (modes/Hz)
Plate 1	1.25	1.4×1.2	2	0.4286
Plate 2	1.25	1.4×1.3	2	0.4643

Table 1: Properties of the plates.

The spring connections in the interior of the first plate have stiffness $\hat{k}_u^1 = 2\times10^6$ N/m, $(u = 1,2,3)$ and attachment points $(0.3, 0.8), (0.6, 0.4)$, and $(0.8, 0.6)$ measured in metres along the x and y directions and relative to point the o_1. The second plate is connected via springs of stiffness $\hat{k}_l^2 = 2\times10^4$ N/m, $(l = 1,2,3)$ attached at points $(0.4, 0.4), (0.5, 0.9)$, and $(0.9, 0.7)$ measured in metres along the x and y directions and relative to the point o_2.

The hybrid FE/SEA model of the system comprises two SEA subsystems (the plates), which are highly random, and a mass/spring system (FE component) with two uncertain parameters, namely k_1 and k_2. k_1 is described by a lognormal pdf with mean value 6×10^6 N/m and variance $10^{11}\,(\text{N/m})^2$. k_2 is imprecisely known and it is specified in terms of bounds on statistical expectations as summarised in Table 2 and depicted in Figure 4.

1	2	3	4
$(18\times10^5, 14.27)$	$(22\times10^5, 14.52)$	$(22\times10^5, 14.50)$	$(18\times10^5, 14.24)$

Table 2: Coordinates of the vertices of the m-domain.

The system is forced by a unit force applied to the first mass of the mass/spring system (as shown in Figure 3). The design target is the energy level of plate 1 at 145 Hz, and a limiting value of $E_0 = 0.02\times10^{-4}\,J$ is considered.

The initial step of the analysis consists of evaluating the probability density function of the uncertain parameter k_2. This is achieved by using the procedure described in Subsection 2.1. The pdf of k_2 has the form

$$p(k_2\,|\,\mathbf{a}) = \exp[a_1 - a_2 x - a_3 \ln(x)], \qquad (22)$$

where a_1 is obtained by using the normalization condition as:

$$a_1 = -\ln\left(a_2^{(a_3-1)}\Gamma(1-a_3)\right), \qquad (23)$$

where $\Gamma(\bullet)$ is the gamma function.

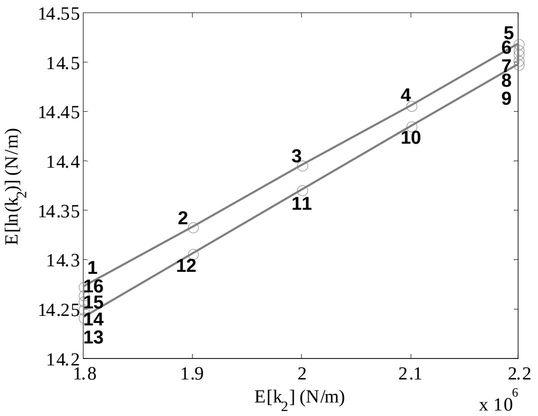


Figure 4: Moment domain for k_2

The a-domain is then calculated by using the strategy summarized in Subsection 2.2. In particular, the quadratic approximation of statistical expectations (Eq. (5)) was employed and 16 points along the m-domain (as shown in Figure 4) were mapped into the a-domain. The resulting approximate domain is shown in Figure 5.

Each point of the a-domain defines a single pdf. Some of the pdfs corresponding to the a-domain are shown in Figure 6.

The second step of the analysis consists of approximating the bounds on the failure probability as described in Subsection 4.2.

The a-domain was overlaid with a grid of 50×50 equally-spaced points. The 16 points along the domain and 414 points internal to the domain were considered (for a total of 430 pdfs). For each grid point (a_2, a_3) the procedure illustrated in Subsection 4.1 was applied. In particular, for fixed $\left(\hat{\mathbf{b}}_j, \mathbf{b}_{imp,j}\middle|\mathbf{a}\right)$ the hybrid FE/SEA method was applied to estimate the mean and variance of the response. These were used, under the assumption of a lognormal distribution of the vibrational energy of plate 1, to evaluate $p\left(w\middle|\left(\hat{\mathbf{b}}_j, \mathbf{b}_{imp,j}\middle|\mathbf{a}\right)\right)$. $P_f\left(\hat{\mathbf{b}}_j, \mathbf{b}_{imp,j}\middle|\mathbf{a}\right)$ was then calculated by using Eq. (15). The minimum point(s) of $-\ln\left[P_f\left(\hat{\mathbf{b}}_j^*, \mathbf{b}_{imp,j}^*\middle|\mathbf{a}\right)p\left(\hat{\mathbf{b}}_j^*\right)p\left(\mathbf{b}_{imp,j}^*\middle|\mathbf{a}\right)\right]$ was calculated by using the Matlab function fminunc. The Hessian matrix was approximated by using third order Lagrange polynomials. Finally, the failure probability conditional on the basic variables was computed by using Eq. (20).

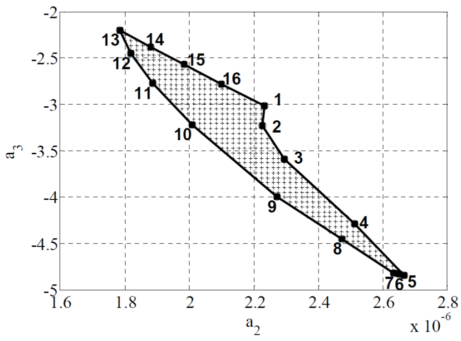


Figure 5: Approximate a-domain.

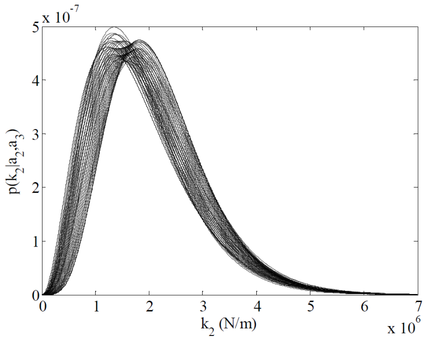


Figure 6: Pdfs generated from the a-domain.

The results obtained for each grid point are shown in Figure 7.

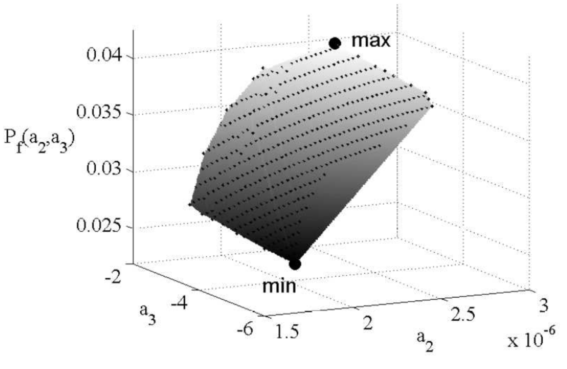


Figure 7: Failure probability as a function of the basic variables. The lower and upper bounds of the failure probability are labeled as "min" and "max".

The bounds on the failure probability are (by using Eq. (21)): $0.02192 \le P_f \le 0.04245$ (respectively, at point 1 and 9 of the a-domain), meaning that the uncertainty in the input parameters significantly affects the failure probability estimates. The computational time required by the proposed approach was of about 3 minutes.

The failure probability obtained for the MAXENT distribution (corresponding to the point 10 of the a-domain in Figure 7) is 0.03976. The MAXENT distribution would therefore underestimate the maximum failure probability.

The results obtained with the proposed approach were validated against direct numerical integration of Eq. (19), which took about 6 hours, showing differences less than 1%. Full FE Monte Carlo simulations for the present system considering a single point (and therefore a single pdf) of the a-domain requires about 45 hours. Full FE Monte Carlo simulations are therefore unfeasible even for this example system. It can be concluded that the proposed approach provides a very efficient tool for the reliability analysis of system with uncertain properties.

6 Summary and Conclusions

An imprecise probability model based on expressing the pdf of a random variable in the form of a maximum entropy distribution with bounded parameters was used to describe the parametric uncertainty of the FE components of a hybrid FE/SEA model. The hybrid FE/SEA + Laplace method, which fully accounts for both parametric (FE components) and non-parametric (SEA components) uncertainties, was applied to establish bounds on the failure probability. These bounds give an indication of the sensitivity of the system reliability to the uncertain input parameters and allow establishing the highest failure probability expected.

This approach provides a very useful tool for evaluating the reliability of complex engineering systems given that:

- The partition of the system in SEA and FE components leads to a large reduction of the number of degrees of freedom employed in the model (potentially thousand of finite elements nodes are substituted with a single degree of freedom SEA subsytem) and a large gain in numerical efficiency.
- The SEA subsystem ensemble is dealt with analytically (without using MCS) leading to a further reduction in computational costs.
- The uncertainty in FE components is dealt with using the Laplace asymptotic method instead of MCS.
- The bounds on the failure probability can be efficiently established when the imprecise probability model is employed.

The method has been illustrated by application to built-up plate systems, showing a large reduction of the computational cost when compared to a direct integration procedure and to Full FE Monte Carlo simulations.

Acknowledgements

Financial support from the Schlumberger Gould Research Center is gratefully acknowledged.

References

[1] R. S. Langley. Unified approach to probabilistic and possibilistic analysis of uncertain systems. *Journal of Engineering Mechanics*, 126:1163 – 1172, 2000.

[2] S. Ferson, C. A. Joslyn, J. C. Helton, W. L. Oberkampf and K. Sentz. Summary from the epistemic uncertainty workshop: consensus amid diversity. *Reliability Engineering and System Safety*, 85:335 – 369, 2004.

[3] J. C. Helton, J. D. Johnson and W. L. Oberkampf. An exploration of alternative approaches to representation of uncertainty in model predictions. *Reliability Engineering and System Safety*, 85:39 – 71, 2004.

[4] B. Moller and M. Beer. Engineering computation under uncertainty – Capabilities of non-traditional models. *Computer and Structures*, 86:1024 – 1041, 2008.

[5] C. Soize. Random matrix theory for modeling uncertainties in computational mechanics, *Computer Methods in Applied Mechanics and Engineering*, 194, 28 – 30, 3301-3315, 2005.

[6] R. S. Langley and V. Cotoni. Response variance prediction in the statistical energy analysis of built-up systems. *Journal of the Acoustical Society of America*, 115:706 – 718, 2004.

[7] R. S. Langley and V. Cotoni. Response variance prediction for uncertain vibro-acoustic system using a hybrid deterministic-statistical method. *Journal of the Acoustical Society of America*, 122 3445 – 3463, 2007.

[8] A. Cicirello and R. S. Langley. The vibro-acoustic analysis of built-up systems using a hybrid method with parametric and non-parametric uncertainties. *Journal of Sound and Vibration*, 332: 2165 – 2178, 2013.

[9] C. Soize. Stochastic modeling of uncertainties in computational structural dynamics—Recent theoretical advances. *Journal of Sound and Vibration*, 332: 2379 – 2395, 2013.

[10] R. E. Melchers. *Structural reliability analysis and prediction* (second edtion). Wiley, England, 1999.

[11] E. Nikolaidis, D. M. Ghiocel and S. Singhal. *Engineering Design Reliability Applications*. CRC Press, 2008.

[12] F. P. A. Coolen. On the use of imprecise probabilities in reliability. *Quality and Reliability Engineering International*, 20:193 – 202, 2004E.

[13] T. Jaynes, *Probability theory – The logic of Science*, Cambridge University Press, 2003.

[14] L. V. Utkin and F. P. A. Coolen. Imprecise reliability: An introductory overview. *Computational Intelligence in Reliability Engineering*, 40:261 – 306, 2007.

[15] C. Jiang, W. X. Li, X. Han, L. X. Liu and P. H. Le. Structural reliability analysis based on random distributions with interval parameters. *Computers and Structures*, 89:2292 – 2302, 2011.

[16] Z. Qiu, D. Yang and I. Elishakoff. Probabilistic interval reliability of structural systems. *International Journal of Solids and Structures*, 45:2850 – 2860, 2008.

[17] A. P. Dempster. Upper and lower probabilities induced by a multi-valued mapping. *Annals of Mathematical Statistics*, 38:325-339, 1967.

[18] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[19] S. Ferson and J. C. Hajagos. Arithmetic with uncertain numbers: rigorous and (often) best possible answer. *Reliability Engineering and System Safety*, 85:135 – 152, 2004.

[20] S. Ferson, V. Kreinovich, L. Ginzburg, D. S. Myers and K. Sentz. Constructing probability boxes and Dempster-Shafer structures. Sandia National Laboratories SAND2002-4015, 2003.

[21] M. Bruns, C. J. J. Paredis and S. Ferson. Computational methods for decision making based on imprecise information. Reliable Engineering Computing Workshop (REC 2006), 2006.

[22] M. A. S. Guth. A Probabilistic Foundation for Vagueness and Imprecision in Fault-Tree analysis. *IEEE Transactions on Reliability*, 40: 573 – 571, 1991.

[23] I. O. Kozine and Y. V. Filimonov. Imprecise reliabilities: experiences and advances. *Reliability Engineering and System Safety*, 67:75 – 83, 2000.

[24] K. Sentz and S. Ferson. Probabilistic bounding analysis in the quantification of margins and uncertainties. *Reliability Engineering and System Safety*, 96:1126 – 1136, 2011.

[25] H. Zhang, R. L. Mullen and R. L. Muhanna. Interval Monte Carlo methods for structural reliability. *Structural Safety*, 32:183 – 190, 2010.

[26] D. A. Alvarez. On the calculation of the bounds of probability of events using infinite random sets. *International Journal of Approximate Reasoning*, 43
(3):241 – 267, 2006

[27] F. Tonon, H. Bae, R. V. Grandhi and C. L. Pettit. Using random set theory to calculate reliability bounds for a wing structure. *Structure and Infrastructure Engineering*, 2(3-4):191 – 200, 2006.

[28] M. Beer, M. Zhang, S. T. Quek and S. Ferson. Structural reliability assessment with fuzzy probabilities. In Proceeding of ISIPTA'11 – Seventh International Symposium on Imprecise Probability: Theory and Applications, 2011.

[29] I. O. Kozine and L. V. Utkin. An approach to combining unreliable pieces of evidence and their propagation in a system response analysis. Reliability *Engineering and System Safety*, 85:103 – 112, 2004.

[30] L. V. Utkin and I. Kozine. On new cautious structural reliability models in the framework of imprecise probabilities. *Structural Safety*, 32:411 – 416, 2010.

[31] H. Zhang. Interval importance sampling method for finite element-based structural reliability assessment under parameter uncertainties. *Structural Safety*, 38:1 – 10, 2012.

[32] O. C. Zienkiewicz. *The Finite Element Method* (third edition). McGraw-Hill Company (UK) Limited, 1977.

[33] R. H. Lyon and R. G. DeJong. *Theory and Application of Statistical Energy Analysis* (Second Edition). Butterworth-Heinemann, 1995.

[34] P. J. Shorter, R. S. Langley. Vibro-acoustic analysis of complex systems. *Journal of Sound and Vibration*, 288:669 – 700, 2005.

[35] A. Cicirello, R. S. Langley. Reliability analysis of complex built-up structures via the hybrid FE-SEA method. Proceedings of the 4th International Conference on Noise and Vibration: Emerging Methods (NOVEM 2012), 2012.

[36] R. S. Langley and A. Cicirello. Uncertain probability density functions. Internal note. Engineering Department, University of Cambridge, 2010.

[37] P. Ishwar and P. Moulin. On the existence and characterization of the maxent distribution under general moment inequality constraints. *Information Theory, IEEE Transactions*, 51(9):3322 – 3333, 2005.

[38] A. Cicirello. Reliability analysis of vibro-acoustic systems with combined parametric and non-parametric probabilistic uncertainty models. Proceedings of the International Conference on Noise and Vibration Engineering (ISMA2012), 2012.

[39] P. J. Shorter and R. S. Langley. On the reciprocity relationship between direct field radiation and diffuse reverberant loading. *Journal of Sound and Vibration*, 117:85 – 95, 2005.

[40] R. S. Langley. On the diffuse field reciprocity relationship and vibrational energy variance in a random system at high frequencies. *Journal of the Acoustical Society of America*, 121:913 – 921, 2007.

[41] N. Bleistein, and R. A. Handelsman. *Asymptotic Expansions of Integrals*, Dover Publications, 1986.

[42] R. S. Langley, J. Legault, J. Woodhouse and E Reynders. On the applicability of the lognormal distribution in random dynamical systems. *Journal of Sound and Vibration*, 332:3289 – 3302. 2013.

[43] A. J. Grime and R. S. Langley. Lifetime reliability based design of an offshore vessel mooring. *Applied Ocean Research*, 30:221 – 234. 2008.

Bayesian-like inference, complete disintegrability and complete conglomerability in coherent conditional possibility theory

Giulianella Coletti and **Davide Petturiti**
Dip. di Matematica e Informatica
Università degli Studi di Perugia
{coletti,davide.petturiti}@dmi.unipg.it

Abstract

In this paper we consider Bayesian-like inference processes involving coherent T-conditional possibilities assessed on infinite sets of conditional events. For this, a characterization of coherent assessments of possibilistic prior and likelihood is carried on. Since we are working in a finitely maxitive setting, the notions of complete disintegrability and of complete conglomerability are also studied and their relevance in the infinite version of the possibilistic Bayes formula is highlighted.

Keywords. Complete disintegrability, complete conglomerability, finite maxitivity, T-conditional possibility, possibilistic likelihood function, coherence.

1 Introduction

This paper deals with finitely maxitive T-conditional possibilities (with T any continuous t-norm) and focuses the attention on problems related to the updating of possibility by Bayesian-like procedures.

In the first part of the paper we mainly deal with the characterization of coherent T-conditional possibility assessments, both for arbitrary families of conditional events and for particular families of the type $\{H_i, E|H_i\}_{i\in I}$, with I infinite, where the H_i's form a partition of the sure event while E is an arbitrary event. For these last assessments we also characterize the set of coherent values for their extension to E, in the case T is the minimum or a strict t-norm and E is logically independent of the H_i's.

In the second part we take into consideration two concepts: complete disintegrability and complete conglomerability for events, defined in analogy to those introduced in probability theory (originally given for countable partitions [18, 29, 1]), considering infinite partitions with arbitrary cardinality. As it is well-known, in probability theory the two properties (see, e.g., [17, 21, 29, 30, 31, 4]) are strictly related to σ-additivity. In fact for finitely additive conditional probabilities it is possible to have examples which, contrary to intuition, show that a P needs not be conglomerative (and so disintegrable). In Bayesian literature, the phenomenon of nonconglomerability has emerged in the so-called marginalization paradoxes [7]. In this paper we show similarities and differences between the probabilistic and possibilistic contexts about complete disintegrability and complete conglomerability, moreover we investigate their connection with complete maxitivity. In particular, we find that, for a fixed infinite partition $\mathcal{L}$, complete disintegrability w.r.t. $\mathcal{L}$ implies both complete maxitivity w.r.t. $\mathcal{L}$ and complete conglomerability w.r.t. $\mathcal{L}$ but the implications are not invertible. Furthermore, complete conglomerability w.r.t. $\mathcal{L}$ and complete maxitivity w.r.t. $\mathcal{L}$ are independent.

2 Coherent T-conditional possibility

In this section we recall the definition of conditional possibility given in [5, 6, 13, 14], that can be obtained as a particular instance of the one introduced in [10].

An *event* E is singled out by a Boolean proposition, that is a statement that can be either true or false. Since in general it is not known whether E is true or not, we are uncertain on E, which is said to be *possible.* Two particular events are the *certain event* Ω and the *impossible event* $\emptyset$, that coincide with, respectively, the top and the bottom of every Boolean algebra $\mathcal{B}$ of events, i.e., a set of events closed w.r.t. the familiar Boolean operations of *contrary* ^{c}, *conjunction* $\wedge$ and *disjunction* $\vee$ and equipped with the partial order $\subseteq$. Recall that due to Stone's theorem, events can be represented as subsets of a universe set that is identified with Ω: in this case we continue to use ^{c}, $\wedge$ and $\vee$ in place of set-theoretic operations.

A *conditional event* $E|H$ is an ordered pair (E, H), with $H \neq \emptyset$, where E and H are events of the same "nature", but with a different role (in fact H acts as a "possible hypothesis"). In particular any event E can be seen as the conditional event $E|\Omega$.

In what follows, $\mathcal{B} \times \mathcal{H}$ denotes a set of conditional events with $\mathcal{B}$ a Boolean algebra and $\mathcal{H}$ an additive set (i.e., closed with respect to finite disjunctions) such that $\mathcal{H} \subseteq \mathcal{B}^0 = \mathcal{B} \setminus \{\emptyset\}$. Moreover, given an arbitrary set $\mathcal{G} = \{E_j|H_j\}_{j\in J}$, denote with $\langle\{E_j, H_j\}_{j\in J}\rangle$ the Boolean algebra generated by the events $\{E_j, H_j\}_{j\in J}$.

We recall that a *t-norm* T is a commutative, associative, increasing, binary operation on $[0,1]$, having 1 as neutral element. A t-norm is called *continuous* (analogously, *left-continuous* or *right-continuous*) if it is continuous as a function, in the usual interval topology on $[0,1]^2$. Prototypical examples of continuous t-norms are the minimum, the algebraic product and the Łukasiewicz t-norm, moreover, any continuous t-norm is isomorphic to an ordinal sum of previous t-norms (see for instance [24]). A t-norm is called *strict* if it is continuous and strictly monotone: strict t-norms are isomorphic to the algebraic product through an order automorphism of the unit interval.

Definition 1. *Let T be any t-norm. A function $\Pi : \mathcal{B} \times \mathcal{H} \to [0,1]$ is a* **T-conditional possibility** *if it satisfies the following properties:*

- *(i)* $\Pi(E|H) = \Pi(E \wedge H|H)$*, for every* $E \in \mathcal{B}$ *and* $H \in \mathcal{H}$*;*
- *(ii)* $\Pi(\cdot|H)$ *is a finitely maxitive possibility on* $\mathcal{B}$*, for any* $H \in \mathcal{H}$*;*
- *(iii)* $\Pi(E \wedge F|H) = T(\Pi(E|H), \Pi(F|E \wedge H))$*, for any* $H, E \wedge H \in \mathcal{H}$ *and* $E, F \in \mathcal{B}$*.*

Let us stress that condition *(ii)* requires that, for every $H \in \mathcal{H}$, $\Pi(\emptyset|H) = 0$, $\Pi(\Omega|H) = 1$ and for every $E_1, \ldots, E_n \in \mathcal{B}$, $\Pi\left(\bigvee_{i=1}^n E_i|H\right) = \max\limits_{i=1,\ldots,n} \Pi(E_i|H)$, which is called *finite maxitivity axiom* [33]. Moreover conditions *(i)* and *(ii)* imply that $\Pi(H|H) = 1$ for every $H \in \mathcal{H}$.

Notice that in this paper we do not postulate the stronger condition of *complete maxitivity*, which requires that for every $\{E_i\}_{i\in I} \subseteq \mathcal{B}$ with $\bigvee_{i\in I} E_i \in \mathcal{B}$ and arbitrary I, $\Pi\left(\bigvee_{i\in I} E_i|H\right) = \sup\limits_{i\in I} \Pi(E_i|H)$, thus we always mean finitely maxitive T-conditional possibilities even when not explicitly stated.

Remark 1. *Every finitely maxitive unconditional possibility $\Pi(\cdot)$ on $\mathcal{B}$ can be seen as a T-conditional possibility on $\mathcal{B} \times \{\Omega\}$, where T is an arbitrary t-norm. In particular, for a T-conditional possibility Π on $\mathcal{B} \times \mathcal{H}$, we will write $\Pi(E)$ for $\Pi(E|\Omega)$, provided that $\Omega \in \mathcal{H}$.*

For every finite set of incompatible events $H_1, \ldots, H_n \in \mathcal{H}$ with $H = \bigvee_{i=1}^n H_i$ and for every $E \in \mathcal{B}$, axioms *(ii)* and *(iii)* imply a possibilistic counterpart of the well-known *disintegration formula*

$$\Pi(E|H) = \max_{i=1,\ldots,n} \{T(\Pi(E|H_i), \Pi(H_i|H))\}. \tag{1}$$

Definition 1 does not require any particular property for the t-norm T. The only constraint is the distributivity over the maximum operation used in condition *(ii)*, but this constraint is vacuous since every t-norm is distributive over max.

Nevertheless, continuity of the t-norm T is fundamental [14, 27] in order to guarantee the extendability (generally not in a unique way) of a T-conditional possibility on $\mathcal{B} \times \mathcal{H}$ to a *full* T-conditional possibility on $\mathcal{B}$ (i.e., with domain $\mathcal{B} \times \mathcal{B}^0$). For this, in the rest of the paper we will always assume T is continuous when not explicitly stated.

Differently from other common notions of conditioning in possibility theory [36, 23, 22, 15], a full T-conditional possibility $\Pi(\cdot|\cdot)$ is not singled out by a single unconditional possibility measure $\Pi(\cdot)$, in general, but one needs a class of finitely maxitive measures [33] defined on a family of ideals linearly ordered by proper set inclusion.

Remark 2. *We notice that in the particular case where the t-norm T is the usual product, $\Omega \in \mathcal{H}$ and $\Pi(H) = \Pi(H|\Omega) > 0$, for every $H \in \mathcal{H}$, the definition of T-conditional possibility coincides with Dempster's rule [20]:*

$$\Pi_D(E|H) = \frac{\Pi(E \wedge H)}{\Pi(H)}.$$

We recall that the conditional possibility Π_D is not necessarily a coherent conditional upper probability (see [16, 35]), vice versa a conditional possibility obtained as upper envelope of a class of conditional probabilities in general does not satisfy condition (iii) of Definition 1.

Definition 2. *Let $\mathcal{B}$ be a Boolean algebra and T a continuous t-norm. A family $\{(\mathcal{I}_i, \pi_i) : i \in I\}$ is a* **T-nested class** *if:*

- *(a) for every $i \in I$, $\mathcal{I}_i$ is a Boolean ideal of $\mathcal{B}$ and the family $\{\mathcal{I}_i : i \in I\}$ is linearly ordered by proper set inclusion;*
- *(b) for every $E \in \mathcal{B}^0$, there exists $i \in I$ such that $E \in \mathcal{I}_i \setminus \bigcup\{\mathcal{I}_j : \mathcal{I}_j \subset \mathcal{I}_i\}$;*
- *(c) for every $i \in I$, π_i is a (non-identically equal to 0) finitely maxitive measure on $\mathcal{I}_i$ ranging in $[0,1]$, such that for every $E \in \mathcal{I}_i$, $\pi_i(E) < 1$ if and only if $E \in \bigcup\{\mathcal{I}_j : \mathcal{I}_j \subset \mathcal{I}_i\}$;*

(d) *for every $i, j \in I$ such that $\mathcal{I}_i \subset \mathcal{I}_j$ and every $E, F \in \mathcal{I}_i$, all the solutions of equation $\pi_i(E \wedge F) = T(x, \pi_i(F))$ are solutions of the equation $\pi_j(E \wedge F) = T(x, \pi_j(F))$;*

(e) *for every $i, j \in I$ such that $\mathcal{I}_i \subset \mathcal{I}_j$, $\pi_{j|\mathcal{I}_i} \leq \pi_i$.*

Notice that, Definition 2 is equivalent in the finite case to the notion of T-nested class introduced in [14]. In particular, each finitely maxitive measure π_i on $\mathcal{I}_i$ is a restriction of a finitely maxitive possibility measure on $\mathcal{B}$.

The algebraic requirement on the domain of the function Π in Definition 1 cannot be relaxed, indeed axioms *(i)–(iii)* are no more sufficient to characterize Π if it is defined on an arbitrary set of conditional events $\mathcal{G}$. Hence, in order to deal with this eventuality, the axiomatic system must be reinforced going back to the concept of *coherence*, originally introduced by de Finetti [19] in the context of (finitely additive) probabilities.

Definition 3. *Let T be any continuous t-norm. A function $\Pi : \mathcal{G} \to [0,1]$ is a* **coherent T-conditional possibility (assessment)** *if there exists a T-conditional possibility $\Pi' : \mathcal{B} \times \mathcal{H} \to [0,1]$ such that $\Pi'_{|\mathcal{G}} = \Pi$, where $\mathcal{B} \times \mathcal{H} \supseteq \mathcal{G}$ with $\mathcal{B}$ a Boolean algebra and $\mathcal{H} \subseteq \mathcal{B}^0$ and additive class.*

Remark 3. *Previous definition can be equivalently formulated by requiring that Π can be extended as a full T-conditional possibility on $\mathcal{B}$. In fact in [27] the extendability of any T-conditional possibility on $\mathcal{B} \times \mathcal{H}$ to a full T-conditional possibility on $\mathcal{B}$ has been proved.*

Coherent T-conditional possibility assessments on finite domains have been characterized in [14]. Such characterization has been extended to the infinite case in [27], where the coherence of an assessment Π on $\mathcal{G}$ is expressed in terms of coherence of $\Pi_{|\mathcal{F}}$ on every finite $\mathcal{F} \subseteq \mathcal{G}$. The following Theorem 1 provides also a characterization in terms of a T-nested class agreeing with the assessment.

Theorem 1. *Let T be a continuous t-norm, $\mathcal{G} = \{E_j|H_j\}_{j \in J}$ an arbitrary set of conditional events and $\mathcal{B}$ the Boolean algebra generated by $\{E_j, H_j\}_{j \in J}$. For any $\mathcal{F} = \{E_1|H_1, \ldots, E_n|H_n\} \subseteq \mathcal{G}$, let $\mathcal{B}_{\mathcal{F}}$ be the Boolean algebra generated by $\{E_i, H_i\}$ whose set of atoms is $\mathcal{C}_{\mathcal{F}}$, and $\mathcal{H}_{\mathcal{F}} \subseteq \mathcal{B}^0_{\mathcal{F}}$ an additive set such that $\{H_i\} \subseteq \mathcal{H}_{\mathcal{F}}$. For a function $\Pi : \mathcal{G} \to [0,1]$, the following statements are equivalent:*

(i) *Π is a coherent T-conditional possibility on $\mathcal{G}$;*

(ii) *for any $\mathcal{F} = \{E_1|H_1, \ldots, E_n|H_n\} \subseteq \mathcal{G}$, if $\mathcal{C}_{\mathcal{F}0} = \{C_r \in \mathcal{C}_{\mathcal{F}} : C_r \subseteq H_0^0\}$ and $H_0^0 = \bigvee_{H \in \mathcal{H}_{\mathcal{F}}} H$, there exists a sequence of compatible systems $\mathcal{S}^{\Pi}_{\mathcal{F}\alpha}$, for $\alpha = 0, \ldots, k$, with unknowns $x_r^\alpha \geq 0$ for $C_r \in \mathcal{C}_{\mathcal{F}\alpha}$,*

$$\mathcal{S}^{\Pi}_{\mathcal{F}\alpha} : \begin{cases} \max\limits_{C_r \subseteq E_i \wedge H_i} x_r^\alpha = T\left(\Pi(E_i|H_i), \max\limits_{C_r \subseteq H_i} x_r^\alpha\right) \\ \left[\textit{for } E_i|H_i \in \mathcal{F} \textit{ s.t. } \max\limits_{C_r \subseteq H_i} \xi_r^{\alpha-1} < 1\right] \\ x_r^\alpha \geq \xi_r^{\alpha-1}, \ \textit{if } C_r \in \mathcal{C}_{\mathcal{F}\alpha} \\ \xi_r^{\alpha-1} = T\left(x_r^\alpha, \max\limits_{C_s \in \mathcal{C}_{\mathcal{F}\alpha}} \xi_s^{\alpha-1}\right), \ \textit{if } C_r \in \mathcal{C}_{\mathcal{F}\alpha} \\ \max\limits_{C_r \in \mathcal{C}_{\mathcal{F}\alpha}} x_r^\alpha = 1 \end{cases} \tag{2}$$

where $\overline{\xi}^\alpha$ (with r-th component ξ_r^α) is the solution of the system $\mathcal{S}^{\Pi}_{\mathcal{F}\alpha}$ and $\mathcal{C}_{\mathcal{F}\alpha}$ is the set of atoms $\{C_r \in \mathcal{C}_{\mathcal{F}\alpha-1} : C_r \subseteq H_0^\alpha\}$ with $H_0^\alpha = \bigvee\left\{H \in \mathcal{H}_{\mathcal{F}} : \max\limits_{C_r \subseteq H} \xi_r^\beta < 1, \beta \leq \alpha - 1\right\}$, moreover $\xi_r^{-1} = 0$ for any C_r in $\mathcal{C}_{\mathcal{F}0}$;

(iii) *there exists a T-nested class $\{(\mathcal{I}_i, \pi_i) : i \in I\}$ on $\mathcal{B}$ such that for every $E_j|H_j \in \mathcal{G}$ there exists $i \in I$ such that $H_j \in \mathcal{I}_i$ and $\pi_i(H_j) = 1$ and $\pi_i(E_j \wedge H_j) = \Pi(E_j|H_j)$.*

Proof. The equivalence between *(i)* and *(ii)* has been proved in [27]. To prove the equivalence between *(i)* and *(iii)* we follow the line of the construction introduced by Krauss in [25] for full conditional probabilities. Due to space limitations we give here just a sketch of the proof. For this aim, consider that for any full T-conditional possibility Π' on $\mathcal{B}$ it is possible to define a total preorder $\preceq$ on $\mathcal{B}^0$, setting $E \preceq F$ if and only if $\Pi'(F|E \vee F) = 1$, for every $E, F \in \mathcal{B}^0$. For every $E \in \mathcal{B}^0$, the relation $\preceq$ determines the Boolean ideal $\mathcal{I}_E = \{F \in \mathcal{B}^0 : F \preceq E\} \cup \{\emptyset\}$, and the family $\{\mathcal{I}_E : E \in \mathcal{B}^0\}$ results to be linearly ordered by set inclusion. For every $E \in \mathcal{B}^0$, define $\pi_E(F) = \Pi'(F|E \vee F)$ for every $F \in \mathcal{I}_E$, which results to be a finitely maxitive measure on the ideal $\mathcal{I}_E$. The family $\{(\mathcal{I}_E, \pi_E) : E \in \mathcal{B}^0\}$ is such that if $\mathcal{I}_E = \mathcal{I}_F$ then $\pi_E = \pi_F$. Thus, up to equal ideals, we can obtain a unique T-nested class $\{(\mathcal{I}_i, \pi_i) : i \in I\}$ which uniquely represents the full T-conditional possibility Π' on $\mathcal{B}$, since for every $E|H \in \mathcal{B} \times \mathcal{B}^0$, there exists an index $i \in I$ such that $\pi_i(H) = 1$ and $\pi_i(E \wedge H) = \Pi'(E|H)$. Now, since by Remark 3 the coherence of the assessment Π is equivalent to the existence of a full T-conditional possibility Π' on $\mathcal{B}$ extending Π, this is equivalent, in turn, to the existence of a T-nested class on $\mathcal{B}$ agreeing with the assessment Π. □

Remark 4. *In condition (ii) of previous theorem, for* *any finite $\mathcal{F} \subseteq \mathcal{G}$, the sequence of solutions $\overline{\xi}^0, \dots, \overline{\xi}^k$ gives rise to a class of possibilities $\mathcal{P}^\Pi = \{\Pi_0, \dots, \Pi_k\}$ on $\mathcal{B}_\mathcal{F}$ representing a T-conditional possibility on $\mathcal{B}_\mathcal{F} \times \mathcal{H}_\mathcal{F}$ extending $\Pi_{|\mathcal{F}}$ [27]. The choice of $\mathcal{H}_\mathcal{F}$ essentially impacts on the number of systems to solve [2, 3]. Let us notice that for the sake of convenience one can always take for $\mathcal{H}_\mathcal{F}$ the minimal additive set containing $\{H_i\}$, that is, the additive set generated by the H_i's. In the particular case $\mathcal{H}_\mathcal{F}$ is taken equal to $\mathcal{B}^0_\mathcal{F}$, then the solutions $\overline{\xi}^0, \dots, \overline{\xi}^k$ correspond exactly to a finite T-nested class $\{(\mathcal{I}_0, \pi_0), \dots, (\mathcal{I}_k, \pi_k)\}$ with $\mathcal{I}_\alpha \subset \mathcal{I}_{\alpha-1}$, $\alpha = 1, \dots, k$.*

Remark 5. *The characterization of coherence given in Theorem 1 implies that if $\Pi : \mathcal{G}' \to [0,1]$ is coherent, then for any subset $\mathcal{G} \subset \mathcal{G}'$ also $\Pi_{|\mathcal{G}}$ is coherent.*

Now we focus on the main t-norms used for conditioning in possibility theory, i.e., the minimum and strict t-norms. Under this choice, the coherence of an assessment is a sufficient (and necessary) condition for the extendability to any superset of conditional events, as stated in next theorem [27], which is a possibilistic counterpart of the celebrated de Finetti's fundamental theorem for conditional probabilities.

Theorem 2. *Let T be the minimum or a strict t-norm. Let $\mathcal{G}$ be an arbitrary set of conditional events and $\Pi : \mathcal{G} \to [0,1]$ a coherent T-conditional possibility. Then Π can be extended as a coherent T-conditional possibility Π' to any superset $\mathcal{G}' \supset \mathcal{G}$. Moreover, if $\mathcal{G}' = \mathcal{G} \cup \{E|H\}$ then the coherent values for $\Pi'(E|H)$ lie in a closed interval $[\pi_*, \pi^*]$.*

Previous theorem, whose proof relies on Zorn's lemma, generalizes to the infinite case a result proved in [14] for finite domains. In particular, the extension interval $[\pi_*, \pi^*]$ is computed as the intersection of all the intervals $[\pi_{\mathcal{F}*}, \pi_\mathcal{F}{}^*]$ expressing the coherent extensions of $\Pi_{|\mathcal{F}}$ on $E|H$, for any finite subfamily $\mathcal{F} \subseteq \mathcal{G}$.

Remark 6. *Let $\Pi : \mathcal{G}' \to [0,1]$ be a coherent T-conditional possibility and $\mathcal{G} \subset \mathcal{G}'$. If we denote with $[\pi'_*, \pi'^*]$ the extension interval of Π on $E|H$ and with $[\pi_*, \pi^*]$ the extension interval of $\Pi_{|\mathcal{G}}$ on $E|H$, then it holds $[\pi'_*, \pi'^*] \subseteq [\pi_*, \pi^*]$.*

Example 1. *Take $\mathbb{N}$ as universe, let $\mathcal{E} = \{E_i = \{i\}\}_{i\in\mathbb{N}}$, and $\mathcal{H} = \{H_1 = \{1\}^c, \mathbb{N}\}$. Consider the assessment Π defined for every $E_i \in \mathcal{E}$ and $H \in \mathcal{H}$ as*

$$\Pi(E_i|H) = \begin{cases} \frac{1}{i} & \text{if } E_i \wedge H \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

The function Π is a coherent min*-conditional possibility as it can be extended as a* min*-conditional possibility on $\mathcal{B} \times \mathcal{H}$, where $\mathcal{B}$ is the field of finite-cofinite subsets of $\mathbb{N}$. For example, a possible extension is the function Π' defined for $H \in \mathcal{H}$ putting $\Pi'(E|H) = 1$ if E is cofinite, while if E is finite we set*

$$\Pi'(E|H) = \begin{cases} \frac{1}{\min\{i \,:\, i \in E\wedge H\}} & \text{if } E \wedge H \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

Actually, Π' turns out to be a T-conditional possibility for every continuous t-norm T. Indeed, conditions (i) and (ii) are easily verified, while condition (iii) reduces to

$$\Pi'(E \wedge H_1) = T(\Pi'(E|H_1), \Pi'(H_1)),$$

for every $E \in \mathcal{B}$, which trivially holds since $\Pi'(H_1) = 1$ and $\Pi'(E \wedge H_1) = \Pi'(E|H_1)$ for every $E \in \mathcal{B}$.

We want to determine the coherent extension interval of the coherent min*-conditional possibility Π to the new event $H_1 = H_1|\mathbb{N}$. By previous discussion we know that 1 is the upper bound, thus we only need to compute the lower bound. Recalling that $\mathcal{E} \times \mathcal{H}$ is a countable set, for every $\{i_1, \dots, i_n\} \subseteq \mathbb{N}$ we can focus on the family $\mathcal{F} = \{E_{i_j}, E_{i_j}|H_1, : j = 1, \dots, n\}$. Indeed, by virtue of Remark 6 every finite subset of $\mathcal{F}$ gives rise to a larger extension interval, thus it can be ignored.*

Denote with $C_{i_j} = E_{i_j} \wedge H_1$ and $C'_{i_j} = E_{i_j} \wedge H_1^c$, $j = 1, \dots, n$, and $C_{i_{n+1}} = \bigwedge_{j=1}^n E^c_{i_j} \wedge H_1$ and $C'_{i_{n+1}} = \bigwedge_{j=1}^n E^c_{i_j} \wedge H_1^c$, the atoms generated by $\{E_{i_j}, H_1 : j = 1, \dots, n\}$, where only possible ones are considered.

The lower bound of the extension interval of $\Pi_{|\mathcal{F}}$ on H_1 is computed solving the following optimization problem under the system $\mathcal{S}^\Pi_{\mathcal{F}_0}$ [27], which has unknowns $x^0_{i_j}, x^0_{i_j}{}' \geq 0$ for atoms C_{i_j}, C'_{i_j}, $j = 1, \dots, n+1$, and results to be

$$\text{minimize} \left[\max_{j=1,\dots,n+1} \{x^0_{i_j}\} \right]$$

$$\mathcal{S}^\Pi_{\mathcal{F}_0} : \begin{cases} \max\{x^0_{i_j}, x^0_{i_j}{}'\} = \frac{1}{i_j} \\ [j = 1, \dots, n] \\ \\ x^0_{i_j} = \min\left\{ \frac{1}{i_j}, \max\limits_{j=1,\dots,n+1} \{x^0_{i_j}\} \right\} \\ [j = 1, \dots, n] \\ \\ \max\limits_{j=1,\dots,n+1} \{x^0_{i_j}, x^0_{i_j}{}'\} = 1 \end{cases}$$

where equations of the second kind in which $C_{i_j} = \emptyset$ are neglected as well as unknowns corresponding to $C_{i_j} = \emptyset$ or $C'_{i_j} = \emptyset$.

The lower bound can be written as $m_{\{i_1,\dots,i_n\}} = \max\left\{ \frac{1}{i_j} : j = 1, \dots, n, i_j \neq 1 \right\}$.

Hence, the coherent min*-conditional possibility values* *for H_1 range in the closed interval*

$$\bigcap_{\{i_1,\ldots,i_n\}\subseteq\mathbb{N}} [m_{\{i_1,\ldots,i_n\}}, 1] = \left[\frac{1}{2}, 1\right].$$

3 Possibilistic likelihood functions and possibilistic priors on infinite partitions

Theorem 1 and 2 deal with coherence and extension in their most general form. Nevertheless, there are situations in which coherence is immediately implied by some conditions and the extension on a new conditional event is easily computed.

This is the case of Bayesian-like inference processes in which one considers a *prior possibility* $\pi(\cdot)$ on a partition $\{H_i\}_{i\in I}$ and a *possibilistic likelihood* $f(E|\cdot)$ on the set $\{E|H_i\}_{i\in I}$, where E is the *evidence* event. The aim is to evaluate the *posterior possibility* of the conditional events $\{H_i|E\}_{i\in I}$.

To accomplish this task it is fundamental to establish whether the two assessments π and f are coherent *per se* and moreover whether the global assessment $\{f,\pi\}$ is coherent.

A complete characterization of the coherence of previous assessments has been given for a finite $I = \{1,\ldots,n\}$ in [9]. In this case, the coherence of $\{f,\pi\}$ allows to regard the global assessment as a $\Pi(\cdot|\cdot)$ on the set $\mathcal{G} = \{H_i, E|H_i\}_{i\in I}$ and to apply the following possibilistic counterpart of the *Bayes formula* (where we denote with Π also the posterior) for $i = 1,\ldots,n$,

$$\begin{aligned} T\left(\Pi(H_i|E), \max_{j=1,\ldots,n}\{T(\Pi(E|H_j),\Pi(H_j))\}\right) = \\ = T(\Pi(E|H_i),\Pi(H_i)). \qquad (3) \end{aligned}$$

Notice that, differently from the probabilistic case, depending on the particular t-norm T, the posterior possibility $\Pi(\cdot|E)$ could be non-unique on some H_i even requiring $\Pi(E) > 0$. In particular, if we consider $T = \min$ or a strict t-norm, Theorem 2 implies that each posterior $\Pi(H_i|E)$ lies in a (possibly degenerate) closed interval. Hence, in case of non-uniqueness, an arbitrary value in each interval can be chosen: the only constraint we have is that $\max\limits_{i=1,\ldots,n} \Pi(H_i|E) = 1$.

Example 2. *Consider the finite partition $\mathcal{L} = \{H_1,H_2,H_3\}$ together with the event E such that $E\wedge H_1 = \emptyset$. The following global assessment $\Pi(H_1) = 1$, $\Pi(H_2) = \Pi(H_3) = \frac{1}{3}$, $\Pi(E|H_1) = 0$, $\Pi(E|H_2) = \frac{1}{2}$ and $\Pi(E|H_3) = \frac{1}{3}$, is a coherent* min-*conditional possibility.*

In order to get the posterior (that we still denote with Π) we compute

$$\max_{j=1,2,3}\{\min\{\Pi(E|H_j),\Pi(H_j)\}\} = \frac{1}{3},$$

thus for $i = 1,2,3$ we need to solve

$$\min\left\{\Pi(H_i|E), \frac{1}{3}\right\} = \min\{\Pi(E|H_i),\Pi(H_i)\},$$

that implies $\Pi(H_1|E) = 0$, $\Pi(H_2|E),\Pi(H_3|E) \in \left[\frac{1}{3},1\right]$ such that $\max\{\Pi(H_2|E),\Pi(H_3|E)\} = 1$.

Our goal in this section is to generalize previous results to the case of an infinite index set I with $\operatorname{card} I \geq \operatorname{card}\mathbb{N}$.

Next theorem puts in evidence that every function defined on an infinite partition $\mathcal{L} = \{H_i\}_{i\in I}$ and ranging in $[0,1]$ (in particular the null function) is a coherent finitely maxitive possibility (i.e., it can be extended as a finitely maxitive possibility on $\langle\mathcal{L}\rangle$), and so, by Remark 1, a coherent T-conditional possibility, for any continuous t-norm T.

Theorem 3. *Let $\mathcal{L} = \{H_i\}_{i\in I}$ be a partition of Ω with $\operatorname{card} I \geq \operatorname{card}\mathbb{N}$. Then any function $\pi : \mathcal{L} \to [0,1]$ is a coherent T-conditional possibility (for every continuous t-norm T).*

Proof. We use condition *(ii)* of Theorem 1. Then for every $\{i_1,\ldots,i_n\} \subseteq I$, take the set $\mathcal{F} = \{H_{i_j} : j = 1,\ldots,n\}$ and denote $C_{i_j} = H_{i_j}$ for $j = 1,\ldots,n$, and $C_{i_{n+1}} = \bigwedge_{j=1}^n H_{i_j}^c$, the atoms generated by $\mathcal{F}$.

Consider the sequence of systems $\mathcal{S}^{\Pi}_{\mathcal{F}\alpha}$ with $\mathcal{H}_{\mathcal{F}} = \{\Omega\}$. The first (and unique) system of the sequence has unknowns $x^0_{i_j} \geq 0$ for C_{i_j}, $j = 1,\ldots,n+1$, and results to be

$$\mathcal{S}^{\Pi}_{\mathcal{F}0} : \begin{cases} x^0_{i_j} = \pi(H_{i_j}) & j = 1,\ldots,n \\ \max\limits_{j=1,\ldots,n+1}\{x^0_{i_j}\} = 1. \end{cases}$$

System $\mathcal{S}^{\Pi}_{\mathcal{F}0}$ admits the solution $x^0_{i_j} = \pi(H_{i_j})$, for $j = 1,\ldots,n$, and $x^0_{i_{n+1}} = 1$, and so π is coherent. □

Let $\mathcal{L} = \{H_i\}_{i\in I}$ be an arbitrary partition of Ω and E an arbitrary event, in the following we call *likelihood function* any function $f : \{E\}\times\mathcal{L} \to [0,1]$ defined as:

$$f(E|H_i) = \begin{cases} 0 \text{ when } E\wedge H_i = \emptyset, \\ 1 \text{ when } H_i \subseteq E, \\ \text{a value } \gamma_i \in [0,1] \text{ otherwise.} \end{cases} \qquad (4)$$

We underline that for the values γ_i's the only constraint is to be between 0 and 1.

Theorem 4. *Let $\mathcal{L} = \{H_i\}_{i\in I}$ be a partition of Ω with $\operatorname{card} I \geq \operatorname{card}\mathbb{N}$ and E an arbitrary event. For a likelihood function $f : \{E\}\times\mathcal{L} \to [0,1]$, defined by (4), the following statements hold:*

(i) f is a coherent conditional probability;

(ii) f is a coherent T-conditional possibility (for every continuous t-norm T).

Proof. In [9] this theorem has been proved for a finite partition $\mathcal{L}$, we prove it for the infinite case. Condition *(i)* follows by Proposition 1 in [8] and Theorem 4 in [11]. To prove *(ii)*, by condition *(ii)* of Theorem 1, for every $\{i_1, \ldots, i_n\} \subseteq I$, take the set $\mathcal{F} = \{E|H_{i_j} : j = 1, \ldots, n\}$ and denote $C_{i_j} = E \wedge H_{i_j}$ and $C'_{i_j} = E^c \wedge H_{i_j}$ for $j = 1, \ldots, n$, and $C_{i_{n+1}} = E \wedge \bigwedge_{j=1}^n H^c_{i_j}$ and $C'_{i_{n+1}} = E^c \wedge \bigwedge_{j=1}^n H^c_{i_j}$, the atoms generated by $\{E, H_{i_j} : j = 1, \ldots, n\}$, where only possible ones are considered.

Consider the sequence of systems $\mathcal{S}^{\Pi}_{\mathcal{F}_\alpha}$ with $\mathcal{H}_{\mathcal{F}}$ equal to the additive set generated by the H_{i_j}'s. The first (and unique) system of the sequence has unknowns $x^0_{i_j}, {x^0_{i_j}}' \geq 0$ for C_{i_j}, C'_{i_j}, $j = 1, \ldots, n$, and results to be

$$\mathcal{S}^{\Pi}_{\mathcal{F}0} : \begin{cases} x^0_{i_j} = T\left(f(E|H_{i_j}), \max\{x^0_{i_j}, {x^0_{i_j}}'\}\right) \\ [j = 1, \ldots, n] \\ \\ \max\limits_{j=1,\ldots,n} \{x^0_{i_j}, {x^0_{i_j}}'\} = 1 \end{cases}$$

where equations in which $C_{i_j} = \emptyset$ are neglected as well as unknowns corresponding to $C_{i_j} = \emptyset$ or $C'_{i_j} = \emptyset$. A solution for $\mathcal{S}^{\Pi}_{\mathcal{F}0}$ is $x^0_{i_j} = f(E|H_{i_j})$ and ${x^0_{i_j}}' = 1$ for $j = 1, \ldots, n$, implying that f is coherent. □

Previous theorem highlights that no significant property characterizes a likelihood function (defined by (4)) regarded either as coherent conditional probability or as coherent T-conditional possibility.

Remark 7. *We notice that Theorem 4 is related to a function defined only on a set of events $\{E\} \times \mathcal{L}$, (the conditioned event E is only one). Obviously, if we have a family of likelihood functions $\{f_j : j \in J\}$ each defined on $\{E_j\} \times \mathcal{L}$, where $\mathcal{E} = \{E_j\}_{j \in J}$ is an arbitrary set, the assessment could be non-globally coherent. In particular if $\mathcal{E}$ is a finite partition we must take into account additivity in the probabilistic case and maxitivity in the possibilistic case, as the following Theorem 5 shows.*

Theorem 5. *Let $\mathcal{E} = \{E_j\}_{j=1,\ldots,m}$ and $\mathcal{L} = \{H_i\}_{i \in I}$ be two partitions and let $\mathcal{F}$ be a (finite) class $\{f_j : j = 1, \ldots, m\}$ of likelihood functions, where each f_j is defined by (4) on $\{E_j\} \times \mathcal{L}$, for $j = 1, \ldots, m$. Then the following statements hold:*

(i) the global assessment $\mathcal{F}$ is a coherent conditional probability if and only if $\sum\limits_{j=1}^m f_j(E_j|H_i) = 1$ for every H_i;

(ii) the global assessment $\mathcal{F}$ is a coherent T-conditional possibility (for every continuous t-norm T) if and only if $\max\limits_{j=1,\ldots,m} f_j(E_j|H_i) = 1$ for every H_i.

Proof. In [9] this theorem has been proved for a finite partition $\mathcal{L}$, we prove it for the infinite case. Condition *(i)* follows by Theorem 4 in [11]. Condition *(ii)* follows by Theorem 1 on the same line of the proof of Theorem 4. □

Next theorem focuses on a likelihood function taking into account also a probabilistic or possibilistic prior.

Theorem 6. *Let $\mathcal{L} = \{H_i\}_{i \in I}$ be a partition of Ω with* $\operatorname{card} I \geq \operatorname{card} \mathbb{N}$ *and E an arbitrary event. Consider a likelihood function $f : \{E\} \times \mathcal{L} \to [0, 1]$, defined by (4), a coherent probability assessment $p : \mathcal{L} \to [0, 1]$ and a coherent possibility assessment $\pi : \mathcal{L} \to [0, 1]$. The following statements hold:*

(i) the global assessment $\{f, p\}$ is a coherent conditional probability;

(ii) the global assessment $\{f, \pi\}$ is a coherent T-conditional possibility (for every continuous t-norm T).

Proof. In [9] this theorem has been proved for a finite partition $\mathcal{L}$, we prove it for the infinite case. Condition *(i)* follows by Proposition 2 in [8] and Theorem 4 in [11] (see also [28, 32]). Condition *(ii)* follows by Theorem 1 in analogy to the proof of Theorem 4, and taking into account Remark 5. □

Example 3. *Consider $\mathbb{N}$ as universe and take the partition $\mathcal{L} = \{H_i = \{2i - 1, 2i\}\}_{i \in \mathbb{N}}$, together with $E = \{2i : i \in \mathbb{N}\}$. Consider the assessments $f(E|H_i) = \frac{1}{i}$, $p(H_i) = \pi(H_i) = 0$ for $i \in \mathbb{N}$. We have that $f(E|\cdot)$ verifies condition (4), moreover $p(\cdot)$ and $\pi(\cdot)$ are, respectively, a coherent probability and a coherent possibility. This implies $\{f, p\}$ and $\{f, \pi\}$ are, respectively, a coherent conditional probability and a coherent T-conditional possibility (for every continuous T-norm).*

4 Complete disintegrability and complete conglomerability

In this section we consider a T-conditional possibility Π on $\mathcal{B} \times \mathcal{H}$, with $\mathcal{H}$ containing Ω and a partition $\mathcal{L} = \{H_i\}_{i \in I}$, where I is arbitrary. Moreover, we say that an event $E \in \mathcal{B}$ is *logically independent* of the elements of $\mathcal{L}$ if $\emptyset \neq E \wedge H_i \neq H_i$, for $i \in I$.

Definition 4. *A T-conditional possibility Π on $\mathcal{B}\times\mathcal{H}$ is* **completely maxitive on** *$\mathcal{L}$ if it holds*

$$\sup_{i\in I}\Pi(H_i)=1. \tag{5}$$

Definition 5. *Given an event $E\in\mathcal{B}$, and a T-conditional possibility Π on $\mathcal{B}\times\mathcal{H}$, we say that Π is* **completely $\mathcal{L}$-disintegrable on** *E if it holds*

$$\Pi(E)=\sup_{i\in I}T(\Pi(E|H_i),\Pi(H_i)). \tag{6}$$

We introduce now a notion of conglomerability analogous the one introduced by de Finetti [17, 18, 19] (see also [29, 7, 30, 31, 1]), involving only events. We recall that in probability theory a stronger notion of conglomerability involving linear spaces of bounded random variables is present (see for instance [21, 28, 4]).

Definition 6. *Given an event $E\in\mathcal{B}$, and a T-conditional possibility Π on $\mathcal{B}\times\mathcal{H}$, we say that Π is* **completely $\mathcal{L}$-conglomerative on** *E if it holds*

$$\inf_{i\in I}\Pi(E|H_i)\leq\Pi(E)\leq\sup_{i\in I}\Pi(E|H_i). \tag{7}$$

Remark 8. *Definitions 5 and 6 actually involve only a family $\mathcal{G}=\{E,H_i,E|H_i\}_{i\in I}$ contained in $\mathcal{B}\times\mathcal{H}$, so they can be given for a coherent T-conditional possibility assessment on $\mathcal{G}$, if we are interested only on complete $\mathcal{L}$-conglomerability or complete $\mathcal{L}$-disintegrability on E (for instance in Bayesian-like updating). In fact, these properties are satisfied (for the given E and $\mathcal{L}$) by all the possible extensions on $\mathcal{B}\times\mathcal{H}$. Nevertheless, as discussed in the following, the above properties required only for one event E are not particularly meaningful, so we use a Π on $\mathcal{B}\times\mathcal{H}$ to enforce the properties to all the events of $\mathcal{B}$.*

In the case the partition $\mathcal{L}$ is finite, it is readily verified that complete maxitivity on $\mathcal{L}$ collapses into finite maxitivity and complete $\mathcal{L}$-disintegrability and complete $\mathcal{L}$-conglomerability always hold for every $E\in\mathcal{B}$, as simple implications of Definition 1. Nevertheless, previous properties could not be verified when the partition is infinite. In particular, in analogy with finitely additive conditional probability [18, 29], there can exist events $E\in\mathcal{B}$ on which Π is completely $\mathcal{L}$-disintegrable but not completely $\mathcal{L}$-conglomerative and vice versa, as shown in next example.

Example 4. *Let T be a continuous t-norm and consider the countable set $\mathcal{G}=\{E,H_i,E|H_i\}_{i\in\mathbb{N}}$ with E logically independent of the elements of the partition $\mathcal{L}=\{H_i\}_{i\in\mathbb{N}}$. Recall that the coherence of an assessment on $\mathcal{G}$ implies its extendability on $\mathcal{B}\times\mathcal{H}$, where $\mathcal{B}=\langle\{E\}\cup\mathcal{L}\rangle$ and $\mathcal{H}$ is the additive set generated by $\mathcal{L}$.*

The coherent T-conditional possibility assessment $\Pi(E)=\frac{1}{2}$, $\Pi(E|H_i)=\frac{1}{i}$ and $\Pi(H_i)=0$ for $i\in\mathbb{N}$ is completely $\mathcal{L}$-conglomerative on E, but not completely $\mathcal{L}$-disintegrable on E. In fact, we have $\Pi(E)=\frac{1}{2}\neq 0=\sup_{i\in I}T(\Pi(E|H_i),\Pi(H_i))$

On the other hand, the coherent assessment $\Pi(E)=\Pi(H_i)=0$ and $\Pi(E|H_i)=\frac{1}{2}$ for $i\in\mathbb{N}$ is completely $\mathcal{L}$-disintegrable on E, but it is not completely $\mathcal{L}$-conglomerative on E, since we have $\Pi(E)=0<\frac{1}{2}=\inf_{i\in I}\Pi(E|H_i)$.

Previous claim suggests to give a definition of complete $\mathcal{L}$-disintegrability and complete $\mathcal{L}$-conglomerability which is not dependent on the event E.

Definition 7. *A T-conditional possibility Π on $\mathcal{B}\times\mathcal{H}$ is* **completely $\mathcal{L}$-disintegrable** *if it is completely $\mathcal{L}$-disintegrable on E, for every $E\in\mathcal{B}$.*

Definition 8. *A T-conditional possibility Π on $\mathcal{B}\times\mathcal{H}$ is* **completely $\mathcal{L}$-conglomerative** *if it is completely $\mathcal{L}$-conglomerative on E, for every $E\in\mathcal{B}$.*

Let us note that the notion of conglomerability given in previous definition differs from the ones proposed for coherent lower and upper previsions (see for instance [34, 16, 26]). The difference is essentially due to the different concepts of conditioning adopted (see Remark 2).

Remark 9. *Suppose to have a possibilistic prior π on a partition $\mathcal{L}$ and two likelihood functions f_j on $\{E_j\}\times\mathcal{L}$, with $E_j\in\mathcal{B}$, ($j=1,2$), such that each $\{f_j,\pi\}$ admits a completely $\mathcal{L}$-conglomerative extension on $\mathcal{B}\times\mathcal{H}$. Even in the case $\{f_1,f_2,\pi\}$ is globally coherent there could not exist a completely $\mathcal{L}$-conglomerative extension on $\mathcal{B}\times\mathcal{H}$ (similarly for complete $\mathcal{L}$-disintegrability). Previous discussion generalizes to a larger class of likelihood functions.*

It is well-known that, in the probabilistic framework (see for instance [18, 21, 29]), for a countable I, $\mathcal{L}$-disintegrability and σ-additivity on $\mathcal{L}$ are equivalent. Nevertheless, since in the case of probability the equivalence is implied by the subtractive property, the same equivalence does not hold in the case of possibility, as shown by next example.

Example 5. *Let T be a continuous t-norm and I an index set s.t.* $\operatorname{card} I\geq\operatorname{card}\mathbb{N}$. *Consider the set $\mathcal{G}=\{E,H_i,E|H_i\}_{i\in I}$, where the H_i's form a partition $\mathcal{L}$ of Ω and E is logically independent of the H_i's.*

The assessment $\Pi(E)=\Pi(H_i)=1$ and $\Pi(E|H_i)=0$ for $i\in I$, is a coherent T-conditional possibility.

We have that Π is completely maxitive on the partition $\mathcal{L}$ since $\sup_{i\in I}\Pi(H_i)=1$, while it is not completely $\mathcal{L}$-disintegrable on E since $\Pi(E)=1\neq 0=$

$\sup_{i\in I} T(\Pi(E|H_i), \Pi(H_i))$.

In the possibilistic setting, complete maxitivity on $\mathcal{L}$ is only a necessary condition for complete $\mathcal{L}$-disintegrability.

Proposition 1. *If a coherent T-conditional possibility Π on $\mathcal{B}\times\mathcal{H}$ is completely $\mathcal{L}$-disintegrable, then it is completely maxitive on $\mathcal{L}$.*

Proof. It holds

$$1 = \Pi(\Omega) = \sup_{i\in I} T(\Pi(\Omega|H_i), \Pi(H_i)) = \sup_{i\in I} \Pi(H_i).$$

□

We notice that if Π is not completely maxitive on $\mathcal{L}$ then, if there exists an $E \in \mathcal{B}$ such that Π is completely $\mathcal{L}$-disintegrable on E then Π is not completely $\mathcal{L}$-disintegrable on E^c.

Next theorem shows that, analogously to the probabilistic case [19], complete $\mathcal{L}$-disintegrability implies the complete $\mathcal{L}$-conglomerative property.

Theorem 7. *If a T-conditional possibility Π on $\mathcal{B}\times\mathcal{H}$ is completely $\mathcal{L}$-disintegrable, then it is completely $\mathcal{L}$-conglomerative.*

Proof. For every $E \in \mathcal{B}$, complete $\mathcal{L}$-disintegrability implies that

$$\Pi(E) = \sup_{i\in I} T(\Pi(E|H_i), \Pi(H_i)) \leq \sup_{i\in I} \Pi(E|H_i),$$

moreover, setting $\kappa = \inf_{i\in I} \Pi(E|H_i)$ and recalling Proposition 1 and that any left-continuous t-norm commutes with the supremum, we get

$$\begin{aligned}\Pi(E) &= \sup_{i\in I} T(\Pi(E|H_i), \Pi(H_i)) \\ &\geq \sup_{i\in I} T(\kappa, \Pi(H_i)) = T\left(\kappa, \sup_{i\in I} \Pi(H_i)\right) = \kappa.\end{aligned}$$

□

Nevertheless, as it is shown in [29] for probability theory in the case of a countable partition, complete $\mathcal{L}$-disintegrability and complete $\mathcal{L}$-conglomerability are not equivalent. The next example in fact shows that complete $\mathcal{L}$-disintegrability is just a sufficient condition for the complete $\mathcal{L}$-conglomerative property.

Example 6. *Take $\mathbb{N}$ as universe, let $\mathcal{B}$ be the field of finite-cofinite subsets of $\mathbb{N}$ and $\mathcal{L} = \{H_i = \{i\}\}_{i\in\mathbb{N}}$. Consider on $\mathcal{B}\times\mathcal{B}^0$ the function Π defined for any $E|H \in \mathcal{B}\times\mathcal{B}^0$ putting if H is cofinite*

$$\Pi(E|H) = \begin{cases} 0 & \text{if } E\wedge H \text{ is finite,} \\ 1 & \text{otherwise,} \end{cases}$$

while if H is finite

$$\Pi(E|H) = \begin{cases} 0 & \text{if } E\wedge H = \emptyset, \\ 1 & \text{otherwise.} \end{cases}$$

First we show that Π is a full T-conditional possibility on $\mathcal{B}$ for any continuous t-norm T. For this, it is sufficient to show that axiom (iii) of Definition 1 is satisfied, since axioms (i) and (ii) are easily seen to be verified. At this aim, for any $H, E\wedge H \in \mathcal{B}^0$ and $E, F \in \mathcal{B}$ we consider the following cases.

(Case 1). If $E\wedge H$ and H are cofinite then we have $\Pi(E|H) = 1$, thus axiom (iii) is verified both when $E\wedge F\wedge H$ is cofinite (in this case we have $\Pi(E\wedge F|H) = \Pi(F|E\wedge H) = 1$) and when $E\wedge F\wedge H$ is finite (in this case we have $\Pi(E\wedge F|H) = \Pi(F|E\wedge H) = 0$).

(Case 2). If $E\wedge H$ is finite and H is cofinite then we have $\Pi(E|H) = 0$, thus axiom (iii) is verified for every value of $\Pi(F|E\wedge H)$, since $E\wedge F\wedge H$ is finite and so we have $\Pi(E\wedge F|H) = 0$.

(Case 3). If $E\wedge H$ and H are finite then we have $\Pi(E|H) = 1$, thus axiom (iii) is verified both when $E\wedge F\wedge H \neq \emptyset$ (in this case we have $\Pi(E\wedge F|H) = \Pi(F|E\wedge H) = 1$) and when $E\wedge F\wedge H = \emptyset$ (in this case we have $\Pi(E\wedge F|H) = \Pi(F|E\wedge H) = 0$).

It is easily seen that Π is not completely maxitive on $\mathcal{L}$, since

$$\Pi(\mathbb{N}) = 1 > 0 = \sup_{i\in\mathbb{N}} \Pi(H_i),$$

thus by virtue of Proposition 1, Π is not completely $\mathcal{L}$-disintegrable. On the contrary, we have that Π is completely $\mathcal{L}$-conglomerative. Indeed, if E is cofinite we have $\Pi(E) = 1 \geq \inf_{i\in\mathbb{N}} \Pi(E|H_i)$, and there must exist $j \in \mathbb{N}$ such that $E\wedge H_j \neq \emptyset$, thus $\sup_{i\in\mathbb{N}} \Pi(E|H_i) = 1$. Moreover, if E is finite we have $\Pi(E) = 0 \leq \sup_{i\in\mathbb{N}} \Pi(E|H_i)$, and there must exist $j \in \mathbb{N}$ such that $E\wedge H_j = \emptyset$, thus $\inf_{i\in\mathbb{N}} \Pi(E|H_i) = 0$.

Since complete $\mathcal{L}$-disintegrability and complete $\mathcal{L}$-conglomerability refer to a partition $\mathcal{L}\subset\mathcal{H}$, it is natural to ask if their validity w.r.t. an infinite $\mathcal{L}$ implies the validity w.r.t. any other infinite partition $\mathcal{L}'\subset\mathcal{H}$. In next example, inspired to the well-known Lévy's paradox [19, 30, 31, 7], we show that it is not the case.

Example 7. *Take $\mathbb{N}^2$ as universe, let $\mathcal{B}$ be the power set of $\mathbb{N}^2$ and take the two partitions $\mathcal{L}_1 = \{H_i = \{i\}\times\mathbb{N}\}_{i\in\mathbb{N}}$ and $\mathcal{L}_2 = \{K_i = \mathbb{N}\times\{i\}\}_{i\in\mathbb{N}}$. Consider on $\mathcal{B}\times(\mathcal{L}_1\cup\mathcal{L}_2)$ the function Π defined for any $E|H \in \mathcal{B}\times(\mathcal{L}_1\cup\mathcal{L}_2)$ putting*

$$\Pi(E|H) = \begin{cases} 0 & \text{if } E\wedge H \text{ is finite,} \\ 1 & \text{otherwise.} \end{cases}$$

It is possible to show that the assessment Π *is a coherent* T*-conditional possibility for any continuous* t*-norm* T.

The coherence of Π *implies its extendability to* $\mathcal{B} \times \mathcal{H}$, *where* $\mathcal{H}$ *is the additive set generated by* $\mathcal{L}_1 \cup \mathcal{L}_2$. *In particular, taking* $E = \{(i,j) \in \mathbb{N}^2 \ : \ i \geq j\}$ *we have* $\Pi(E|H_i) = \Pi(E^c|K_i) = 0$, *for any* $i \in \mathbb{N}$, *which implies that no extension* Π' *can be simultaneously completely* $\mathcal{L}_1$*-conglomerative and completely* $\mathcal{L}_2$*-conglomerative.*

Finally, by virtue of Theorem 7 it follows that no extension Π' *can be simultaneously completely* $\mathcal{L}_1$*-disintegrable and completely* $\mathcal{L}_2$*-disintegrable.*

Complete $\mathcal{L}$-disintegrability and complete $\mathcal{L}$-conglomerability are particularly relevant in the context of Bayesian-like inference processes since they constrain the set of coherent values for the posterior possibility. Anyway, when they are not satisfied, one needs to go back to the general enlargement procedure in which the posterior values are determined by Theorem 2.

For this, we are interested in the coherent extensions Π' on $\mathcal{G} \cup \{E\}$ of a coherent T-conditional possibility Π assessed on a family $\mathcal{G} = \{H_i, E|H_i\}_{i \in I}$, $\operatorname{card} I \geq \operatorname{card} \mathbb{N}$, where the set $\mathcal{L} = \{H_i\}_{i \in I}$ is a partition of Ω and E is an arbitrary event. Let us stress that Π is nothing else than the global assessment corresponding to a likelihood f and a possibilistic prior π (coherent by Theorem 6).

Next theorem characterizes the set of coherent values for the possibility $\Pi'(E)$ in the case E is logically independent of the H_i's and T is the minimum or a strict t-norm. Notice that if $H_i \subseteq E$ for every $i \in I$, then it must be $\Pi(E|H_i) = 1$ for every $i \in I$ and so $\Pi'(E) = 1$; similarly, if $H_i \wedge E = \emptyset$ for every $i \in I$, then it must be $\Pi(E|H_i) = 0$ for every $i \in I$ and so $\Pi'(E) = 0$. Thus in this two trivial situations complete $\mathcal{L}$-conglomerability on E holds compulsorily.

Theorem 8. *Let* Π *be a coherent* T*-conditional possibility on* $\mathcal{G}$ *(with* $T = \min$ *or strict) such that for* $i \in I$ *it is* $\emptyset \neq E \wedge H_i \neq H_i$, $\Pi(E|H_i) = \pi_i$ *and* $\Pi(H_i) = \pi'_i$, *with* $\operatorname{card} I \geq \operatorname{card} \mathbb{N}$. *Then the set of coherent values for* $\Pi'(E)$ *is*

$$\bigcap_{\{i_1,\ldots,i_n\} \subseteq I} \left[M_{\{i_1,\ldots,i_n\}}, 1\right], \qquad (8)$$

where $M_{\{i_1,\ldots,i_n\}} = \max_{j=1,\ldots,n} T(\pi_{i_j}, \pi'_{i_j})$.

Proof. By Theorem 2 the coherent values for $\Pi'(E)$ are a closed interval $[\pi_*, \pi^*]$, that is obtained as the intersection of all the intervals $[\pi_{\mathcal{F}*}, \pi_{\mathcal{F}}{}^*]$ expressing the coherent extensions of $\Pi_{|\mathcal{F}}$ on E, for any finite subfamily $\mathcal{F} \subseteq \mathcal{G}$.

Thus, for every $\{i_1, \ldots, i_n\} \subseteq I$ take the set $\mathcal{F} = \{H_{i_j}, E|H_{i_j} : j = 1, \ldots, n\}$. Notice that by Remark 6 every finite subset of $\mathcal{F}$ gives rise to a larger extension interval than the one induced by $\mathcal{F}$ and thus can be ignored. Denote with $C_{i_j} = E \wedge H_{i_j}$ and $C'_{i_j} = E^c \wedge H_{i_j}$, $j = 1, \ldots, n$, and $C_{i_{n+1}} = E \wedge \bigwedge_{j=1}^n H_{i_j}^c$ and $C'_{i_{n+1}} = E^c \wedge \bigwedge_{j=1}^n H_{i_j}^c$, the atoms generated by $\{E, H_{i_j} : j = 1, \ldots, n\}$.

The endpoints of the extension interval of $\Pi_{|\mathcal{F}}$ on E are computed solving the following two optimization problems under the system $\mathcal{S}_{\mathcal{F}_0}^{\Pi}$, which has unknowns $x_{i_j}^0, x_{i_j}^0{}' \geq 0$ for atoms C_{i_j}, C'_{i_j}, $j = 1, \ldots, n+1$, and result to be

$$\text{minimize} \Big/ \text{maximize} \left[\max_{j=1,\ldots,n+1} \{x_{i_j}^0\}\right]$$

$$\mathcal{S}_{\mathcal{F}0}^{\Pi} : \begin{cases} \max\{x_{i_j}^0, x_{i_j}^0{}'\} = \pi'_{i_j} & j = 1, \ldots, n \\ x_{i_j}^0 = T\left(\pi_{i_j}, \max\{x_{i_j}^0, x_{i_j}^0{}'\}\right) & j = 1, \ldots, n \\ \max_{j=1,\ldots,n+1} \{x_{i_j}^0, x_{i_j}^0{}'\} = 1 & \end{cases}$$

for which any solution is such that $x_{i_j}^0 = T(\pi_{i_j}, \pi'_{i_j})$, for $j = 1, \ldots, n$, thus the possibility of E is determined by the value assigned to $x_{i_{n+1}}^0$ which is only asked to belong to $[0,1]$. This implies the extension of $\Pi_{|\mathcal{F}}$ on E ranges in $\left[M_{\{i_1,\ldots,i_n\}}, 1\right]$ with $M_{\{i_1,\ldots,i_n\}} = \max_{j=1,\ldots,n} T(\pi_{i_j}, \pi'_{i_j})$, and the conclusion follows. □

In particular, previous theorem implies that if $\Pi(E|H_i) = \pi$ for $i \in I$, then the extension Π' on $\mathcal{G} \cup \{E\}$ of every coherent T-conditional possibility Π on $\mathcal{G}$ is generally not completely $\mathcal{L}$-conglomerative on E if $\pi < 1$, since the value $\Pi'(E) = 1$ is always coherent. Theorem 8 also implies the coherence of the posterior (that we still denote with Π) defined as:

$$\Pi(H_i|E) = T(\Pi(E|H_i), \Pi(H_i)) \quad \text{for } i \in I. \qquad (9)$$

5 Conclusions

In probability theory, in particular in modern Bayesian analysis, concepts of conglomerability and disintegrability have been deeply studied, especially with respect to finitely additive probability, where many famous examples of nonconglomerative conditional probability assessments are proposed. We studied the analogous concepts in possibility theory, starting from the definition of finitely maxitive T-conditional possibility, with T any continuous t-norm. We put in evidence analogies and differences between the two frameworks.

References

[1] T.E. Armstrong. Conglomerability of Probability Measures on Boolean Algebras. *Journal of Mathematical Analysis and Applications*, 150:335–358, 1990.

[2] M. Baioletti, G. Coletti, D. Petturiti, B. Vantaggi. Inferential models and relevant algorithms in a possibilistic framework. *International Journal of Approximate Reasoning*, 52(5):580-598, 2011.

[3] M. Baioletti and D. Petturiti. Algorithms for possibility assessments: Coherence and extension. *Fuzzy Sets and Systems*, 169(1):1–25, 2011.

[4] P. Berti and P. Rigo. Weak disintegrability as a form of preservation of coherence. *Journal of the Italian Statistical Society*, 1(2):161–181, 1992.

[5] B. Bouchon-Meunier, G. Coletti, and C. Marsala. *Conditional Possibility and Necessity*, volume 2 of *Technologies for Constructing Intelligent Systems*, pages 59–71. Springer, Berlin, 2001.

[6] B. Bouchon-Meunier, G. Coletti, and C. Marsala. Independence and possibilistic conditioning. *Annals of Mathematics and Artificial Intelligence*, 35:107–123, 2002.

[7] D.M. Cifarelli and E. Regazzini. De Finetti's Contribution to Probability and Statistics. *Statistical Science*, 11(4):253–282, 1996.

[8] G. Coletti, O. Gervasi, S. Tasso, and B. Vantaggi. Generalized Bayesian inference in a fuzzy context: From theory to a virtual reality application. *Computational Statistics & Data Analysis*, 56(4):967–980, 2012.

[9] G. Coletti, D. Petturiti, and B. Vantaggi. Possibilistic and probabilistic likelihood functions and their extensions: Common features and specific characteristics. *Fuzzy Sets and Systems.* (Under review).

[10] G. Coletti and R. Scozzafava. From conditional events to conditional measures: A new axiomatic approach. *Annals of Mathematics and Artificial Intelligence*, 32(1):373–392, 2001.

[11] G. Coletti and R. Scozzafava. *Probabilistic Logic in a Coherent Setting.* Vol. 15 of Trends in Logic, Kluwer Academic Publisher, Dordrecht/Boston/London, 2002.

[12] G. Coletti, R. Scozzafava, and B. Vantaggi. Integrated Likelihood in a Finitely Additive Setting. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, C. Sossai and G. Chemello (eds.), Lecture Notes in Computer Science, 5590:554–565, 2009.

[13] G. Coletti and B. Vantaggi. Possibility theory: Conditional independence. *Fuzzy Sets and Systems*, 157(11):1491–1513, 2006.

[14] G. Coletti and B. Vantaggi. T-conditional possibilities: Coherence and inference. *Fuzzy Sets and Systems*, 160(3):306–324, 2009.

[15] G. de Cooman. Possibility theory II: Conditional possibility. *International Journal of General Systems*, 25:325–351, 1997.

[16] G. de Cooman. Integration and conditioning in numerical possibility theory. *Annals of Mathematics and Artificial Intelligence*, 32:87–123, 2001.

[17] B. de Finetti. Sulla proprietà conglomerativa delle probabilità subordinate. *Atti Reale Accademia Nazionale dei Lincei*, Serie VI, Rend. 12 , 278–282, 1930.

[18] B. de Finetti. Sull'impostazione assiomatica del calcolo delle probabilità and Aggiunta alla nota sull'assiomatica della probabilità. (Two articles). *Annali Triestini* 19:29–81 and 20:3–20, 1949.

[19] B. de Finetti. *Probability, Induction and Statistics: The art of guessing.* John Wiley & Sons, London, New York, Sydney, Toronto, 1972.

[20] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics* 38:325–339, 1967

[21] L.E. Dubins. Finitely additive conditional probabilities, conglomerability and disintegrations. *Annals of Probability* 3(1):89–99, 1970.

[22] D. Dubois and H. Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty.* Plenum Press, New York and London, 1988.

[23] E. Hisdal. Conditional possibilities independence and noninteraction. *Fuzzy Sets and Systems*, 1(4):283–297, 1978.

[24] E.P. Klement, R. Mesiar, and E. Pap. *Triangualr Norms.* Vol. 8 of Trends in Logic, Kluwer Academic Publishers, Dordrecht/Boston/London, 2000.

[25] P.H. Krauss. Representation of conditional probability measures on Boolean algebras. *Acta Mathematica Academiae Scientiarum Hungaricae*, 19(3-4):229–241, 1968.

[26] E. Miranda, M. Zaffalon, and G. de Cooman. Conglomerable natural extension. *International Journal of Approximate Reasoning*, 53(8):1200–1227, 2012.

[27] D. Petturiti. Coherent Conditional Possibility Theory and Possibilistic Graphical Modeling in a Coherent Setting. *PhD thesis*, Università degli Studi di Perugia, 2013.

[28] E. Regazzini. De Finetti's Coherence and Statistical Inference. *Annals of Statistics*, 15(2):845–864, 1987.

[29] R. Scozzafava. Probabilità σ-additive e non. *Bollettino U.M.I.*, 1-A(6):1–33, 1982.

[30] M.J. Schervish, T. Seidenfeld, and J.B. Kadane. The Extent of Non-Conglomerability of Finitely Additive Probabilities. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 66:205–226, 1984.

[31] T. Seidenfeld, M.J. Schervish, and J.B. Kadane. Non-conglomerability for finite-valued, finitely additive probability. *The Indian Journal of Statistics*, Special issue on Bayesian Analysis, Vol. 60, Series A, 476–491, 1998.

[32] B. Vantaggi. Statistical matching of multiple sources: A look through coherence. *International Journal of Approximate Reasoning*, 49(3):701–711, 2008.

[33] N. Shilkret. Maxitive measure and integration. *Indagationes Mathematicae (Proceedings)*, 74(0):109–116, 1971.

[34] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.

[35] P. Walley and G. de Cooman. Coherence of rules for defining conditional possibility. *International Journal of Approximate Reasoning*, 21(1):63–107, 1999.

[36] L.A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1):3–28, 1978.

Conditional non–additive measures and fuzzy sets

Giulianella Coletti
Dip. Matematica e Informatica,
University of Perugia, Italy
coletti@dmi.unipg.it

Barbara Vantaggi
Dip. S.B.A.I.
University “La Sapienza”, Rome, Italy
barbara.vantaggi@sbai.uniroma1.it

Abstract

Consistency of partial assessments with different frameworks (probability, possibility, plausibility) is studied. We are interested in inferential processes like the Bayesian one, with particular attention when a part of the information is expressed in natural language and can be modeled by a possibilistic or a plausibilistic likelihood.

Keywords. Natural extension, conditional Plausibilities, T-conditional possibilities, generalized Bayesian inference, fuzzy sets.

1 Introduction

Fuzzy set theory, introduced by Zadeh [42], has become very popular and it provides a formalization of some concepts expressed by means of natural language. Different interpretations of fuzzy sets have been given [35, 26, 38] in terms of (conditional) probabilities, we refer to that given in [9, 10, 8], where the membership function of a fuzzy subset is interpreted in terms of a coherent conditional probability assessment. This interpretation, as shown in [5, 14, 13], is particularly useful when fuzzy and statistical information is simultaneously available.

Nevertheless sometimes the statistical information is related to a family of events different from that of interest and in which the fuzzy information is available (as a particular case we can have two partitions such that the elements of one are finite conjunctions of the element of the others): by extending the probabilistic assessment a la de Finetti [20, 41] we obtain a family of probabilities, whose upper envelope, which is in general only an upper probability, could be a plausibility [23, 31, 40, 11] or a possibility [28, 15, 22].

In this paper we consider the above problems by focusing mainly on plausibility and possibility measures, for which many proposals of conditioning are present. We adopt the definition of T-conditional possibility, with T any t-norm (introduced in [3] for minimum and generalized in [17] for any t-norm): this class of conditional measures includes as a particular case the conditional possibilities obtained by using the Dubois and Prade rule based on minimum specificity principle [27]. For conditional plausibility we adopt a definition generalizing Dempster rule, introduced in [6, 36], also if, as it is well known, it cannot be obtained as the lower envelope of a class of conditional probabilities. Nevertheless it assures a “weak disintegration rule” and admits as particular case T-conditional possibility, with T the usual product.

In the first part (Section 2 and 3) of the paper, in order to consider a generalized Bayesian inferential procedure, by using the concept of coherence (that is the consistency of a partial assessment with a conditional possibility or plausibility), we study the properties of likelihood functions, both as point and set functions, in the different frameworks. Moreover, we study the coherence of a likelihood with a plausibility (or possibility) measure having the role of “a prior”.

In Section 4 we give an interpretation of the membership of fuzzy sets as a possibilistic or a plausibilistic likelihood function and we study which properties of fuzzy set theory are maintained. In both cases the semantic of the interpretation seem to be very similar: if φ is a property, related to a variable X, the meaning associated to the membership $\mu_\varphi(x)$ on x consists into the possibility [plausibility] that You claim that X is φ under the hypothesis that X assumes the value x. We show that from a syntactical point of view many differences and common features can occur. About the specific feature the most relevant is that the membership $\mu_{\varphi\vee\psi}$ of the union of two fuzzy sets, with memberships μ_φ and μ_ψ, is not linked to $\mu_{\varphi\wedge\psi}$ by the Frank equation ([30]), as in probability theory. On the contrary, in the case of possibilistic setting $\mu_{\varphi\vee\psi}$ is univocally determined by μ_φ and μ_ψ independently of $\mu_{\varphi\wedge\psi}$. While in the case of plausibilistic framework it is not univocally determined, but $\mu_{\varphi\vee\psi}(x)$ must be between $\max\{\mu_\varphi(x),\mu_\psi(x)\}$ and $\min\{\mu_\varphi(x)+\mu_\psi(x)\}-\mu_{\varphi\wedge\psi}(x),1\}$.

In this interpretation the fuzzy membership μ_φ coincides with a likelihood and the fuzzy event E_φ is the Boolean event "You claim that X is φ"; moreover for the measure of uncertainty of E_φ when the prior on X is a plausibility we get an upper bound, while when the prior is a possibility we give an analytic formula depending on the chosen t-norm.

2 Conditional measures

Usually in literature a conditional measure is presented as a derived notion of the unconditional one, by introducing a law involving the joint measure and its marginal. Nevertheless, this could be restrictive, since for some pair of events the solution of the equation (the conditional measure) can either not exists or to be not unique. So, in analogy with conditional probability [21], it is preferable to define conditional measures in an axiomatic way, directly as a function defined on a suitable set of conditional events. We recall here the notion of T-conditional possibility (with T any t-norm)[3, 17]

Definition 1. *Let T be any t-norm. Given a Boolean algebra $\mathcal{B}$ and an additive set (closed under finite disjunctions) $\mathcal{H}$ with $\mathcal{H}\subseteq\mathcal{B}^0=(\mathcal{B}\setminus\{\emptyset\})$, a function $\Pi:\mathcal{B}\times\mathcal{H}\to[0,1]$ is a T-conditional possibility if it satisfies the following properties:*

(i) $\Pi(E|H)=\Pi(E\wedge H|H)$, *for every $E\in\mathcal{B}$ and $H\in\mathcal{H}$;*

(ii) $\Pi(\cdot|H)$ *is a (finitely maxitive) possibility on $\mathcal{B}$, for any $H\in\mathcal{H}$;*

(iii) $\Pi(E\wedge F|H)=T(\Pi(E|H),\Pi(F|E\wedge H))$, *for any $H,E\wedge H\in\mathcal{H}$ and $E,F\in\mathcal{B}$.*

Condition *(ii)* of previous definition requires that $\Pi(\Omega|H)=1$, $\Pi(\emptyset|H)=0$ and for every $H\in\mathcal{H}$, $\Pi(\bigvee_{i=1,\dots,n}A_i|H)=\max_{i=1,\dots,n}\Pi(A_i|H)$, for every $A_1,\dots,A_n\in\mathcal{B}$ [37]. Moreover from *(i)* and *(ii)* $\Pi(H|H)=1$ for every $H\in\mathcal{H}$.

Actually, conditional possibility (according to Definition 1) cannot be in general induced by a unique possibility, but by a class of possibilities (for more details, see [17]). Nevertheless, by using some principle, conditional possibility could be defined by means of a unique possibility measure. Obviously some principles can give rise to assessments inconsistent with axioms *(i) – (iii)*, see [16, 17].

Taken the minimum t-norm, by considering the minimum specificity principle the following notion of conditioning [27] arises (in the following called DP-conditional possibility, where DP stands for Dubois and Prade):

for any $E|H$ in $\mathcal{B}\times\mathcal{H}^0$, $\Pi(E|H)=1$, when $\Pi(E\wedge H)=\Pi(H)$ and $E\wedge H\neq\emptyset$, $\Pi(E|H)=\Pi(E\wedge H)$ otherwise.

It is easy to see that a DP-conditional possibility is a conditional possibility in the sense of Definition 1. More generally, for a continuous t-norm, the T-conditional possibility $\Pi(E|H)$ can be seen as the residuum $\to_T$ of the t-norm T

$x\to_T y=\sup\{z\in[0,1]:T(x,z)=y\}$

that means $\Pi(H)\to_T\Pi(E\wedge H)$ whenever $E\wedge H\neq\emptyset$ (see [19]). In [2] a link between these kinds of conditioning and Jeffrey's rule is studied, while in [25] connections between conditioning in possibility and belief function context are studied.

In [17] we proved that if T is a continuous t-norm, a conditional possibility can be extended on any other set $\mathcal{B}'\times\mathcal{H}'$ with $\mathcal{B}'$ a Boolean algebra and $\mathcal{H}'$ an additive set ($\mathcal{H}'\subseteq\mathcal{B}^0$) with $\mathcal{B}\times\mathcal{H}\subset\mathcal{B}'\times\mathcal{H}'$. Moreover, for any $E|H$ in $\mathcal{B}'\times\mathcal{H}'\setminus\mathcal{B}\times\mathcal{H}$ the admissible values lay on a closed interval.

Analogously, conditional plausibility can be defined axiomatically as follows (see [6, 11]):

Definition 2. *Let $\mathcal{B}$ be a Boolean algebra and $\mathcal{H}\subseteq\mathcal{B}^0$ an additive set. A function Pl defined on $\mathcal{C}=\mathcal{B}\times\mathcal{H}$ is a conditional plausibility if it satisfies the following conditions*

i) $Pl(E|H)=Pl(E\wedge H|H)$;

ii) $Pl(\cdot|H)$ *is a plausibility function* $\forall H\in\mathcal{H}$;

iii) For every $E\in\mathcal{B}$ and $H,K\in\mathcal{H}$

$$Pl(E\wedge H|K)=Pl(E|H\wedge K)\cdot Pl(H|K).$$

Moreover, given a conditional plausibility, a conditional belief function $Bel(\cdot|\cdot)$ is defined by duality as follows: for every event $E|H\in\mathcal{C}$

$$Bel(E|H)=1-Pl(E^c|H).$$

Condition *i)* and *ii)* requires that $Pl(\Omega|H)=Pl(H|H)=1$ and $Pl(\emptyset|H)=0$ and moreover, for any n, $Pl(\cdot|H)$ is n-alternating [23]:

$$Pl(A|H)\leq\sum(-1)^{|I|+1}Pl(\wedge_{i\in I}A_i|H) \qquad (1)$$

for any $A_1,\dots,A_n,A\in A$ with $A=\vee_{i=1}^n A_i$. Then, $Bel(\cdot|H)$ is n-monotone for any n.

This axiomatization extends the Dempster's rule, i.e.

$$Bel(F|H)=1-\frac{Pl(F^c\wedge H)}{Pl(H)},$$

for all conditioning events H such that $Pl(H) > 0$. When all the conditioning events have positive plausibility, i.e. $Pl(H|H^0) > 0$ for any $H \in \mathcal{H}$ (with $H^0 = \vee_{H\in\mathcal{H}}H$), the above notions of conditional plausibility and conditional belief coincide with that given in [24]. In fact, if $Pl(H) > 0$ it follows

$$Bel(F|H) = \frac{Bel(F \vee H^c) - Bel(H^c)}{Pl(H)}. \qquad (2)$$

An easy consequence of Definition 2 is a weak form of disintegration formula for the plausibility of an event $E|H$ with respect to a partition $H_1, ..., H_N$ of H

$$Pl(E|H) \leq \sum_{k=1}^{N} Pl(H_k|H)Pl(E|H_k) \qquad (3)$$

Taking into the following definition of conditioning (see [29, 33, 40, 41]):

$$Pl(F|H) = \frac{Pl(F \wedge H)}{Pl(F \wedge H) + Bel(F^c \wedge H)} \qquad (4)$$

the obtained conditional plausibility Pl does not satisfy axiom *iii)* of Definition 2. Therefore conditional plausibilities given trough equation (4) does not satisfy equation (3).

Note that for T equal to the usual product every T-conditional possibility is a conditional plausibility.

In the next result we show that every conditional plausibility on $\mathcal{B} \times \mathcal{H}$ can be extended (not uniquely) to a full conditional plausibility on $\mathcal{B}$ (i.e., a conditional plausibility on $\mathcal{B} \times \mathcal{B}^0$).

Theorem 1. *Let $\mathcal{B}$ be a finite algebra. If Pl on $\mathcal{B} \times \mathcal{H} \to [0, 1]$ is a conditional plausibility, then there exists a conditional plausibility $Pl' : \mathcal{B} \times \mathcal{B}^0 \to [0, 1]$ such that $Pl'_{|\mathcal{B}\times\mathcal{H}} = Pl$.*

Proof. Denote $H_0^0 = \bigvee_{H\in\mathcal{H}} H$. If H_0^0 coincides with the certain event Ω, $Pl(\cdot|\Omega)$ defines univocally $Pl'(E|H)$ for $Pl(H|\Omega) > 0$. Let $\mathcal{H}_0^1 = \{H \in \mathcal{B}^0 : Pl(H|\Omega) = 0\}$, $H_0^1 = \bigvee_{\mathcal{H}_0^1} H$ belongs to $\mathcal{B}^0$ and $Pl'(H_0^1|\Omega) = 0$ since $Pl(H_0^1|\Omega) \leq \sum_{H\in\mathcal{H}_0^1} Pl(H|\Omega)$. If $H_0^1 \in \mathcal{H}$ again for $Pl(H|H_0^1) > 0$ $Pl'(\cdot|H)$ is univocally defined, so proceed as before.

While for $H_0^1 \notin \mathcal{H}$ check whether the set $\mathcal{K} = \{H \in \mathcal{H} : Pl(H|H_0^1)\}$ is not empty. If it is not empty, consider the event $K_1 = \bigvee_{H\in\mathcal{K}} H$ in $\mathcal{H}$ and $K_1 \subseteq H_0^1$. Define $Pl'(E|H_0^1) = Pl(E|K_1)$ for any $E \in \mathcal{B}$. Note that $Pl'(K|H_0^1) = 1$, $Pl'(K^c|H_0^1) = 0$ and $Pl'(\cdot|H_0^1)$ is a plausibility since $Pl(\cdot|K_1)$ is. Otherwise if $\mathcal{K}$ is empty define $Pl'(E|H_o^1) = 1$ for any $E \in \mathcal{B}$ such that $E \wedge H_0^1 \neq \emptyset$. It is easy to check that even in this case $Pl'(\cdot|H_0^1)$ is a plausibility.

Now, define $\mathcal{H}_0^2 = \{H \in \mathcal{B}^0 : Pl(H|H_0^1) = 0\}$ and proceed as before.

It is easy to check that Pl' satisfies the axioms *iii)* of Definition 2 and so it is a conditional plausibility. □

Now we show that every full conditional plausibility on $\mathcal{B}$ can be extended as a full conditional plausibility on every finite superalgebra $\mathcal{B}' \supseteq \mathcal{B}$.

Theorem 2. *Let $\mathcal{B}$ be a finite algebra and $\mathcal{B}' \supseteq \mathcal{B}$ a finite superalgebra. If $Pl : \mathcal{B} \times \mathcal{B}^0 \to [0, 1]$ is a full conditional plausibility, then there exists a full conditional plausibility $Pl' : \mathcal{B}' \times \mathcal{B}'^0 \to [0, 1]$ such that $Pl'_{|\mathcal{B}\times\mathcal{B}^0} = Pl$.*

Proof. For any $A' \in \mathcal{B}'$ consider the smallest event $A \in \mathcal{B}$ containing A', $A = \vee_{C\in\mathcal{B}:C\wedge A'\neq\emptyset}C$ and define $Pl'(A') = Pl(A)$.
Since for any $A', B' \in \mathcal{B}'$, $Pl(A \wedge B) = Pl'(A' \wedge B')$ the function Pl' is a plausibility and induces a full conditional plausibility on $\mathcal{B}'$. By construction for any $A|B \in \mathcal{B} \times \mathcal{B}^0$ it holds $Pl'(A|B) = Pl(A|B)$. □

Note that the full conditional plausibility on $\mathcal{B}'$ extending the given conditional plausibility is not unique, that one given in the proof of Theorem 2 is just an example.

2.1 Coherent conditional plausibility

Analogously to probability theory, it is possible to introduce a notion of coherence in the context of plausibility functions, as done for conditional probabilities [21] and also for T-conditional possibilities [17].

Definition 3. *A function (or assessment) $\gamma : \mathcal{C} \to [0, 1]$, on a set of conditional events $\mathcal{C}$, is a coherent conditional plausibility (T-conditional possibility) iff there exists a full conditional plausibility Pl (full T-conditional possibility Π) on an algebra $\mathcal{B}$ such that $\mathcal{C} \subseteq \mathcal{B} \times \mathcal{B}^0$ and the restriction of Pl (Π) on $\mathcal{C}$ coincides with γ.*

For a characterization of (coherent) conditional possibility, with T-continuous t-norm, see [17, 1]. Theorem 3 characterizes (coherent) conditional plausibility functions in terms of a class of plausibilities.

Theorem 3. *Let $\mathcal{F} = \{E_1|F_1, E_2|F_2, \dots, E_m|F_m\}$ and denote by $\mathcal{B}$ the algebra generated by $\{E_1, \dots, E_m, F_1, \dots, F_m\}$, $H_0^0 = \vee_{j=1}^m F_j$. For $Pl : \mathcal{F} \to [0, 1]$ the following statements are equivalent:*

(a) Pl is a coherent conditional plausibility;

(b) there exists a class $\mathcal{P} = \{Pl_\alpha\}$ of plausibility functions such that $Pl_\alpha(H_0^\alpha) = 1$ and $H_0^\alpha \subset H_0^\beta$

for all $\beta < \alpha$, where H_0^α is the greatest (with respect to the inclusion) element of $\mathcal{K}$ for which $Pl_{(\alpha-1)}(H_0^\alpha) = 0$.

Moreover, for every $E_i|F_i$, there exists a unique index α such that $Pl_\beta(F_i) = 0$ for all $\alpha > \beta$, $Pl_\alpha(F_i) > 0$ and

$$Pl(E_i|F_i) = \frac{Pl_\alpha(E_i \wedge F_i)}{Pl_\alpha(F_i)}, \qquad (5)$$

(c) all the following systems (S^α), with $\alpha = 0, 1, 2, ..., k \leq n$, admit a solution $\mathbf{X}^\alpha = (\mathbf{x}_\mathbf{1}^\alpha, ..., \mathbf{x}_{\mathbf{j}_\alpha}^\alpha)$ with $\mathbf{x}_\mathbf{j}^\alpha = m_\alpha(H_j)$ $(j = 1, ..., j_\alpha)$:

$$(S^\alpha) = \begin{cases} \sum\limits_{H_k \wedge F_i \neq \emptyset} x_k^\alpha \cdot Pl(E_i|F_i) = \sum\limits_{H_k \wedge E_i \wedge F_i \neq \emptyset} x_k^\alpha, \ \forall F_i \subseteq H_0^\alpha \\ \sum\limits_{H_k \in H_0^\alpha} x_k^\alpha = 1 \\ x_k^\alpha \geq 0, \qquad\qquad \forall H_k \subseteq H_0^\alpha \end{cases}$$

where H_0^α is the greatest element of $\mathcal{K}$ such that $\sum\limits_{H_i \wedge H_0^\alpha \neq \emptyset} m_{(\alpha-1)}(H_i) = 0$.

In particular, conditions *(b)* and *(c)* stress that this conditional measure can be written in terms of a suitable class of basic assignments, instead of just one as in the classical case, where all the conditioning events have positive plausibility.

Note that every class $\mathcal{P}$ (condition *(b)* of Theorem 3) is said to be agreeing with conditional plausibility Pl. Whenever there are events in $\mathcal{K}$ with zero plausibility the class of unconditional plausibilities contains more than one element and we can say that Pl_1 gives a refinement of those events judged with zero plausibility under Pl_0.

For an example showing the construction of the class $\mathcal{P}$ characterizing (in the sense of the above result) a conditional plausibility see [36].

3 Likelihood functions

This section is devoted to a comparative analysis of likelihood functions under different frameworks: probability, possibility, plausibility.

Given an event E and a partition $\mathcal{L}$, a likelihood function is an assessment on $\{E|H_i : H_i \in \mathcal{L}\}$ (that is a function $f : \{E\} \times \mathcal{L} \to [0, 1]$) satisfying only the following trivial condition:

(L1) for every H_i such that $E \wedge H_i = \emptyset$ one has $f(E|H_i) = 0$ and for every H_i such that $H_i \subseteq E$ one has $f(E|H_i) = 1$

Theorem 4. *Let $\mathcal{L} = \{H_1, \ldots, H_n\}$ be a finite partition of Ω and E an event. For every likelihood function f on $\{E\} \times \mathcal{L}$ the following statements hold:*

a) f is a coherent conditional probability;

b) f is a coherent T-conditional possibility (for every continuous t-norm T);

c) f is a coherent conditional plausibility.

Proof. Condition *a)* and *b)* have been proved in [10] and [7], respectively.

Condition *c)* derives from *a)* and the fact that any coherent conditional probability is a coherent conditional plausibility (or equivalently from condition *b)* and the fact that any coherent T-conditional possibility, with T the usual product, is a coherent conditional plausibility). □

Theorem 5. *Let $\mathcal{L} = \{H_1, \ldots, H_n\}$ be a finite partition of Ω and E an event. If the only coherent conditional plausibility (possibility) f takes values in $\{0, 1\}$, then it is $H_i \wedge E = \emptyset$ for every H_i such that $f(E|H_i) = 0$ and it is $H_i \subseteq E$ for every H_i such that $f(E|H_i) = 1$.*

Proof. It follows directly from Theorem 3 and the characterization theorem for T-conditional possibilities [17]. □

The above results put in evidence that (in all contexts) no significant property characterizes likelihood as point function (i.e. an assessment on a partition).

This implies that since two likelihoods

$$f_i : \{E_i\} \times \mathcal{L}_i \to [0, 1]$$

$(i = 1, 2)$, related to events logically independent E_i are coherent with a conditional probability, then they should be coherent also with a conditional plausibility.

It is easy to show that $\{f_1, f_2\}$ are coherent also with a T-conditional possibility.

3.1 Likelihood and prior

The aim now is to make inference with a Bayesian-like procedure, so we have to deal with an initial assessment consisting of a "prior" φ on a partition $\mathcal{L}$ and a "likelihood function" f related to the set of conditional events $E|H_i$'s, with E an arbitrary event and $H_i \in \mathcal{L}$. This topic has been deeply discussed in [40, 41] by considering several interesting examples.

First of all we need to test the consistency of the global assessment

$$\{f, \varphi\} = \{f(E|H_i), \varphi(A) : H_i \in \mathcal{L}, A \in \langle\mathcal{L}\rangle\}$$

with respect to the framework of reference ($\langle\mathcal{L}\rangle$ denotes the algebra generated by $\mathcal{L}$). The choice of the framework of reference is essentially decided by the prior, since as shown in Theorem 4, a likelihood can be re-read in any framework. This can happen also when the prior comes from a previous inferential process such as the enlargement of an uncertainty assessment (see [15, 22, 28, 41]).

Theorem 6. *Let $\mathcal{L}$ be a partition of Ω, consider a likelihood f related to an event E on $\mathcal{L}$ and consider a probability P, a plausibility Pl and a possibility Π on the algebra $\langle\mathcal{L}\rangle$. Then, the following conditions hold:*

a) the global assessment $\{f, P\}$ is a coherent conditional probability;

b) the global assessment $\{f, Pl\}$ is a coherent conditional plausibility;

c) the global assessment $\{f, \Pi\}$ is a coherent T-conditional possibility (for every continuous t-norm T);

Proof. Condition *a)* has been proved in [39], while condition *c)* has been proved in [1].

Concerning condition *b)* note that Pl on $\langle\mathcal{L}\rangle$ defines a unique basic assignment function m_0 on $\langle\mathcal{L}\rangle$ that is the unique solution of S^0_{Pl} concerning the coherence of Pl. Then, we need to establish whether the assessment $\{f, Pl\}$ is coherent inside conditional plausibility, so we need to check whether the relevant system $S^0_{Pl,f}$ has solution and so whether there is a class of basic assignment $\{m'_\alpha\}$ on $\langle E, \mathcal{L}\rangle$. Notice if the system $S^0_{Pl,f}$ has a solution then coherence with respect to conditional plausibility follows from Theorem 5.

Actually, the atoms in $\langle E, \mathcal{L}\rangle$ are all the events $E \wedge H_i, E^c \wedge H_i$ with $H_i \in \mathcal{L}$. From [18] any plausibility on $\langle\mathcal{L}\rangle$ induces a unique function, called basic plausibility assignment, ν (possibly taking also negative values) on $\langle\mathcal{L}\rangle$ such that $\sum_{A\in\langle\mathcal{L}\rangle} \nu(A) = 1$ and $\sum_{A\in\langle\mathcal{L}\rangle:A\subseteq B} \nu(A) = Pl(B)$.

Let μ be on $\langle\mathcal{L}\rangle$ be the plausibility assignment induced by Pl, consider μ' defined on $\langle E, L\rangle$ as $\mu'(H_i) = 0$, $\mu'(E \wedge H_i) = f(E|H_i)Pl(H_i)$, $\mu'(E^c \wedge H_i) = \mu(H_i) - \mu'(E \wedge H_i)$, and, for any $A \in \langle\mathcal{L}\rangle \setminus \mathcal{L}$, $\mu(A) = \mu'(A)$. By construction $\sum_{A\in\langle E,\mathcal{L}\rangle} \mu'(A) = 1$. For any B in $\langle E, \mathcal{L}\rangle$, but not in $(\langle\mathcal{L}\rangle \cup \{E \wedge H_i, E^c \wedge H_i : H_i \in \mathcal{L}\})$ one has $\mu'(B) = 0$. Then, the function f on $\langle E, L\rangle$ defined as $\sum_{A\in\langle E,\mathcal{L}\rangle:A\subseteq B} \mu'(A) = f(B)$ is such that by construction, for any $B \in \langle\mathcal{L}\rangle$,

$$f(B) = \sum_{A\in\langle E,\mathcal{L}\rangle:A\subseteq B} \mu'(A) =$$

$$\sum_{A\in\langle\mathcal{L}\rangle:A\subseteq B} \mu'(E \wedge A) + \mu'(E^c \wedge A) + \mu'(A) =$$

$$\sum_{A\in\langle\mathcal{L}\rangle:A\subseteq B} \mu(A) = Pl(B)$$

then f extends Pl.

We need to prove that f is a plausibility: the proof can be made by induction, we prove here that is 2-alternating, the proof that it is n-alternating under the hyphothesis that is $(n-1)$-alternating is similar.

For any event $A \in \langle E, \mathcal{L}\rangle$ there is an event $\bar{A} \in \langle\mathcal{L}\rangle$ such that $\bar{A} \subseteq A$ and no event $B \in \langle\mathcal{L}\rangle$ such that $\bar{A} \subset B \subseteq A$, that is the maximal event of $\langle\mathcal{L}\rangle$ contained in A. Then, given any pair of events $A, B \in \langle E, \mathcal{L}\rangle$ let $\bar{A}, \bar{B} \in \langle\mathcal{L}\rangle$ be the two maximal events contained, respectively in A and B. Thus,

$$f(A\vee B) = \sum_{C\in\langle E,\mathcal{L}\rangle:C\subseteq A\vee B} \mu'(C) = \sum_{E\wedge H_i\subseteq A\vee B} \mu'(E\wedge H_i)+$$

$$\sum_{E^c\wedge H_i\subseteq A\vee B} \mu'(E^c \wedge H_i) + \sum_{C\in\langle\mathcal{L}\rangle\setminus\mathcal{L},C\subseteq A\vee B} \mu'(C) =$$

$$\sum_{H_i\subseteq A\vee B} \mu(H_i) + \sum_{E\wedge H_i\subseteq A\vee B,E^c\wedge H_i\not\subseteq A\vee B} \mu'(E \wedge H_i)+$$

$$\sum_{E^c\wedge H_i\subseteq A\vee B,E\wedge H_i\not\subseteq A\vee B} \mu'(E^c \wedge H_i) + \sum_{C\in\langle\mathcal{L}\rangle\setminus\mathcal{L},C\subseteq A\vee B} \mu(C)$$

$$= Pl(\overline{A \vee B}) + \sum_{E\wedge H_i\subseteq A\vee B,E^c\wedge H_i\not\subseteq A\vee B} \mu'(E \wedge H_i)+$$

$$\sum_{E^c\wedge H_i\subseteq A\vee B,E\wedge H_i\not\subseteq A\vee B} \mu'(E^c \wedge H_i) =$$

$$Pl(\bar{A} \vee \bar{B}) + \sum_{H_i\subseteq\overline{A\vee B},H_i\not\subseteq\bar{A}\vee\bar{B}} \mu(H_i)+$$

$$\sum_{E\wedge H_i\subseteq A\vee B,E^c\wedge H_i\not\subseteq A\vee B} \mu'(E \wedge H_i)+$$

$$\sum_{E^c\wedge H_i\subseteq A\vee B,E\wedge H_i\not\subseteq A\vee B} \mu'(E^c \wedge H_i).$$

Note that $A = \bar{A}\vee\bigvee_{H_i\in\mathcal{L}:H_i\not\subseteq A}((E\wedge H_i\wedge A)\vee(E^c\wedge H_i\wedge A))$ and analogously for B. Obviously, $\bar{A} \vee \bar{B} \subseteq A \vee B$ and $\bar{A} \wedge \bar{B}$ coincides with $\overline{A \wedge B}$. Moreover, $\bar{A} \vee \bar{B}$ is included into $\overline{A \vee B}$, but does not coincide with it, in fact $H_i \in \mathcal{L}$ could be included in $A \vee B$, but H_i is not included neither in A nor in B (e.g. $E \wedge H_i \subseteq A$ and $E^c \wedge H_i \subseteq B$). Hence,

$$f(A\vee B) \leq Pl(\bar{A})+Pl(\bar{B})-Pl(\bar{A}\wedge\bar{B})+ \sum_{H_i\subseteq A\bar{\vee}B,H_i\not\subseteq\bar{A}\vee\bar{B}} \mu(H_i)+$$

$$\sum_{E\wedge H_i\subseteq A\vee B,E^c\wedge H_i\not\subseteq A\vee B} \mu'(E \wedge H_i)+$$

$$\sum_{E^c\wedge H_i\subseteq A\vee B,E\wedge H_i\not\subseteq A\vee B} \mu'(E^c \wedge H_i)$$

$$\leq Pl(\bar{A}) + Pl(\bar{B}) - Pl(\bar{A} \wedge \bar{B})+$$

$$\sum_{H_i\subseteq\overline{A\vee B},H_i\not\subseteq\bar{A}\vee\bar{B}} (\mu'(E \wedge H_i) + \mu'(E^c \wedge H_i))+$$

$$\sum_{E\wedge H_i\subseteq A\vee B,E^c\wedge H_i\not\subseteq A\vee B} \mu'(E \wedge H_i)+$$

$$\begin{aligned}
&\sum_{E^c\wedge H_i\subseteq A\vee B, E\wedge H_i\not\subseteq A\vee B} \mu'(E^c\wedge H_i)\\
&= f(A)+f(B)-Pl(\bar{A}\wedge\bar{B})\\
&-\sum_{E\wedge H_i\subseteq A\wedge B, E^c\wedge H_i\not\subseteq A\vee B} \mu'(E\wedge H_i)\\
&-\sum_{E^c\wedge H_i\subseteq A\wedge B, E\wedge H_i\not\subseteq A\vee B} \mu'(E^c\wedge H_i)\\
&= f(A)+f(B)-f(A\wedge B)
\end{aligned}$$

Finally, f induces a conditional plausibility, that we continue to denote by f, on $\langle E,\mathcal{L}\rangle\times\mathcal{H}$ where H is the additive set generated by $H_i\in\mathcal{L}$ such that $f(H_i)>0$. For any $H_i\in\mathcal{L}$ one has
$f(E|H_i)=\frac{f(E\wedge H_i)}{f(H_i)}=\frac{\mu'(E\wedge H_i)}{Pl(H_i)}=f(E|H_i)$.
This implies that the system $S^0_{Pl,f}$ admits a solution and so for the above consideration the assessment $\{Pl,f\}$ is a coherent conditional plausibility. □

3.2 Aggregated likelihoods

Now we study the properties of *aggregated likelihood functions*, that is all the coherent extensions g of the assessment $\{f(E|H_i)\ :\ H_i\in\mathcal{L}\}$ to the events $E|K$, with K belonging to the additive set $\mathcal{H}=\langle\mathcal{L}\rangle^0=(\langle\mathcal{L}\rangle\setminus\{\emptyset\})$.

The interest derives from inferential problems in which the available information consists of a (probabilistic or plausibilistic or possibilistic) "prior" on a partition $\{K_j\}$ and a likelihood related to the events of another partition refining the previous one. So first of all we need to aggregate the likelihood function preserving coherence with the framework of reference.

In what follows $g:\{E\}\times\mathcal{H}\to[0,1]$ denotes a function such that its restriction to $\{E\}\times\mathcal{L}$ coincides with f.

We recall a common feature of probabilistic and possibility framework: any aggregated likelihood g, regarded as a coherent conditional probability or a coherent T-conditional possibility, satisfies the following condition for every $K\in\mathcal{H}$:

$$\min_{H_i\subseteq K} f(E|H_i)\leq g(E|K)\leq \max_{H_i\subseteq K} f(E|H_i). \qquad (6)$$

Now the question is to investigate whether an aggregated likelihood seen as a coherent conditional plausibility must satisfy the same constraints.

In the following example we show that, for a coherent conditional plausibility, the value $\max\limits_{H_i\subseteq K} f(E|H_i)$ is not an upper bound.

Example 1. *Let $\mathcal{L}=\{H_1,H_2\}$ be a partition and E an event logically independent of the events $H_i\in\mathcal{L}$. Consider the following likelihood on $\mathcal{L}$*

$$f(E|H_1)=\frac{1}{4};\ f(E|H_2)=\frac{1}{2}$$

and let g be a function extending f on $\{E\}\times\mathcal{H}$ such that $g(E|H_1\vee H_2)=\frac{3}{4}=f(E|H_1)+f(E|H_2)$.

From equation (6) it follows that g is not a coherent T-conditional possibility or conditional probability; we prove that it is indeed a coherent conditional plausibility. For that let us consider the following system with unknowns $m_0(C)$, where $C\in\langle E,\mathcal{L}\rangle$

$$(S^0)=\begin{cases}
1/4\cdot\sum\limits_{H_1\wedge C\neq\emptyset} m_0(C)=\sum\limits_{H_1\wedge E\wedge C\neq\emptyset} m_0(C),\\
1/2\cdot\sum\limits_{H_2\wedge C\neq\emptyset} m_0(C)=\sum\limits_{H_2\wedge E\wedge C\neq\emptyset} m_0(C),\\
3/4\cdot\sum\limits_{(H_1\vee H_2)\wedge C\neq\emptyset} m_0(C)=\sum\limits_{(H_1\vee H_2)\wedge E\wedge C\neq\emptyset} m_0(C),\\
\sum\limits_{C\subseteq H_1\vee H_2} m_0(C)=1\\
m_0(C)\geq 0, \qquad \forall C\in\langle E,\mathcal{L}\rangle
\end{cases}$$

It is easy to see that the basic assignment:

$$m_0((E\wedge H_1)\vee(E^c\wedge H_2))=m_0(H_1\vee(E^c\wedge H_2))=\frac{1}{8},$$

$$m_0((E^c\wedge H_1)\vee(E\wedge H_2))=m_0((E^c\wedge H_1)\vee H_2)=$$
$$m_0(E^c\wedge(H_1\vee H_2))=\frac{1}{4}$$

and $m_0(C)=0$ for any other event $C\in\langle E,\mathcal{L}\rangle$, is a solution of S_0, giving positive plausibility to both the events H_i.

The following example shows that also the lower bound of condition (6) can be violated in the plausibility framework.

Example 2. *Let $\mathcal{L}=\{H_1,H_2\}$ be a partition and E an event logically independent of all the events H_i.*

Consider the following aggregated likelihood on $\mathcal{H}$

$$f(E|H_1)=f(E|H_2)=\frac{2}{3},\ f(E|H_1\vee H_2)=\frac{1}{2}.$$

To prove that the assessment is coherent within a conditional plausibility, we consider the following system with unknowns $m_0(C)$, where $C\in\langle E,\mathcal{L}\rangle$

$$(S^0)=\begin{cases}
2/3\cdot\sum\limits_{H_1\wedge C\neq\emptyset} m_0(C)=\sum\limits_{H_1\wedge E\wedge C\neq\emptyset} m_0(C),\\
2/3\cdot\sum\limits_{H_2\wedge C\neq\emptyset} m_0(C)=\sum\limits_{H_2\wedge E\wedge C\neq\emptyset} m_0(C),\\
1/2\cdot\sum\limits_{(H_1\vee H_2)\wedge C\neq\emptyset} m_0(C)=\sum\limits_{(H_1\vee H_2)\wedge E\wedge C\neq\emptyset} m_0(C),\\
\sum\limits_{C\subseteq H_1\vee H_2} m_0(C)=1\\
m_0(C)\geq 0, \qquad \forall C\in\langle E,\mathcal{L}\rangle
\end{cases}$$

The following basic assignment on $\langle E,\mathcal{L}\rangle$:

$$m_0=(E^c\wedge H_1)=m_0(E^c\wedge H_2)=m_0(E)=m_0(\Omega)=\frac{1}{4}$$

and $m_0(C) = 0$ for any other event $C \in \langle E, \mathcal{L} \rangle$, is a solution of S_0, giving positive plausibility to both the events H_i.

The fact that the lower bound of coherent values of $Pl(E|H_i \vee H_j)$ can be less than $\inf\{Pl(E|H_i), Pl(E|H_j)\}$ is an indirect proof that a conditional plausibility (Definition 2) is not an upper envelope of a set of conditional probabilities.

Theorem 7. *Any coherent conditional plausibility Pl, extending a likelihood $f : E \times \mathcal{L} \to [0, 1]$ on $E \times \mathcal{H}$, satisfies the following inequality for every $K \in \mathcal{H}$:*

$$(L2) \quad 0 \leq Pl(E|K) \leq \min\{\sum_{H_i \subseteq K} f(E|H_i), 1\}.$$

Proof. Since f is a coherent conditional plausibility assessment, then there is a coherent conditional plausibility Pl on $\mathcal{B} \times \mathcal{H}$ with $\mathcal{B} = \langle \mathcal{H} \cup \{E\} \rangle$, extending f. The restriction of Pl to $E \times \mathcal{H}$ is a coherent conditional plausibility and for every $K \in \mathcal{H}$, satisfies (3) and $Pl(E|K) \geq 0$. So we have $0 \leq g(E|K) \leq \sum_{H_i \subseteq K} f(E|H_i) g(H_i|K)$, and then the thesis. □

Theorem 7 shows that in plausibility framework there is much more freedom than in both probabilistic and possibilistic ones, where aggregated likelihood functions are monotone, with respect to $\subseteq$, only if the extension is obtained, for every K, as $\max_{H_i \subseteq K} f(E|H_i)$ and they are anti-monotone if and only if their extensions are obtained as $\min_{H_i \subseteq K} f(E|H_i)$.

Since any likelihood (see Theorem 4) is also a coherent conditional probability and in [10, 12] it is proved that an aggregated likelihood coherent within conditional probability can be obtained by taking the minimum (maximum), this extension is obviously also a coherent conditional plausibility.

In the following Proposition we prove that we could take the sum of likelihoods.

Theorem 8. *Let f be a likelihood on $\mathcal{L}$ related to an event E and consider the function g on $\{E\} \times \mathcal{H}$ defined as follows: for all $K_1, K_2 \in \mathcal{H}$ with $K_1 \wedge K_2 = \emptyset$*

$$g(E|K_1 \vee K_2) = g(E|K_1) + g(E|K_2).$$

If $\sum_{H_i \in \mathcal{L}} f(E|H_i) \leq 1$, then g

is a coherent conditional plausibility extending f.

Proof. To prove the result it is enough to consider the following basic assignment m on $\langle E, \mathcal{L} \rangle$:

$$m((E \wedge H_i) \vee \bigvee_{j \neq i} (E^c \wedge H_j)) +$$

$$m(H_i \vee \bigvee_{j \neq i} (E^c \wedge H_j)) = f(E|H_i)$$

for $H_i \in \mathcal{L}$ and $m(E^c) = 1 - \sum_{H_i \in \mathcal{L}} f(E|H_i)$.

It is easy to show that this basic assignment m is agreeing with g (see Theorem 3) and the plausibility of H_i is positive. □

4 Fuzzy sets

The aim of this sections is to apply the results of the previous section to an inferential problem, starting from linguistic information (fuzzy sets) and statistical information. We refer to the interpretation of fuzzy sets in terms of coherent conditional probabilities [8, 9, 5]: the idea behind such interpretation is related to that given in the seminal work [32], and we extend it inside imprecise probabilities.

Let X be a (not necessarily numerical) variable, with range $\mathcal{C}_X$, and, for any $x \in \mathcal{C}_X$, let us indicate by A_x the event $\{X = x\}$. Let φ be any *property* related to the variable X and let us refer to the state of information of a real (or fictitious) person that will be denoted by "You". A coherent conditional probability (possibility) [plausibility] $f(E_\varphi|A_x)$ measures (in different frameworks) the degree of belief of You in E_φ, when X assumes the different values x in $\mathcal{C}_X$.

Then $f(E_\varphi|\cdot)$ comes out to be a natural interpretation of the membership function $\mu_\varphi(\cdot)$, analogously to the probabilistic case [9] (see also [8, 5]).

Definition 4. *For any variable X with range $\mathcal{C}_X$ and a related property φ, the fuzzy subset E^*_φ of $\mathcal{C}_X$ is the pair*

$$E^*_\varphi = \{E_\varphi\, , \, \mu_{E_\varphi}\},$$

with $\mu_{E_\varphi}(x) = f(E_\varphi|A_x)$ for every $x \in \mathcal{C}_X$ (f stands for a coherent conditional probability or plausibility or possibility).

Theorem 4 assures that any assessment $\{f(E|A_x)\}_{x \in \mathcal{C}_X}$ is coherent within conditional probability, plausibility and possibility: so we have no syntactical restriction for f; Theorem 5 assures that in all the three frameworks the notion of fuzzy subsets, defined by a likelihood, is a generalization of crisp subsets.

Now denote by $\varphi \vee \psi$, $\varphi \wedge \psi$, respectively, the properties "φ or ψ" , "φ and ψ", and define

$$E_{\varphi \vee \psi} = E_\varphi \vee E_\psi \, ,$$

$$E_{\varphi \wedge \psi} = E_\varphi \wedge E_\psi \, .$$

Let us consider two fuzzy subsets E^*_φ, E^*_ψ, corresponding to the same variable X, with the events

E_φ, E_ψ logically independent with respect to X. As proved in [9], for any given x in the range of X, the assessment $P(E_\varphi \wedge E_\psi|A_x) = v$ is coherent within a conditional probability if and only if takes values in the interval

$$\max\{P(E_\varphi|A_x) + P(E_\psi|A_x) - 1, 0\} \leq v \leq$$
$$\leq \min\{P(E_\varphi|A_x), P(E_\psi|A_x)\}.$$

It is easy to see that the assessment $f(E_\varphi \wedge E_\psi|A_x) = v$ is coherent within a conditional plausibility or possibility if and only if takes values in the interval

$$0 \leq v \leq \min\{f(E_\varphi|A_x), f(E_\psi|A_x)\}.$$

Then, the lower bound of conditional probability does not continue to be valid.

While probability rules imply that given a value to $f(E_\varphi \wedge E_\psi|A_x)$, we get also the value of $f(E_\varphi \vee E_\psi|A_x)$, in the case of possibility we have that the value of $f(E_\varphi \vee E_\psi|A_x)$ is univocally determined by $f(E_\varphi|A_x)$ and $f(E_\psi|A_x)$ without taking into account the value of $f(E_\varphi \wedge E_\psi|A_x)$.

In the case of plausibility we have that the value of $f(E_\varphi \vee E_\psi|A_x)$ is not univocally determined but it must be

$$\max\{f(E_\varphi|A_x), f(E_\psi|A_x)\} \leq f(E_\varphi \vee E_\psi|A_x) \leq$$
$$\min\{f(E_\varphi|A_x) + f(E_\psi|A_x) - f(E_\varphi \wedge E_\psi|A_x), 1\}$$

Then we can put

$$E^*_\varphi \cup E^*_\psi = \{E_{\varphi\vee\psi}\,,\, \mu_{\varphi\vee\psi}\}\ ,$$
$$E^*_\varphi \cap E^*_\psi = \{E_{\varphi\wedge\psi}\,,\, \mu_{\varphi\wedge\psi}\}\ ,$$

with

$$\mu_{\varphi\vee\psi}(x) = f(E_\varphi \vee E_\psi|A_x)\ ,$$
$$\mu_{\varphi\wedge\psi}(x) = f(E_\varphi \wedge E_\psi|A_x)\,.$$

Moreover, denoting by $E^*_{\neg\varphi}$ the complementary fuzzy set of E^*_φ, the relation $E_{\neg\varphi} \neq (E_\varphi)^c$ holds, since the propositions "You *claim* $\neg\varphi$" and "You *do not claim* φ" are logically independent. In fact, we can claim both "X has the property φ" and "X has the property $\neg\varphi''$, or only one of them or finally neither of them; similarly are logical independent E_φ and E_ψ, where ψ is the superlative of φ.

Then, while $E_\varphi \vee (E_\varphi)^c = C_X$, we have instead $E_\varphi \vee E_{\neg\varphi} \subset C_X$, and, if we consider the union of a fuzzy subset and its complement

$$E^*_\varphi \cup (E^*_\varphi)' = \{E_{\varphi\vee\neg\varphi}\,,\, \mu_{\varphi\vee\neg\varphi}\}$$

we obtain in general a *fuzzy subset* of C_X.

The constraints on the function f depend, as shown before, on the framework of reference.

The concept of fuzzy event, as introduced by Zadeh, can be seen an ordinary event of the kind

$$E_\varphi = \text{“You claim that } X \text{ is } \varphi\text{”}.$$

and for any uncertainty measure (probability, possibility and plausibility) on the events related to X the assessment together μ_φ is coherent with respect the relative measure (see Theorem 6) and so coherently extendible to E_φ (Theorem 2 for plausibilities, [17] for conditional possibilities).

In the case of probability and possibility it is easily to see that the only coherent value for the probability or possibility of E_φ is

$$g(E_\varphi) = \bigoplus_{x\in\mathcal{C}_X} \mu_{\varphi_i}(x) \bigodot g(x)\,,$$

where $\bigoplus$ and $\bigodot$ are the sum and the product in the case of probability, while they are the maximum and minimum in the case of possibility.

Obviously, only in the case of probability it coincides with Zadeh's definition of the probability of a "fuzzy event" [42].

5 Conclusion

The first part of the paper is devoted into studying likelihood functions seen as assessment on a set of conditional events $E|H_i$, with E the evidence and H_i varying on a partition $\mathcal{L}$. It is shown that likelihood functions are assessment coherent with respect probability, possibility and plausibility. Then, inferential processes, like Bayesian one, are studied in the different setting taking a likelihood function and a prior, that could be a probability or a possibility or a plausibility. I particular we prove that any likelihood function on $E \times \mathcal{L}$ and any plausibility on $\mathcal{L}$, with $\mathcal{L}$ a partition, are globally coherent within conditional plausibility. Then, a comparison of aggregated likelihoods, that are coherent extensions of a likelihood function on $E \times \mathcal{L}$ to $E \times \langle\mathcal{L}\rangle^0$ is studied in the different setting by showing the common characteristic and the specific features.

Finally, by using the above results we give an interpretation of fuzzy sets in terms of likelihood function in the different setting: by starting from the interpretation in the probabilistic setting given in [9] we give a similar interpretation in plausibility and possibilistic settings.

References

[1] M. Baioletti, G. Coletti, D. Petturiti and B. Vantaggi. Inferential models and relevant algorithms in a possibilistic framework. *International Journal of Approximate Reasoning*, 52: 580–598, 2011.

[2] Benferhat Salem, Tabia Karim and Sedki Karima. Jeffrey's rule of conditioning in a possibilistic framework. *Annals of Mathematics and Artificial Intelligence*, 66(3): 185–202, 2011.

[3] B. Bouchon-Meunier, G. Coletti and C. Marsala. Independence and Possibilistic Conditioning. *Annals of Mathematics and Artificial Intelligence*, 35: 107–123, 2002.

[4] A. Chateauneuf and J.Y. Jaffray. Some characterizations of lower probabilities and other monotone capacities through the use of Mobius inversion. *Mathematical Social Sciences* 17(3): 263-283, 1989.

[5] G. Coletti, O. Gervasi, S. Tasso and B. Vantaggi. Generalized Bayesian inference in a fuzzy context: From theory to a virtual reality application. *Computational Statistics & Data Analysis*, 56: 967–980, 2012.

[6] G. Coletti and M. Mastroleo Conditional belief functions: a comparison among different definitions. *Proc. of 7^{th} Workshop on Uncertainty Processing*, 2006.

[7] G. Coletti, D. Petturiti and B. Vantaggi. Possibilistic and probabilistic likelihood functions and their extensions: Common features and specific characteristics. *Fuzzy Sets and Systems*, submitted.

[8] G. Coletti and R. Scozzafava, *Probabilistic logic in a coherent setting.* (Trends in logic n.15), Kluwer, Dordrecht, 2002.

[9] G. Coletti and R. Scozzafava. Conditional Probability, Fuzzy Sets, and Possibility: a Unifying View. *Fuzzy Sets and Systems*, 144: 227–249, 2004.

[10] G. Coletti and R. Scozzafava. Conditional Probability and Fuzzy Information. *Computational Statistics & Data Analysis* 51:115–132, 2006.

[11] G. Coletti and R. Scozzafava. Toward a general theory of conditional beliefs. *Int. J. of Intelligent Systems*, 21: 229–259, 2006.

[12] G. Coletti, R. Scozzafava and B. Vantaggi. Integrated Likelihood in a Finitely Additive Setting. *Lecture Notes in Computer Science*: LNAI 5590: 554–565, 2009.

[13] G. Coletti and B. Vantaggi. Probabilistic Reasoning in a Fuzzy Context. *Proc. of the Second World Conference on Soft Computing*, Baku, Azerbaijan, 65–72, 2012.

[14] G. Coletti and B. Vantaggi. Hybrid models: probabilistic and fuzzy information. *Synergies of Soft Computing and Statistics for Intelligent Data Analysis.* In: Kruse, R.; Berthold, M.R.; Moewes, C.; Gil, M.A.; Grzegorzewski, P.; Hryniewicz, O. (Eds.) Advances in Intelligent Systems and Computing. 63-72, 2013.

[15] G. Coletti, R. Scozzafava and B. Vantaggi. Inferential processes leading to possibility and necessity *Information Sciences*, in press (doi: 10.1016/j.ins.2012.10.034).

[16] G. Coletti and B. Vantaggi. Possibility theory: conditional independence. *Fuzzy Sets and Systems*, 157(11): 1491–1513, 2006.

[17] G. Coletti and B. Vantaggi. T-conditional possibilities: coherence and inference. *Fuzzy Set and Systems* 160: 306–324, 2009.

[18] F. Cuzzolin. Three alternative combinatorial formulations of the theory of evidence. *Journal of Intelligent Data Analysis*, 14: 439-464 , 2010.

[19] G. de Cooman. Possibility theory II: Conditional Possibility. International Journal of General Systems, 25: 325-351, 1997.

[20] B. de Finetti. *Teoria della probabilitá.* Torino: Einaudi, (1970) (Engl. Transl. (1974) Theory of probability vol.I,II, London: Wiley & Sons).

[21] B. de Finetti. Sull'impostazione assiomatica del calcolo delle probabilità. *Annali Univ. Trieste*, 19, 3–55 1949 - *Engl. transl.*: Ch.5 in *Probability, Induction, Statistics*, Wiley, London, 1972.

[22] G. de Cooman, E. Miranda, and I. Couso. Lower previsions induced by multi-valued mappings. *J. of Statistical Planning and Inference*, 133: 173–197, 2005.

[23] A.P. Dempster. A generalizatin of Bayesian Inference. *The Royal Stat. Soc. B*, 50: 205–247, 1968.

[24] T. Denoeux, P. Smets, Classification using Belief Functions: the Relationship between the Case-based and Model-based Approaches, *IEEE Transactions on Systems, Man and Cybernetics B*, 36(6): 1395-1406, 2006.

[25] S. Destercke and D. Dubois. Idempotent conjunctive combination of belief functions: Extending the minimum rule of possibility theory. *Information Sciences*, 181: 3925–3945, 2011.

[26] D. Dubois, S. Moral and H. Prade. A semantics for possibility theory based on likelihoods. *J. Math. Anal. Appl.*, 205: 359-380, 1997.

[27] D. Dubois and H. Prade. Possibility theory. Plenum Press, New-York, 1988

[28] D. Dubois and H. Prade. When upper probabilities are possibility measures. *Fuzzy Sets and Systems*, 49: 65–74, 1992.

[29] R. Fagin and J. Y. Halpern. A New Approach to Updating Beliefs. In P. P. Bonissone, M. Henrion, L. N. Kanal, J. F. Lemmer (eds.), *Uncertainty in Artificial Intelligence* 6: 347-374, 1991.

[30] M. J. Frank. On the simultaneous associativity of $F(x, y)$ and $x+y-F(x, y)$. *Aequationes Math.*, 19: 194–226, 1979.

[31] J. Halpern. Reasoning about uncertainty. The MIT Press, Boston, 2003.

[32] E. Hisdal. Are grades of membership probabilities. *Fuzzy Sets and Systems*, 25: 325–348, 1988.

[33] J.Y. Jaffray. Bayesian Updating and Belief Functions. *IEEE Transactions on Systems, Man, and Cybernetics*, 22: 1144-1152, 1992.

[34] C. Kraft, J. Pratt, A. Seidenberg. Intuitive probability on finite sets, *Annals of Mathematical Statistics* 30, 408-419, 1959.

[35] V.I. Loginov. Probability treatment of Zadeh membership functions and theris use in patter recognition. *Engineering Cybernetics*, 68-69, 1966.

[36] M. Mastroleo and B. Vantaggi. An independence concept under plausibility function. *Proceeding of 5th International Symposium on Imprecise Probabilities and their Applications*, 287–296, 2007.

[37] , N. Shilkret. Maxitive measure and integration. Indagationes Mathematicae, 74: 109–116, 1971.

[38] N.D. Singpurwalla and J.M. Booker. Membership functions and probability measures of fuzzy sets (with discussion). *Journal of America Statist. Association* 99: 867–889, 2004.

[39] B. Vantaggi. Statistical matching of multiple sources: A look through coherence. *International Journal of Approximate Reasoning*, 49: 701–711, 2008.

[40] P. Walley. Belief function representations of statistical evidence. *Annals of Statistics*, 4, 1439–1465, 1987.

[41] P. Walley. Statistical reasoning with Imprecise Probabilities. Chapman and Hall, London 1991.

[42] L. Zadeh. Fuzzy sets. *Information and Control*, 8: 338–353, 1965.

Is the mode a lower prevision?

Inés Couso
Dep. Statistics and O.R.
Universidad de Oviedo
couso@uniovi.es

Luciano Sánchez
Dep. Computer Sciences and A.I.
Universidad de Oviedo
luciano@uniovi.es

Abstract

We introduce the notion of mode-desirability of a gamble, that generalizes the idea of non-negativeness of the mode of a random variable. The lower and upper previsions derived from this new definition coincide with the minimum and maximum values of the set of modes of a gamble, when the credal set is a singleton, but they only bound them in the general case. The reason why the minimum and the maximum of the set of modes can not be written, in general, by means of a pair of lower and upper previsions is discussed.

Keywords. Expectation, median, mode, desirability, preference.

1 Introduction

In Decision Making Literature, several criteria of preference between random variables have been proposed within the setting of classical Probability Theory, like for instance stochastic dominance [10], dominance in the sense of expected utility [13], or statistical preference [7, 14], the last one being based on Condorcet's voting criterion ([2]). The above mentioned criteria share a commonality: the joint probability distribution induced by the pair of variables is assumed to be known in order to define each preference criterion, which is expressed in terms of it. Some generalizations of the aforementioned preference criteria have been recently reviewed ([3]) to the case where the joint distribution is not completely determined. Some of those generalizations had been previously introduced in the literature: Denoeux ([8]) generalized first-stochastic dominance to the case of belief-plausibility measures and Destercke ([9]) and Troffaes ([15]), for instance, consider several generalizations of Savage dominance criterion. We have shown that many of those preference generalizations can be expressed in terms of a general formulation that is related to the expectation of a function of both random variables, increasing in the first component and decreasing in the second one.

Differently, in Walley's setting, first hand information is expressed by means of a family of ordered pairs of variables (or "gambles"), the first one in the pair being preferred to the second one. This kind of knowledge can be equivalently represented by means of a coherent family of "desirable" gambles (those preferred to the null one). The family of desirable gambles induces a closed and convex set of linear previsions (also called a "credal set"). Each of those linear previsions is defined on the initial space and induces, for each pair of gambles, a (finitely-additive) joint probability. Thus, what is primary information in this framework is secondary information in the previous setting and vice versa. Notwithstanding, from a purely formal point of view, Walley's almost preference can be seen as a particular case of the general formula introduced in [3], if we consider the function that assigns, to each pair, the difference between both components. With those ideas in mind, we proposed in [6] a generalization of the notion of statistical preference from the setting of classical Probability Theory to the framework of Imprecise Probabilities. It leaded us naturally to a new desirability criterion that we called "signed-desirability". We say that X is signed-desirable if its sign (the gamble that takes the value 1 when X takes a positive value and -1, when it is negative) is desirable, according to Walley's framework. In [5], a set of axioms characterizing the family of signed-desirable gambles induced by a coherent set of desirable gambles is provided. Furthermore, we have found an interesting connection with the notion of median: the infimum and supremum of the set of medians of a gamble, when we range an arbitrary credal set, can be respectively expressed as the lower and upper previsions, according to this new desirability definition.

In this paper, we will propose a new desirability condition very closely related to the notion of mode. The minimum and maximum values of the family of modes of a gamble associated to a single prevision do coincide with the lower and upper previsions of this new desirability condition. However, when we consider an arbitrary credal set, those lower and upper previsions bound the set of modes, but do not necessarily coincide with their minimum and maximum values. We will explore in Section 4 the reasons why those pairs of values do not coincide in general.

2 Preliminaries

The basics on Imprecise Probabilities are assumed to be known by the reader. Notwithstanding we will introduce here the formal notation used in the rest of the paper, and specify the axioms that characterize a coherent family of desirable gambles ([16]). Those axioms have not been stable along the literature in what concerns the inclusion of the null gamble (see [4] for a detailed discussion). In this paper, we will assume it to be non-desirable.

Let Ω denote the set of outcomes of an experiment. $\mathcal{L}$ will denote the set of all gambles (bounded mappings from Ω to $\mathbb{R}$). For $X, Y \in \mathcal{L}$ let $X \geq Y$ mean that $X(\omega) \geq Y(\omega)$, $\forall \omega \in \Omega$ and let $X > Y$ mean that $X \geq Y$ and $X(\omega) > Y(\omega)$ for some $\omega \in \Omega$. A subset $\mathcal{D}$ of $\mathcal{L}$ is said to be a *coherent set of desirable gambles* [16] when it satisfies the following four axioms:

D1. If $X \leq 0$ then $X \notin \mathcal{D}$. *(Avoiding partial loss).*

D2. If $X > 0$, then $X \in \mathcal{D}$. *(Accepting partial gain).*

D3. If $X \in \mathcal{D}$ and $c \in \mathbb{R}^+$, then $cX \in \mathcal{D}$. *(Positive homogeneity).*

D4. If $X \in \mathcal{D}$ and $Y \in \mathcal{D}$, then $X + Y \in \mathcal{D}$. *(Addition).*

The *lower prevision induced by a set of desirable gambles* $\mathcal{D}$ is the set function $\underline{P} : \mathcal{L} \to \mathbb{R}$ defined as follows:

$$\underline{P}(X) = \sup\{c : X - c \in \mathcal{D}\}.$$

The *upper prevision induced by* $\mathcal{D}$ is the set function $\overline{P} : \mathcal{L} \to \mathbb{R}$ defined as follows:

$$\overline{P}(X) = \inf\{c : c - X \in \mathcal{D}\}.$$

The set of linear previsions induced by a coherent set of gambles $\mathcal{D}$ is defined as:

$$\mathcal{P}_{\mathcal{D}} = \{P : P(X) \geq 0 \text{ for all } X \in \mathcal{D}\}.$$

$\mathcal{P}_{\mathcal{D}}$ is always a *credal set* (a closed and convex set of linear previsions, whose restrictions to events are finitely additive probability measures). $\underline{P}$ and $\overline{P}$ are dual and they respectively coincide with the minimum and the maximum of $\mathcal{P}_{\mathcal{D}}$, which can be defined in turn, as the set of linear previsions that dominate $\underline{P}$. On the other hand, a subset $\mathcal{D}^- \subset \mathcal{L}$ satisfying Axioms D2–D4 and

D1'. If $\sup X < 0$ then $X \notin \mathcal{D}^-$. *(Avoiding sure loss).*

D5. If $X + \delta \in \mathcal{D}^-$, for all $\delta > 0$ then $X \in \mathcal{D}^-$. *(Closure).*

is called a coherent set of *almost desirable gambles.* A set of almost desirable gambles $\mathcal{D}^-$ determines a pair of lower and upper previsions, and a credal set, by means of expressions analogous to the case of desirable gambles. Conversely, a credal set univocally determines a coherent set of almost desirable gambles via the formula:

$$\mathcal{D}^-_{\mathcal{P}} = \{X \in \mathcal{L} : P(X) \geq 0, \ \forall P \in \mathcal{P}\}.$$

Finally, a set $\mathcal{D}^+ \subset \mathcal{L}$ is said to be a coherent set of *strict desirable gambles* if it is a coherent set of desirable gambles, and it satisfies, in addition, the following axiom:

D6. If $X \in \mathcal{D}^+$, then either $X > 0$ or $X - \delta \in \mathcal{D}^+$, for some $\delta > 0$. (Openness).

A coherent set of strict desirable gambles can be derived from a credal set as follows:

$$\mathcal{D}^+_{\mathcal{P}} = \{X : X > 0 \text{ or } P(X) > 0 \ \forall P \in \mathcal{P}\}.$$

Let the reader notice that $\mathcal{D}^+_{\mathcal{P}}$ can be alternatively expressed in terms of the lower prevision $\underline{P}$ as follows:

$$\mathcal{D}^+_{\mathcal{P}} = \{X : X > 0 \text{ or } \underline{P}(X) > 0\}. \tag{1}$$

In Walley's theory, the notion of *preference* between two gambles is dual to the above notion of desirability: X is said to be preferred to Y when their difference $X - Y$ is desirable. Conversely, if our primary information is described by means of a partial preference ordering, we will say that X is desirable when it is preferred to the null gamble. Furthermore, there exists a formal connection between preference criteria in classical Probability literature and Walley's notion of preference: in the particular situation where the credal set associated to a preference ordering (according to Walley's view) is a singleton, $\{P\}$, Walley's almost preference of X over Y, $P(X - Y) \geq 0$, is equivalent to dominance according to the expectation, i.e., X is almost preferred to Y if and only if $E_P(X) \geq E_P(Y)$. (In the last expression, P is considered as a probability defined on the set of events, instead of a linear prevision defined in the set of gambles.)

In [3], some known notions of dominance in the (classical) probabilistic setting were reviewed, and it was shown that all of these orderings can be expressed by means of the formula $E_P[g(X,Y)] \geq 0$, where $g : \mathbb{R}^2 \to \mathbb{R}$ is increasing in the first component, and decreasing in the second one. It was also clarified that some generalizations of the above notions considered in the recent literature (see, for instance, [8, 9, 12, 15]) are very closely related to the formula $E_P[g(X,Y)] \geq 0$. This idea made possible to connect Walley's framework, where the initial information is expressed in terms of a partial ordering and the alternative setting considered in those reviewed papers, where the initial information is represented by means of a lower prevision. Therefore, we can join both frameworks and say that X is *g-preferred* to Y if $g(X,Y)$ is desirable according to Walley's framework. With this idea in mind we introduced the notion of sign-desirability in ([6]). X is said to be sign-preferred to Y if $\mathrm{sgn(X-Y)} = 1_{\mathrm{X>Y}} - 1_{\mathrm{Y>X}}$ is desirable, where 1_A denotes the indicator function of $A \subseteq \Omega$, and $X > Y$ and $Y > X$ respectively denote the subsets of Ω where X and Y satisfy each of those inequalities. According to this new preference condition, X is said to be sign-desirable when $\mathrm{sgn(X)} = 1_{\mathrm{X>0}} - 1_{\mathrm{X<0}}$ is desirable. In words, X is said to be sign-desirable when we are disposed to pay one probability currency unit if X takes a negative value in return for the gamble $1_{X>0}$ (receiving 1 unit if X takes a -strictly- positive value.). In [5] an axiomatic characterization of "coherent" sets of sign-desirable gambles is provided. The associated pair of lower and upper previsions can be defined as follows:

$$\underline{P}_S(X) = \sup\{c : X - c \text{ is strictly sign-desirable}\}$$

$$\overline{P}_S(X) = \inf\{c : c - X \text{ is strictly sign-desirable}\}.$$

We have checked in [6] that those lower and upper previsions do coincide, in fact, with the infimum and the supremum of the set of medians of X when we range the credal set associated to the initial coherent set of desirable gambles.

In this paper, we will explore the generalization of the notion of mode, and its connections with Walley's desirability theory. We will introduce a new notion of desirability, but it will not be expressed in terms of the desirability of an increasing function of the considered gamble, as it happens with the notion of sign-desirability. We will also consider the pair of lower and upper previsions of a gamble, according to the new desirability condition. The infimum of the set of modes associated to a credal set will be bounded by the lower prevision, but it will not coincide in general with it.

3 The notion of mode-desirability

Let $\mathcal{L}_F$ denote the family of "simple gambles" (those with a finite number of different possible values). Let us consider an arbitrary but fixed probability measure P on Ω. According to the classical definition, the set of modes of a gamble $X \in \mathcal{L}_F$ with a finite image $Im(X) = \{x_1, \ldots, x_n\}$ is defined as follows:

$$\begin{aligned}&Mo_P(X) = \\ &\{x_i \in Im(X) : P(X = x_j) \leq P(X = x_i),\ \forall j \neq i\} = \\ &\{x_i \in Im(X) : \nexists x_j \neq x_i \text{ with } P(X = x_j) > P(X = x_i)\} = \\ &\{x_i \in Im(X) : \nexists j \neq i \text{ s.t. } E_P(1_{X=x_j} - 1_{X=x_i}) > 0\}.\end{aligned}$$

Let us now consider the credal set, $\mathcal{P}_{\mathcal{D}}$, associated to an arbitrary coherent set of desirable gambles $\mathcal{D}$. Let $\underline{P}$ denote the induced lower prevision. A natural way to extend the classical notion of mode seems to be the following one:

$$\begin{aligned}&Mo_{\underline{P}}(X) = \\ &\{x_i \in Im(X) : \underline{P}(1_{X=x_j} - 1_{X=x_i}) \leq 0,\ \forall j \neq i\} = \\ &\{x_i \in Im(X) \quad \nexists j \neq i \text{ s.t. } \underline{P}(1_{X=x_j} - 1_{X=x_i}) > 0\}.\end{aligned}$$

We will prove the following result, in order to connect this definition with Walley's desirability framework.

Lemma 1 *Let $\underline{P}$ be the lower prevision induced by a coherent set of gambles $\mathcal{D}$. Let $\mathcal{D}^+_{\mathcal{P}}$ be the set associated set of strictly desirable gambles, according to Equation 1. Let $X \in \mathcal{L}_F$. For every $x \in Im(X)$ and all $y \in \mathbb{R}$:*

$$\underline{P}(1_{X=y} - 1_{X=x}) > 0 \quad \textit{iff} \quad 1_{X=y} - 1_{X=x} \in \mathcal{D}^+.$$

Proof: By definition, the gamble $1_{X=y} - 1_{X=x}$ is strictly desirable if and only if it is some of the following conditions are fulfilled:

$$\underline{P}(1_{X=y} - 1_{X=x}) > 0 \text{ or } 1_{X=y} - 1_{X=x} > 0.$$

But $1_{X=y} - 1_{X=x} > 0$ implies that x does not belong to the set of outcomes of X, what is a contradiction. □

According to the above lemma, we can alternatively express the set of modes as follows:

$$\begin{aligned}&Mo_{\underline{P}}(X) = \\ &\{x_i \in Im(X) : \nexists j \neq i \text{ s.t. } 1_{X=x_j} - 1_{X=x_i} \in \mathcal{D}^+\} = \\ &\{x_i \in Im(X) : \nexists j \neq i \text{ s.t. } (1_{\{x_j\}} - 1_{\{x_i\}}) \circ X \in \mathcal{D}^+\},\end{aligned}$$

where the symbol "$\circ$" stands for the composition of functions.

Furthermore, we can skip our reference to the set of outcomes of X by taking into account the following result.

Lemma 2 *Let us consider a credal set $\mathcal{P}$, and let $\mathcal{D}^+$ denote the set of strictly desirable gambles induced by it. Let $X \in \mathcal{L}$. Then:*

1. *If $y \notin Im(X)$, and $x \in \mathbb{R}$, $1_{X=y} - 1_{X=x} \notin \mathcal{D}^+$.*
2. *$A_X^+ = \{x : \nexists y \neq x$ s.t. $(1_{\{y\}} - 1_{\{x\}}) \circ X \in \mathcal{D}^+\}$ is included in $Im(X)$.*

Proof:

1. If $y \notin Im(X)$, then $(1_{X=y} - 1_{X=x}) = -1_{X=x} \leq 0$. According to Axiom D1, this gamble does not belong to $\mathcal{D}^+$.
2. The second part is also straightforward: if $x \notin Im(X)$, then $(1_{\{y\}} - 1_{\{x\}}) \circ X > 0, \forall\, y \in Im(X)$, and therefore, the gamble $(1_{\{y\}} - 1_{\{x\}}) \circ X$ belongs to $\mathcal{D}^+$ for every $y \in Im(X) \subseteq \mathbb{R} \setminus \{x\}$. □

According to the above lemma, the set of modes associated to the credal set, $Mo_{\underline{P}}(X)$, can be alternatively expressed as:

$$Mo_{\underline{P}}(X) = A_X^+ =$$
$$\{x : \nexists y \neq x \text{ s.t. } (1_{\{y\}} - 1_{\{x\}}) \circ X \in \mathcal{D}^+\}.$$

This new expression suggests us to consider the following new desirability condition. We will say that X is mode-desirable when $Mo_{\underline{P}}(X) = A_X^+$ does not contain any negative number:

Definition 1 *A gamble $X \in \mathcal{L}_F$ is said to be* mode-desirable, *if*

$$[\forall\, x < 0,\ \exists y \neq x \text{ s.t. } (1_{\{y\}} - 1_{\{x\}}) \circ X \in \mathcal{D}^+].$$

We will denote it $X \in \mathcal{D}_{Mo}$.

Remark 3.1 *There is an alternative equivalent definition for the notion of mode-desirability of simple gambles. In fact we can check that X is mode-desirable if and only if:*

$$[\forall\, x < 0,\ \exists y > x \text{ s.t. } (1_{\{y\}} - 1_{\{x\}}) \circ X \in \mathcal{D}^+].$$

One of the implications is straightforward, so we just need to check the second one: Let us suppose that $X \in \mathcal{D}_{Mo}$ and let us consider an arbitrary but fixed value $x \leq 0$. According to the definition of $\mathcal{D}_{Mo}$, there exists $y_1 \neq x$ such that $(1_{\{y_1\}} - 1_{\{x\}}) \circ X$. Furthermore, we can assure that y_1 belongs to $Im(X)$. If $y_1 > x$, the proof is finished. Otherwise, there will exist $y_2 \neq y_1$, $y_2 \in Im(X)$ such that $(1_{\{y_2\}} - 1_{\{y_1\}}) \circ X \in \mathcal{D}^+$. According to the additivity of $\mathcal{D}^+$ (Axiom D4), we can easily check that $(1_{\{y_2\}} - 1_{\{x\}}) \circ X \in \mathcal{D}^+$. According to this procedure, after a finite number of steps, $k \leq \#Im(X)$, we will get $y_{k+1} > x$ such that $(1_{y_{k+1}} - 1_{y_k}) \circ X \in \mathcal{D}^+$. Otherwise, we would need to assume that y_n is less than or equal to x, and it would lead us to a contradiction, because, there would need to exist $y \notin Im(X)$ with $(1_y - 1_{y_n}) \circ X \in \mathcal{D}^+$.

If X is mode-desirable, then, for every $x < 0$, there exists some $y \neq x$ such that we are disposed to exchange the gamble $1_{X=x}$ in return for the gamble $1_{X=y}$. The new desirability condition induces a pair of lower and upper previsions as follows:

Definition 2 *Let $\mathcal{D}$ be a coherent family of desirable gambles, and let $\mathcal{D}_{Mo}$ denote the family of mode-desirable gambles induced by it. Let $X \in \mathcal{L}_F$. The lower prevision of X is defined as follows:*

$$\underline{P}_{Mo}(X) = \sup\{c \in \mathbb{R} : X - c \in \mathcal{D}_{Mo}\}$$

Analogously, the upper prevision is:

$$\overline{P}_{Mo}(X) = \inf\{c \in \mathbb{R} : c - X \in \mathcal{D}_{Mo}\}.$$

Now we will prove that the minimum and the maximum values of the set A_X^+ do coincide with the pair of lower and upper previsions defined above. Let us first prove the following supporting result:

Lemma 3

- *The set $C = \{c : X - c \in \mathcal{D}_{Mo}\}$ can be alternatively expressed as:*

  $$\{c : [x < c \Rightarrow \exists\, y \neq x \ \text{ with } \ (1_{\{y\}} - 1_{\{x\}}) \circ X \in \mathcal{D}^+]\} =$$
  $$\{c : [x < c \Rightarrow x \notin A_X^+]\} = (-\infty, \min A_X^+].$$

- *The set $D = \{d : d - X \in \mathcal{D}_{Mo}\}$ can be alternatively written as:*

  $$\{d : [x > d \Rightarrow \exists\, y \neq x \ \text{ with } \ (1_{\{y\}} - 1_{\{x\}}) \circ X \in \mathcal{D}^+]\} =$$
  $$\{d : [x > d \Rightarrow x \notin A_X^+]\} = [\max A_X^+, \infty).$$

Proof: The proof is almost immediate, if we take into account that $1_{\{y\}} \circ (X - c) = 1_{\{y+c\}} \circ X$, and $1_{\{y\}} \circ (d - X) = 1_{\{d-y\}} \circ X\ \forall\, c, d, y \in \mathbb{R}$. □

The next result is straightforward, according to the above lemma:

Proposition 4 *The following equalities hold: $\min A_X^+ = \underline{P}_{Mo}(X)$ and $\max A_X^+ = \overline{P}_{Mo}(X)$.*

Remark 3.2 *According to the proof of Lemma 3, the supremum of C and the infimum of D are, indeed, maximum and minimum values, respectively, and they do coincide with the minimum and the maximum of A_X^+, respectively.*

Let us now consider the set of mode values associated to the credal set:

$$Mo_{\mathcal{P}_{\mathcal{D}}}(X) = \cup_{P \in \mathcal{P}_{\mathcal{D}}}\{Mo_P(X)\}.$$

If it coincided with A^+_X, the minimum and the maximum of the family of modes associated to the credal set would coincide with the lower and upper previsions of X, according to the notion of mode-desirability. Nevertheless, those lower and upper previsions just bound, but they do not coincide in general with the minimum and maximum of the set of modes of X. More specifically, we can check that:

Proposition 5 *The set of mode values associated to the credal set $\mathcal{P}_{\mathcal{D}}$, $Mo_{\mathcal{P}_{\mathcal{D}}}(X)$ is included in A^+_X. Furthermore, if the credal set is a singleton, both sets of values do coincide.*

Proof: The set of modes can be expressed as follows:

$$Mo_{\mathcal{P}_{\mathcal{D}}}(X) = \cup_{P \in \mathcal{P}_{\mathcal{D}}}\{Mo_P(X)\} =$$

$$\cup_{P \in \mathcal{P}_{\mathcal{D}}}\{x \,:\, \forall y \neq x, P(1_{\{y\}} - 1_{\{x\}}) \circ X \leq 0\} =$$

$$\{x \,:\, \exists P \in \mathcal{P}_{\mathcal{D}} \text{ s.t. } \; \forall y \neq x P(1_{\{y\}} - 1_{\{x\}}) \circ X \leq 0\}.$$

On the other hand,

$$A^+_X = \{x \,:\, \forall y \neq x, \underline{P}(1_{\{y\}} - 1_{\{x\}}) \circ X) \leq 0\}.$$

According to the above expressions, and taking into account that $\underline{P}$ is the minimum of the credal set, we can easily derive the thesis of this proposition. □

According to the last results, A^+_X is a finite set containing the set of modes, $Mo_{\mathcal{P}_{\mathcal{D}}}(X)$, and included in the set of images of X. Under some additional constraints ($\mathcal{P}_{\mathcal{D}}$ being a singleton or, contrarily, expressing vacuous information, or A^+_X being included in the set of images with maximum upper probability, etc.) they do coincide. But they do not in general, as we illustrate in the following example.

Example 1 *Let Ω be a finite set with four elements, $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ and let us consider the credal set $\mathcal{P} = \{(\frac{3}{8}-\alpha, \frac{1}{8}-\frac{\alpha}{4}, \frac{1}{8}+\frac{\alpha}{4}, \frac{3}{8}+\alpha) \,:\, \alpha \in [-\frac{3}{8}, \frac{3}{8}]\}$. In the above formula, each vector of the form (p_1, p_2, p_3, p_4) represents the linear prevision P defined as:*

$$P(X) = \sum_{i=1}^{4} p_i X(\omega_i), \; \forall X \in \mathcal{L}.$$

Let $\mathcal{D}^+_{\mathcal{P}}$ denote the set of strictly desirable gambles associated to $\mathcal{P}$: $\mathcal{D}^+_{\mathcal{P}} = \{Y \,:\, Y > 0 \;\; or \; \underline{P}(Y) > 0\}$. Let us now consider the gamble X defined as $X(\omega_i) = i$, $i = 1, 2, 3, 4$. Let A^+_X denote the collection of numbers:

$$A^+_X = \{x \,:\, \nexists y \neq x \text{ with } (1_{\{y\}} - 1_{\{x\}}) \circ X \in \mathcal{D}^+\} =$$

$$\{i \in \{1, \ldots, 4\} \,:\, \forall j \neq i, \; \underline{P}(1_{\{\omega_j\}} - 1_{\{\omega_i\}}) \leq 0\}.$$

$A^+_X = \{1, 2, 3, 4\}$, but $Mo_{\mathcal{P}}(X) = \{1, 4\}$. In order to check it, Tables 1 and 2 respectively display, for each pair (j, i), the value that the linear prevision $P_\alpha \equiv (\frac{3}{8} - \alpha, \frac{1}{8} - \frac{\alpha}{4}, \frac{1}{8} + \frac{\alpha}{4}, \frac{3}{8} + \alpha)$ and the lower prevision $\underline{P} = \min_{\alpha \in [-\frac{3}{8}, \frac{3}{8}]} P_\alpha$ assign to the gamble $(1_{\{x_j\}} - 1_{\{x_i\}}) \circ X = 1_{\{\omega_j\}} - 1_{\{\omega_i\}}$.

$j \setminus i$	1	2	3	4
1	0	$\frac{1}{4} - \frac{3\alpha}{4}$	$\frac{1}{4} - \frac{5\alpha}{4}$	2α
2	$\frac{3\alpha}{4} - \frac{1}{4}$	0	$-\frac{\alpha}{2}$	$-\frac{1}{4} - \frac{5\alpha}{4}$
3	$\frac{5\alpha}{4} - \frac{1}{4}$	$\frac{\alpha}{2}$	0	$-\frac{1}{4} - \frac{3\alpha}{4}$
4	-2α	$\frac{1}{4} - \frac{5\alpha}{4}$	$\frac{1}{4} + \frac{3\alpha}{4}$	0

Table 1: It displays $P_\alpha(1_{\{\omega_j\}} - 1_{\{\omega_i\}})$, for each (j, i).

$j \setminus i$	1	2	3	4
1	0	$-\frac{1}{32}$	$-\frac{7}{32}$	$-\frac{3}{4}$
2	$-\frac{17}{32}$	0	$-\frac{3}{16}$	$-\frac{3}{8}$
3	$-\frac{23}{32}$	$-\frac{3}{16}$	0	$-\frac{1}{32}$
4	$-\frac{3}{4}$	$-\frac{7}{32}$	$-\frac{1}{32}$	0

Table 2: It displays $\underline{P}(1_{\{\omega_j\}} - 1_{\{\omega_i\}})$, for each (j, i).

None of the values in Table 2 is strictly positive, and this means that A^+_X coincides with the set of possible outcomes of the gamble X, $\{1, 2, 3, 4\}$. On the other hand, there does not exist any $\alpha \in [-\frac{3}{8}, \frac{3}{8}]$ such that the values 2 or 3 belong to the set of modes of X associated to the linear prevision P_α, $Mo_{P_\alpha}(X)$. Thus, the set of modes associated to the credal set, $Mo_{\mathcal{P}}(X)$, is strictly included in A^+_X.

We can ask ourselves what happens if we replace, $\mathcal{D}^+$ by $\mathcal{D}$ or $\mathcal{D}^-$ in the construction of the set of values:

$$\{x \,:\, \nexists y \neq x \text{ with } \; (1_{\{y\}} - 1_{\{x\}}) \circ X \in \mathcal{D}^+\}.$$

Let us consider the pair of sets:

$$A_X = \{x \,:\, \nexists y \neq x \text{ with } \; (1_{\{y\}} - 1_{\{x\}}) \circ X \in \mathcal{D}\}$$

and

$$A^-_X = \{x \,:\, \nexists y \neq x \text{ with } \; (1_{\{y\}} - 1_{\{x\}}) \circ X \in \mathcal{D}^-\} =$$

$$\{x \,:\, \underline{P}((1_{\{y\}} - 1_{\{x\}}) \circ X) < 0, \; \forall y \neq x\},$$

and let us compare them with A^+_X.

Lemma 6 *$A^-_X \subseteq A_X \subseteq A^+_X$. Furthermore, if $\mathcal{P}_{\mathcal{D}}$ is a singleton, $\mathcal{P}_{\mathcal{D}} = \{P\}$, then $A^-_X = \emptyset$, unless the distribution of X is unimodal. In that case, $A^-_X = A_X = A^+_X = Mo_P(X)$.*

Proof: The first part is easy to prove if we take into account the chain of inclusions $\mathcal{D}^+ \subseteq \mathcal{D} \subseteq D^-$. Secondly, if $\mathcal{P}_{\mathcal{D}} = \{P\}$, we can easily check that x belongs to A^-_X if and only if $P(1_{X=y}) < P(1_{X=x})$, $\forall\, y \neq x$. This only happens when x is the only mode of X, with respect to the linear prevision P. □

Remark 3.3 *Using expressions analogous to those considered in Lemma 3, we can easily prove that the minimum and maximum of A^-_X do respectively coincide with* $\sup\{c : X - c \in \mathcal{D}^-_{Mo}\}$ *and* $\inf\{d : d - X \in \mathcal{D}^-_{Mo}\}$*, where $\mathcal{D}^-_{Mo}$ is defined as:*

$$\{X \in \mathcal{L}_F : \forall\, x < 0\ \exists\, y \neq x \text{ s.t. } (1_{\{y\}} - 1_{\{x\}}) \circ X \in \mathcal{D}^-\}.$$

Furthermore, we have seen that A^-_X is included in A^+_X, and that the last one coincides with the set of modes, when the credal set is a singleton. We can ask ourselves whether A^-_X is, in general a subset of $Mo_{\mathcal{P}}(X)$, and therefore it approximates it from below. But we can easily check that this does not happen. In Example 1, we have shown that none of the lower previsions displayed in Table 2 was strictly positive. Furthermore, we observe that all of them are negative (except for those in the diagonal). This means that A^-_X also coincides with the whole family of possible outcomes of X, $A^-_X = \{1, 2, 3, 4\}$ and therefore, it strictly includes the set of mode values associated to the credal set.

4 What's the problem with mode-desirability?

In Walley's framework ([16]), any coherent set of gambles satisfies Axioms D2 and D4. The following property can be easily derived from both axioms:

$$Y \in \mathcal{D}, \text{ and } \ X > Y \Rightarrow X \in \mathcal{D}. \tag{2}$$

On the other hand, the set of sign-desirable gambles induced by a coherent set of gambles $\mathcal{D}$ satisfies Axiom D2, but it does not necessarily satisfy Axiom D4. However we can easily check that it fulfills the property mentioned in Equation 2, since it is connected to $\mathcal{D}^+$ through the function sgn : $\mathbb{R} \to \mathbb{R}$, that is increasing. More explicitly:

Definition 3 *Let $\mathcal{D}$ be a coherent set of desirable gambles, and let $f : \mathbb{R} \to \mathbb{R}$ be an increasing function. We will say that X is f-desirable if and only if $f(X)$ belongs to $\mathcal{D}$. We will denote it $X \in \mathcal{D}_f$.*

Lemma 7 *Let $\mathcal{D}$ be a coherent set of desirable gambles, and let $f : \mathbb{R} \to \mathbb{R}$ be an increasing function. The set of f-desirable gambles satisfies the property:*

$$X \in \mathcal{D}_f, Y > X \Rightarrow Y \in \mathcal{D}_f.$$

A "coherent" set of mode-desirable gambles does not necessarily satisfy the property considered in Equation 2 as we illustrate in Example 2:

Example 2 *Let Ω be the unit interval, and let P denote the uniform probability distribution defined on it. Let Y denote the gamble defined as follows:*

$$Y(\omega) = \begin{cases} -1 & \text{if } \omega \in [0, 1/3) \\ 1 & \text{if } \omega \in [1/3, 5/6) \\ 2 & \text{if } \omega \in [5/6, 1] \end{cases}$$

Y takes the values -1, 1 and 2 with respective probabilities $1/3$, $1/2$ and $1/6$. Thus, we can easily check that Y is mode-desirable, since $P(1_{\{1\}} - 1_{\{x\}} \circ Y) > 0,\ \forall\, x < 0$. Let us now consider the gamble:

$$X(\omega) = \begin{cases} -1 & \text{if } \omega \in [0, 1/3) \\ 1 & \text{if } \omega \in [1/3, 1/2) \\ 2 & \text{if } \omega \in [1/2, 2/3) \\ 3 & \text{if } \omega \in [2/3, 5/6) \\ 4 & \text{if } \omega \in [5/6, 1] \end{cases}$$

We clearly see that $Y \geq X$, but it is not mode-desirable. In fact, for $x = -1$ there does not exist any $y > x$ such that $P(1_{\{y\}} - 1_{\{x\}} \circ X) > 0$.

From this example, and according to Lemma 7, a "coherent" sets of mode-desirable gambles can not be expressed, in general, as the family of f-desirable gambles, according to some increasing function $f : \mathbb{R} \to \mathbb{R}$ and some coherent set of desirable gambles $\mathcal{D}$. This fact seems to be essential in relation with the properties of the lower and upper previsions derived from it, as we show below.

Lemma 8 *Let $\mathcal{D}$ be a coherent set of desirable gambles, and let us consider an increasing function $f : \mathbb{R} \to \mathbb{R}$. The set $C = \{c : f(X - c) \in \mathcal{D}\}$ satisfies the following property: $c \in C, c' \leq c \Rightarrow c' \in C$.*

Proof: Let us suppose that $c \in C$ and $c' \leq c$. By definition, $f(X - c) \in \mathcal{D}$. According to the properties of f, $f(X - c') \geq f(X - c)$ and, therefore, according to the coherence of $\mathcal{D}$, $f(X - c')$ belongs to it. □

Proposition 9 *Let $\mathcal{D}$ be a coherent set of desirable gambles, and let us consider an increasing function $f : \mathbb{R} \to \mathbb{R}$. Let $\mathcal{D}^+_f$ denote the set of $f-$desirable gambles with respect to the coherent set $\mathcal{D}^+$, $\mathcal{D}^+_f = \{X : f(X) \in \mathcal{D}^+\}$. Let us also consider, for every $P \in \mathcal{P}_{\mathcal{D}}$, the set of f-desirable gambles with respect to $\mathcal{D}^+_{\{P\}}$, i.e.: $\mathcal{D}^+_{f,\{P\}} = \{X : f(X) > 0 \ \text{ or } P(f(X)) > 0\}$. Then:*

$$\sup\{c : X - c \in \mathcal{D}^+_f\} = \inf_{P \in \mathcal{P}_{\mathcal{D}}} \sup\{c : X - c \in \mathcal{D}^+_{f,\{P\}}\}.$$

Proof: First of all, let us take into account that $\mathcal{D}^+ \subseteq \mathcal{D}^+_{\{P\}}$, and therefore $\mathcal{D}^+_f \subseteq \mathcal{D}^+_{f,\{P\}}$, $\forall\, P \in \mathcal{P}$. Thus, the set $\{c : X - c \in \mathcal{D}^+_f\}$ is included in $\{c : X - c \in \mathcal{D}^+_{f,\{P\}}\}$, $\forall\, P \in \mathcal{P}$, and therefore

$$\sup\{c : X - c \in \mathcal{D}^+_f\} \leq \inf_{P \in \mathcal{P}_{\mathcal{D}}} \sup\{c : X - c \in \mathcal{D}^+_{f,\{P\}}\}.$$

Let us now prove the reverse inequality.Let c_P denote the supremum of the set $\{c : f(X - c) \in \mathcal{D}^+_{f,\{P\}}\}$ and let $c = \inf_{P \in \mathcal{P}} c_P$. Let us consider an arbitrary $c' < c$. It will suffice to check that, $c' \in \{c : X - c \in \mathcal{D}^+_f\}$. Let us consider the difference $\epsilon = c - c' > 0$. According to the definition of supremum, for every $P \in \mathcal{P}$ there exists $c'_P \in \{c : X - c \in \mathcal{D}^+_{f,\{P\}}$ such that $c_P - \epsilon < c'_P \leq c_P$. Therefore, $c' \leq \inf_{P \in \mathcal{P}} c'_P$ and thus, according to Lemma 8, $f(X - c') \in \mathcal{D}^+_{\{P\}}$, $\forall\, P \in \mathcal{P}$. Having into account that $\mathcal{D}^+ = \cap_{P \in \mathcal{P}} \mathcal{D}^+_{\{P\}}$, we have that $c' \in \{c : X - c \in \mathcal{D}^+_f\}$, and the result is proved. □

According to the last result, when we consider an increasing function $f : \mathbb{R} \to \mathbb{R}$, and the supremum $\sup\{c : f(X - c) \in \mathcal{D}^+_{\{P\}}\}$ coincides with some well-known parameter, $\theta_P(X)$ induced by the probability distribution P_X (like, for instance, the expectation for $f(\cdot) = \cdot$, or the infimum of the interval of medians, for $f = \text{sgn}$, the supremum $\sup\{c : f(X-c) \in \mathcal{D}^+\}$ coincides with the infimum of the values of the parameter, when we range the credal set, $\inf_{P \in \mathcal{P}_{\mathcal{D}}} \theta_P(X)$.

The condition of mode-desirability cannot be expressed in terms of an increasing function. According to Example 2, it is something inherent to the standard definition of mode, and it does not depend on the particular definition we have introduced in order to extend the idea of non-negativity of the mode to the Imprecise Probabilities framework. Even for the family of single-pointed credal sets, we cannot find an increasing function $f : \mathbb{R} \to \mathbb{R}$ such that $\sup\{c : f(X - c) \in \mathcal{D}^+_{\{P\}}\} = \min Mo_P(X)$, for every linear prevision, P.

5 Alternative definitions of mode desirability

As we have mentioned in the introduction, [3] reviews several classical stochastic preference criteria and shows that many of them can be written according to the general formulation:

$$X \text{ is preferred to } Y \text{ iff } E_P(g(X, Y)) \geq 0,$$

where $g : \mathbb{R}^2 \to \mathbb{R}$ is increasing in the first component and decreasing in the second one. Furthermore, in most cases, g can be expressed in terms of an increasing point-to-point function $f : \mathbb{R} \to \mathbb{R}$ as $g(x, y) = f(x) - f(y)$, $\forall\, (x, y) \in \mathbb{R}^2$. As we clarify in [3], some extensions of those stochastic orderings introduced in the recent literature ([6, 8, 9, 11, 15]) can be written in terms of the non-negativity of the lower prevision of $g(X, Y)$. Some others, instead, take into account the pairs of lower and upper previsions of $f(X)$ and $f(Y)$, $(\underline{E}(f(X)), \overline{E}(f(X)))$ and $(\underline{E}(f(Y)), \overline{E}(f(Y)))$. Based on both pairs, we can generate four different preference relations, that, for the sake of shortness, will be called min-max, max-max, max-min and min-min.

In Section 3, we considered the following generalization of the notion of mode:

$$Mo_{\underline{P}}(X) = \{x_i : \underline{P}(1_{X=x_j} - 1_{X=x_i}) \leq 0,\ \forall\, j \neq i\}.$$

Instead of the lower prevision of gambles of the form $(1_{\{x_j\}} - 1_{\{x_i\}}) \circ X$, we can alternatively consider the pairs of lower and upper previsions of the gambles $1_{\{x_j\}} \circ X$ and $1_{\{x_i\}} \circ X$ and compare them, according to the four criteria mentioned in the last paragraph. In this section we will briefly discuss these four alternative definitions.

Min-max criterion

Let $\underline{P}$ and $\overline{P}$ respectively denote the lower and upper previsions induced by a credal set $\mathcal{P}$. Let $X \in \mathcal{L}_F$ be an arbitrary simple gamble. We will define the *min-max-mode* of X with respect to $\mathcal{P}$ as the set:

$${}^{M}_{m}Mo_{\mathcal{P}}(X) = \{x_i : \underline{P}(1_{X=x_j}) \leq \overline{P}(1_{X=x_i}),\ \forall\, j \neq i\}.$$

According to the super-additivity of $\underline{P}$, and the duality between $\underline{P}$ and $\overline{P}$, the following inequality holds:

$$\underline{P}(1_{X=x_j}) - \overline{P}(1_{X=x_i}) \geq \underline{P}(1_{X=x_j}) - \overline{P}(1_{X=x_i}),$$

and therefore, we can easily check that the max-min-mode of X contains the set $Mo_{\underline{P}}(X)$, that is, in turn, a superset of the family of modes of X, when we range the credal set. Therefore, the max-min-mode is even less precise than our initial generalization of the mode.

Max-max criterion

We will define the *max-max-mode* of X with respect to $\mathcal{P}$ as follows:

$${}^{M}_{M}Mo_{\mathcal{P}}(X) = \{x_i : \overline{P}(1_{X=x_j}) \leq \overline{P}(1_{X=x_i}),\ \forall\, j \neq i\}.$$

This set is included in the set of modes of X, when we range the credal set. In fact, according to the coherence of $\overline{P}$, it is the maximum of the credal set, $\mathcal{P}$, and that means that there exists, for every $i \in$

${}^{M}_{M}Mo_{\mathcal{P}}(X)$, some $P_i \in \mathcal{P}$ that satisfies the equality $P_i(1_{X=x_i}) = \overline{P}(1_{X=x_i})$, that satisfies, by definition, the inequalities $\overline{P}(1_{X=x_i}) \geq \overline{P}(1_{X=x_j})$, $\forall\, j$. Thus, we get the inequalities:

$$P_i(1_{X=x_i}) = \overline{P}(1_{X=x_i}) \geq \overline{P}(1_{X=x_j}) \geq P_i(1_{X=x_j}),\ \forall\, j.$$

Therefore, the max-max-mode approximates the set of modes from below.

Max-min criterion

We will define the *max-min-mode* of X with respect to $\mathcal{P}$ as follows:

$${}^{m}_{M}Mo_{\mathcal{P}}(X) = \{x_i\, :\, \overline{P}(1_{X=x_j}) \leq \underline{P}(1_{X=x_i}),\ \forall\, j \neq i\}.$$

This set of values is clearly included in the max-max-mode, and therefore, it is a less precise approximation of the family of modes $Mo_{\mathcal{P}}(X)$.

Min-min criterion

We will define the *min-min-mode* of X with respect to $\mathcal{P}$ as the set:

$${}^{m}_{m}Mo_{\mathcal{P}}(X) = \{x_i\, :\, \underline{P}(1_{X=x_j}) \leq \underline{P}(1_{X=x_i}),\ \forall\, j \neq i\}.$$

$\underline{P}((1_{\{x_j\}} - 1_{\{x_i\}}) \circ X) \leq \overline{P}((1_{\{x_j\}} - 1_{\{x_i\}}) \circ X) \leq \overline{P}(1_{\{x_j\}} \circ X) - \underline{P}(1_{\{x_i\}} \circ X),\ \forall\, i, j.$

The above set does not necessarily include, nor is it necessarily included in the family of modes, $Mo_{\mathcal{P}}(X)$. Both sets may even be disjoint, as it happens in the following example.

Example 3 *Let us consider again the credal set of Example 1,* $\mathcal{P} = \{(\frac{3}{8} - \alpha, \frac{1}{8} - \frac{\alpha}{4}, \frac{1}{8} + \frac{\alpha}{4}, \frac{3}{8} + \alpha) : \alpha \in [-\frac{3}{8}, \frac{3}{8}]\}$. *The lower previsions of the gambles of the form* $1_{X=x_i}$, $i = 1, 2, 3, 4$, *are, respectively* 0, $\frac{1}{32}$, $\frac{1}{32}$ *and* 0. *Thus, the min-min-mode,* ${}^{m}_{m}Mo_{\mathcal{P}}(X) = \{2, 3\}$ *is the complementary of the set of modes of* X, $Mo_{\mathcal{P}}(X) = \{1, 4\}$.

6 Concluding remarks and open problems

We have introduced the notion of mode-desirability, and connected the classical notion of mode to Walley's desirability framework. The lower and upper previsions of a gamble bound, but do not necessarily coincide with the minimum and the maximum of the set of modes, when we consider an arbitrary credal set. In Section 4, we have discussed the reason why there does not seem to exist a way to express the pair of minimum and maximum values as the pair of lower and upper previsions, according to some desirability condition.

We have also studied four alternative generalizations of the notion of mode. The "min-max" approach leads to a pair of bounds that are even less precise than the lower and upper previsions induced from the notion of mode-desirability. Notwithstanding, the number of comparisons needed to calculate the outer approximation A^+_X is greater than the number needed in order to calculate the min-max mode. It will be the expert that uses those approximations in practical problems who has to decide what is the most convenient procedure in each specific situation. On the other hand, the min-min mode does not seem to be related in general with the set of modes. Finally, the max-min and the max-max modes are included in the family of modes, the last one being the most precise of the two. In a specific problem, we can consider the outer and inner approximations of $Mo_{\mathcal{P}}(X)$ respectively derived from the notions of mode-desirability (or, alternatively, the min-max mode, when the calculation of A^+_X is non-viable) and max-max mode. According to the notion of upper prevision, the max-max mode can be alternatively expressed as:

$$\left\{x_i\, :\, \cup_{j=1}^{n}\{d\, :\, d - 1_{X=x_j} \not\in \mathcal{D}\} \subseteq \{d\, :\, d - 1_{X=x_i} \not\in \mathcal{D}\}\right\}. \tag{3}$$

The max-max mode and the set A^+_X approximate the set of bounds, respectively from below and above. At first sight, the problem of characterizing the set of modes associated to a credal set seems to be more complicated: the mode of a linear convex combination is not between the modes of both extremes. Therefore, the set of modes associated to a credal set does not seem to be easily characterized by the modes of the extremes, as it happens with other parameters, like the entropy (see [1], for instance). At least, the fact of departing from a pair of inner and outer approximations can simplify the process of characterizing the set of modes in some specific problems.

In the future, we plan to study the properties of the desirability condition that matches with the generalization of the notion of mode considered in Equation 3, as well as for the notion of mode-desirability. According to the definition introduced in this paper, a gamble is mode-desirable if and only if $A^+_X \cap (-\infty, 0) \neq \emptyset$. The set of mode-desirable gambles does not satisfy, in general, Axiom D1 ("avoiding partial loss"). In order to overcome this inconvenient, we could have alternatively considered X to be mode-desirable if and only if $A^+_X \cap (-\infty, 0] = \emptyset$. But this would not entail a substantial improvement, since the set of mode-desirable gambles would no longer satisfy Axiom D2 ("accepting partial gain"). We plan to study other alternatives in order to find a new definition that simultaneously satisfies both axioms.

We also plan to study necessary and sufficient conditions for a credal set $\mathcal{P}$ in order to satisfy the equality $Mo_{\mathcal{P}}(X) = A_X^+$, so that the minimum and the maximum of the set of modes do coincide with the lower and upper previsions induced by the set of mode-desirable gambles.

In the paper, we have assumed that the outcomes of the gambles were numbers, but we could easily extended this framework to a non-necessarily numerical setting. The definitions of mode-desirability and lower and upper prevision would require, anyway, the universe being an ordered set including a "neutral" element that plays the role of the value 0 in the real line.

Acknowledgements

We thank very much three anonymous reviewers for their detailed reports and insightful comments. This work has been jointly supported by the Spanish Ministry of Education and Science and the European Regional Development Fund (FEDER), under projects TIN2011-24302 and TIN2010-20900-C04.

References

[1] J. Abellán and S. Moral, Maximum of Entropy for Credal Sets. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 11 (2003) 587-598.

[2] M. de Condorcet, Essay sur le application de l'analyse à la probabilité des decisions rendues à la pluralité des voix, Imprimirie Royale, Paris, 1785.

[3] I. Couso and D. Dubois, An Imprecise Probability Approach to Joint Extensions of Stochastic and Interval Orderings, Lecture Notes in Artificial Intelligence / Communications in Computer and Information Science 299, Part III, p.388-399, Springer, 2012.

[4] I. Couso and S. Moral, Sets of desirable gambles: Conditioning, representation, and precise probabilities, International Journal of Approximate Reasoning, 52 (2011) 1034-1055.

[5] I. Couso, S. Moral and L. Sánchez, The median as the lower prevision with respect to sign-desirability *In preparation*, 2013.

[6] I. Couso and L. Sánchez, The behavioral meaning of the median, in *C. Borgelt et al., Eds., Combining Soft Computing and Statistical Methods in Data Analysis*, Springer, 2010.

[7] H. David, The method of paired comparisons, Griffin's Statistical Monographs & Courses, vol. 12, Charles Griffin & D. Ltd., London, 1963.

[8] T. Denoeux, Extending stochastic ordering to belief functions on the real line. Inf. Sci. 179 (2009) 1362-1376.

[9] S. Destercke, A decision rule for imprecise probabilities based on pair-wise comparison of expectation bounds, in *C. Borgelt et al., Eds., Combining Soft Computing and Statistical Methods in Data Analysis*, Springer, 2010.

[10] J. Hadar, W. Russell, Rules for Ordering Uncertain Prospects, American Economic Review 59 (1969) 25-34.

[11] I. Montes, E. Miranda, S. Montes, Stochastic dominance with imprecise information. Computational Statistics and Data Analysis, doi.org/10.1016/j.csda.2012.07.030 (In press).

[12] L. Sánchez, I. Couso, J. Casillas, Genetic Learning of Fuzzy Rules based on Low Quality Data, Fuzzy Sets and Systems, 160 (2009) 2524-2552.

[13] L.J. Savage. The Foundations of Statistics. Wiley (1954); 2nd edition, Dover Publications Inc., New York, 1972.

[14] B. De Schuymer, H. De Meyer, B. De Baets and S. Jenei, On the cycle-transitivity of the dice model, Theory and Decision 54 (2003) 261-285.

[15] M.C.M. Troffaes, Decision making under uncertainty using imprecise probabilities. Int. J. Approx. Reasoning 45 (2007) 17-29.

[16] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, London, 1991.

Independence for Sets of Full Conditional Probabilities, Sets of Lexicographic Probabilities, and Sets of Desirable Gambles

Fabio Gagliardi Cozman
Escola Politecnica - Universidade de Sao Paulo
Av. Prof. Mello Moraes 2231 - Cidade Universitaria, Sao Paulo, SP - Brazil

Abstract

In this paper we examine concepts of independence for sets of full conditional probabilities; that is, for sets of set-functions where conditional probability is the primitive concept, and where conditioning can be considered on events of probability zero. We also discuss the related issue of independence for (sets of) lexicographic probabilities and for sets of desirable gambles.

Keywords. Sets of probability measures, full conditional probabilities, lexicographic probability, sets of desirable gambles, independence concepts, graphoids.

1 Introduction

This paper examines concepts of independence for sets of full conditional probabilities and related models. We study the behavior of several concepts of independence in the literature, and propose a number of possible additional concepts. The results should be of interest to anyone concerned with representations of uncertainty that allow indeterminacy and imprecision in probability values, and that allow conditioning on every nonempty event.

The motivation for this paper is the following.

The use of a single standard probability measure fails to encode indeterminacy and imprecision about probability values. Belief functions, interval-valued probability, and sets of probability measures have been proposed to handle such indeterminacy and imprecision. It is not obvious how to generalize the concept of stochastic independence when one deals with sets of probability measures; accordingly, there have been many proposed concepts of independence in the literature.

Another problem with standard probability measures is that they do not handle conditioning on events of probability zero; that is, if $P(B) = 0$, then $P(A|B)$ does not exist, regardless of the event A. Indeed, standard conditional probability is merely a derived, incompletely specified concept, while one might argue that conditional probability should be the primitive object of interest. Full conditional probabilities offer an account of conditional probability as primitive objects that can be specified even if conditioning events have probability zero. As standard stochastic independence is quite weak when applied to full conditional probabilities, there have been several proposals for concepts of independence that are appropriate for a single full conditional probability.

However, there is still much to be understood about concepts of independence for *sets* of full conditional probabilities. This paper tries to partially fill this gap, by examining a number of concepts of independence and deriving their graphoid properties (these properties are often taken as abstract properties that any "sensible" concept of independence should satisfy). We also discuss concepts of independence for (sets of) lexicographic probabilities and sets of desirable gambles, as they share several features with full conditional probabilities.

Section 2 describes existing and novel concepts of independence for credal sets and full conditional probabilities. It does not seem that a similar analysis can be found in the literature. Section 3 examines a number of new concepts of independence for sets of full conditional probabilities. Section 4 then examines concepts of independence that resort to lexicographic probabilities and to sets of desirable gambles.

2 Concepts of independence

We assume throughout that the possibility space Ω is finite, so there are no issues of measurability. Throughout the paper we use W, X, Y and Z to denote random variables. Then w denotes a possible value of W, x denotes a possible value of X, y denotes a possible value of Y, z denotes a possible value of Z. And $\{x\}$ denotes the event $\{\omega \in \Omega : X(\omega) = x\}$; likewise for $\{w\}$, $\{y\}$ and $\{z\}$. The letters A and C will always denote nonempty events in the algebra generated by X. Likewise, the letters B and D will always denote nonempty events in the algebra generated by Y. The letter f will always denote a function of X, and the letter g will always denote a function of Y.

The intersection of events G and H is written either as GH or as G, H. When the event $\{x\}$ appears in an intersection, we remove braces whenever possible; for instance, xG denotes the event $\{x\} \cap G$. Sometimes we add braces to enhance clarity; for instance, we may write $\{y, z\}$ instead of simply y, z.

Finally, when w, x, y, z appear in expressions, they are universally quantified unless explicitly noted. Likewise, when functions f and g appear in expressions, they are universally quantified unless explicitly noted.

Conditional stochastic independence of random variables X and Y given random variable Z obtains when $P(x, y|z) = P(x|z)\, P(y|z)$ whenever $P(z) > 0$.

Throughout, if Z is any constant function, we remove the expression "given Z" and in that case we have "unconditional" independence of X and Y (for any concept of independence of interest). Often we just write "independence" to mean both conditional and unconditional independence.

Concepts of independence can be evaluated by their graphoid properties [14, 34]. For any three-place relation $(\cdot \perp\!\!\!\perp \cdot \,|\, \cdot)$, we are interested in the following properties, all of them satisfied by stochastic independence:

Symmetry: $(X \perp\!\!\!\perp Y \,|\, Z) \Rightarrow (Y \perp\!\!\!\perp X \,|\, Z)$

Redundancy: $(X \perp\!\!\!\perp Y \,|\, X)$

Decomposition: $(X \perp\!\!\!\perp (W, Y) \,|\, Z) \Rightarrow (X \perp\!\!\!\perp Y \,|\, Z)$

Weak union: $(X \perp\!\!\!\perp (W, Y) \,|\, Z) \Rightarrow (X \perp\!\!\!\perp Y \,|\, (W, Z))$

Contraction:
$(X \perp\!\!\!\perp Y \,|\, Z) \wedge (X \perp\!\!\!\perp W \,|\, (Y, Z)) \Rightarrow (X \perp\!\!\!\perp (W, Y) \,|\, Z)$.

2.1 Independence for sets of standard probability measures

A set of standard (Kolmogorovian-style) probability measures, not assumed to be closed and convex, is referred to as a *credal set*. Denote by $K(X)$ the set of probability distributions for variable X. Given a function $f(X)$, its lower and upper expectations are, respectively $\underline{E}[f(X)] = \inf_{P \in K} E_P[f(X)]$ and $\overline{E}[f(X)] = \sup_{P \in K} E_P[f(X)]$, where $E_P[f(X)]$ is the expectation of $f(X)$ with respect to P. Similarly, given an event A, its lower and upper probabilities are, respectively $\underline{P}(A) = \inf_{P \in K} P(A)$ and $\overline{P}(A) = \sup_{P \in K} P(A)$.

Given a credal set $K(X)$, we define the conditional credal set

$$K(X|A) = \{P(\cdot|A) : P \in K(X)\} \quad \text{if } \underline{P}(A) > 0;$$

otherwise, $K(X|A)$ is left undefined [21]. Another option is to define a conditional credal set that focuses on those probability measures that assign positive probability to A:

$$K^{>}(X|A) = \{P(\cdot|A) : P \in K(X) \text{ and } P(A) > 0\} \quad \text{if } \overline{P}(A) > 0; \tag{1}$$

otherwise $K^{>}(X|A)$ is left undefined [44, 45]. Obviously, if $\underline{P}(A) > 0$, then $K(X|A) = K^{>}(X|A)$. The set $K^{>}(X|A)$ is convex when $K(X)$ is convex, but it may be open even when $K(X)$ is closed. We define $\underline{E}^{>}[f(X)|A] = \inf_{P(\cdot|A) \in K^{>}(X|A)} E_P[f(X)|A]$ and $\overline{E}^{>}[f(X)|A] = \sup_{P(\cdot|A) \in K^{>}(X|A)} E_P[f(X)|A]$.

For a moment, assume that all lower probabilities are positive.

Following Levi [29], say that Y is *confirmationally irrelevant* to X given Z when

$$K(X|y, z) = K(X|z)\,. \tag{2}$$

Walley has proposed a similar concept [41, 42]: Y is *epistemically irrelevant* to X given Z when

$$\underline{E}[f(X)|y, z] = \underline{E}[f(X)|z] \tag{3}$$

(recall our conventions: by implicit quantification, this equality is required for all f, for all y, z).

Both confirmational and epistemic irrelevance fail Symmetry. Walley's clever solution, borrowed from the work of Keynes, was to "symmetrize" irrelevance to obtain *epistemic independence*: X and Y are epistemically independent given Z when X is epistemically irrelevant to Y given Z and Y is epistemically irrelevant to X given Z [42]. Take *confirmational independence* to be a likewise symmetrized version of confirmational irrelevance.

If all credal sets are closed and convex, then confirmational and epistemic independence are equivalent. Now even if all lower probabilities are positive and all credal sets are closed and convex, epistemic independence (and confirmational independence) fails Contraction [7]. And if credal sets are not required to be convex, then confirmational independence fails Decomposition, Weak Union and Contraction even when all lower probabilities are positive [9].

Matters become more complicated if lower probabilites are allowed to be zero. Suppose first that Y is taken to be confirmationally irrelevant to X if

$$K(X|y,z) = K(X|z) \text{ whenever } \underline{P}(y,z) > 0.$$

We are surely flirting with disaster here, because it is not difficult to have a variable Z such that every value of Z has zero lower probability, and yet $K(Z)$ is not a vacuous credal set (that is, it does not contain every possible distribution for Z). Now given such a variable Z, every two other variables are confirmationally independent! This is not reasonable.

The other path to handle events of zero lower probability within confirmational independence is to say that Y is confirmationally irrelevant to X given Z when

$$K^{>}(X|y,z) = K(X|z) \text{ whenever } \overline{P}(y,z) > 0. \quad (4)$$

The symmetrized concept of independence fails Decomposition, Weak Union and Contraction (as noted before, these properties fail even when all lower probabilities are positive [9]).

Another possibility is to define epistemic irrelevance of Y to X given Z by requiring:

$$\underline{E}^{>}[f(X)|y,z] = \underline{E}[f(X)|z] \text{ whenever } \overline{P}(y,z) > 0. \quad (5)$$

The resulting symmetrized concept of independence fails Contraction (as noted before, this property fails even when all lower probabilities are positive [7]). It is an open question whether Decomposition and Weak Union hold when Expression (5) is used to define independence; Decomposition and Weak Union hold for epistemic independence when all lower probabilities are positive [12].

Note: Expressions (4) and (5) impose different constraints, as $K^{>}(X|A)$ may be open even when $K(X)$ is closed.

Yet another path has been followed by de Campos and Moral [15]: they say Y is *type-5* irrelevant to X if

$$K^{>}(X|B) = K(X) \text{ whenever } \overline{P}(B) > 0$$

(recall: B is an event in the algebra generated by Y). Accordingly, say that Y is *type-5 irrelevant* to X given Z if

$$K^{>}(X|B,z) = K(X|z) \text{ whenever } \overline{P}(B,z) > 0.$$

Now we might also modify epistemic irrelevance, and say that Y is *type-5 epistemically irrelevant* to X given Z if

$$\underline{E}^{>}[f(X)|B,z] = \underline{E}[f(X)|z] \text{ whenever } \overline{P}(B,z) > 0.$$

And we can symmetrize type-5 irrelevance and type-5 epistemic irrelevance to obtain corresponding concepts of independence. Now, Contraction fails for type-5 independence and for type-5 epistemic independence (Contraction fails already when all lower probabilities are positive [7]). It is an open question whether Weak Union holds for these concepts of independence. As for Decomposition:

Proposition 1 *Both type-5 independence and type-5 epistemic independence satisfy Decomposition.*

Proof. Assume X and (W,Y) are type-5 independent given Z. Then $K(Y|A,z) = K(Y|z)$ by marginalization, and $K(X|B,z) = K(X|z)$ because any B belongs to the algebra generated by (W,Y). Likewise, assume type-5 epistemic independence holds for X and (W,Y). Then $\underline{E}[g(Y)|A,z] = \underline{E}[g(Y)|z]$ because any function of Y is a function of (W,Y), and $\underline{E}[f(X)|B,z] = \underline{E}[f(X)|z]$. □

Type-5 irrelevance may seem very attractive at first, but the following example, due to de Campos and Moral [15], displays rather weird behavior when lower probabilities are zero. Take binary variables X and Y, and $K(X,Y)$ with two distributions, one that assigns probability one to (x_0,y_0) and another that assigns probability one to (x_1,y_1) (if $K(X,Y)$ must be convex, take the convex hull of these two distributions). Both distributions satisfy stochastic independence, but X and Y fail to be type-5 independent! In general, type-5 independence may fail even when all elements of the credal set $K(X,Y)$ factorize.

This discussion suggests that concepts of independence for credal sets must handle conditioning carefully. We now describe a few concepts of independence that require no discussion about conditioning.

Strong independence was also proposed by Levi [29], initially with the name *strong confirmational irrelevance*: X and Y are strongly independent when $K(X,Y)$ is the convex hull of a set of probability measures that satisfy stochastic independence. Strong independence is an attempt to stay close to stochastic independence while assuming convexity (given that imposing stochastic independence over a set of probability measures may generate a nonconvex set of measures). Strong independence can be derived from assumptions of infinite exchangeability [9] or finite exchangeability together with epistemic independence [16]. Note that strong independence, and slight variants of it, have received several names in the literature, such as *type-1 product*, *type-2 product*, *type-2 independence*, *independence in the selection*, *repetition independence* [9].

Complete independence abandons convexity and imposes stochastic independence directly: X and Y are completely independent when every joint distribution in $K(X, Y)$ satisfies stochastic independence [9]. Complete independence satisfies all graphoid properties previously mentioned.

The last notable concept of independence we mention for credal sets is due to Kuznetsov [28]: X and Y are Kuznetsov-independent if

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)] \boxtimes \mathbb{E}[g(Y)]$$

for all functions $f(X)$ and $g(Y)$, where $\mathbb{E}[\cdot]$ denotes the interval from lower to upper expectations, and $\boxtimes$ denotes interval multiplication. Kuznetsov-independence satisfies Symmetry, Redundancy and Decomposition; it fails Contraction even when all probabilities are positive [8], and it is an open question whether it satisfies Weak Union or not.

2.2 Independence for full conditional probabilities

A full conditional probability [20] $P : \mathcal{B} \times (\mathcal{B} \backslash \emptyset) \to \Re$, where $\mathcal{B}$ is a Boolean algebra, is a two-place set-function such that for every event $H \neq \emptyset$:
(1) $P(H|H) = 1$;
(2) $P(G|H) \geq 0$ for all G;
(3) $P(G_1 \cup G_2|H) = P(G_1|H) + P(G_2|H)$
whenever $G_1 \cap G_2 = \emptyset$;
(4) $P(G_1, G_2|H) = P(G_1|G_2, H) \times P(G_2|H)$
whenever $G_2H \neq \emptyset$.

This fourth axiom is often stated as $P(G_1|H) = P(G_1|G_2)\, P(G_2|H)$ when $G_1 \subseteq G_2 \subseteq H$ and $G_2 \neq \emptyset$ [13, Section 2].

Define the "unconditional" probability $P(G)$ of an event G to be $P(G|\Omega)$. That is, whenever the conditioning event H is equal to Ω, we suppress it and write the "unconditional" probability $P(G)$.

There are other names for full conditional probabilities in the literature, such as *conditional probabilities* [27] and *complete conditional probability systems* [33]. We simplify to *full probability* whenever possible. Full probabilities have found applications in several fields, notably economy, philosophy, and statistics [5, 19, 26, 30, 32, 35, 38].

We can partition Ω into events $L_0, \ldots, L_K$ as follows. First, take L_0 to be the set of elements of Ω that have positive unconditional probability. Then take L_1 to be the set of elements of Ω that have positive probability conditional on $\Omega \backslash L_0$. And then take L_i, for $i \in \{2, \ldots, K\}$, to be the set of elements of Ω that have positive probability conditional on $\Omega \backslash \cup_{j=0}^{i-1} L_j$. The events L_i are called the *layers* of the full probability. Note that some authors use a different terminology, using instead the sequence $\cup_{j=i}^{K} L_j$ rather than L_i [5, 27].

	y_0	y_1
x_0	$\lfloor 1 \rfloor_0$	$\lfloor 1-\alpha \rfloor_1$
x_1	$\lfloor \alpha \rfloor_1$	$\lfloor 1 \rfloor_2$

	y_0	y_1
x_0	$\lfloor 1 \rfloor_0$	$\lfloor 1 \rfloor_i$
x_1	$\lfloor 1 \rfloor_j$	$\lfloor 1 \rfloor_3$

Table 1: Joint full distributions for binary variables X and Y. The right table stands for two full distributions: one for $i = 1, j = 2$; another for $i = 2, j = 1$.

Any full probability can be represented by a sequence of probability measures $P_0, \ldots, P_K$, where P_i is positive over L_i. This useful result that has been derived by several authors [3, 5, 23, 27].

For nonempty G, denote by L_G the first layer such that $P(G|L_G) > 0$, and refer to it as the *layer of* G. We then have $P(G|H) = P(G|H \cap L_H)$ [2, Lemma 2.1a].

We often write $\lfloor \alpha \rfloor_i$ to denote a probability value α that belongs to the ith layer L_i. Table 1 shows three full distributions using this compact notation.

Given a full probability and a nonempty event H, the two-place function $P(\cdot|\cdot \cap H)$ is also a full probability from which a partition of H consisting of layers $L_{0|H}, L_{1|H}, \ldots, L_{K|H}$ can be built. Given an event G such that $G \cap H \neq \emptyset$, denote by $L_{G|H}$ the first layer of $P(\cdot|\cdot \cap H)$ such that $P\big(G|L_{G|H}\big) > 0$.

For a nonempty event G, the index i of the first layer L_i of the full probability P such that $P(G|L_i) > 0$ is the *layer number* of G. Layer numbers have been studied by Coletti and Scozzafava [5], who refer to them as *zero-layers*. The layer number of G is denoted by $\circ(G)$. Inspired by Coletti and Scozzafava [5], we define the layer number of G given nonempty H as $\circ(G|H) = \circ(G \cap H) - \circ(H)$, and we adopt $\circ(\emptyset) = \infty$.

Now consider concepts of independence for full probabilities.

Stochastic independence satisfies all graphoid properties we have mentioned previously, when applied to full probabilities. Unfortunately, it may happen that X and Y are stochastically independent and yet $P(A|B) \neq P(A)$ when $P(B) = 0$. Table 2 shows an extreme example. To avoid this embarrassment, more stringent notions of independence have been proposed for full probabilities [3, 5, 23, 39].

Say that Y is *epistemically irrelevant* to X given Z if $P(A|y, z) = P(A|z)$ whenever $\{y, z\} \neq \emptyset$, and then say that X and Y are *epistemically independent* given Z if X is epistemically irrelevant to Y given Z and vice-versa. Epistemic independence satisfies Sym-

	y_0	y_1
x_0	$\lfloor 1 \rfloor_0$	$\lfloor 1 \rfloor_3$
x_1	$\lfloor 1 \rfloor_1$	$\lfloor 1 \rfloor_2$

Table 2: Joint full distributions for stochastically independent binary variables, where $P(x_0) = 1 \neq 0 = P(x_0|y_1)$.

	$w_0 y_0$	$w_1 y_0$	$w_0 y_1$	$w_1 y_1$
x_0	$\lfloor \alpha \rfloor_0$	$\lfloor \beta \rfloor_2$	$\lfloor 1-\alpha \rfloor_0$	$\lfloor 1-\beta \rfloor_2$
x_1	$\lfloor \alpha \rfloor_1$	$\lfloor \gamma \rfloor_3$	$\lfloor 1-\alpha \rfloor_1$	$\lfloor 1-\gamma \rfloor_3$

Table 3: Full distribution for W, X, Y, with distinct $\alpha \in (0,1)$, $\beta \in (0,1)$, $\gamma \in (0,1)$.

metry, Redundancy, Decomposition and Contraction, but it fails Weak Union [11, Proposition 4.2]. The full distribution in Table 3 displays failure of Weak Union for epistemic independence.

As proposed by Hammond [23], say that Y is *h-irrelevant* to X given Z when

$$P(A|B,C,z) = P(A|C,z) \text{ whenever } \{B,C,z\} \neq \emptyset,$$

and say that X and Y are *h-independent* given Z when X is h-irrelevant to Y given Z and vice-versa (recall our conventions: this equality must hold for every A and C in the algebra generated by X, and for every B in the algebra generated by Y).

If X and Y are h-independent given Z, then

$$\begin{aligned} P(A,B|C,D,z) &= P(A|C,z)\,P(B|D,z) \\ &\quad \text{whenever } \{C,D,z\} \neq \emptyset. \end{aligned}$$

H-independence satisfies Symmetry, Redundancy, Decomposition and Weak Union, but it fails Contraction [11, Theorem 5.4]. The full distribution in Table 3 displays failure of Contraction for h-independence.

Coletti and Scozzafava [5] have proposed conditions on zero-layers to characterize independence. Say that event H is *cs-irrelevant* to event G, where $H \neq \emptyset \neq H^c$, if $P(G|H) = P(G|H^c)$, $\circ(G|H) = \circ(G|H^c)$, and $\circ(G^c|H) = \circ(G^c|H^c)$. To understand the motivation for these conditions on layer numbers, suppose that GH, GH^c, G^cH are nonempty, but $G^cH^c = \emptyset$. Hence observation of H^c does provide information about G. However, the indicator functions of G and H can be epistemically/h-independent! Coletti and Scozzafava eliminate such difficulties using their conditions on layer numbers; other authors, such as Hammond [23] and Battigalli [2], explicitly require the possibility space to be the product of the possibility spaces for each of the variables.

Vantaggi [39, 40] has extended Coletti and Scozzafava conditions to independence of variables. Say that Y is *cs-irrelevant* to X given Z when event $\{y\}$ is cs-irrelevant to event $\{x\}$ given event $\{z\}$, whenever $\{y,z\} \neq \emptyset \neq \{\{y\}^c, z\}$ [39, Definition 7.3]. Call the symmetrized concept *cs-independence* of X and Y given Z. Besides Symmetry, cs-independence satisfies Redundancy, Decomposition and Contraction, and it fails Weak Union [39, Section 9].

The conditions on layer numbers imposed by cs-independence can be written as [11, Corollary 4.11]:

$$\circ(x,y|z) = \circ(x|z) + \circ(y|z) \qquad \text{for } \{z\} \neq \emptyset. \qquad (6)$$

Condition (6) can be used to generate additional concepts of independence. For instance, say that Y is *fully irrelevant* to X given Z if Y is h-irrelevant to X given Z and if they satisfy Condition (6); say that X and Y are *fully independent* given Z if they are h-independent given Z and satisfy Condition (6) [11].

Full independence satisfies Symmetry, Redundancy, Decomposition and Weak Union, but it fails Contraction [11, Theorem 5.7]. Table 3 displays failure of Contraction for full independence.

A different concept of independence has been proposed by Kohlberg and Reny [26], essentially as follows. Say that X and Y are kr-independent given Z when both:

- if $\{x,z\} \neq \emptyset$ and $\{y,z\} \neq \emptyset$, then $\{x,y,z\} \neq \emptyset$;
- if, whenever conditioning events are nonempty,
$$\frac{P\big(x,y|L_{x,y|z} \cup L_{x',y'|z}\big)}{P\big(x',y'|L_{x,y|z} \cup L_{x',y'|z}\big)} = \lim_{n\to\infty} \frac{P_n(x|z)P_n(y|z)}{P_n(x'|z)P_n(y'|z)}$$
for some sequence of product probability measures $P_n(\cdot|z)$.

Relatively little is known about kr-independence; we only note that it satisfies Symmetry, Redundancy, Decomposition and Weak Union, and it fails Contraction as can be seen in Table 3 [10, Theorem 1].

We now introduce a new concept of independence for full probabilities where we require factorization across layers of the full probability [10]. Consider:

Definition 1 *X and Y are* layer independent *given Z if, for each layer L_i of the underlying full probability P, and each z such that $\{L_i, z\} \neq \emptyset$, we have both*

$$P(x,y|L_i,z) = P(x|L_i,z)\,P(y|L_i,z)\,,$$

$$\circ(x,y|z) = \circ(x|z) + \circ(y|z)\,.$$

This concept of independence satisfies Symmetry, Redundancy, Decomposition, Weak Union and Contraction; in fact, this seems to be the only known concept of independence for full probabilities that satisfies all these five properties.

We conclude this section by commenting on an aspect of full probabilities that has not received the deserved attention so far; namely, failure of uniqueness (some comments about it appear in the work of Battigalli [1] and Kohlberg and Reny [26]). The issue is this. Suppose one is given marginal probabilities $P(x_0) = P(y_0) = 1$ for binary variables X and Y. Now *every* full distribution in Table 1 (for *every* $\alpha \in (0, 1)$) satisfies these marginal assessments and epistemic/h-/cs-/full/kr-independence; moreover, the two full distributions encoded by the right table satisfy layer independence. In general, one cannot uniquely determine a single full probability by specifying marginal assessments and judgments of independence. Once assessments are to be combined with existing concepts of independence, one must be prepared to consider a set of joint full probabilities that satisfies all constraints.

3 Full credal sets and independence

We now focus on sets of full probabilities, and investigate the graphoid properties of several concepts of independence. We refer to such sets as *full credal sets*; we do not assume the sets to be convex and closed.

As already noted, a concept of independence that relies on product factorizations is too weak in the context of full probabilities. Indeed we have that Kuznetsov, strong, complete and type-5 independence declare X and Y independent for the full credal set containing only the full distribution in Table 2.

Complete independence can be adapted to full credal sets as follows. Define *elementwise epistemic/h-/cs-/full/kr-/layer independence* of X and Y given Z to hold when every element of the full credal set $K(X, Y|z)$ satisfies respectively epistemic/h-/cs-/full/kr-/layer independence whenever $\{z\} \neq \emptyset$. We note that Coletti and Scozzafava's concept of independence for lower probabilities [4, Definition 6], extended to variables by Vantaggi [40, Definition 7], is quite similar to elementwise cs-independence.

Given the results mentioned in the previous section:

Proposition 2 *Elementwise epistemic/cs-independence satisfy Symmetry, Redundancy, Decomposition and Contraction (and fail Weak Union). Elementwise h-/full/kr-independence satisfy Symmetry, Redundancy, Decomposition and Weak Union (and fail Contraction). Elementwise layer independence satisfies Symmetry, Redundancy, Decomposition, Weak Union and Contraction.*

A challenge that merits future work is to justify these concepts of independence from behavioral or decision-theoretic arguments. Even though complete independence has an intuitive justification using choice functions [9, 37], the interaction between choice functions and full probabilities is yet to be explored.

Consider now confirmational and epistemic independence as defined in Section 2.1, but applied to full credal sets. The resulting concepts were originally proposed by Levi [29] and by Walley [42] within theories that adopt full probabilities.

Confirmational independence fails Decomposition, Weak Union and Contraction when applied to general full credal sets (even when all lower probabilities are positive [9]).

Epistemic independence fails Decomposition and Weak Union when applied to full credal sets [12], as can be seen in Example 1, and fails Contraction even when all lower probabilities are positive [7].

Example 1 Consider a full credal set with the two distributions depicted in Table 4, where $\alpha \in (0, 1/2)$. We have $P(w_0) \in [\alpha, 1-\alpha]$ and $P(w_0|x, y) \in [\alpha, 1-\alpha]$ for all possible x, y: (X, Y) is epistemically irrelevant to W. The reader can verify that both distributions yield identical values of $P(x, y|w)$ and $P(x, y)$ such that $P(x, y|w) = P(x, y)$, for all possible (x, y, z). Hence W is epistemically irrelevant to (X, Y). Thus we have epistemic independence of W and (X, Y). However, $P(w_0|x_1) = 1/2$; consequently, X is not epistemically irrelevant to W (Decomposition fails), and Y is not epistemically irrelevant to W given X (Weak Union fails). □

So, at least from the point of view of graphoid properties, both confirmational and epistemic independence fare rather poorly.

Note that the motivation behind confirmational/epistemic irrelevance of Y to X is that observation of Y does not change beliefs about X. However, for a full probability the beliefs about X are encoded not just by expectations $E[f(X)]$ but rather by conditional expectations $E[f(X)|A]$ for events A in the algebra generated by X. This is indeed the rationale behind h-independence; for this reason, the combination of h-independence and full credal sets seems very attractive.

Consider then adapting h-independence to full credal sets as follows:

P_1	w_0y_0	w_0y_1	w_1y_0	w_1y_1
x_0	$\lfloor \frac{\alpha}{2} \rfloor_0$	$\lfloor \frac{\alpha}{2} \rfloor_0$	$\lfloor \frac{1-\alpha}{2} \rfloor_0$	$\lfloor \frac{1-\alpha}{2} \rfloor_0$
x_1	$\lfloor \frac{\alpha}{2} \rfloor_1$	$\lfloor \frac{1-\alpha}{2} \rfloor_1$	$\lfloor \frac{1-\alpha}{2} \rfloor_1$	$\lfloor \frac{\alpha}{2} \rfloor_1$

P_2	w_0y_0	w_0y_1	w_1y_0	w_1y_1
x_0	$\lfloor \frac{1-\alpha}{2} \rfloor_0$	$\lfloor \frac{1-\alpha}{2} \rfloor_0$	$\lfloor \frac{\alpha}{2} \rfloor_0$	$\lfloor \frac{\alpha}{2} \rfloor_0$
x_1	$\lfloor \frac{1-\alpha}{2} \rfloor_1$	$\lfloor \frac{\alpha}{2} \rfloor_1$	$\lfloor \frac{\alpha}{2} \rfloor_1$	$\lfloor \frac{1-\alpha}{2} \rfloor_1$

Table 4: Extreme points of the full credal set in Example 1.

	y_0	y_1
x_0	$\lfloor \alpha \rfloor_0$	$\lfloor 1-\alpha \rfloor_0$
x_1	$\lfloor \alpha \rfloor_1$	$\lfloor 1-\alpha \rfloor_1$

Table 5: Marginal probabilities from Table 3.

	y_0	y_1
x_0	$\lfloor \alpha \rfloor_0$, $\lfloor \beta \rfloor_2$	$\lfloor 1-\alpha \rfloor_0$, $\lfloor 1-\beta \rfloor_2$
x_1	$\lfloor \alpha \rfloor_1$, $\lfloor \gamma \rfloor_3$	$\lfloor 1-\alpha \rfloor_1$, $\lfloor 1-\gamma \rfloor_3$

Table 6: Lexicographic marginal probabilities from Table 3.

Definition 2 *Y is h-irrelevant to X given Z if*

$$\begin{aligned}\underline{E}[f(X)|A,B,z] &= \underline{E}[f(X)|A,z] \\ &\quad \textit{whenever } \{A,B,z\} \neq \emptyset.\end{aligned}$$

X and Y are h-independent given Z when X is h-irrelevant to Y given Z and vice-versa.

We have:

Theorem 1 *H-independence satisfies Symmetry, Redundancy, Decomposition, and Weak Union.*

Proof. Symmetry holds by definition; Redundancy is trivial. From the assumed h-independence of X and (W,Y), we have: $\underline{E}[f(X)|A,B,z] = \underline{E}[f(X)|A,z]$, and $\underline{E}[g(Y)|A,B,z] = \underline{E}[g(Y)|B,z]$ (Decomposition). Weak Union follows from $\underline{E}[g(Y)|A,B,w,z] = \underline{E}[g(Y)|B,w,z]$, and then, using Decomposition, $\underline{E}[f(X)|A,w,z] = \underline{E}[f(X)|A,z] = \underline{E}[f(X)|A,B,w,z]$. □

Note that h-independence fails Contraction (Table 3).

In the next section we examine two other representations that are closely related to full conditional measures and full credal sets.

4 Lexicographic probabilities and sets of desirable gambles

Consider again Table 3. For this full distribution we have X and Y epistemic/h-/cs-/full/kr-/layer independent. One might argue that there is something strange about this "independence". For take a function $g(Y)$ such that $g(y_0) = -(1-\alpha)$ and $g(y_1) = \alpha$. This function has expected utility zero. But if $\beta < \alpha$ one might argue that g is better than the zero function; after all, if $\{w_1\}$ happens to be observed, then the expected value of g given $\{w_1\}$ is $\alpha - \beta$, and g should then be considered better than the zero function. And if $\gamma > \alpha$, then conditional on $\{w_1, x_1\}$ the zero function should be considered better than g. Hence conditioning on $\{x_1\}$ seems to change opinions about a function of Y.

One way to understand this example is to look at the marginal full probability for (X,Y), shown in Table 5. Note that when the full probability in Table 3 is marginalized over W, the content of layers L_2 and L_3 disappear: in Table 5 one sees neither β nor γ. Preferences about g that might depend on deeper layers can only be exposed by observing W. In a sense, the direct marginalization of Table 3 loses important information about the joint full probability. It would make more sense to say that the marginal probabilities obtained from Table 3 should be given by the overlapping layers in Table 6, so as to conclude that X and Y are *not* independent.

We are then moving into *lexicographic probabilities* that assign probability measures to various layers with possibly overlapping support. Due to the lack of space, we omit detailed background on lexicographic probabilities, and refer the reader to the work of Blume et al. [3] for all necessary definitions. We assume their axiomatization of the non-Archimedean preference relation $\succeq$, and use the fact that this preference relation can be represented by a sequence of probability measures over Ω; each one of these measures is a "layer" of the lexicographic probability. [3, Corollary 3.1]. Two functions $f_1(X)$ and $f_2(X)$ are compared with respect to a lexicographic rule in the sense that $f_1 \succeq f_2$ if and only if

$$\left[\sum_x f_1(x)P_i(x)\right]_{i=0}^{K} \geq_{\mathrm{L}} \left[\sum_x f_2(x)P_i(x)\right]_{i=0}^{K},$$

(for $a, b \in \Re^K$, $a \geq_{\mathrm{L}} b$ iff whenever $b_j > a_j$, there exists a $k < j$ such that $a_k > b_k$). These probabilities are unique only up to linear transformations, so there is some intrinsic non-uniqueness associated with lexicographic probabilities:

Example 2 Suppose that a binary variable Y is associated with two layers such that $P_0(y_0) = 1 - P_0(y_1) = \alpha$ and $P_1(y_0) = 1 - P_1(y_1) = \beta$. For fixed α, every $\beta \in [0, \alpha)$ yields identical preferences; likewise, every $\beta \in (\alpha, 1]$ yields identical preferences. So the specific value of β cannot be fixed by resorting to lexicographic preferences. □

Conditional lexicographic probabilities given nonempty event H are obtained by conditioning every layer of the lexicographic probability on H, after discarding those layers that do not intersect H. These conditional probabilities encode the preferences $f_1(X)I_H \succeq f_2(X)I_H$ [3, Theorem 4.3], denoted by $[f_1(X) \succeq f_2(X)|H]$.

The close proximity between full probabilities and lexicographic probabilities is apparent. A full probability can be represented by a lexicographic probability with disjoint layers [22, 23]. And for any lexicographic probability, the function $P(A|B) = P_i(A|B)$, where P_i the the first measure such that $P_i(B) > 0$, is a full probability. However, as indicated by the discussion of marginalization concerning Tables 3, 5 and 6, full probabilities and lexicographic probabilities do not behave identically.

Now consider defining a concept of independence for lexicographic probabilities. We might try to define a "product" for lexicographic probabilities. Here difficulties abound due to non-uniqueness. First, probabilities in various layers can be modified so as to break factorization. Additionally, probability values are not tied to specific layer numbers. For instance, if we have a lexicographic probability with three overlapping layers, each with probability measures p_0, p_1 and p_2, we can generate an *equivalent* representation with four layers p_0, p_0, p_1 and p_2. Therefore a condition such as layer factorization seems rather fragile as we cannot control layer numbers just by looking at marginal lexicographic probabilities.

Indeed the difficulties with product lexicographic probabilities have already been discussed by several authors [3, 23, 24]. Solutions based on factorization of nonstandard measures have been advanced by these authors; the interpretation and the manipulation of such concepts do not seem easy, and we leave that to future work.

Hence we are led, in our study of lexicographic probabilities, to concepts of independence that rely on conditioning. Blume et al. [3] say that X and Y are independent when we have both

$$[f_1(X) \succeq f_2(X)|y_1] \Leftrightarrow [f_1(X) \succeq f_2(X)|y_2],$$

$$[g_1(Y) \succeq g_2(Y)|x_1] \Leftrightarrow [g_1(Y) \succeq g_2(Y)|x_2]$$

whenever conditioning events are nonempty. Say that X and Y are independent given Z when the expressions above are satisfied conditional on any $\{z\}$ such that conditioning events are nonempty.

	w_0y_0	w_1y_0	w_0y_1	w_1y_1
x_0	$\lfloor\alpha\rfloor_0$	$\lfloor\beta\rfloor_2$	$\lfloor 1-\alpha\rfloor_0$	$\lfloor 1-\beta\rfloor_2$
x_1	$\lfloor\alpha\rfloor_1$	$\lfloor\beta\rfloor_3$, $\lfloor\gamma\rfloor_4$	$\lfloor 1-\alpha\rfloor_1$	$\lfloor 1-\beta\rfloor_3$, $\lfloor 1-\gamma\rfloor_4$

Table 7: Lexicographic distribution for W, X, Y, with distinct $\alpha \in (0, 1)$, $\beta \in (0, 1)$, $\gamma \in (0, 1)$.

	y_0	y_1
x_0	$\lfloor\alpha\rfloor_0$, $\lfloor\beta\rfloor_2$	$\lfloor 1-\alpha\rfloor_0$, $\lfloor 1-\beta\rfloor_2$
x_1	$\lfloor\alpha\rfloor_1$, $\lfloor\beta\rfloor_3$	$\lfloor 1-\alpha\rfloor_1$, $\lfloor 1-\beta\rfloor_3$

$P(W, X|Y = y)$

	w_0	w_1
x_0	$\lfloor 1\rfloor_0$	$\lfloor 1\rfloor_2$
x_1	$\lfloor 1\rfloor_1$	$\lfloor 1\rfloor_3$

Table 8: Marginal (left) and conditional (right) lexicographic probabilities from Table 7.

Even though Table 3 no longer fails Contraction if we use this concept of independence (because X and Y are no longer independent), consider Table 7. The distributions for (X, Y), for (X, W) given $\{y_0\}$, and for (X, W) given $\{y_1\}$ are shown in Table 8. Here X and Y are independent and X and W are independent given Y; yet X and (W, Y) are not independent. Contraction fails. The fourth layer "vanishes" when one marginalizes out W as preferences are decided already at the third layer. To understand this, consider Example 2: once α and β are fixed, every preference about Y is fixed, and there is no need to examine further layers.

Now suppose we have a *set* of lexicographic probabilities, where preference is given by unanimity amongst lexicographic comparisons [36]. Example 1 shows that Decomposition and Weak Union can fail for Blume et al.'s concept of independence (just consider each full probability a lexicographic probability, and take their convex hull if a convex set is desired).

We suggest that a more promising concept of independence for (sets of) lexicographic probabilities is obtained by symmetrizing the following concept: Y is irrelevant to X given Z when

$$[f_1(X) \succeq f_2(X)|A, B, z] \Leftrightarrow [f_1(X) \succeq f_2(X)|A, z],$$

for all functions, whenever conditioning events are nonempty. And X and Y are independent given Z when Y is irrelevant to X given Z and vice-versa.

This concept of independence satisfies Symmetry, Redundancy, Decomposition and Weak Union; Contraction fails (Table 7). Redundancy obtains because

$$\begin{aligned}[f_1(X) \succeq f_2(X)|A, B, x] \quad &\Leftrightarrow \quad f_1(x) \geq f_2(x) \\ &\Leftrightarrow \quad [f_1(X) \succeq f_2(X)|B, x].\end{aligned}$$

Decomposition holds because any event B belongs to the algebra generated by (W, Y), and any function $g(Y)$ is also a function of (W, Y) (hence independence of X and (W, Y) given Z implies independence of X and Y given Z). Weak Union holds because, assuming X and (W, Y) independent given Z, we have

$$[g_1(Y) \succeq g_2(Y)|A, B, w, z] \Leftrightarrow [g_1(Y) \succeq g_2(Y)|B, w, z],$$

and, using Decomposition,

$$\begin{aligned}[f_1(X) \succeq f_2(X)|A, w, z] \quad &\Leftrightarrow \quad [f_1(X) \succeq f_2(X)|A, z] \\ &\Leftrightarrow \quad [f_1(X) \succeq f_2(X)|A, B, w, z].\end{aligned}$$

Sets of lexicographic probabilities are equivalent, from the point of view of preference representations, to *sets of desirable gambles*, a representation that has received considerable attention [6, 17, 18, 31, 43]. Indeed the derivation of lexicographic representations for sets of desirable gambles appears already in the work of Seidenfeld et al. [36], who show that a partially ordered set of preferences (that encodes a set of desirable gambles) can be represented by a set of complete orderings, each one of which can be represented by a lexicographic probability (either using results by Kee [25] or the more direct results by Blume et al. [3]). In recent work, Couso and Moral [6] have studied the representation of sets of desirable gambles through lexicographic probabilities.

A set of desirable gambles $\mathbb{D}$ is a set of variables not containing the zero function and containing all nonnegative variables that are different from zero, and such that $\lambda X \in \mathbb{D}$ if $X \in \mathbb{D}$ and $\lambda > 0$, and $X + Y \in \mathbb{D}$ if $X, Y \in \mathbb{D}$ [17, Definition 1]. The set of desirable gambles conditional on event A, denoted by $[\mathbb{D}|A]$, contains all desirable gambles X such that $XI_A = X$, where I_A is the indicator function of A [18, Section 3.2]. Following notation by Moral [31], denote by $\mathbb{D}^{\downarrow X}$ the set of desirable gambles that are functions of X (that is, $\mathbb{D}^{\downarrow X}$ is the "marginal" set of gambles with respect to X). A natural concept of independence for sets of desirable gambles is [17, Definition 3]: Y is irrelevant to X given Z if

$$[\mathbb{D}|y, z]^{\downarrow X} = [\mathbb{D}|z]^{\downarrow X} \;\text{ whenever } \{y, z\} \neq \emptyset.$$

And then: X and Y are independent given Z if X is irrelevant to Y given Z and vice-versa. (Note that there are other concepts of independence for sets of desirable gambles in the literature [31].)

Mimicking our proposal for (sets of) lexicographic probabilities, consider the following definition of independence for sets of desirable gambles: Y is irrelevant to X if

$$[\mathbb{D}|A, B, z]^{\downarrow X} = [\mathbb{D}|A, z]^{\downarrow X} \;\text{ whenever } \{A, B, z\} \neq \emptyset.$$

And then define independence of X and Y given Z by symmetrizing this concept of irrelevance.

5 Conclusion

This paper has studied concepts of independence for sets of full probabilities, and for their close relatives, sets of lexicographic probabilities, and sets of desirable gambles. We have tried to offer a commented and organized review of the literature in Section 2. We have then analyzed a large number of concepts of independence in Sections 3 and 4.

At this point the only concept of independence for full credal sets that satisfy Symmetry, Redundancy, Decomposition, Weak Union and Contraction is elementwise layer independence. The concepts of confirmational and epistemic independence seem particularly weak when applied to full credal sets. The concept of h-independence fares considerably better but still fails Contraction. The extent to which one can adopt concepts that fail various graphoid properties is yet to be fully analyzed.

Concerning lexicographic probabilities: they do add flexibility, but they introduce significant complexity in dealing with non-uniqueness and marginalization. Sets of desirable gambles also require some care in dealing with marginalization. The new concepts of independence suggested here for sets of lexicographic probabilities and sets of desirable gambles should be helpful in future work.

Acknowledgements

Thanks to reviewers, who suggested many improvements to the original submission.

The author is partially supported by CNPq.

References

[1] Pierpaolo Battigalli. Strategic independence and perfect Bayesian equilibria. *Journal of Economic Theory*, 70:201–234, 1996.

[2] Pierpaolo Battigalli and Pietro Veronesi. A note on stochastic independence without Savage-null

events. *Journal of Economic Theory*, 70(1):235–248, 1996.

[3] Lawrence Blume, Adam Brandenburger, and Eddie Dekel. Lexicographic probabilities and choice under uncertainty. *Econometrica*, 58(1):61–79, January 1991.

[4] Giulianella Coletti and Romano Scozzafava. Stochastic independence in a coherent setting. *Annals of Mathematics and Artificial Intelligence*, 35:151-176, 2002.

[5] Giulianella Coletti and Romano Scozzafava. *Probabilistic Logic in a Coherent Setting*. Trends in logic, 15. Kluwer, Dordrecht, 2002.

[6] Inés Couso and Serafín Moral. Sets of desirable gambles: Conditioning, representation, and precise probabilities. *International Journal of Approximate Reasoning*, 52:1034–1055, 2011.

[7] Fabio Gagliardi Cozman. Irrelevance and independence axioms in quasi-Bayesian theory. In Anthony Hunter and Simon Parsons, editors, *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU)*, pages 128–136. Springer, London, England, 1999.

[8] Fabio Gagliardi Cozman. Computing lower expectations with Kuznetsov's independence condition. In *Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications*, pages 177–187, Lugano, Switzerland, 2003. Carleton Scientific.

[9] Fabio Gagliardi Cozman. Sets of probability distributions, independence, and convexity. *Synthese*, 186(2):577–600, 2012.

[10] Fabio Gagliardi Cozman. Independence for Full Conditional Probabilities: Structure, Nonuniqueness, Factorization, and Bayesian Networks. Technical Report Decision Making Lab 2013-001, São Paulo, Brazil, 2013.

[11] Fabio Gagliardi Cozman and Teddy Seidenfeld. Independence for full conditional measures and their graphoid properties. In Benedikt Lowe, Eric Pacuit, and Jan-Willem Romeijn, editors, *Reasoning about Probabilities and Probabilistic Reasoning*, volume 16 of *Foundations of the Formal Sciences VI*, pages 1–29. College Publications, London, 2009.

[12] Fabio Gagliardi Cozman and Peter Walley. Graphoid properties of epistemic irrelevance and independence. *Annals of Mathematics and Artificial Intelligence*, 45:173–195, 2005.

[13] Ákos Császár. Sur la structure des espaces de probabilité conditionnelle. *Acta Mathematica Academiae Scientiarum Hungarica*, 6(3-4):337–361, 1955.

[14] A. Philip Dawid. Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society B*, 41:1–31, 1979.

[15] L. de Campos and Serafín Moral. Independence concepts for convex sets of probabilities. In Phillippe Besnard and Steve Hanks, editors, *XI Conference on Uncertainty in Artificial Intelligence*, pages 108–115, San Francisco, California, United States, 1995. Morgan Kaufmann.

[16] Jasper De Bock and Gert De Cooman. Imprecise Bernoulli processes. In *Communications in Computer and Information Science*, volume 299, pages 400–409. Springer, 2012.

[17] Gert de Cooman and Enrique Miranda. Irrelevant and independent natural extension for sets of desirable gambles. *Journal of Artificial Intelligence Research*, 45:601–640, 2012.

[18] Gert de Cooman and Erik Quaeghebeur. Exchangeability and sets of desirable gambles. *International Journal of Approximate Reasoning*, 53:363–395, 2012.

[19] Bruno de Finetti. *Theory of Probability, vol. 1-2.* Wiley, New York, 1974.

[20] Lester E. Dubins. Finitely additive conditional probability, conglomerability and disintegrations. *Annals of Statistics*, 3(1):89–99, 1975.

[21] F. J. Giron and S. Rios. Quasi-Bayesian behaviour: A more realistic approach to decision making? In J. M. Bernardo, J. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*, pages 17–38. University Press, Valencia, Spain, 1980.

[22] Joseph Y. Halpern. Lexicographic probability, conditional probability, and nonstandard probability. *Games and Economic Behavior*, 68:155–179, 2010.

[23] Peter J. Hammond. Elementary non-Archimedean representations of probability for decision theory and games. In P. Humphreys, editor, *Patrick Suppes: Scientific Philosopher; Volume 1*, pages 25–59. Kluwer, Dordrecht, The Netherlands, 1994.

[24] Peter J. Hammond. Non-Archimedean subjective probabilities in decision theory and games. Technical Report Working Paper No. 97-038, Stanford University Department of Economics, 1997.

[25] V. L. Klee Jr. The structure of semispaces. *Mathematica Scandinavia*, 4:54-64, 1956.

[26] Elon Kohlberg and Philip J. Reny. Independence on relative probability spaces and consistent assessments in game trees. *Journal of Economic Theory*, 75:280–313, 1997.

[27] Peter Krauss. Representation of conditional probability measures on Boolean algebras. *Acta Mathematica Academiae Scientiarum Hungaricae*, 19(3-4):229–241, 1968.

[28] V. P. Kuznetsov. *Interval Statistical Methods.* Radio i Svyaz Publ., (in Russian), 1991.

[29] Isaac Levi. *The Enterprise of Knowledge.* MIT Press, Cambridge, Massachusetts, 1980.

[30] V. McGee. Learning the impossible. In E. Bells and B. Skyrms, editors, *Probability and Conditionals*, pages 179–199. Cambridge University Press, 1994.

[31] Serafin Moral. Epistemic irrelevance on sets of desirable gambles. *Annals of Mathematics and Artificial Intelligence*, 45(1-2):197–214, October 2005.

[32] Roger B. Myerson. *Game Theory: Analysis of Conflict.* Harvard University Press, Cambridge, MA, 1991.

[33] Roger B. Myerson. Multistage games with communication. *Econometrica*, 54(2):323–358, 1986.

[34] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, California, 1988.

[35] Karl Raimund Popper. *The Logic of Scientific Discovery.* Hutchinson, London, 1975.

[36] Teddy Seidenfeld, Mark J. Schervish, and Joseph B. Kadane. Decisions without ordering. In W. Sieg, editor, *Acting and Reflecting*, pages 143–170. Kluwer Academic Publishers, 1990.

[37] Teddy Seidenfeld, Mark J. Schervish, and Joseph Kadane. Coherent choice functions under uncertainty. In *International Symposium on Imprecise Probability: Theories and Applications.* Prague, Czech Republic, 2007.

[38] Bas Cornelis van Fraassen. Fine-grained opinion, probability, and the logic of full belief. *Journal of Philosophical Logic*, 24:349–377, 1995.

[39] Barbara Vantaggi. Conditional independence in a coherent finite setting. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):287–313, 2001.

[40] Barbara Vantaggi. Conditional independence structures and graphical models. *International Journal of Uncertainty, Fuzziness and Knoweledge-Based Systems*, 11(5):545-571, 2003.

[41] Peter Walley. Coherent lower (and upper) probabilities. Technical Report Statistics Report 23, University of Warwick, Coventry, 1981.

[42] Peter Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.

[43] Peter Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24:125–148, 2000.

[44] Kurt Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24(2-3):149–170, 2000.

[45] Kurt Weichselberger, T. Augustin (assistant), and A. Wallner (assistant). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept.* Physica-Verlag Heidelberg, 2001.

Credal networks under epistemic irrelevance using sets of desirable gambles

Jasper De Bock
Ghent University, Belgium
jasper.debock@ugent.be

Gert de Cooman
Ghent University, Belgium
gert.decooman@ugent.be

Abstract

We present a new approach to credal networks, which are graphical models that generalise Bayesian nets to deal with imprecise probabilities. Instead of applying the commonly used notion of strong independence, we replace it by the weaker notion of epistemic irrelevance. We show how assessments of epistemic irrelevance allow us to construct a global model out of given local uncertainty models, leading to an intuitive expression for the so-called irrelevant natural extension of a network. In contrast with Cozman [2], who introduced this notion in terms of credal sets, our main results are presented using the language of sets of desirable gambles. This has allowed us to derive a number of useful properties of the irrelevant natural extension. It has powerful marginalisation properties and satisfies all graphoid properties but symmetry, both in their direct and reverse forms.

Keywords. Credal networks, epistemic irrelevance, sets of desirable gambles, graphoid properties, irrelevant natural extension, lower previsions, coherence.

1 Introduction

In his overview paper [2], Cozman discussed and compared a number of different extensions for so-called credal networks, which generalise standard Bayesian networks to allow for imprecise probability assessments.

One of these extensions is the so-called irrelevant natural extension, which captures that the non-parent non-descendants of any variable in the network are epistemically irrelevant to that variable given the value of its parents. Cozman argues that of all the possible extensions, this irrelevant natural extension is perhaps the most appealing one. Nevertheless, it has thus far received little attention.

The present paper tries to remedy this situation by providing a firm theoretical foundation for the irrelevant natural extension of a network, leading to, amongst other things, a powerful marginalisation property and a proof that it satisfies all graphoid properties but symmetry.

The main results are stated using the theory of sets of desirable gambles, which we introduce in Section 2. We go on to introduce and discuss important concepts such as directed acyclic graphs and epistemic irrelevance in Section 3, and use these in Section 4 to show how assessments of epistemic irrelevance can be combined with given local sets of desirable gambles to construct a joint model. We call this the *irrelevant natural extension* of the credal network and prove that it is the most conservative coherent model that extends the local models and expresses all conditional irrelevancies encoded in the network. In Section 5 we present a powerful marginalisation property, and in Section 6, we use an asymmetric version of D-separation to show that the irrelevant natural extension satisfies all graphoid properties except symmetry, both in their direct and reverse forms. Finally, Section 7 establishes a connection between the sets of desirable gambles approach to credal networks under epistemic irrelevance that we presented in this paper, and a similar approach using coherent lower previsions.

2 Sets of desirable gambles

Consider a variable X taking values in some non-empty and finite set $\mathscr{X}$. Beliefs about the possible values this variable may assume can be modelled in various ways: probability mass functions, credal sets and coherent lower previsions are only a few of the many options. We choose to adopt a different approach, using sets of desirable gambles. We will model a subject's beliefs regarding the value of a variable X by means of his behaviour: which gambles (or bets) on the unknown value of X would our subject strictly prefer to the status quo (the zero gamble).

Although they are not as well known as other (imprecise) probability models, sets of desirable gambles have definite advantages. To begin with, they are more expressive than both credal sets and lower previsions. For example, they are easily able to deal with such things as conditioning on events with probability zero, which tends to be much more involved when using other imprecise probability models. Secondly, they have the advantage of being operational, meaning that there is a practical way of constructing a model that represents the subject's beliefs. For sets of desirable gambles this can be done by offering the subject certain gambles and asking him whether or not he strictly prefers them to the status quo. And finally, our experience tells us that it is usually easier to construct proofs in the language of coherent sets of desirable gambles than in other, perhaps more familiar languages. We give a brief survey of the basics of sets of desirable gambles and refer to Refs. [7, 1, 12] for more details and further discussion.

2.1 Desirable gambles

A gamble f is a real-valued map on $\mathscr{X}$ that is interpreted as an uncertain reward. If the value of the variable X turns out to be x, the (possibly negative) reward is $f(x)$. A non-zero gamble is called *desirable* to a subject if he strictly prefers to zero the transaction in which (i) the actual value x of the variable is determined, and (ii) he receives the reward $f(x)$. The zero gamble is therefore not considered to be desirable.

We model a subject's beliefs regarding the possible values $\mathscr{X}$ that a variable X can assume by means of a set $\mathscr{D}$ of desirable gambles—some subset of the set $\mathscr{G}(\mathscr{X})$ of all gambles on $\mathscr{X}$. For any two gambles f and g in $\mathscr{G}(\mathscr{X})$, we say that $f \geq g$ if $f(x) \geq g(x)$ for all x in $\mathscr{X}$ and $f > g$ if both $f \geq g$ and $f \neq g$. We use $\mathscr{G}(\mathscr{X})_{>0}$ to denote the set of all gambles $f \in \mathscr{G}(\mathscr{X})$ for which $f > 0$ and $\mathscr{G}(\mathscr{X})_{\leq 0}$ to denote the set of all gambles $f \in \mathscr{G}(\mathscr{X})$ for which $f \leq 0$. As a special kind of gambles we consider *indicators* $\mathbb{I}_A$ of events $A \subseteq \mathscr{X}$. $\mathbb{I}_A$ is equal to 1 if the event A occurs—the variable X assumes a value in A—and zero otherwise.

2.2 Coherence

In order to represent a rational subject's beliefs about the values a variable can assume, a set $\mathscr{D} \subseteq \mathscr{G}(\mathscr{X})$ of desirable gambles should satisfy some rationality requirements. If these requirements are met, we call the set $\mathscr{D}$ *coherent*. We require that for all f, f_1, $f_2 \in \mathscr{G}(\mathscr{X})$ and all real $\lambda > 0$:

D1. if $f \leq 0$ then $f \notin \mathscr{D}$;

D2. if $f > 0$ then $f \in \mathscr{D}$;

D3. if $f \in \mathscr{D}$ then $\lambda f \in \mathscr{D}$; [scaling]

D4. if $f_1, f_2 \in \mathscr{D}$ then $f_1 + f_2 \in \mathscr{D}$. [combination]

Requirements D3 and D4 turn $\mathscr{D}$ into a convex cone: $\operatorname{posi}(\mathscr{D}) = \mathscr{D}$, where we use the positive hull operator 'posi' that generates the set of finite strictly positive linear combinations of elements of its argument set:

$$\operatorname{posi}(\mathscr{D}) := \left\{ \sum_{k=1}^{n} \lambda_k f_k \colon f_k \in \mathscr{D}, \lambda_k \in \mathbb{R}_0^+, n \in \mathbb{N}_0 \right\}.$$

Here $\mathbb{R}_0^+$ is the set of all (strictly) positive real numbers, and $\mathbb{N}_0$ the set of all natural numbers (zero not included).

3 Credal networks

3.1 Directed acyclic graphs

A directed acyclic graph (DAG) is a graphical model that is well known for its use in Bayesian networks. It consists of a finite set of nodes (vertices), joined into a network by a set of directed edges, each edge connecting one node with another. Since this directed graph is assumed to be acyclic, it is not possible to follow a sequence of edges from node to node and end up at the same node one started out from.

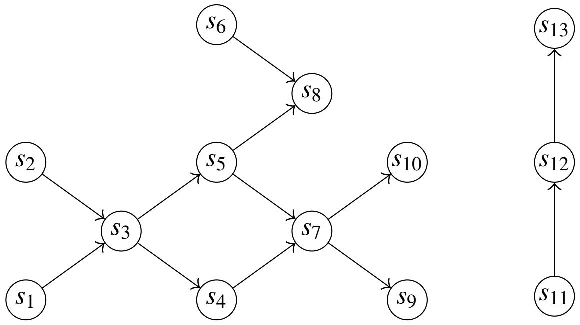
We will call G the set of nodes s associated with a given DAG. For two nodes s and t, if there is a directed edge from s to t, we denote this as $s \to t$ and say that s is a *parent* of t and t is a *child* of s. A single node can have multiple parents and multiple children. For any node s, its set of parents is denoted by $P(s)$ and its set of children by $C(s)$. If a node s has no parents, $P(s) = \emptyset$, and we call s a *root node*. If $C(s) = \emptyset$, then we call s a *leaf*, or *terminal node*.

Two nodes s and t are said to have a *path* between them if one can start from s, follow the edges of the DAG regardless of their direction and end up in t. In other words: one can find a sequence of nodes $s = s_1, \ldots, s_n = t$, $n \geq 1$, such that for all $i \in \{1, \ldots, n-1\}$ either $s_i \to s_{i+1}$ or $s_i \leftarrow s_{i+1}$. If this sequence is such that $s_i \to s_{i+1}$ for all $i \in \{1, \ldots, n-1\}$ (all edges in the path point away from s), we say that there is a *directed path* from s to t and write $s \sqsubseteq t$. In that case we also say that *s precedes t*. If $s \sqsubseteq t$ and $s \neq t$, we say that *s strictly precedes t* and write $s \sqsubset t$. For any node s, we denote its set of *descendants* by $D(s) := \{t \in G \colon s \sqsubset t\}$ and its set of *non-parent non-descendants* by $N(s) := G \setminus (P(s) \cup \{s\} \cup D(s))$. We also use the shorthand notation $PN(s) := P(s) \cup N(s) = G \setminus (\{s\} \cup D(s))$ to refer to the so-called *non-descendants* of s.

We extend these notions to subsets of G in the following way. For any $K \subseteq G$, $P(K) := (\bigcup_{s \in K} P(s)) \setminus K$ is its set of parents and $D(K) := (\bigcup_{s \in K} D(s)) \setminus K$ is its set of descendants. The non-parent non-descendants of K are given by $N(K) := G \setminus (P(K) \cup K \cup D(K)) = \bigcap_{s \in K} N(s)$, and we also define $PN(K) := P(K) \cup N(K)$. This last set cannot be referred to as the non-descendants of K since $P(K)$ and $D(K)$ are not necessarily disjoint.

Special subsets of G that we will consider, are the closed ones: we call a set $K \subseteq G$ *closed* if for all $s, t \in K$ and any $k \in G$ such that $s \sqsubseteq k \sqsubseteq t$, it holds that $k \in K$. For closed $K \subseteq G$, $P(K) \cap D(K) = \emptyset$ and therefore $PN(K) = G \setminus (K \cup D(K))$, which means that for closed K, $PN(K)$ can rightfully be referred to as the non-descendants of K.

With any subset K of G, we can associate a so-called *sub-DAG* of the DAG that is associated with G. The nodes of this sub-DAG are the elements of K and the directed edges of this sub-DAG are those edges in the original DAG that



$$G = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}\}$$

Figure 1: Example of a directed acyclic graph (DAG)

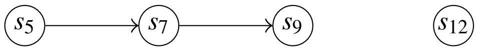


Figure 2: Example of a sub-DAG

connect elements in K. For a sub-DAG that is associated with some subset K of G, we will use similar definitions as those for the original DAG, adding the subset K as an index. As an example: for all $k \in K$, we denote by $P_K(k)$ the parents of k in the sub-DAG that is associated with the nodes in K. For all $K \subseteq G$ and $k \in K$, we have $P_K(k) = P(k) \cap K$ and $P(k) \setminus P_K(k) = P(k) \cap P(K)$.

Example 1. *Consider the DAG in Figure 1. For the node $s_7 \in G$, we find that $P(s_7) = \{s_4, s_5\}$, $D(s_7) = \{s_9, s_{10}\}$ and $N(s_7) = \{s_1, s_2, s_3, s_6, s_8, s_{11}, s_{12}, s_{13}\}$. For the closed subset $K = \{s_5, s_7, s_9, s_{12}\} \subset G$, we have $P(K) = \{s_3, s_4, s_{11}\}$, $D(K) = \{s_8, s_{10}, s_{13}\}$ and $N(K) = \{s_1, s_2, s_6\}$. The sub-DAG that corresponds to K is drawn in Figure 2. We find that $P_K(s_7) = \{s_5\}$, $D_K(s_7) = \{s_9\}$ and $N_K(s_7) = \{s_{12}\}$.* ◊

3.2 Variables and gambles on them

With each node s of the network, we associate a variable X_s assuming values in some non-empty finite set $\mathscr{X}_s$. We denote by $\mathscr{G}(\mathscr{X}_s)$ the set of all gambles on $\mathscr{X}_s$. We extend this notation to more complicated situations as follows. If S is any subset of G, then we denote by X_S the tuple of variables whose components are the X_s for all $s \in S$. This new joint variable assumes values in the finite set $\mathscr{X}_S := \times_{s \in S} \mathscr{X}_s$ and the corresponding set of gambles is denoted by $\mathscr{G}(\mathscr{X}_S)$. When $S = \emptyset$, we let $\mathscr{X}_\emptyset$ be a singleton. The corresponding variable $X_\emptyset$ can then only assume this single value, so there is no uncertainty about it. $\mathscr{G}(\mathscr{X}_\emptyset)$ can then be identified with the set $\mathbb{R}$ of real numbers. Generic elements of $\mathscr{X}_s$ are denoted by x_s or z_s and similarly for x_S and z_S in $\mathscr{X}_S$. Also, if we mention a tuple z_S, then for any $t \in S$, the corresponding element in the tuple will be denoted by z_t. We assume all variables in the network to be logically independent, meaning that the variable X_S may assume *all* values in $\mathscr{X}_S$, for all $\emptyset \subseteq S \subseteq G$.

We will use the simplifying device of identifying a gamble f_S on $\mathscr{X}_S$ with its *cylindrical extension* to $\mathscr{X}_U$, where $S \subseteq U \subseteq G$: the gamble f_U on $\mathscr{X}_U$ defined by $f_U(x_U) := f_S(x_S)$ for all $x_U \in \mathscr{X}_U$. For instance, if $\mathscr{K} \subseteq \mathscr{G}(\mathscr{X}_G)$, this allows us to consider $\mathscr{K} \cap \mathscr{G}(\mathscr{X}_S)$ as the set of those gambles in $\mathscr{K}$ that depend only on the variable X_S.

3.3 Modelling our beliefs about the network

Throughout, we consider sets of desirable gambles as models for a subject's beliefs about the values that certain variables in the network may assume. One of the main contributions of this paper, further on in Section 4, will be to show how to construct a joint model for our network, being a coherent set $\mathscr{D}_G$ of desirable gambles on $\mathscr{X}_G$.

From such a joint model, one can derive both conditional and marginal models [7, 6]. Let us start by explaining how to condition the global model $\mathscr{D}_G$. Consider an event $A_I \subseteq \mathscr{X}_I$, with $I \subseteq G$, and assume that we want to update the model $\mathscr{D}_G$ with the information that $X_I \in A_I$. This leads to the following updated set of desirable gambles:

$$\mathscr{D}_G \rfloor A_I := \{f \in \mathscr{G}(\mathscr{X}_{G \setminus I}) \colon \mathbb{I}_{A_I} f \in \mathscr{D}_G\},$$

which represents our subject's beliefs about the value of the variable $X_{G \setminus I}$, conditional on the observation that X_I assumes a value in A_I. This definition is very intuitive, since $\mathbb{I}_{A_I} f$ is the unique gamble that is called off (is equal to zero) if $X_I \notin A_I$ and equal to f if $X_I \in A_I$. Since $\mathbb{I}_{\{x_\emptyset\}} = 1$, the special case of conditioning on the certain variable $X_\emptyset$ yields no problems: it amounts to not conditioning at all.

Marginalisation too is very intuitive in the language of sets of desirable gambles. Suppose we want to derive a marginal model for our subject's beliefs about the variable X_O, where O is some subset of G. This can be done by using the set of desirable gambles that belong to $\mathscr{D}_G$ but only depend on the variable X_O:

$$\operatorname{marg}_O(\mathscr{D}_G) := \{f \in \mathscr{G}(\mathscr{X}_O) \colon f \in \mathscr{D}_G\} = \mathscr{D}_G \cap \mathscr{G}(\mathscr{X}_O).$$

Now let I and O be *disjoint* subsets of G and let A_I be any subset of $\mathscr{X}_I$. By sequentially applying the process of conditioning and marginalisation we can obtain conditional marginal models for our subject's beliefs about the value of the variable X_O, conditional on the observation that X_I assumes a value in A_I:

$$\operatorname{marg}_O(\mathscr{D}_G \rfloor A_I) = \{f \in \mathscr{G}(\mathscr{X}_O) \colon \mathbb{I}_{A_I} f \in \mathscr{D}_G\}. \quad (1)$$

Conditioning and marginalisation are special cases of Eq. (1); they can be obtained by letting $O = G \setminus I$ or $I = \emptyset$. If A_I is a singleton $\{x_I\}$, with $x_I \in \mathscr{X}_I$, we will use the shorthand notation $\operatorname{marg}_O(\mathscr{D}_G \rfloor x_I) := \operatorname{marg}_O(\mathscr{D}_G \rfloor \{x_I\})$.

Since coherence is trivially preserved under both conditioning and marginalisation, we find that if the joint model $\mathscr{D}_G$ is coherent, all the derived models will also be coherent.

3.4 Epistemic irrelevance

We now have the necessary tools to introduce one of the most important concepts for this paper, that of epistemic irrelevance. We describe the case of conditional irrelevance, as the unconditional version of epistemic irrelevance can easily be recovered as a special case.

Consider three disjoint subsets C, I, and O of G. When a subject judges X_I to be *epistemically irrelevant to* X_O *conditional on* X_C, denoted as $\mathrm{IR}(I,O|C)$, he assumes that if he knew the value of X_C, then learning in addition which value X_I assumes in $\mathscr{X}_I$ would not affect his beliefs about X_O. More formally put, he assumes for all $x_C \in \mathscr{X}_C$ and $x_I \in \mathscr{X}_I$ that:

$$\mathrm{marg}_O(\mathscr{D}_G \rfloor x_{C\cup I}) = \mathrm{marg}_O(\mathscr{D}_G \rfloor x_C).$$

Alternatively, a subject can make the even stronger statement that he judges X_I to be epistemically *subset-irrelevant* to X_O conditional on X_C, denoted as $\mathrm{SIR}(I,O|C)$. In that case, he assumes that if he knew the value of X_C, then receiving the additional information that X_I is an element of any non-empty subset A_I of $\mathscr{X}_I$ would not affect his beliefs about X_O. In other words, he assumes for all $x_C \in \mathscr{X}_C$ and all non-empty $A_I \subseteq \mathscr{X}_I$ that:

$$\mathrm{marg}_O(\mathscr{D}_G \rfloor \{x_C\} \times A_I) = \mathrm{marg}_O(\mathscr{D}_G \rfloor x_C).$$

Making a subset-irrelevance statement $\mathrm{SIR}(I,O|C)$ implies the corresponding irrelevance statement $\mathrm{IR}(I,O|C)$. Even stronger, it implies for all $I' \subseteq I$ that $\mathrm{IR}(I',O|C)$. The converse does not hold in general. However, as we will show further on, credal networks under epistemic irrelevance are a useful exception: although we define the joint model by imposing irrelevance, it will also satisfy subset-irrelevance. For the unconditional irrelevance case it suffices, in the discussion above, to let $C=\emptyset$. This makes sure the variable X_C has only one possible value, so conditioning on that variable amounts to not conditioning at all.

Irrelevance and subset-irrelevance can also be extended to cases where I, O and C are not disjoint, but $I \setminus C$ and $O \setminus C$ are. We then call X_I epistemically (subset-)irrelevant to X_O conditional on X_C provided that $X_{I\setminus C}$ is epistemically (subset-)irrelevant to $X_{O\setminus C}$ conditional on X_C. Although these cases are admittedly artificial, they will help us state and prove some of the graphoid properties further on.

3.5 Local uncertainty models

We now add *local uncertainty models* to each of the nodes s in our network. These local models are assumed to be given beforehand and will be used further on in Section 4 as basic building blocks for constructing a joint model for a given network.

If s is not a root node of the network, i.e. has a non-empty set of parents $P(s)$, then we have a conditional local model for every instantiation of its parents: for each $x_{P(s)} \in \mathscr{X}_{P(s)}$, we have a coherent set $\mathscr{D}_{s\rfloor x_{P(s)}}$ of desirable gambles on $\mathscr{X}_s$. It represents our subject's beliefs about the variable X_s conditional on its parents $X_{P(s)}$ assuming the value $x_{P(s)}$.

If s is a root node, i.e. has no parents, then our subject's local beliefs about the variable X_s are represented by an unconditional local model. It should be a coherent set of desirable gambles and will be denoted by $\mathscr{D}_s$. As was explained in Section 3.3, we can also use the common generic notation $\mathscr{D}_{s\rfloor x_{P(s)}}$ in this unconditional case, since for a root node s, its set of parents $P(s)$ is equal to the empty set $\emptyset$.

3.6 The interpretation of the graphical model

In classical Bayesian nets, the graphical structure is taken to represent the following assessments: for any node s, conditional on its parent variables, the associated variable is independent of its non-parent non-descendant variables

When generalising this interpretation to credal networks, the classical notion of independence gets replaced by a more general, imprecise-probabilistic notion of independence, which in the existing literature is usually chosen to be strong independence; see Ref. [3] for an overview of different approaches, including relevant references. Here, we will not do so: we choose to use the weaker, asymmetric notion of epistemic irrelevance, introduced in Section 3.4. In the special case of precise uncertainty models, both epistemic irrelevance and strong independence reduce to the classical notion of independence and the corresponding interpretations of the graphical network are equivalent to the one used in classical Bayesian networks.

In the present context, we therefore assume that the graphical structure of the network embodies the following conditional irrelevance assessments, turning the network into a *credal network under epistemic irrelevance*. Consider any node s in the network, its set of parents $P(s)$ and its set of non-parent non-descendants $N(s)$. Then *conditional on* $X_{P(s)}$, $X_{N(s)}$ *is assumed to be epistemically irrelevant to* X_s:

$$\mathrm{IR}(N(s),\{s\}|P(s)).$$

For a coherent set of desirable gambles $\mathscr{D}_G$ that describes our subject's global beliefs about all the variables in the network, this has the following consequences. For every $s \in G$ and all $x_{PN(s)} \in \mathscr{X}_{PN(s)}$, $\mathscr{D}_G$ must satisfy:

$$\mathrm{marg}_s(\mathscr{D}_G \rfloor x_{PN(s)}) = \mathrm{marg}_s(\mathscr{D}_G \rfloor x_{P(s)}). \tag{2}$$

4 Constructing a joint model

We now show how to construct a joint model for the variables in the network, and argue that it is the most conservative coherent model that extends the local models and expresses all conditional irrelevancies encoded in the network. But before we do so, let us provide some motivation. Suppose we have a global set of desirable gambles $\mathscr{D}_G$, how do we express that such a model is compatible with the assessments encoded in the network?

4.1 Defining properties of the joint model

We will require our joint model to satisfy the following four properties. First of all, we require that our global model should extend the local ones. This means that the local models derived from the global one by marginalisation should be equal to the given local models:

G1. The joint model $\mathscr{D}_G$ marginalises to the given local uncertainty models: $\operatorname{marg}_s(\mathscr{D}_G \rfloor x_{P(s)}) = \mathscr{D}_{s\rfloor x_{P(s)}}$ for all $s \in G$ and $x_{P(s)} \in \mathscr{X}_{P(s)}$.

The second requirement is that our model should reflect all epistemic irrelevancies encoded in the graphical structure of the network:

G2. $\mathscr{D}_G$ satisfies all equalities that are imposed by Eq. (2). In these equalities, the right hand side can be replaced by $\mathscr{D}_{s\rfloor x_{P(s)}}$ due to requirement G1.

The third requirement is that our model should be coherent:

G3. $\mathscr{D}_G$ is coherent (satisfies requirements D1–D4).

Since requirements G1–G3 do not determine a unique global model, we impose a final requirement to ensure that all inferences we make on the basis of our global models are as conservative as possible, and are therefore based on no other considerations than what is encoded in the network:

G4. $\mathscr{D}_G$ is the smallest set of desirable gambles on $\mathscr{X}_G$ satisfying requirements G1–G3: it is a subset of any other set that satisfies them.

We will now show how to construct the unique global model $\mathscr{D}_G$ that satisfies all of the four requirements G1–G4.

4.2 An intuitive expression for the joint model

Let us start by looking at a single given marginal model $\mathscr{D}_{s\rfloor z_{P(s)}}$ and investigate some of its implications for the joint model $\mathscr{D}_G$. Consider any node s and fix values $z_{P(s)}$ and $z_{N(s)}$ for its parents and non-parent non-descendants. Due to requirements G1 and G2, any gamble $f \in \mathscr{D}_{s\rfloor z_{P(s)}}$ should also be an element of $\operatorname{marg}_s(\mathscr{D}_G \rfloor z_{PN(s)})$, which by definition means that $\mathbb{I}_{\{z_{PN(s)}\}} f \in \mathscr{D}_G$. Inspired by this observation, we introduce the following set of gambles on $\mathscr{X}_G$:

$$\mathscr{A}_G^{\mathrm{irr}} := \left\{ \mathbb{I}_{\{z_{PN(s)}\}} f \colon s \in G,\, z_{PN(s)} \in \mathscr{X}_{PN(s)},\, f \in \mathscr{D}_{s\rfloor z_{P(s)}} \right\}.$$

It should now be clear that $\mathscr{A}_G^{\mathrm{irr}}$ must be a subset of our joint model $\mathscr{D}_G$.

Proposition 1. *$\mathscr{A}_G^{\mathrm{irr}}$ is a subset of any joint model $\mathscr{D}_G$ that satisfies requirements* G1 *and* G2.

Since our eventual joint model should also be coherent (satisfy requirement G3), and thus in particular should be a convex cone, we can derive the following corollary.

Corollary 2. $\operatorname{posi}(\mathscr{A}_G^{\mathrm{irr}})$ *is a subset of any joint model $\mathscr{D}_G$ that satisfies requirements* G1–G3.

We now suggest the following expression for the joint model describing our subject's beliefs about the variables in the network:

$$\mathscr{D}_G^{\mathrm{irr}} := \operatorname{posi}(\mathscr{A}_G^{\mathrm{irr}}). \tag{3}$$

We will refer to $\mathscr{D}_G^{\mathrm{irr}}$ as the *irrelevant natural extension* of the local models $\mathscr{D}_{s\rfloor x_{P(s)}}$. Since we know from Corollary 2 that it is guaranteed to be a subset of the joint model we are looking for, we propose it as a candidate for the joint model itself. In the next section, we set out to prove that $\mathscr{D}_G^{\mathrm{irr}}$ is indeed the unique joint model satisfying all four requirements G1–G4.

We would like to point out that $\mathscr{D}_G^{\mathrm{irr}}$ is a generalisation of the so-called *independent natural extension* of a number of unconditional marginal models [6, Section 7]. This special case corresponds to a DAG that has no edges, consisting of a finite amount of disconnected nodes [6, Section 10]. Quite a few of the results obtained further on can therefore be regarded as generalisations of those in Ref. [6].

4.3 Justifying our expression for the joint model

We start by proving a number of useful properties of $\mathscr{D}_G^{\mathrm{irr}}$.

Proposition 3. *A gamble $f \in \mathscr{G}(\mathscr{X}_G)$ is an element of $\mathscr{D}_G^{\mathrm{irr}}$ if and only if it can be written as:*

$$f = \sum_{s \in G} \sum_{z_{PN(s)} \in \mathscr{X}_{PN(s)}} \mathbb{I}_{\{z_{PN(s)}\}} f_{s, z_{PN(s)}},$$

where $f_{s,z_{PN(s)}} \in \mathscr{D}_{s\rfloor z_{P(s)}} \cup \{0\}$ for every $s \in G$ and all $z_{PN(s)} \in \mathscr{X}_{PN(s)}$, and at least one of them is non-zero.

Proposition 4. $\mathscr{G}(\mathscr{X}_G)_{>0}$ *is a subset of $\mathscr{D}_G^{\mathrm{irr}}$.*

These two propositions serve as a first step towards the following coherence result, which states that our joint model $\mathscr{D}_G^{\mathrm{irr}}$ satisfies requirement G3.

Proposition 5. *$\mathscr{D}_G^{\mathrm{irr}}$ satisfies requirement* G3*: it is a coherent set of desirable gambles.*

Our proof for this result has an interesting feature that deserves to be borne out. The crucial step hinges on the assumption that if the local models of our network were precise probability mass functions, we would be able to construct a joint probability mass function that satisfies all irrelevancies (in that case independencies) that are encoded in our network. Since the precise version of a credal net under epistemic irrelevance is a classical Bayesian network, this assumption is indeed true. What we believe is useful about this approach, is that it can be extended to credal networks with irrelevance assumptions that differ from the ones we impose in the present article, as long as the assumption above is satisfied. In this way, it enables us to use existing coherence results for precise networks to prove their counterparts for credal networks.

Next, we turn to an important factorisation result that is essential in order to prove that our joint model extends the local models and expresses all conditional irrelevancies encoded in the network, and therefore satisfies G1 and G2.

Proposition 6. *Fix arbitrary $s \in G$, $x_{P(s)} \in \mathscr{X}_{P(s)}$ and $g \in \mathscr{G}(\mathscr{X}_{N(s)})_{>0}$. For every $f \in \mathscr{G}(\mathscr{X}_s)$:*

$$g\mathbb{I}_{\{x_{P(s)}\}}f \in \mathscr{D}_G^{\mathrm{irr}} \Leftrightarrow f \in \mathscr{D}_{s\rfloor x_{P(s)}}.$$

Corollary 7. *$\mathscr{D}_G^{\mathrm{irr}}$ satisfies requirements* G1 *and* G2*: it holds for every $s \in G$ and all $x_{PN(s)} \in \mathscr{X}_{PN(s)}$ that*

$$\mathrm{marg}_s(\mathscr{D}_G^{\mathrm{irr}}\rfloor x_{PN(s)}) = \mathrm{marg}_s(\mathscr{D}_G^{\mathrm{irr}}\rfloor x_{P(s)}) = \mathscr{D}_{s\rfloor x_{P(s)}}.$$

We now have all tools necessary to formulate our first important result. It is one of the main contributions of this paper and provides a justification for the joint model $\mathscr{D}_G^{\mathrm{irr}}$ that was proposed in Eq. (3).

Theorem 8. *The irrelevant natural extension $\mathscr{D}_G^{\mathrm{irr}}$ is the unique set of desirable gambles on $\mathscr{X}_G$ that satisfies all four requirements* G1–G4.

It is already apparent from Proposition 6 that the properties of the irrelevant natural extension $\mathscr{D}_G^{\mathrm{irr}}$ are not limited to G1–G4. As a first example, Proposition 6 implies that for any node s, conditional on its parent variables $X_{P(s)}$, the non-parent non-descendant variables $X_{N(s)}$ are not only epistemically irrelevant, but also subset-irrelevant to X_s.

Corollary 9. *All nodes $s \in G$ satisfy the subset-irrelevance statement* $\mathrm{SIR}(N(s),\{s\}|P(s))$*: for any $x_{P(s)} \in \mathscr{X}_{P(s)}$ and non-empty $A_{N(s)} \subseteq \mathscr{X}_{N(s)}$, it holds that*

$$\mathrm{marg}_s(\mathscr{D}_G^{\mathrm{irr}}\rfloor \{x_{P(s)}\} \times A_{N(s)}) = \mathrm{marg}_s(\mathscr{D}_G^{\mathrm{irr}}\rfloor x_{P(s)}).$$

In the next two sections, we establish a number of even stronger properties of $\mathscr{D}_G^{\mathrm{irr}}$.

5 Additional marginalisation properties

As explained in Section 3.1, a subset K of G can be associated with a so-called sub-DAG of the original DAG. Similarly to what we have done for the original DAG, we can use Eq. (3) to construct a joint model for this sub-DAG. All we need to do is provide, for every $s \in K$ and $z_{P_K(s)} \in \mathscr{X}_{P_K(s)}$, a local model $\mathscr{D}_{s\rfloor z_{P_K(s)}}$.

One particular way of providing these local models is to derive them from the ones of the original DAG. The starting point to do so is fixing a value $x_{P(K)} \in \mathscr{X}_{P(K)}$ for the parent variables of K. This provides us, for every $s \in K$, with a value $x_{P(s)\setminus P_K(s)} \in \mathscr{X}_{P(s)\setminus P_K(s)}$ because $P(s)\setminus P_K(s) \subseteq P(K)$. For every $s \in K$ and $z_{P_K(s)} \in \mathscr{X}_{P_K(s)}$, we can then identify the local model $\mathscr{D}_{s\rfloor z_{P_K(s)}}$ of the sub-DAG with the local model $\mathscr{D}_{s\rfloor z_{P(s)}}$ of the original DAG, where $z_{P(s)\setminus P_K(s)} = x_{P(s)\setminus P_K(s)}$. In other words, for every $s \in K$ and $z_{P_K(s)} \in \mathscr{X}_{P_K(s)}$

$$\mathscr{D}_{s\rfloor z_{P_K(s)}} = \mathscr{D}_{s\rfloor (z_{P_K(s)}, x_{P(s)\setminus P_K(s)})}.$$

Example 2. *Consider again the DAG in Figure 1 and the sub-DAG in Figure 2 that corresponds to the closed subset $K = \{s_5, s_7, s_9, s_{12}\} \subset G$. In order to provide this sub-DAG with local models, we fix a value $x_{P(K)} \in \mathscr{X}_{P(K)}$. Using Eq.* (5)*, this provides us with unconditional local models $\mathscr{D}_{s_5} = \mathscr{D}_{s_5\rfloor x_{s_3}}$ and $\mathscr{D}_{s_{12}} = \mathscr{D}_{s_{12}\rfloor x_{s_{11}}}$, for all $z_{s_5} \in \mathscr{X}_{s_5}$, a conditional local model $\mathscr{D}_{s_7\rfloor z_{s_5}} = \mathscr{D}_{s_7\rfloor (z_{s_5}, x_{s_4})}$ and, for all $z_{s_7} \in \mathscr{X}_{s_7}$, a conditional local model $\mathscr{D}_{s_9\rfloor z_{s_7}}$.* ◊

For every $K \subseteq G$ and all $x_{P(K)} \in \mathscr{X}_{P(K)}$, the resulting joint model for the sub-DAG that is associated with K is given by

$$\mathscr{D}_{K\rfloor x_{P(K)}}^{\mathrm{irr}} := \mathrm{posi}(\mathscr{A}_{K\rfloor x_{P(K)}}^{\mathrm{irr}}),$$

where

$$\mathscr{A}_{K\rfloor x_{P(K)}}^{\mathrm{irr}} := \Big\{\mathbb{I}_{\{z_{PN_K(s)}\}}f \colon s \in K, z_{PN_K(s)} \in \mathscr{X}_{PN_K(s)}, \\ f \in \mathscr{D}_{s\rfloor (z_{P_K(s)}, x_{P(s)\setminus P_K(s)})}\Big\}.$$

A question that now naturally arises is whether these joint models for sub-DAGs can be related to the original joint model $\mathscr{D}_G^{\mathrm{irr}}$. It turns out that, for subsets K of G that are closed, this is indeed the case.

Theorem 10. *If K is a closed subset of G, then for any $x_{P(K)} \in \mathscr{X}_{P(K)}$, $g \in \mathscr{G}(\mathscr{X}_{N(K)})_{>0}$ and $f \in \mathscr{G}(\mathscr{X}_K)$:*

$$g\mathbb{I}_{\{x_{P(K)}\}}f \in \mathscr{D}_G^{\mathrm{irr}} \Leftrightarrow f \in \mathscr{D}_{K\rfloor x_{P(K)}}^{\mathrm{irr}}.$$

The proof, although complex and elaborate, is essentially a simple separating hyperplane argument. We consider this result to be the main technical achievement of this paper. It is a significant generalisation of Proposition 6 [with $K = \{s\}$] and has a number of interesting consequences. As a first example, it implies the following generalisations of Corollaries 7 and 9.

Corollary 11. *For all closed $K \subseteq G$, $x_{P(K)} \in \mathscr{X}_{P(K)}$ and non-empty $A_{N(K)} \subseteq \mathscr{X}_{N(K)}$, we have that*

$$\mathrm{marg}_K(\mathscr{D}_G^{\mathrm{irr}} \rfloor \{x_{P(K)}\} \times A_{N(K)}) = \mathscr{D}_{K \rfloor x_{P(K)}}^{\mathrm{irr}}.$$

Corollary 12. *All closed sets $K \subseteq G$ satisfy the subset-irrelevance statement* $\mathrm{SIR}(N(K), K|P(K))$*: for any $x_{P(K)} \in \mathscr{X}_{P(K)}$ and non-empty $A_{N(K)} \subseteq \mathscr{X}_{N(K)}$, it holds that*

$$\mathrm{marg}_K(\mathscr{D}_G^{\mathrm{irr}} \rfloor \{x_{P(K)}\} \times A_{N(K)}) = \mathrm{marg}_K(\mathscr{D}_G^{\mathrm{irr}} \rfloor x_{P(K)}).$$

In the next section, we will extend this subset-irrelevance result to even more general cases.

6 AD-Separation and graphoid properties

In credal networks that are defined by means of a symmetrical independence concept, the notion of D-separation is a very powerful tool [9]. For asymmetrical independence concepts such as epistemic (subset-)irrelevance, D-separation has been modified to take this asymmetry into account. Moral [8] speaks of *asymmetrical D-separation* (AD-separation) and Vantaggi [10] has introduced the very similar *L-separation* criterion. Here, we choose not to use one of these existing concepts, but to introduce a slightly modified version of AD-separation. We do so because our definition is weaker (more general) than both Moral's AD-separation and L-separation and yet has stronger properties.

Consider any path $s_1, \dots, s_n$ in G, with $n \geq 1$. We say that this path is *blocked* by a set of nodes $C \subseteq G$ whenever at least one of the following four conditions holds:

B1. $s_1 \in C$;

B2. there is some $1 < i < n$ such that $s_i \to s_{i+1}$ and $s_i \in C$;

B3. there is some $1 < i < n$ such that $s_{i-1} \to s_i \leftarrow s_{i+1}$, $s_i \notin C$ and $D(s_i) \cap C = \emptyset$;

B4. $s_n \in C$.

Now consider (not necessarily disjoint) subsets I, O and C of G. We say that O is *AD-separated* from I by C, denoted as $\mathrm{AD}(I,O|C)$, if every path $i = s_1, \dots, s_n = o$, $n \geq 1$, from a node $i \in I$ to a node $o \in O$, is blocked by C. Our version of AD-separation satisfies a number of useful properties.

Theorem 13. *For any subsets I, O, S and C of G, the following properties hold:*

Direct redundancy: $\mathrm{AD}(I,O|I)$

Reverse redundancy: $\mathrm{AD}(I,O|O)$

Direct decomposition: $\mathrm{AD}(I,O \cup S|C) \Rightarrow \mathrm{AD}(I,O|C)$

Reverse decomposition: $\mathrm{AD}(I \cup S,O|C) \Rightarrow \mathrm{AD}(I,O|C)$

Direct weak union: $\mathrm{AD}(I,O \cup S|C) \Rightarrow \mathrm{AD}(I,O|C \cup S)$

Reverse weak union: $\mathrm{AD}(I \cup S,O|C) \Rightarrow \mathrm{AD}(I,O|C \cup S)$

Direct contraction:

$$\mathrm{AD}(I,O|C) \ \& \ \mathrm{AD}(I,S|C \cup O) \Rightarrow \mathrm{AD}(I,O \cup S|C)$$

Reverse contraction:

$$\mathrm{AD}(I,O|C) \ \& \ \mathrm{AD}(S,O|C \cup I) \Rightarrow \mathrm{AD}(I \cup S,O|C)$$

Direct intersection: *if $O \cap S = \emptyset$, then*

$$\mathrm{AD}(I,O|C \cup S) \ \& \ \mathrm{AD}(I,S|C \cup O) \Rightarrow \mathrm{AD}(I,O \cup S|C)$$

Reverse intersection: *if $I \cap S = \emptyset$, then*

$$\mathrm{AD}(I,O|C \cup S) \ \& \ \mathrm{AD}(S,O|C \cup I) \Rightarrow \mathrm{AD}(I \cup S,O|C)$$

This result (and our proof for it) is very similar to, and heavily inspired by, the work of Vantaggi [10, Theorem 7.1]. The main difference is that Vantaggi does not include the two redundancy properties, since L-separation is defined only for *disjoint* subsets I, O and C of G. Moral's version of AD-separation [8] does not require I, O and C to be disjoint, but it does not satisfy direct redundancy, and proofs for a number of other properties are not given [8, Theorem 4]. We therefore prefer our version of AD-separation.

Example 3. *Consider the sets of nodes $I = \{s_2, s_3, s_4, s_{11}\}$, $O = \{s_5, s_6, s_9, s_{13}\}$, $C = \{s_4, s_6, s_{12}\}$, $S_{\mathrm{d}} = \{s_8, s_{10}\}$ and $S_{\mathrm{r}} = \{s_1\}$ in the DAG that is depicted in Figure 1. The direct properties in Theorem 13 are illustrated by I, O, C and S_{d} and the reverse ones by I, O, C and S_{r}.* ◊

Theorem 10 implies a very general factorisation result.

Theorem 14. *If $I, O, C \subseteq G$ are such that* $\mathrm{AD}(I,O|C)$ *then for all $x_C \in \mathscr{X}_C$, $g \in \mathscr{G}(\mathscr{X}_{I \setminus C})_{>0}$ and $f \in \mathscr{G}(\mathscr{X}_{O \setminus C})$:*

$$g\mathbb{I}_{\{x_C\}} f \in \mathscr{D}_G^{\mathrm{irr}} \Leftrightarrow \mathbb{I}_{\{x_C\}} f \in \mathscr{D}_G^{\mathrm{irr}}.$$

This result can be combined with Theorem 13 to derive a collection of (subset-)irrelevance statements that are fulfilled by the irrelevant natural extension $\mathscr{D}_G^{\mathrm{irr}}$.

Corollary 15. *For any $I, O, C \subseteq G$ such that* $\mathrm{AD}(I,O|C)$ *we have that* $\mathrm{SIR}(I,O|C)$ *(and thus also* $\mathrm{IR}(I,O|C)$*): for all $x_C \in \mathscr{X}_C$ and non-empty $A_{I \setminus C} \subseteq \mathscr{X}_{I \setminus C}$ it holds that*

$$\mathrm{marg}_{O \setminus C}(\mathscr{D}_G^{\mathrm{irr}} \rfloor \{x_C\} \times A_{I \setminus C}) = \mathrm{marg}_{O \setminus C}(\mathscr{D}_G^{\mathrm{irr}} \rfloor x_C).$$

This family of subset-irrelevance statements satisfies all graphoid properties except symmetry: it satisfies redundancy, decomposition, weak union, contraction and intersection, both in their direct and reverse form.

We leave it to the reader to show that Theorem 14 is a generalisation of Theorem 10 and that Corollary 15 generalises the first part of Corollary 12. In other words: for any closed subset K of G, it holds that $\mathrm{AD}(N(K), K|P(K))$.

Readers who are familiar with the work in Ref. [8] might have noticed the similarity between Ref. [8, Theorem 5] and the first part of Corollary 15. The main difference between our approach and Moral's approach [8], besides the fact that we use a slightly different separation criterion, is that he enforces a more stringent version of epistemic irrelevance than we do. He calls X_I epistemically irrelevant to X_O if and only if the joint model $\mathscr{D}_{I\cup O}$ is the so-called irrelevant natural extension of $\mathscr{D}_I$ and $\mathscr{D}_O$ and refers to our concept of irrelevance as weak epistemic irrelevance. Consequently, if we understand his work correctly, his results are not applicable to all directed acyclic networks. As a simple example: his concept of irrelevance does not seem to allow for two variables to be mutually irrelevant, except in some degenerated uninformative cases. Therefore, it appears to us his results cannot be applied to a network consisting of two unconnected nodes.

As far as the second part of Corollary 15 is concerned, some clarification is perhaps in order. We do not claim that epistemic irrelevance satisfies the graphoid axioms that are stated in Theorem 13. As was proven in Ref. [4], epistemic irrelevance can violate direct contraction and both direct and reverse intersection. In fact, we believe that this negative result might even be one of the main reasons why a result such as Corollary 15 has thus far not appeared in any literature.

Indeed, in Bayesian networks, proving the counterpart to Corollary 15—with AD-separation replaced by D-separation and epistemic irrelevance replaced by stochastic independence—is usually done by using the fact that stochastic independence satisfies the graphoid axioms [9]. By applying these axioms to the independence assessements that are used to define a Bayesian network, one can infer new independencies, namely those that correspond to D-separations in the DAG of that network.

If one tries to mimic this approach in our context, then since epistemic irrelevance can fail some of the graphoid axioms, one might suspect that Corollary 15 cannot be proven. However, it is not necessary to use the axioms: our proof for Theorem 14—of which the the first part of Corollary 15 is a straightforward consequence—uses only Theorem 10 and a number of properties of AD-separation. At no point does it invoke graphoid properties of epistemic irrelevance. The second part of Corollary 15 is then but a mere consequence of the first part and Theorem 13. It states that the family of irrelevance statements that are proven to hold in the first part, are closed under the graphoid properties in Theorem 13.

So in order to conclude this section: epistemic irrelevance can fail a number of graphoid axioms, which implies that the irrelevance statements that are proven in Corollary 15 do not necessarily hold for every joint model $\mathscr{D}_G$ that satisfies requirements G1–G3. However for the unique one that also satisfies G4, being the irrelevant natural extension $\mathscr{D}_G^{\text{irr}}$ of the network, this family of irrelevance statements does hold, the reason being that for this specific model, one can provide a direct proof that does not invoke any graphoid axioms of epistemic irrelevance.

7 Credal nets under epistemic irrelevance using coherent lower previsions

Credal networks under epistemic irrelevance can also be defined using imprecise probability concepts other than coherent sets of desirable gambles. In this section, we describe an approach that uses coherent lower previsions, and we show how it is related to the desirable gambles approach of the previous sections.

7.1 Coherent lower previsions

For any subset O of G, we define a *coherent lower prevision* $\underline{P}_O$ as a real-valued functional on $\mathscr{G}(\mathscr{X}_O)$ that satisfies the following three conditions. For all $f,g \in \mathscr{G}(\mathscr{X}_O)$ and all real $\lambda \geq 0$:

C1. $\underline{P}_O(f) \geq \min f$;

C2. $\underline{P}_O(\lambda f) = \lambda \underline{P}_O(f)$; [non-negative homogeneity]

C3. $\underline{P}_O(f+g) \geq \underline{P}_O(f) + \underline{P}_O(g)$. [super-additivity]

Now consider two disjoint subsets O and I of G and suppose that we have, for all $x_I \in \mathscr{X}_I$, a coherent lower prevision $\underline{P}_O(\cdot|x_I)$ on $\mathscr{G}(\mathscr{X}_O)$. The corresponding *coherent conditional lower prevision* $\underline{P}_{O\cup I}(\cdot|X_I)$ is then a special two-place function that is defined, for all $f \in \mathscr{G}(\mathscr{X}_{O\cup I})$ and $x_I \in \mathscr{X}_I$, by $\underline{P}_{O\cup I}(f|x_I) := \underline{P}_O(f(\cdot,x_I)|x_I)$.

7.2 Defining a credal network

Suppose now that the local models of our credal network under epistemic irrelevance are coherent lower previsions: for all $s \in G$ and $x_{P(s)} \in \mathscr{X}_{P(s)}$, we have a coherent lower prevision $\underline{P}_{s\rfloor x_{P(s)}}$ on $\mathscr{G}(\mathscr{X}_s)$.

The irrelevance assessments that are encoded in the network can then be expressed as follows. For all $s \in G$, $I \subseteq N(s)$, $x_{P(s)\cup I} \in \mathscr{X}_{P(s)\cup I}$ and $f \in \mathscr{G}(\mathscr{X}_s)$, we require that:

$$\underline{P}_{\{s\}}(f|x_{P(s)\cup I}) := \underline{P}_{s\rfloor x_{P(s)}}(f).$$

For all $s \in G$ and $I \subseteq N(s)$, the corresponding conditional lower prevision $\underline{P}_{\{s\}\cup P(s)\cup I}(\cdot|X_{P(s)\cup I})$ is then given, for all $f \in \mathscr{G}(\mathscr{X}_{\{s\}\cup P(s)\cup I})$ and $x_{P(s)\cup I} \in \mathscr{X}_{P(s)\cup I}$, by

$$\underline{P}_{\{s\}\cup P(s)\cup I}(f|x_{P(s)\cup I}) := \underline{P}_{s\rfloor x_{P(s)}}(f(\cdot,x_{P(s)\cup I})).$$

We will denote the set consisting of all these conditional lower previsions as $\mathscr{I}(\underline{P}_{s\rfloor x_{P(s)}}, s \in G, x_{P(s)} \in \mathscr{X}_{P(s)})$.

The global model $\underline{E}_G^{\mathrm{irr}}$ is now defined as the smallest coherent lower prevision on $\mathscr{G}(\mathscr{X}_G)$ that is (strongly) coherent with this set of conditional lower previsions. We will refer to it as the *irrelevant natural extension* of the local models $\underline{P}_{s\rfloor x_{P(s)}}$. We will not get into the details of what strong coherence means, but one can very roughly think of it as requiring that the conditional lower previsions in the set $\mathscr{I}(\underline{P}_{s\rfloor x_{P(s)}}, s\in G, x_{P(s)}\in\mathscr{X}_{P(s)})$ (i) are compatible with one another and (ii) can be obtained by conditioning the global model $\underline{E}_G^{\mathrm{irr}}$; see Ref. [5, Section 2.4] for more details on strong coherence.

We know from Walley's Finite Extension Theorem [11, Theorem 8.1.9] that if $\underline{E}_G^{\mathrm{irr}}$ exists, then it is equal to the *natural extension* of the collection $\mathscr{I}(\underline{P}_{s\rfloor x_{P(s)}}, s\in G, x_{P(s)}\in\mathscr{X}_{P(s)})$ to an unconditional lower prevision on $\mathscr{G}(\mathscr{X}_G)$. In that case, by applying a derivation that is similar to the one for [5, Eq.(10), Section 5.2], we find for all $f\in\mathscr{G}(\mathscr{X}_G)$ that

$$\underline{E}_G^{\mathrm{irr}}(f) = \sup_{\substack{g_{\{s\}\cup P(s)\cup I}\\ \in\mathscr{G}(\mathscr{X}_{\{s\}\cup P(s)\cup I})}} \Bigg\{ \min_{z_G\in\mathscr{X}_G} \Bigg[f(z_G) - \sum_{s\in G, I\subseteq N(s)} \Big[g_{\{s\}\cup P(s)\cup I}(z_s, z_{P(s)\cup I}) - \underline{P}_{s\rfloor z_{P(s)}}\big(g_{\{s\}\cup P(s)\cup I}(\cdot, z_{P(s)\cup I})\big)\Big]\Bigg]\Bigg\}. \tag{4}$$

7.3 Connections with our approach

For every $s\in G$ and $x_{P(s)}\in\mathscr{X}_{P(s)}$, the local coherent set of desirable gambles $\mathscr{D}_{s\rfloor x_{P(s)}}$ uniquely defines a corresponding coherent lower prevision $\underline{P}_{s\rfloor x_{P(s)}}$. For all $f\in\mathscr{G}(\mathscr{X}_s)$

$$\underline{P}_{s\rfloor x_{P(s)}}(f) := \sup\{\mu\in\mathbb{R}\colon f-\mu\in\mathscr{D}_{s\rfloor x_{P(s)}}\}. \tag{5}$$

Conversely, every local coherent lower prevision $\underline{P}_{s\rfloor x_{P(s)}}$ has at least one coherent set of desirable gambles $\mathscr{D}_{s\rfloor x_{P(s)}}$ from which it can be derived by Eq. (5). These sets are however not unique since coherent sets of desirable gambles are generally more expressive than coherent lower previsions. Using any such family of corresponding local sets of desirable gambles, we can then apply Eq. (3) to obtain their irrelevant natural extension $\mathscr{D}_G^{\mathrm{irr}}$. This joint set also has a corresponding coherent lower prevision. It is denoted as $\underline{P}_G^{\mathrm{irr}}$ and given for all $f\in\mathscr{G}(\mathscr{X}_G)$ by

$$\underline{P}_G^{\mathrm{irr}}(f) := \sup\{\mu\in\mathbb{R}\colon f-\mu\in\mathscr{D}_G^{\mathrm{irr}}\}. \tag{6}$$

The coherent lower prevision $\underline{P}_G^{\mathrm{irr}}$ that is constructed in this way from given local models $\underline{P}_{s\rfloor x_{P(s)}}$ might depend on the particular choice for the sets $\mathscr{D}_{s\rfloor x_{P(s)}}$ in its construction. We will show in Theorem 17 that such is not the case, however.

Proposition 16. *Choose, for all $s\in G$ and $x_{P(s)}\in\mathscr{X}_{P(s)}$, any coherent local set of desirable gambles $\mathscr{D}_{s\rfloor x_{P(s)}}$ on $\mathscr{X}_s$ such that the given local coherent lower prevision $\underline{P}_{s\rfloor x_{P(s)}}$ satisfies Eq.* (5). *Construct the irrelevant natural extension $\mathscr{D}_G^{\mathrm{irr}}$ by applying Eq.* (3) *and let $\underline{P}_G^{\mathrm{irr}}$ be the coherent lower prevision on $\mathscr{G}(\mathscr{X}_G)$ as given by Eq.* (6). *Then $\underline{P}_G^{\mathrm{irr}}$ is strongly coherent with $\mathscr{I}(\underline{P}_{s\rfloor x_{P(s)}}, s\in G, x_{P(s)}\in\mathscr{X}_{P(s)})$.*

Proposition 16 shows that it is possible to construct at least one coherent lower prevision $\underline{P}_G^{\mathrm{irr}}$ on $\mathscr{G}(\mathscr{X}_G)$ that is strongly coherent with $\mathscr{I}(\underline{P}_{s\rfloor x_{P(s)}}, s\in G, x_{P(s)}\in\mathscr{X}_{P(s)})$, implying that the irrelevant natural extension $\underline{E}_G^{\mathrm{irr}}$ is always well defined and given by Eq. (4).

The following result now establishes the final connection between the irrelevant natural extensions $\mathscr{D}_G^{\mathrm{irr}}$ and $\underline{E}_G^{\mathrm{irr}}$ that were outlined in this paper. We show that $\underline{P}_G^{\mathrm{irr}}$ is always equal to the irrelevant natural extension $\underline{E}_G^{\mathrm{irr}}$, regardless of the local sets $\mathscr{D}_{s\rfloor x_{P(s)}}$ that are chosen to construct it.

Theorem 17. *Let $\mathscr{D}_G^{\mathrm{irr}}$ be the irrelevant natural extension of local coherent sets of desirable gambles $\mathscr{D}_{s\rfloor x_{P(s)}}$, $s\in G$ and $x_{P(s)}\in\mathscr{X}_{P(s)}$, as given by Eq.* (3). *Construct local coherent lower previsions $\underline{P}_{s\rfloor x_{P(s)}}$ by applying Eq.* (5) *and let $\underline{E}_G^{\mathrm{irr}}$ be their irrelevant natural extension, as given by Eq.* (4). *It then holds for all $f\in\mathscr{G}(\mathscr{X}_G)$ that*

$$\underline{E}_G^{\mathrm{irr}}(f) = \sup\{\mu\in\mathbb{R}\colon f-\mu\in\mathscr{D}_G^{\mathrm{irr}}\} = \underline{P}_G^{\mathrm{irr}}(f).$$

We believe that this connection between the two approaches can be used to translate at least some of our results for sets of desirable gambles into the language of coherent lower previsions. We intend to explore this further in future work.

8 Summary and conclusions

This paper has developed the notion of a credal network under epistemic irrelevance using sets of desirable gambles. We have proven that the resulting irrelevant natural extension of a network has a number of interesting properties. It marginalises in an intuitive way and satisfies all graphoid properties except symmetry. Finally, we have established a connection with an approach to credal networks under epistemic irrelevance that uses coherent lower previsions.

Future goals that we intend to pursue are to derive counterparts to the marginalisation and graphoid properties in this paper, expressed in terms of coherent lower previsions rather than sets of desirable gambles. By exploiting these properties, we would like to develop algorithms for credal networks under epistemic irrelevance that are able to perform inferences in an efficient manner.

Acknowledgements

Jasper De Bock is a Ph.D. Fellow of the Fund for Scientific Research – Flanders (FWO) and wishes to acknowledge the financial support of the FWO. The authors also wish to thank three anonymous referees for their helpful comments.

References

[1] I. Couso, S. Moral: Sets of desirable gambles: conditioning, representation, and precise probabilities. *International Journal of Approximate Reasoning*, 52(7):1034–1055, 2011.

[2] F. G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[3] F. G. Cozman. Graphical Models for Imprecise Probabilities. *International Journal of Approximate Reasoning*, 39(2–3):167–184, 2005.

[4] F. G. Cozman and P. Walley. Graphoid properties of epistemic irrelevance and independence. *Annals of Mathematics and Artificial Intelligence*, 45(1–2):173–195, 2005.

[5] G. de Cooman, E. Miranda, M. Zaffalon. Independent natural extension. *Artificial Intelligence*, 175(12–13):1911-1950, 2011.

[6] G. de Cooman, E. Miranda: Irrelevant and Independent Natural Extension for Sets of Desirable Gambles. *Journal of Artificial Intelligence Research*, 45:601–640, 2012.

[7] G. de Cooman, E. Quaeghebeur: Exchangeability and sets of desirable gambles. *International Journal of Approximate Reasoning*, 53(3):363–395, 2012. Special issue in honour of Henry E. Kyburg, Jr.

[8] S. Moral: Epistemic irrelevance on sets of desirable gambles. *Annals of Mathematics and Artificial Intelligence* 45:197–214, 2005.

[9] J. Pearl: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[10] B. Vantaggi: The L-separation criterion for description of cs-independence models. *International Journal of Approximate Reasoning*, 29:291–316, 2002.

[11] P. Walley: *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[12] P. Walley: Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24:125–148, 2000.

Allowing for probability zero in credal networks under epistemic irrelevance

Jasper De Bock
Ghent University, Belgium
jasper.debock@ugent.be

Gert de Cooman
Ghent University, Belgium
gert.decooman@ugent.be

Abstract

We generalise Cozman's concept of a credal network under epistemic irrelevance [2, Section 8.3] to the case where lower (and upper) probabilities are allowed to be zero. Our main definition is expressed in terms of coherent lower previsions and imposes epistemic irrelevance by means of strong coherence rather than element-wise Bayes's rule. We also present a number of alternative representations for the resulting joint model, both in terms of lower previsions and credal sets, a notable example being an intuitive characterisation of the joint credal set by means of linear constraints. We end by applying our method to a simple case: the independent natural extension for two binary variables. This allows us to, for the first time, find analytical expressions for the extreme points of this special type of independent product.

Keywords. Credal networks, epistemic irrelevance, lower previsions, credal sets, coherence, irrelevant natural extension, independent natural extension.

1 Introduction

Standard Bayesian networks can be generalised to allow for imprecise probability assessments in a multitude of ways; see Ref. [3, Section 3] for an overview. One way to do so is by means of a credal network under epistemic irrelevance. It differs from standard Bayesian networks in two ways: beliefs are modelled by means of closed convex sets of probability measures (so-called *credal sets*) rather than single probability measures, and the non-parent non-descendants of a variable are *epistemically irrelevant* to that variable given its parents, rather than independent of it.

Credal networks under epistemic irrelevance were introduced by Cozman in Ref. [2, Section 8.3]. In order to impose the assessment of epistemic irrelevance, he assumed that all conditioning events have strictly positive lower probability. Under this assumption, a credal set can be conditioned by applying Bayes's rule to each of its probability measures. However, we feel this assumption to be rather restrictive since an event with zero lower probability may have strictly positive upper probability. Therefore, in the present paper, we get rid of this positivity assumption. We do so by using coherent lower previsions as an alternative, equivalent representation for credal sets and using the concept of (strong) coherence to impose epistemic irrelevance assessments, even when the conditioning events have lower or upper probability zero. See Ref. [8] for an earlier successful application of this method to the special case of credal trees.

The graphical structure of a credal network is a directed acyclic graph, of which we recall some basic definitions in Section 2. Section 3 goes on to introduce some basic terminology regarding the variables in the network and we explain in Section 4 how to model a subject's beliefs regarding the values of these variables by means of coherent lower previsions. Section 5 introduces the notion of a credal network under epistemic irrelevance. We first recall how it is defined under the positivity assumption, then provide a definition that does not need that assumption, and prove a number of useful properties and alternative characterisations. We explain how to describe the joint model by means of a set of linear constraints in Section 6, and reformulate this approach in Section 7 for the special case of the so-called independent natural extension. Finally, in Section 8, we apply our method to the independent natural extension of two binary variables and use it to, for the first time, obtain analytical expressions for the extreme points of this extension.

2 Directed acyclic graphs

A directed acyclic graph (DAG) is a graphical model that is well known for its use in Bayesian networks. It consists of a finite set of nodes (vertices), which are joined together into a network by a set of directed edges, each edge connecting one node with another. Since this directed graph is assumed to be acyclic, it is not possible to follow a sequence of directed edges from node to node and end up back at the same node you started out from.

We denote the set of nodes associated with a given DAG by G. For two nodes s and t in G, if there is a directed edge from s to t, we denote this as $s \to t$ and say that s is a *parent* of t and t is a *child* of s. A single node can have multiple parents and multiple children. For any node s, its set of parents is denoted by $P(s)$ and its set of children by $C(s)$. If a node s has no parents, $P(s) = \emptyset$, and we call s a *root node*. If $C(s) = \emptyset$, then we call s a *leaf*, or *terminal node*.

Two nodes s and t, are said to have a *directed path* between them if one can start from s, follow the edges of the DAG taking their direction into account, and end up in t. In other words: one can find a sequence of nodes $s = s_1, \dots, s_n = t$, $n \geq 1$, in G such that it holds for all $i \in \{1, \dots, n-1\}$ that $s_i \to s_{i+1}$. In that case we also say that s *precedes* t and write $s \sqsubseteq t$. If $s \sqsubseteq t$ and $s \neq t$, we say that s strictly precedes t and write $s \sqsubset t$. For any node s, we denote its set of *descendants* by $D(s) := \{t \in G : s \sqsubset t\}$, its set of *ascendants* by $A(s) := \{t \in G : t \sqsubset s\}$ and its set of *non-parent non-descendants* by $N(s) := G \setminus (P(s) \cup \{s\} \cup D(s))$.

3 Variables and gambles on them

With each node s in G, we associate a variable X_s taking values in some non-empty finite set $\mathscr{X}_s$. Generic elements of this set are denoted by x_s or z_s. A real-valued function on $\mathscr{X}_s$ is called a gamble and we use $\mathscr{G}(\mathscr{X}_s)$ to denote the set of all of them. Generic gambles are denoted by f, g or γ. As a special kind of gambles we consider *indicators* $\mathbb{I}_A$ of events $A \subseteq \mathscr{X}_s$. $\mathbb{I}_A$ is equal to 1 if the event A occurs (the variable X_s assumes a value in A) and zero otherwise.

We extend this notation to more complicated situations as follows. For any subset S of G, we denote by X_S the tuple of variables (with one component X_s for each $s \in S$) that takes values in the Cartesian product $\mathscr{X}_S := \times_{s \in S} \mathscr{X}_s$. We assume logical independence, meaning that X_S may assume *all* values in $\mathscr{X}_S$. Generic elements of the finite set $\mathscr{X}_S$ are denoted by x_S or z_S. Also, if we mention a tuple x_S, then for any $s \in S$, the corresponding element in the tuple will be denoted by x_s. The set $\mathscr{G}(\mathscr{X}_S)$ contains all gambles on $\mathscr{X}_S$ and $\mathbb{I}_A$ is again used to denote the indicator of an event $A \subseteq \mathscr{X}_S$.

We will frequently use the simplifying device of identifying a gamble f_S on $\mathscr{X}_S$ with its *cylindrical extension* to $\mathscr{X}_U$, where $S \subseteq U \subseteq G$. This is the gamble f_U on $\mathscr{X}_U$ defined by $f_U(x_U) := f_S(x_S)$ for all $x_U \in \mathscr{X}_U$. To give an example, this device allows us to identify the gambles $\mathbb{I}_{\{x_S\}}$ on $\mathscr{X}_S$ and $\mathbb{I}_{\{x_S\} \times \mathscr{X}_{U \setminus S}}$ on $\mathscr{X}_U$, and therefore also the events $\{x_S\}$ and $\{x_S\} \times \mathscr{X}_{U \setminus S}$.

When $S = \emptyset$, we let $\mathscr{X}_\emptyset := \{x_\emptyset\}$ be a singleton. The corresponding variable $X_\emptyset$ can only take this single value $x_\emptyset$, so there is no uncertainty about it. $\mathscr{G}(\mathscr{X}_\emptyset)$ can then be identified with the set $\mathbb{R}$ of real numbers.

4 Modelling beliefs about the network

For two disjoint subsets O and I of G and any $x_I \in \mathscr{X}_I$ we consider two equivalent methods of modelling a subject's beliefs about the value that X_O will assume in $\mathscr{X}_O$, given the observation that $X_I = x_I$.

The first approach is to use a *credal set* $K(X_O|x_I)$, defined as a closed and convex subset of the so-called $\mathscr{X}_O$-simplex $\Sigma_{\mathscr{X}_O}$, which is the set containing all probability mass functions on $\mathscr{X}_O$. A generic element of $K(X_O|x_I)$ is denoted by $p(X_O|x_I)$. It is a probability mass function on $\mathscr{X}_O$ conditional on the observation that $X_I = x_I$

The second approach is to use a *coherent lower prevision* $\underline{P}_O(\cdot|x_I)$, defined as a real-valued functional on $\mathscr{G}(\mathscr{X}_O)$ that satisfies the following three conditions: for all $f, g \in \mathscr{G}(\mathscr{X}_O)$ and all real $\lambda \geq 0$

C1. $\underline{P}_O(f|x_I) \geq \min f$,

C2. $\underline{P}_O(\lambda f|x_I) = \lambda \underline{P}_O(f|x_I)$,

C3. $\underline{P}_O(f + g|x_I) \geq \underline{P}_O(f|x_I) + \underline{P}_O(g|x_I)$.

The conjugate of $\underline{P}_O(\cdot|x_I)$ is called a coherent upper prevision. It is denoted by $\overline{P}_O(\cdot|x_I)$ and defined for all $f \in \mathscr{G}(\mathscr{X}_O)$ by $\overline{P}_O(f|x_I) := -\underline{P}_O(-f|x_I)$. We will focus on coherent lower previsions, but it is useful to keep in mind that all our results can be reformulated in terms of coherent upper previsions by applying this conjugacy property.

Both approaches are equivalent because there is a one-to-one correspondence between them [12, Section 3.3.3]. If we denote by $P_O(\cdot|x_I)$ the expectation operator on $\mathscr{G}(\mathscr{X}_O)$ that corresponds to a probability mass function $p(X_O|x_I)$, then a credal set $K(X_O|x_I)$ defines a unique coherent lower prevision $\underline{P}_O(\cdot|x_I)$ in the following way. For all $f \in \mathscr{G}(\mathscr{X}_O)$:

$$\underline{P}_O(f|x_I) := \min\{P_O(f|x_I) : p(X_O|x_I) \in K(X_O|x_I)\}.$$

Its conjugate coherent upper prevision $\overline{P}_O(\cdot|x_I)$ is given for all $f \in \mathscr{G}(\mathscr{X}_O)$ by

$$\overline{P}_O(f|x_I) := \max\{P_O(f|x_I) : p(X_O|x_I) \in K(X_O|x_I)\}.$$

Conversely, the unique credal set $K(X_O|x_I)$ that corresponds to a coherent lower prevision $\underline{P}_O(\cdot|x_I)$ is given by

$$K(X_O|x_I) := \{p(X_O|x_I) \in \Sigma_{\mathscr{X}_O} : \\ (\forall f \in \mathscr{G}(\mathscr{X}_O)) P_O(f|x_I) \geq \underline{P}_O(f|x_I)\}. \quad (1)$$

If $I = \emptyset$, then $X_I = X_\emptyset$ assumes its only possible value $x_\emptyset$ with certainty, so conditioning on $X_\emptyset = x_\emptyset$ amounts to not conditioning at all. We reflect this in our notation by using $K(X_O)$ and $\underline{P}_O$ as alternative notations for $K(X_O|x_\emptyset)$ and $\underline{P}_O(\cdot|x_\emptyset)$ respectively. A notable example is $I = \emptyset$ and $O = G$, for which we obtain a credal set $K(X_G)$ and coherent lower prevision $\underline{P}_G$ that can be used to model a subject's beliefs about the value that the joint variable X_G will assume in $\mathscr{X}_G$.

When given for all $x_I \in \mathscr{X}_I$, a coherent lower prevision $\underline{P}_O(\cdot|x_I)$ on $\mathscr{G}(\mathscr{X}_O)$, this defines a unique corresponding *coherent conditional lower prevision* $\underline{P}_{O\cup I}(\cdot|X_I)$. It is a special two-place function that is defined, for all $f \in \mathscr{G}(\mathscr{X}_{O\cup I})$ and all $x_I \in \mathscr{X}_I$, by $\underline{P}_{O\cup I}(f|x_I) := \underline{P}_O(f(\cdot,x_I)|x_I)$.

5 Irrelevant natural extension

We will now show how to construct a joint model for the variables in the network in the form of a credal set $K(X_G)$, or equivalently, a coherent lower prevision $\underline{P}_G$.

5.1 Local uncertainty models

We start by adding *local uncertainty models* to each of the nodes $s \in G$. These local models are assumed to be given beforehand and will be used as basic building blocks to construct the joint model.

If s is not a root node of the network, i.e. has a non-empty set of parents $P(s)$, then we have a conditional local model for every instantiation of its parents: for each $x_{P(s)} \in \mathscr{X}_{P(s)}$, we have a credal set $K(X_s|x_{P(s)})$ and a corresponding coherent lower prevision $\underline{P}_s(\cdot|x_{P(s)})$. They represent our subject's beliefs about the variable X_s conditional on the information that its parent variables $X_{P(s)}$ assume the value $x_{P(s)}$.

If s is a root node, i.e. has no parents, then our subject's local beliefs about the variable X_s are represented by an unconditional local model. We are given a credal set $K(X_s)$ and a corresponding coherent lower prevision $\underline{P}_s$. As explained in Section 4, we can also use the common generic notations $K(X_s|x_{P(s)})$ and $\underline{P}_s(\cdot|x_{P(s)})$ in this unconditional case, since for a root node s, its set of parents $P(s)$ is empty.

In order to turn these local uncertainty models into a joint model, we introduce the important concept of epistemic irrelevance.

5.2 Epistemic irrelevance

We discuss conditional epistemic irrelevance, as the unconditional version can easily be recovered as a special case.

Consider three disjoint subsets C, I, and O of G. When a subject judges X_I to be *epistemically irrelevant to X_O conditional on X_C*, he assumes that if he knew the value of X_C, then learning in addition which value X_I assumes in $\mathscr{X}_I$ would not affect his beliefs about X_O. More formally put, he assumes for all $x_C \in \mathscr{X}_C$ and $x_I \in \mathscr{X}_I$ that

$$K(X_O|x_{C\cup I}) = K(X_O|x_C) \text{ and } \underline{P}_O(\cdot|x_{C\cup I}) = \underline{P}_O(\cdot|x_C).$$

It should be clear that it suffices for the unconditional case, in the discussion above, to let $C = \emptyset$. This makes sure the variable X_C has only one possible value, so conditioning on that variable amounts to not conditioning at all.

Using this concept of epistemic irrelevance, we can provide the graphical structure of the network with an interpretation.

5.3 Interpretation of the graphical model

In Bayesian networks, the graphical structure is taken to represent the following assessments: for any node s, the associated variable is independent of its non-parent non-descendant variables, given its parent variables.

When generalising this interpretation to imprecise graphical networks, the classical notion of independence gets replaced by a more general, imprecise-probabilistic notion of independence. In this paper, we choose to use epistemic irrelevance. We provide the graphical structure of the network with the following interpretation: for any node s and all subsets I of its non-parent non-descendants $N(s)$, the variable X_I is judged to be epistemically irrelevant to X_s conditional on $X_{P(s)}$.

More formally put, we assume for all $s \in G$, $I \subseteq N(s)$ and $x_{P(s)\cup I} \in \mathscr{X}_{P(s)\cup I}$ that

$$K(X_s|x_{P(s)\cup I}) := K(X_s|x_{P(s)}) \text{ and } \underline{P}_s(\cdot|x_{P(s)\cup I}) := \underline{P}_s(\cdot|x_{P(s)}).$$

5.4 Non-zero lower probabilities

Together with the local uncertainty models, the irrelevance assessments that are encoded in the network provide us with a number of belief models about the variables in the network: for all $s \in G$, $I \subseteq N(s)$ and $x_{P(s)\cup I} \in \mathscr{X}_{P(s)\cup I}$, we are given a credal set $K(X_s|x_{P(s)\cup I})$, or equivalently, a coherent lower prevision $\underline{P}_s(\cdot|x_{P(s)\cup I})$. In order to arrive at a joint model, we need to provide a method of translating these belief models into constraints on the joint.

An approach that is often used when dealing with assessments of epistemic irrelevance [6, 2], is to assume that all lower probabilities are strictly positive, or equivalently, that for every probability mass function $p(X_G)$ in the joint credal set $K(X_G)$, all events have strictly positive probability. For all $s \in G$, $I \subseteq N(s)$ and $x_{P(s)\cup I} \in \mathscr{X}_{P(s)\cup I}$, this assumption allows us to apply Bayes's rule to every $p(X_G)$ in $K(X_G)$, resulting in a set of conditional probability mass functions $p(X_s|x_{P(s)\cup I})$. This procedure is called applying *element-wise Bayes's rule*. One can now impose that, for all $s \in G$, $I \subseteq N(s)$ and $x_{P(s)\cup I} \in \mathscr{X}_{P(s)\cup I}$, the set of conditional probability mass functions that is obtained in this way must be equal to the given model $K(X_s|x_{P(s)\cup I})$. Any joint credal set $K(X_G)$ that satisfies these constraints is called an *irrelevant product* of the local models.

One particular credal set that was proven to be an irrelevant product in Ref. [2]—under the positivity assumption mentioned above—is the so-called *strong extension* of the network. Its credal set $K^{\text{str}}(X_G)$ is the convex hull of the set $\mathscr{P}$, which contains all joint probability mass functions $p(X_G)$ that, for all $x_G \in \mathscr{X}_G$, satisfy

$$p(x_G) = \prod_{s \in G} p(x_s | x_{P(s)}),$$

where each $p(X_s|x_{P(s)})$ is selected from the local credal set $K(X_s|x_{P(s)})$. The corresponding coherent lower prevision $\underline{P}_G^{\mathrm{str}}$ is given for all $f \in \mathscr{G}(\mathscr{X}_G)$ by

$$\underline{P}_G^{\mathrm{str}}(f) = \min\{P_G(f) : p(X_G) \in \mathscr{P}\}.$$

The strong extension is not the only irrelevant product of the local models. Although it has the advantage of having an intuitive similarity to standard Bayesian networks, it is somewhat arbitrary in that it satisfies more constraints than those needed to be called an irrelevant product. We prefer to use a least committal strategy: to only satisfy those constraints that are imposed by the network, and no others. The resulting model is the largest of all credal sets that are an irrelevant product. We call it the *irrelevant natural extension* of the network an denote it by $K^{\mathrm{irr}}(X_G)$.

This irrelevant natural extension was introduced by Cozman in Ref. [2], but only under the assumption that all lower probabilities are strictly positive. We feel this assumption to be rather restrictive since an event with zero lower probability may occur with a strictly positive upper probability. The first contribution of this paper will therefore be to extend Cozman's definition of the irrelevant natural extension such that it allows for lower (and upper) probabilities to be zero.

5.5 Getting rid of the positivity assumption

If the conditioning event has lower probability zero, the credal set $K(X_s|x_{P(s)\cup I})$ can no longer be uniquely related to the joint model $K(X_G)$ through element-wise Bayes's rule. Therefore, we have to impose our assessments of epistemic irrelevance in some other way. Here, we choose to do so by means of strong coherence, defining the irrelevant natural extension in terms conditional lower previsions, rather than their corresponding credal sets.

As mentioned in the beginning of Section 5.4, the irrelevance assessments, together with the local uncertainty models, provide us with a number of coherent lower previsions: for all $s \in G, I \subseteq N(s)$ and $x_{P(s)\cup I} \in \mathscr{X}_{P(s)\cup I}$ we are given a coherent lower prevision $\underline{P}_s(\cdot|x_{P(s)\cup I}) := \underline{P}_s(\cdot|x_{P(s)})$ on $\mathscr{G}(\mathscr{X}_s)$. As was explained in Section 4, this provides us with a number of coherent *conditional* lower previsions: for all $s \in G$ and $I \subseteq N(s)$, we have a coherent conditional lower prevision $\underline{P}_{\{s\}\cup P(s)\cup I}(\cdot|X_{P(s)\cup I})$, defined for all $f \in \mathscr{G}(\mathscr{X}_{\{s\}\cup P(s)\cup I})$ and $x_{P(s)\cup I} \in \mathscr{X}_{P(s)\cup I}$ by

$$\underline{P}_{\{s\}\cup P(s)\cup I}(f|x_{P(s)\cup I}) := \underline{P}_s(f(\cdot, x_{P(s)\cup I})|x_{P(s)}).$$

We will denote the set consisting of all these conditional lower previsions as $\mathscr{I}(\underline{P}_{\{s\}\cup P(s)}(\cdot|X_{P(s)}), s \in G)$.

In order to turn these coherent conditional lower previsions into constraints on a joint model, given in the form of a coherent lower prevision $\underline{P}_G$ on $\mathscr{G}(\mathscr{X}_G)$, we use the concept of *(strong) coherence* [12, Section 7.1.4]: we require $\underline{P}_G$ to be strongly coherent with the family $\mathscr{I}(\underline{P}_{\{s\}\cup P(s)}(\cdot|X_{P(s)}), s \in G)$ of coherent conditional lower previsions. Any $\underline{P}_G$ that satisfies this property, is called an *irrelevant product*. The least committal—pointwise smallest— irrelevant product is called the *irrelevant natural extension* of the network and will be denoted by $\underline{P}_G^{\mathrm{irr}}$.

As strong coherence is a rather involved requirement, we will not get into the details of what it means. For our present purposes, it suffices to think of it as a generalisation of the element-wise Bayes's rule approach that was explained in Section 5.4. For the interested reader: Ref. [12, Section 7.1.4] provides a general definition and a behavioural interpretation in terms of supremum buying prices, turning strong coherence into a rationality requirement.

We would like to stress that strong coherence is a consistency criterion, rather than a conditioning rule.[1] In fact, it is compatible with a number of fundamentally different conditioning rules, all of which reduce to element-wise Bayes's rule if the conditioning event has positive lower probability. Also, strong coherence regards conditional models as fundamental, rather than deriving them from unconditional ones. In that respect, it shares fundamental ideas with the well-known concept of full conditional measures. See Ref. [1] for a similar, coherence-based approach to stochastic independence, which has been applied to credal networks in Ref. [11].

When it comes to strong coherence, the so-called Reduction Theorem [12, Theorem 7.1.5] is a very useful result; see also Ref. [9, Theorem 2]. It implies that the unconditional coherent lower prevision $\underline{P}_G$ is strongly coherent with the family $\mathscr{I}(\underline{P}_{\{s\}\cup P(s)}(\cdot|X_{P(s)}), s \in G)$ of conditional ones—is an irrelevant product—, if and only if (i) the family $\mathscr{I}(\underline{P}_{\{s\}\cup P(s)}(\cdot|X_{P(s)}), s \in G)$ is strongly coherent on its own and (ii) $\underline{P}_G$ is weakly coherent [12, Section 7.1.4] with $\mathscr{I}(\underline{P}_{\{s\}\cup P(s)}(\cdot|X_{P(s)}), s \in G)$.

Using an approach that uses so-called sets of desirable gambles rather than coherent lower previsions, it is relatively easy to show that requirement (i) is always satisfied [5, Proposition 16].

Proposition 1. *Consider arbitrary coherent lower previsions $\underline{P}_s(\cdot|x_{P(s)})$ on $\mathscr{G}(\mathscr{X}_s)$, $s \in G$ and $x_{P(s)} \in \mathscr{X}_{P(s)}$. Then the family $\mathscr{I}(\underline{P}_{\{s\}\cup P(s)}(\cdot|X_{P(s)}), s \in G)$ is strongly coherent.*

It follows that $\underline{P}_G$ is an irrelevant product if and only if it

[1] Refs. [7, Definition 12] and [4, Section 3.2.4] provide definitions for epistemic irrelevance that are based on a conditioning rule that is similar to Walley's notion of regular extension [12, Appendix J]. These definitions are applicable in the presence of zero lower probabilities as well. It is not clear to us whether they can be used to construct a joint model from conditional ones, as is done in the current paper.

is weakly coherent with $\mathscr{I}(\underline{P}_{\{s\}\cup P(s)}(\cdot|X_{P(s)}), s \in G)$. In its original form [12, Section 7.1.4], weak coherence is still rather involved, but due to Ref. [9, Theorem 1], it can be reformulated in a very elegant manner that leads directly to the following characterisation of an irrelevant product.

Corollary 2. *A coherent lower prevision $\underline{P}_G$ on $\mathscr{G}(\mathscr{X}_G)$ is strongly coherent with $\mathscr{I}(\underline{P}_{\{s\}\cup P(s)}(\cdot|X_{P(s)}), s \in G)$—is an irrelevant product—if and only if for all $s \in G$, $I \subseteq N(s)$, $x_{P(s)\cup I} \in \mathscr{X}_{P(s)\cup I}$ and $g \in \mathscr{G}(\mathscr{X}_s)$:*

$$\underline{P}_G(\mathbb{I}_{x_{P(s)\cup I}}[g - \underline{P}_s(g|x_{P(s)})]) = 0.$$

The condition imposed in this result is called the *Generalised Bayes's Rule* (GBR), and reduces to element-wise Bayes's rule when all conditioning events have strictly positive lower probabilities [12, Theorem 6.4.2]. It should therefore be clear that the definition of an irrelevant product, as it was given in Section 5.4 under the assumption of strictly positive lower probabilities, is a special case of the definition given in the current section.

Proposition 3. *The strong extension is an irrelevant product: the coherent lower prevision $\underline{P}_G^{\mathrm{str}}$ is strongly coherent with $\mathscr{I}(\underline{P}_{\{s\}\cup P(s)}(\cdot|X_{P(s)}), s \in G)$.*

This result guarantees the existence of at least one irrelevant product, making the irrelevant natural extension well defined: since strong coherence is preserved under taking lower envelopes [12, Section 7.1.6], the irrelevant natural extension is the lower envelope of all irrelevant products, implying that it is indeed pointwise dominated by all other irrelevant products. It should be clear that Corollary 2 provides us with an immediate characterisation for this irrelevant natural extension.

Corollary 4. *The irrelevant natural extension of a network is the pointwise smallest coherent lower prevision $\underline{P}_G^{\mathrm{irr}}$ on $\mathscr{G}(\mathscr{X}_G)$ such that for all $s \in G$, $I \subseteq N(s)$, $x_{P(s)\cup I} \in \mathscr{X}_{P(s)\cup I}$ and $g \in \mathscr{G}(\mathscr{X}_s)$:*

$$\underline{P}_G(\mathbb{I}_{x_{P(s)\cup I}}[g - \underline{P}_s(g|x_{P(s)})]) = 0.$$

Similar to what has been shown in Ref. [2, Lemma 13]—under the positivity assumption—most of the constraints in Corollary 4 turn out to be redundant. We find that we only need to impose those constraints for which $I = N(s)$.

Theorem 5. *The irrelevant natural extension of a network is the pointwise smallest coherent lower prevision $\underline{P}_G^{\mathrm{irr}}$ on $\mathscr{G}(\mathscr{X}_G)$ such that for all $s \in G$, $x_{P(s)\cup N(s)} \in \mathscr{X}_{P(s)\cup N(s)}$ and $g \in \mathscr{G}(\mathscr{X}_s)$:*

$$\underline{P}_G(\mathbb{I}_{x_{P(s)\cup N(s)}}[g - \underline{P}_s(g|x_{P(s)})]) = 0.$$

Although we have defined the irrelevant natural extension in terms of coherent (conditional) lower previsions—since strong coherence is not particularly well-suited for a formulation in terms of credal sets—, it is valid for credal sets as well. Due to the correspondence between credal sets and coherent lower previsions, it suffices to consider the credal set that corresponds to the irrelevant natural extension $\underline{P}_G^{\mathrm{irr}}$. We denote it by $K^{\mathrm{irr}}(X_G)$ and will also refer to it as the irrelevant natural extension of the network. Using Eq. (1), we find that

$$\begin{aligned} K^{\mathrm{irr}}(X_G) = \{p(X_G) \in \Sigma_{\mathscr{X}_G} :& \\ &(\forall f \in \mathscr{G}(\mathscr{X}_G)) P_G(f) \geq \underline{P}_G^{\mathrm{irr}}(f)\}. \end{aligned}$$

The following result provides an intuitive characterisation.

Theorem 6. *A probability mass function $p(X_G) \in \Sigma_{\mathscr{X}_G}$ belongs to $K^{\mathrm{irr}}(X_G)$ if and only if for all $s \in G$ and $x_{P(s)\cup N(s)} \in \mathscr{X}_{P(s)\cup N(s)}$ there are a real number $\lambda \geq 0$ and a probability mass function $p(X_s|x_{P(s)}) \in K(X_s|x_{P(s)})$ such that*

$$\sum_{z_{D(s)} \in \mathscr{X}_{D(s)}} p(x_{P(s)\cup N(s)}, X_s, z_{D(s)}) = \lambda p(X_s|x_{P(s)}).$$

5.6 Marginalisation properties

Given a credal network with nodes G and local models $K(X_s|x_{P(s)})$, $s \in G$ and $x_{P(s)} \in \mathscr{X}_{P(s)}$, a *top sub-network* is a network formed by a subset of nodes $S \subseteq G$ such that for all $s \in S$, its ascendants $A(s)$ also belong to S. The underlying graphical structure consists of those edges in the original network that connect nodes in S and the local models $K(X_s|x_{P(s)})$, $s \in S$ and $x_{P(s)} \in \mathscr{X}_{P(s)}$, are taken to be identical to those of the original model. We denote the irrelevant natural extension of such a top sub-network as $K^{\mathrm{irr}}(X_S)$. It turns out to be closely related to the irrelevant natural extension of the original network, a result that was already present in Ref. [2, Theorem 15] under the assumption that all lower probabilities are strictly positive.

Proposition 7. *Consider a credal network with nodes G and a top sub-network with nodes S. Let $K^{\mathrm{irr}}(X_G)$ and $K^{\mathrm{irr}}(X_S)$ be their respective irrelevant natural extensions. Denote by $\mathrm{marg}_S(K^{\mathrm{irr}}(X_G))$ the credal set obtained by element-wise marginalisation to $\mathscr{X}_S$ of the probability mass functions in $K^{\mathrm{irr}}(X_G)$, then*

$$K^{\mathrm{irr}}(X_S) = \mathrm{marg}_S(K^{\mathrm{irr}}(X_G)).$$

We believe that the irrelevant natural extension also satisfies marginalisation properties for sub-networks other than the very specific subclass of top sub-networks, but we defer any formal result to future work. See Ref. [5] to get an idea of what might be possible.

6 A linear programming approach

The goal of the current section is to construct a set of linear constraints that is able to fully characterise the joint credal set $K^{\mathrm{irr}}(X_G)$ of the irrelevant natural extension of a given network.

In order to derive such a representation for the joint model, we start from similar representations for the local models. For all $s \in G$ and $x_{P(s)} \in \mathscr{X}_{P(s)}$, we characterise the local credal set $K(X_s|x_{P(s)})$ as the set of all real-valued functions $p(z_s|x_{P(s)}) \in \mathbb{R}^{\mathscr{X}_s}$ that satisfy the unitary constraint

$$\sum_{z_s \in \mathscr{X}_s} p(z_s|x_{P(s)}) = 1 \tag{2}$$

and a (possibly infinite) set of linear homogeneous inequalities

$$\sum_{z_s \in \mathscr{X}_s} p(z_s|x_{P(s)})\gamma(z_s) \geq 0, \tag{3}$$

where γ takes values in a (possibly infinite) set $\Gamma(s, x_{P(s)})$ of gambles on $\mathscr{X}_s$.

Such a description for $K(X_s|x_{P(s)})$ always exists, as it can be derived from the corresponding coherent lower prevision $\underline{P}_s(\cdot|x_{P(s)})$ by letting

$$\Gamma(s, x_{P(s)}) = \{f - \underline{P}_s(f|x_{P(s)}) : f \in \mathscr{G}(\mathscr{X}_s)\}. \tag{4}$$

Indeed, for this particular choice of $\Gamma(s, x_{P(s)})$, the combination of Eqs. (2) and (3) will always be equivalent with the constraints imposed by Eq. (1), thereby fully characterising $K(X_s|x_{P(s)})$. To understand why this equivalence holds, start by noticing that if $\gamma = f - \underline{P}_s(f|x_{P(s)})$, with $f \in \mathscr{G}(\mathscr{X}_s)$, then due to Eq. (2), Eq. (3) becomes equivalent to

$$\sum_{z_s \in \mathscr{X}_s} p(z_s|x_{P(s)})f(z_s) \geq \underline{P}_s(f|x_{P(s)}). \tag{5}$$

Coherence of $\underline{P}_s(\cdot|x_{P(s)})$ now implies, for all $z_s \in \mathscr{X}_s$, that $\underline{P}_s(\mathbb{I}_{\{z_s\}}|x_{P(s)}) \geq 0$ and therefore, due to Eq. (5), that $p(z_s|x_{P(s)}) \geq 0$. By combining this with Eq. (2), we find that $p(X_s|x_{P(s)}) \in \Sigma_{\mathscr{X}_s}$. This allows us to rewrite the left-hand side of Eq. (5) as $P_s(f|x_{P(s)})$, thereby establishing the equivalence with the constraints imposed by Eq. (1).

Eq. (4) produces an infinite set of constraints that is guaranteed to characterise $K(X_s|x_{P(s)})$, but in practice, most of these constraints will often be redundant. This is especially the case for so-called *finitely generated* local models, for which the corresponding coherent lower prevision $\underline{P}_s(\cdot|x_{P(s)})$ is fully determined by its value in only a finite number of gambles. For such local models, one can easily construct a set $\Gamma(s, x_{P(s)})$ that contains only a finite number of constraints and yet fully characterises $K(X_s|x_{P(s)})$. The credal set of such a finitely generated local model will always be the convex hull of a finite number of probability mass functions. The reason for this equivalence being that a compact convex set can be specified as the intersection of a finite number of closed half spaces if and only if it is the convex hull of a finite number of vertices [10, Theorem 3.1.3].

The importance of these local representations in terms of linear constraints—regardless of whether $\Gamma(s, x_{P(s)})$ is finite or not—is that we can use the local constraints to derive global ones, thereby obtaining the following representation for the irrelevant natural extension of a network.

Proposition 8. *Consider a credal network for which each of the local credal sets $K(X_s|x_{P(s)})$, $s \in G$ and $x_{P(s)} \in \mathscr{X}_{P(s)}$, is fully characterised by means of Eqs.* (2) *and* (3)*. Then $K^{\mathrm{irr}}(X_G)$ consists of those $p(X_G) \in \Sigma_{\mathscr{X}_G}$ for which for all $s \in G$, $x_{P(s)\cup N(s)} \in \mathscr{X}_{P(s)\cup N(s)}$ and $\gamma \in \Gamma(s, x_{P(s)})$:*

$$\sum_{z_s \in \mathscr{X}_s} \sum_{z_{D(s)} \in \mathscr{X}_{D(s)}} p(x_{P(s)\cup N(s)}, z_s, z_{D(s)})\gamma(z_s) \geq 0.$$

When all lower probabilities are strictly positive, this result is fairly straightforward. The global inequalities can then be obtained by imposing all irrelevancies through element-wise Bayes's rule and clearing the denominators, as is done in Ref. [2, Section 8.3]. The importance of our result is that it shows that these inequalities remain valid if lower (and upper) probabilities are allowed to be zero.

Ref. [2] does not explicitly impose $p(X_G) \in \Sigma_{\mathscr{X}_G}$ as a constraint. It seems to assume that it suffices to impose only the unitary constraint $\sum_{z_G \in \mathscr{X}_G} p(z_G) = 1$, making the requirement that $p(z_G) \geq 0$, $z_G \in \mathscr{X}_G$, redundant. Although we agree with this statement, we do not believe it to be trivial and therefore choose to provide it with a proof.

Theorem 9. *Consider a credal network for which each of the local credal sets $K(X_s|x_{P(s)})$, $s \in G$ and $x_{P(s)} \in \mathscr{X}_{P(s)}$, is fully characterised by means of Eqs.* (2) *and* (3)*. Then $K^{\mathrm{irr}}(X_G)$ consists of those real-valued functions $p(X_G) \in \mathbb{R}^{\mathscr{X}_G}$ for which $\sum_{z_G \in \mathscr{X}_G} p(z_G) = 1$ and for all $s \in G$, $x_{P(s)\cup N(s)} \in \mathscr{X}_{P(s)\cup N(s)}$ and $\gamma \in \Gamma(s, x_{P(s)})$:*

$$\sum_{z_s \in \mathscr{X}_s} \sum_{z_{D(s)} \in \mathscr{X}_{D(s)}} p(x_{P(s)\cup N(s)}, z_s, z_{D(s)})\gamma(z_s) \geq 0.$$

Proposition 8 and Theorem 9 are valid for both finite and infinite sets $\Gamma(s, x_{P(s)})$, but in the infinite case, their value is mainly of a theoretical nature. They can only be used in practice—at least in an exact way—if $L(s, x_{P(s)})$ is finite for all $s \in G$ and $x_{P(s)} \in \mathscr{X}_{P(s)}$, or equivalently, if all local credal sets are finitely generated.[2] Indeed, in that case, Proposition 8 and Theorem 9 will provide linear programs with a finite number of constraints. Although the size of these programs is still exponential in the number of variables that define the network, it allows for inference problems in small networks to be solved in an exact manner. Initial ideas on how to reduce this exponential complexity are provided in our conclusions.

7 Independent natural extension

An important special case is obtained when all nodes in the network are unconnected. Every node $s \in G$ is then both

[2] If we allow for non-linear constraints, then local credal sets that are not finitely generated could be practical as well, as they can often be described by means of a finite set of non-linear constraints. We believe that Proposition 8 and Theorem 9 could be adapted easily to allow for such non-linear (homogeneous) constraints, thereby expanding their practical use when combined with non-linear solvers.

a root and a leaf of the network—meaning that $P(s)$ and $C(s)$ are empty—, its non-parent non-descendants are given by $N(s) = G \setminus \{s\}$ and the local model is an unconditional credal set $K(X_s)$, or equivalently, a coherent lower prevision $\underline{P}_s$ on $\mathscr{G}(\mathscr{X}_s)$.

For such a network, the irrelevancies that are encoded by the network are the following. For every $s \in G$ and all $I \subseteq G \setminus \{s\}$, the variable X_I is epistemically irrelevant to X_s, implying that for any two nodes $s,t \in G$, X_s and X_t are mutually epistemically irrelevant and therefore by definition *epistemically independent*. The resulting irrelevant natural extension is called the *many-to-one independent natural extension* and has been treated in full detail in Ref. [9]. That same reference also introduces the so-called *many-to-many independent natural extension*, which requires that for all disjoint subsets O and I of G, X_I is epistemically irrelevant to X_O. The many-to-one and many-to-many independent natural extensions are shown to be equivalent [9, Theorem 23] and we can therefore simply call it the *independent natural extension*. Its coherent lower prevision is denoted by $\otimes_{s\in G}\underline{P}_s$ and its credal set by $\otimes_{s\in G}K(X_s)$. For this special case, Theorem 9 can be reformulated in the following way.

Corollary 10. *Consider a finite number of local credal sets $K(X_s)$, $s \in G$, each of which is fully characterised means of Eqs.* (2) *and* (3). *Then $\otimes_{s\in G}K(X_s)$ consists of those real-valued functions $p(X_G) \in \mathbb{R}^{\mathscr{X}_G}$ for which $\sum_{z_G\in\mathscr{X}_G} p(z_G) = 1$ and for all $s \in G$, $x_{G\setminus\{s\}} \in \mathscr{X}_{G\setminus\{s\}}$ and $\gamma \in \Gamma(s)$:*

$$\sum_{z_s\in\mathscr{X}_s} p(x_{G\setminus\{s\}}, z_s)\gamma(z_s) \geq 0.$$

We leave it to the reader to reformulate some of the other results that were obtained in the two previous sections, taking the simplifications that correspond to the special case of the independent natural extension into account. In fact, Ref. [9, Proposition 14, Corollary 16 and Theorem 20] already provides results that could be regarded as special cases of Proposition 3, Corollary 2 and Proposition 7.

8 Case study of two binary variables

As an example, we apply our results to the very simple case of two unconnected binary variables X_1 and X_2. For all $i \in \{1,2\}$, the variable X_i assumes values in its binary state space $\mathscr{X}_i = \{h_i,t_i\}$ and has a given local uncertainty model in the form of a credal set $K(X_i)$. We set out to construct the independent natural extension $K(X_1)\otimes K(X_2)$ of these two local models. In order to do so, we will describe it by means of linear constraints and then use this characterisation to find analytical expressions for the so-called *extreme points* of $K(X_1)\otimes K(X_2)$, which are those elements of $K(X_1)\otimes K(X_2)$ that cannot be written as a convex combination of the other elements. $K(X_1)\otimes K(X_2)$ is then equal to the convex hull of these extreme points.

For a binary variable X_i, $i \in \{1,2\}$, the credal set $K(X_i)$ is uniquely characterised by the lower and upper probability of h_i, respectively denoted as $\underline{p}(h_i)$ and $\overline{p}(h_i)$. Each of these two probabilities defines a mass function on $\mathscr{X}_i$ and

$$K(X_i) = \left\{p \in \Sigma_{\mathscr{X}_i} : p(h_i) \in [\underline{p}(h_i), \overline{p}(h_i)]\right\}$$

is obtained by taking their convex hull. The corresponding lower and upper probability of t_i is given by $\underline{p}(t_i) := 1 - \overline{p}(h_i)$ and $\overline{p}(t_i) := 1 - \underline{p}(h_i)$.

In order to apply the method described in Section 6, we first need to characterise $K(X_i)$ by means of the unitary constraint and a finite number of linear homogeneous inequalities. In this particular binary case, the following two inequalities suffice:

$$\overline{p}(t_i)p(h_i) - \underline{p}(h_i)p(t_i) \geq 0$$
$$-\underline{p}(t_i)p(h_i) + \overline{p}(h_i)p(t_i) \geq 0.$$

By applying Corollary 10, these local inequalities can be used to obtain eight global inequalities.

$$\overline{p}(t_1)p(h_1,h_2) - \underline{p}(h_1)p(t_1,h_2) \geq 0 \tag{I1}$$
$$-\underline{p}(t_1)p(h_1,h_2) + \overline{p}(h_1)p(t_1,h_2) \geq 0 \tag{I2}$$
$$\overline{p}(t_1)p(h_1,t_2) - \underline{p}(h_1)p(t_1,t_2) \geq 0 \tag{I3}$$
$$-\underline{p}(t_1)p(h_1,t_2) + \overline{p}(h_1)p(t_1,t_2) \geq 0 \tag{I4}$$
$$\overline{p}(t_2)p(h_1,h_2) - \underline{p}(h_2)p(h_1,t_2) \geq 0 \tag{I5}$$
$$-\underline{p}(t_2)p(h_1,h_2) + \overline{p}(h_2)p(h_1,t_2) \geq 0 \tag{I6}$$
$$\overline{p}(t_2)p(t_1,h_2) - \underline{p}(h_2)p(t_1,t_2) \geq 0 \tag{I7}$$
$$-\underline{p}(t_2)p(t_1,h_2) + \overline{p}(h_2)p(t_1,t_2) \geq 0 \tag{I8}$$

Together with the global unitary constraint

$$p(h_1,h_2) + p(h_1,t_2) + p(t_1,h_2) + p(t_1,t_2) = 1,$$

they fully characterise the credal set $K(X_1)\otimes K(X_2)$. If the inequalities in equations (I1)–(I8) are replaced by equalities, we refer to them as (E1)–(E8).

Lemma 11. *Every extreme point of $K(X_1)\otimes K(X_2)$ is the unique solution to the unitary constraint and three of the equations* (E1)–(E8).

The extreme points of the independent natural extension $K(X_1)\otimes K(X_2)$ can therefore be found in the following way. We need to consider every possible subset of three equalities out of (E1)–(E8). For every such combination of three equalities, we need to combine them with the unitary constraint and check whether this results in a unique solution, and if so, whether this unique solution satisfies the inequalities in (I1)–(I8). If so, that unique solution is an extreme point of $K(X_1)\otimes K(X_2)$.

As there are 56 possible ways of choosing three equalities out of eight, one might suspect that this problem cannot be

	$p(h_1,h_2)\Sigma$	$p(h_1,t_2)\Sigma$	$p(t_1,h_2)\Sigma$	$p(t_1,t_2)\Sigma$	Σ
p_{S1}	$\underline{p}(h_1)\underline{p}(h_2)$	$\underline{p}(h_1)\overline{p}(t_2)$	$\overline{p}(t_1)\underline{p}(h_2)$	$\overline{p}(t_1)\overline{p}(t_2)$	1
p_{S2}	$\underline{p}(h_1)\overline{p}(h_2)$	$\underline{p}(h_1)\underline{p}(t_2)$	$\overline{p}(t_1)\overline{p}(h_2)$	$\overline{p}(t_1)\underline{p}(t_2)$	1
p_{S3}	$\overline{p}(h_1)\underline{p}(h_2)$	$\overline{p}(h_1)\overline{p}(t_2)$	$\underline{p}(t_1)\underline{p}(h_2)$	$\underline{p}(t_1)\overline{p}(t_2)$	1
p_{S4}	$\overline{p}(h_1)\overline{p}(h_2)$	$\overline{p}(h_1)\underline{p}(t_2)$	$\underline{p}(t_1)\overline{p}(h_2)$	$\underline{p}(t_1)\underline{p}(t_2)$	1
p_{A1}	$\underline{p}(h_1)\overline{p}(h_1)\underline{p}(h_2)$	$\underline{p}(h_1)\overline{p}(h_1)\overline{p}(t_2)$	$\overline{p}(t_1)\overline{p}(h_1)\underline{p}(h_2)$	$\underline{p}(h_1)\underline{p}(t_1)\overline{p}(t_2)$	$\underline{p}(h_1)\overline{p}(t_2)+\overline{p}(h_1)\underline{p}(h_2)$
p_{A2}	$\underline{p}(h_1)\overline{p}(h_1)\overline{p}(h_2)$	$\underline{p}(h_1)\overline{p}(h_1)\underline{p}(t_2)$	$\underline{p}(h_1)\underline{p}(t_1)\overline{p}(h_2)$	$\overline{p}(t_1)\overline{p}(h_1)\underline{p}(t_2)$	$\underline{p}(h_1)\overline{p}(h_2)+\overline{p}(h_1)\underline{p}(t_2)$
p_{A3}	$\overline{p}(h_1)\overline{p}(t_1)\underline{p}(h_2)$	$\underline{p}(t_1)\underline{p}(h_1)\overline{p}(t_2)$	$\underline{p}(t_1)\overline{p}(t_1)\underline{p}(h_2)$	$\underline{p}(t_1)\overline{p}(t_1)\overline{p}(t_2)$	$\underline{p}(t_1)\overline{p}(t_2)+\overline{p}(t_1)\underline{p}(h_2)$
p_{A4}	$\underline{p}(t_1)\underline{p}(h_1)\overline{p}(h_2)$	$\overline{p}(h_1)\overline{p}(t_1)\underline{p}(t_2)$	$\underline{p}(t_1)\overline{p}(t_1)\overline{p}(h_2)$	$\underline{p}(t_1)\overline{p}(t_1)\underline{p}(t_2)$	$\underline{p}(t_1)\overline{p}(h_2)+\overline{p}(t_1)\underline{p}(t_2)$
p_{B1}	$\underline{p}(h_2)\overline{p}(h_2)\underline{p}(h_1)$	$\overline{p}(t_2)\overline{p}(h_2)\underline{p}(h_1)$	$\underline{p}(h_2)\overline{p}(h_2)\overline{p}(t_1)$	$\underline{p}(h_2)\underline{p}(t_2)\overline{p}(t_1)$	$\underline{p}(h_2)\overline{p}(t_1)+\overline{p}(h_2)\underline{p}(h_1)$
p_{B2}	$\overline{p}(h_2)\overline{p}(t_2)\underline{p}(h_1)$	$\underline{p}(t_2)\overline{p}(t_2)\underline{p}(h_1)$	$\underline{p}(t_2)\underline{p}(h_2)\overline{p}(t_1)$	$\underline{p}(t_2)\overline{p}(t_2)\overline{p}(t_1)$	$\underline{p}(t_2)\overline{p}(t_1)+\overline{p}(t_2)\underline{p}(h_1)$
p_{B3}	$\underline{p}(h_2)\overline{p}(h_2)\overline{p}(h_1)$	$\underline{p}(h_2)\underline{p}(t_2)\overline{p}(h_1)$	$\underline{p}(h_2)\overline{p}(h_2)\underline{p}(t_1)$	$\overline{p}(t_2)\overline{p}(h_2)\underline{p}(t_1)$	$\underline{p}(h_2)\overline{p}(h_1)+\overline{p}(h_2)\underline{p}(t_1)$
p_{B4}	$\underline{p}(t_2)\underline{p}(h_2)\overline{p}(h_1)$	$\underline{p}(t_2)\overline{p}(t_2)\overline{p}(h_1)$	$\overline{p}(h_2)\overline{p}(t_2)\underline{p}(t_1)$	$\underline{p}(t_2)\overline{p}(t_2)\underline{p}(t_1)$	$\underline{p}(t_2)\overline{p}(h_1)+\overline{p}(t_2)\underline{p}(t_1)$

Table 1: Candidates for the extreme points of the independent natural extension of two binary variables

solved manually. However, due to the extreme symmetry—switching X_1 and X_2, h_1 and t_1 or h_2 and t_2 yields an equivalent set of inequalities—, only 7 of those 56 cases need to be considered, as the others can be related to these 7 by an argument of symmetry. In this way, we managed to obtain analytical expressions for the extreme points of $K(X_1)\otimes K(X_2)$.

Theorem 12. *Analytical expressions for the extreme points of $K(X_1)\otimes K(X_2)$ can be found by means of Table 1 and Figure 1. Table 1 contains expressions for 12 probability mass functions, which can be obtained by dividing the numbers in columns 2–5 by the denominator in column 6. The diagram in Figure 1 shows, depending on the particular values of $\underline{p}(h_1)$, $\overline{p}(h_1)$, $\underline{p}(t_1)$, $\overline{p}(t_1)$, $\underline{p}(h_2)$, $\overline{p}(h_2)$, $\underline{p}(t_2)$ and $\overline{p}(t_2)$, which of these 12 probability mass functions are extreme points of $K(X_1)\otimes K(X_2)$. In this diagram, we use the shorthand notation $p_{\mathrm{S1=S2}}$ to denote that p_{S1} and p_{S2} are two coinciding extreme points.*

Although the diagram in Figure 1 considers quite a number of special or degenerate cases, the main result can be summarised quite easily. If one of the local models is precise or vacuous, then the independent natural extension has the same extreme points as—and therefore coincides with—the strong extension. In all other cases, the independent natural extension has up to four additional extreme points.

9 Summary and Conclusions

In this paper, we have developed a definition for credal networks under epistemic irrelevance that allows for zero lower (and upper) probabilities, generalising Cozman's definition [2, Section 8.3], which requires the lower probabilities of conditioning events to be strictly positive. For the resulting joint model, we have derived a number of properties and alternative characterisations. Some of these results were already mentioned by Cozman, but are now proved to remain valid when his positivity requirement is dropped. One particular result is that the joint credal set that corresponds to a credal network under epistemic irrelevance can be described by means of linear constraints. As a first toy example, we have used this approach to obtain analytical expressions for the extreme points of the independent natural extension of two binary variables.

The main future goal that we intend to pursue is to develop algorithms for credal networks under epistemic irrelevance that are able to perform inference in an efficient manner. This problem has been tackled before by Cozman [2, Section 8.4], but we suspect that a more efficient solution can be obtained. The idea would be to derive counterparts to the marginalisation and graphoid properties that are proven in Ref. [5] and combine these with a linear programming approach that builds upon Theorem 9.

Acknowledgements

Jasper De Bock is a Ph.D. Fellow of the Fund for Scientific Research – Flanders (FWO) and wishes to acknowledge the financial support of the FWO. The authors also wish to thank three anonymous referees for their helpful comments.

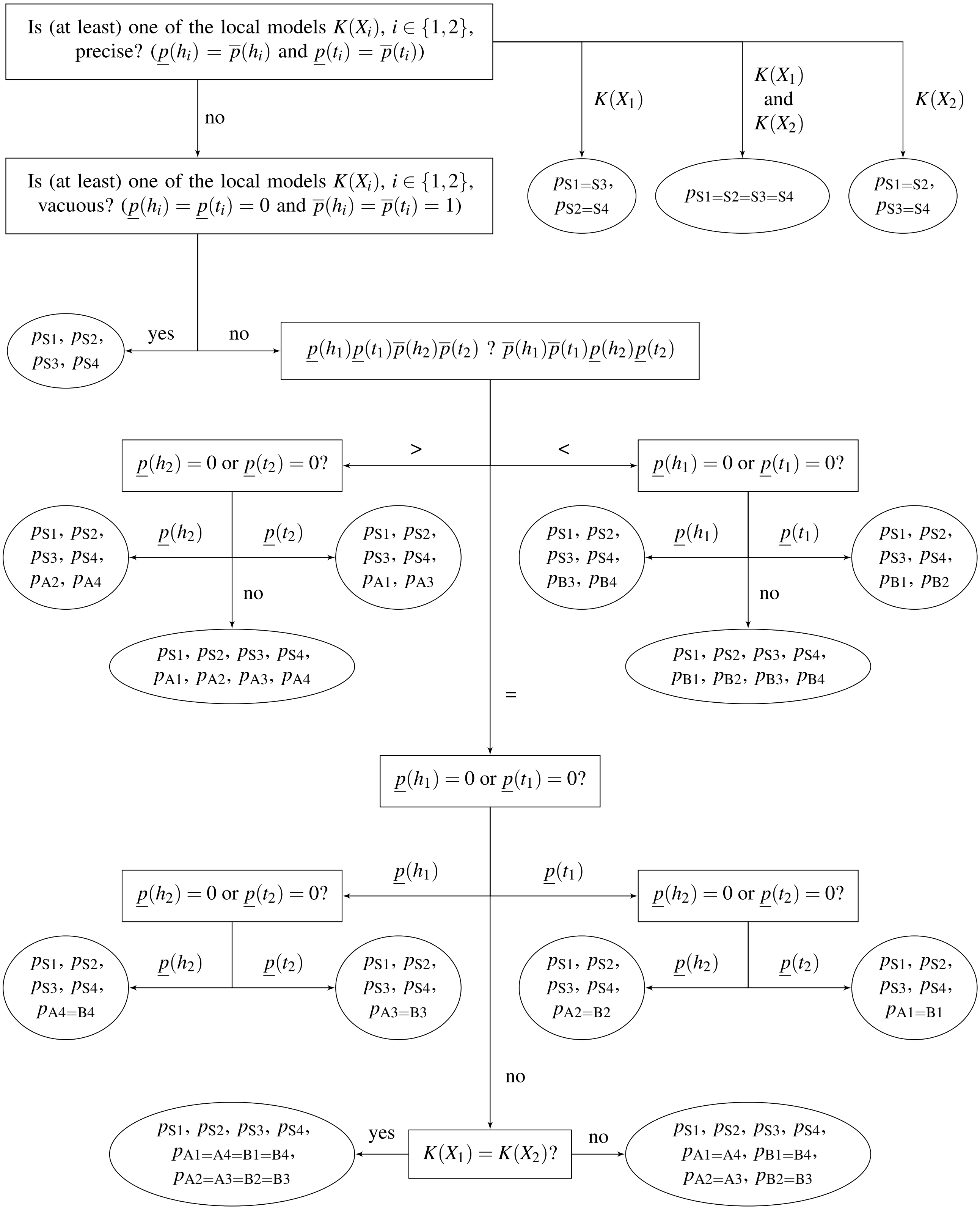

Is (at least) one of the local models $K(X_i)$, $i \in \{1,2\}$, precise? ($\underline{p}(h_i) = \overline{p}(h_i)$ and $\underline{p}(t_i) = \overline{p}(t_i)$)
$K(X_1)$
$K(X_1)$ and $K(X_2)$
$K(X_2)$
no
$p_{S1=S3}$, $p_{S2=S4}$
$p_{S1=S2=S3=S4}$
$p_{S1=S2}$, $p_{S3=S4}$
Is (at least) one of the local models $K(X_i)$, $i \in \{1,2\}$, vacuous? ($\underline{p}(h_i) = \underline{p}(t_i) = 0$ and $\overline{p}(h_i) = \overline{p}(t_i) = 1$)
yes
no
p_{S1}, p_{S2}, p_{S3}, p_{S4}
$\underline{p}(h_1)\underline{p}(t_1)\overline{p}(h_2)\overline{p}(t_2)$? $\overline{p}(h_1)\overline{p}(t_1)\underline{p}(h_2)\underline{p}(t_2)$
>
<
$\underline{p}(h_2) = 0$ or $\underline{p}(t_2) = 0$?
$\underline{p}(h_1) = 0$ or $\underline{p}(t_1) = 0$?
$\underline{p}(h_2)$
$\underline{p}(t_2)$
$\underline{p}(h_1)$
$\underline{p}(t_1)$
p_{S1}, p_{S2}, p_{S3}, p_{S4}, p_{A2}, p_{A4}
p_{S1}, p_{S2}, p_{S3}, p_{S4}, p_{A1}, p_{A3}
p_{S1}, p_{S2}, p_{S3}, p_{S4}, p_{B3}, p_{B4}
p_{S1}, p_{S2}, p_{S3}, p_{S4}, p_{B1}, p_{B2}
no
no
p_{S1}, p_{S2}, p_{S3}, p_{S4}, p_{A1}, p_{A2}, p_{A3}, p_{A4}
p_{S1}, p_{S2}, p_{S3}, p_{S4}, p_{B1}, p_{B2}, p_{B3}, p_{B4}
=
$\underline{p}(h_1) = 0$ or $\underline{p}(t_1) = 0$?
$\underline{p}(h_1)$
$\underline{p}(t_1)$
$\underline{p}(h_2) = 0$ or $\underline{p}(t_2) = 0$?
$\underline{p}(h_2) = 0$ or $\underline{p}(t_2) = 0$?
$\underline{p}(h_2)$
$\underline{p}(t_2)$
$\underline{p}(h_2)$
$\underline{p}(t_2)$
p_{S1}, p_{S2}, p_{S3}, p_{S4}, $p_{A4=B4}$
p_{S1}, p_{S2}, p_{S3}, p_{S4}, $p_{A3=B3}$
p_{S1}, p_{S2}, p_{S3}, p_{S4}, $p_{A2=B2}$
p_{S1}, p_{S2}, p_{S3}, p_{S4}, $p_{A1=B1}$
no
yes
$K(X_1) = K(X_2)$?
no
p_{S1}, p_{S2}, p_{S3}, p_{S4}, $p_{A1=A4=B1=B4}$, $p_{A2=A3=B2=B3}$
p_{S1}, p_{S2}, p_{S3}, p_{S4}, $p_{A1=A4}$, $p_{B1=B4}$, $p_{A2=A3}$, $p_{B2=B3}$

Figure 1: Diagram to obtain the extreme points of the independent natural extension of two binary variables

References

[1] G. Coletti and R. Scozzafava. Stochastic independence in a coherent setting. *Annals of Mathematics and Artificial Intelligence*, 35(1-4):151–176, 2002.

[2] F. G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[3] F. G. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39(2-3):167–184, 2005.

[4] F. G. Cozman. Sets of probability distributions, independence, and convexity. *Synthese*, 186(2):577–600, 2012.

[5] J. De Bock and G. de Cooman. Credal networks under epistemic irrelevance using sets of desirable gambles. Accepted for publication in *Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications*, 2013.

[6] C. P. de Campos and F. G. Cozman. Computing lower and upper expectations under epistemic independence. *International Journal of Approximate Reasoning*, 44(3):244–260, 2007.

[7] L. M. De Campos and S. Moral. Independence concepts for convex sets of probabilities. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, UAI'95, pages 108–115, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[8] G. de Cooman, F. Hermans, A. Antonucci, and M. Zaffalon. Epistemic irrelevance in credal nets: the case of imprecise Markov trees. *International Journal of Approximate Reasoning*, 51:1029–1052, 2010.

[9] G. de Cooman, E. Miranda, and M. Zaffalon. Independent natural extension. *Artificial Intelligence*, 175:1911–1950, 2011.

[10] B. Grünbaum. *Convex polytopes*. Springer, 2nd edition, prepared by V. Kaibel, V. Klee, and G. M. Ziegler, 2003.

[11] B. Vantaggi. Conditional independence structures and graphical models. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(05):545–571, 2003.

[12] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

Sample size determination with imprecise risk aversion

Malcolm Farrow
Newcastle University, UK
malcolm.farrow@newcastle.ac.uk

Abstract

We consider multi-attribute utility functions, particularly applied to the choice of a design and sample sizes for an experiment. We extend earlier work, which allowed imprecision in the trade-offs between attributes, to allow imprecision also in the shape of marginal utility functions. The method is illustrated with a simple example involving a two-group binomial experiment.

Keywords. Design of experiments, imprecise utility, risk aversion, sample size.

1 Introduction

In earlier work [8, 9, 10] a method for decision analysis with multiattribute utilities has been developed which does not require the specification of precise trade-offs between different risks. The original motivation for this work was the design of experiments [7, 8]. Multi-attribute utilities may be imprecisely specified, due to an unwillingness or inability on the part of the client to specify fixed trade-offs or precise marginal utility functions or because of disagreement within a group with responsibility for the decision. In particular this may be so when the decision is the choice of a design or sample size for an experiment. For example, in the design of a medical experiment, participants in the decision-making process may have different viewpoints, may put different weights on such attributes as the information gain and the risks to trial subjects and may be more or less risk averse in terms of these attributes.

An approach to constructing imprecise multi-attribute utility hierarchies and finding the Pareto optimal rules was introduced in [8]. The structure used was based on a utility hierarchy with utility independence at each node and used the notion of imprecise utility trade-offs within such a hierarchy, based on limited collections of stated preferences between outcomes. Pareto optimality, over the set of possible trade-off specifications, was used to reduce the set of alternatives.

Many real decision problems, for example in experimental design, have very large spaces of possible choices. Relaxing the requirement for precise utility specification reduces our ability to eliminate choices by dominance and can leave us with a large class of choices, none of which is dominated by any other over the whole range of possible utility functions allowed by the imprecise specification. Methods were described in [9] to reduce the class of alternatives that must be considered, by eliminating choices which are "ε-dominated" and combining choices which are "ε-equivalent." The effects of different values of ε and of different parts of the hierarchy were explored to see when and why choices were eliminated.

To choose a single alternative d^* from our reduced list, we can use the boundary linear utility approach described in [8], or select the choice which is the last to be eliminated as we increase the value of our ε criterion as described in [9]. We can then find the set D^* of choices which are "almost equivalent" to d^* and perhaps use secondary considerations to choose among them. In [10] methods based on the boundary linear utility for exploring the sensitivity of possible choices to variation in the utility trade-offs were described. This helps us to find a decision which, as far as possible, is a good choice over the whole range of possible trade-offs.

For some other approaches to imprecise utility, see, for example, [12, 2, 13, 16, 17, 5]. A particular feature of the approach used in [8, 9, 10] and this paper is the generality of the form of the utility hierarchy and of the shape of the feasible region.

The purpose of this paper is twofold. Firstly we show how the imprecise utility structure can be extended in a simple way to include imprecision in the shape of the marginal utility functions for attributes, and therefore in the degree of risk aversion, and that this extension preserves all of the results derived for the structure in previous work. Secondly, we return to the original motivation of the work by applying the methods to the choice of design and determination of sample size for experiments.

In Section 2 we briefly outline the Bayesian approach to experimental design, viewing it as a multi-attribute decision problem. In Section 3 we review the earlier work on decisions with imprecise utility trade-offs. In Section 4 we introduce the extension to include imprecision in the shape of the marginal utility functions. Finally, in Section 5, we apply the ideas to sample-size determination for a simple two-group experiment.

2 Bayesian Experimental Design

2.1 Introduction

The problem of experimental design is essentially that of choosing a design for an experiment from a, possibly infinite, set of possibilities. In simple cases this might just be a matter of choosing a sample size. In more complicated cases it may involve choosing several sample sizes, for observations of different types, or even of selecting types of observations to make, for example determining the values of covariates to use. In any case, this is clearly a decision and, usually, the values of various attributes, typically more than one, which are relevant to us, are unknown before the experiment and our distributions for them depend on the choice of design. We therefore formulate experimental design as a multi-attribute decision problem and choose the design which maximises our expectation of a multi-attribute utility function.

A recent, brief, introduction to this view of experimental design is given by [6]. For a more technical introduction to the field of Bayesian experimental design see, for example, [3]. A discussion of sample-size determination in clinical trials is given in Chapter 6 of [19]. See also, for example, [15, 20].

In much published work on Bayesian experimental design, a fixed total number of observations N is assumed. The problem is then to allocate these observations to design points (*ie* types of observation) while keeping the total fixed (sometimes allowing non-integer allocations on the grounds that it is the proportions of the total sample size which are being determined). Often some measure of information gain is used to provide a utility function and costs are assumed to depend only on the total sample size and therefore need not be considered. This is described as the "design problem" (although, perhaps, "allocation problem" might be a better name).

In contrast, in the " sample size problem", the trade-off between costs and benefits is explicitly considered so a utility function is required which involves both, *eg* [20]. Usually, relatively simple designs are considered.

In many real practical problems we need both to determine a total sample size and how the observations should be allocated to different design points. In this paper we do not distinguish between these two types of problem.

Typically, in experimental design we require a multi-attribute utility function where the attributes include costs and benefits. Each of these may be of more than one kind.

In some cases we might represent the "benefit" from an experiment in terms of some measure of information. For example we might use the posterior precision for some quantity of interest. We may, of course, be interested in several different unknown quatities so each would have its respective marginal utility and these utilities need to be combined. In other cases we might base our benefit utility directly on the pay-off from some *terminal decision*, in which our choice is informed by the result of the experiment. In fact the information-measure approach is (usually, at least) a special case of the terminal-decision approach, in which the terminal decision is to declare a value for some unknown (vector) quantity. The benefit utility is then based on the difference between our declared value and the true value.

Figure 1 shows an influence diagram for a typical problem in experimental design. For example this could refer to the design of a clinical trial in which we wish to compare two or more treatments. There could also be several groups of patients, for example divided by age-group, severity-group, sex *etc.* The initial decision D_X consists of the choice of design d_X. Often the set of possible choices will include the option of no experiment at all. In the experiment, we observe data X. The distribution of X depends on d_X and on unknown quantities (parameters) θ. A vector of pay-offs C_X refers to various attributes, for example financial costs or effects on subjects. The distribution of these depends on d_X and X. Having seen the data X we make a terminal decision D_Y. This may well be the choice of treatment for future patients. We choose d_Y. The outcomes Y of this terminal decision may be, for example, the clinical outcomes for some future patients but may also include other attributes such as costs of future treatments. The distribution of these depends on d_Y and on the unknown θ . These outcomes lead to rewards (pay-offs) C_Y which depend on d_Y and Y. (More generally, they may also depend on θ). There may, of course, be a potentially unbounded

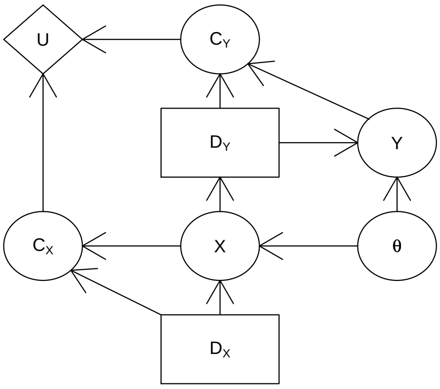


Figure 1: Influence diagram for a typical problem in experimental design.

number of future patients. However, in our utility function, we might discount outcomes as we look further into the future. This might be justified on the grounds that, further into the future, it becomes less likely that our choice of treatment will still be dictated by this experiment. Finally, our overall utility $U = U(C_X, C_Y)$ depends on C_X and on C_Y.

To determine our choice of design $d(X)$ we work our way backwards through the influence diagram. After observing the data $X = x$ in our experiment, we choose

$$\begin{aligned} d_Y &= \arg \max_{d_Y \in D_Y} [\mathrm{E}_{d_Y}\{U(C_X, C_Y) \mid x\}] \\ &= \arg \max_{d_Y \in D_Y} [U(d_Y; C_X, C_Y \mid x)]. \end{aligned}$$

Our expected utility at this stage is

$$\max_{d_Y \in D_Y} [U(d_Y; C_X, C_Y) \mid x].$$

Before observing the data, we choose the design

$$d_X = \arg \max_{d_X \in D_X} \mathrm{E}_{d_X}\{ \max_{d_Y \in D_Y} [U(d_Y; C_X, C_Y) \mid X]\}.$$

A useful variation on this is to use two different prior distributions, an *inference* or *fitting* or *terminal* prior, which is used for choosing d_Y, and a *design* or *sampling* prior which is used for choosing d_X. This approach was suggested by [21]. Similarly we can have different utility functions for the two decisions.

2.2 Risks in Experimental Design

Since we are concerned in this paper with degrees of risk aversion, let us briefly consider some of the many risks associated with experimental design.

We have already mentioned the financial cost of the experiment, which may not be known in advance with certainty, and the effects on experimental subjects. Particularly in the cases of human and animal subjects we are likely to be concerned about the possibility of adverse reactions but, even in other experiments, there might be other costs concerned with effects on valuable material or equipment. We may come to a conclusion, based on our experiment, which is far from the truth. This could lead to a bad choice in a terminal decision and therefore to a bad pay-off. A type of risk which seems to have had little formal consideration is that something may go wrong with the experiment and that this leads to less useful information than expected or perhaps to none at all. In particular we may suffer from missing observations. Some designs, for example those for microarray experiments, could be very sensitive to missingness. See *eg* [1].

In choosing an experimental design we will be seeking to optimise our expectation of a utility function which involves some or all of these risks. Our choice will therefore depend on how we are willing to trade these risks against each other and this, in turn, depends on our attitudes to these risks, including the shapes of our marginal utility functions since these shapes describe our degrees of risk aversion with respect to the various attributes.

2.3 Utility Hierarchy

A hierarchical structure for utilities in a multi-attribute problem was suggested by [14] and [8] adopted such a structure. In [8], an example was used in which there were financial costs of the experiment and also "ethical costs" which related to possible effects on the experimental subjects. The marginal utilities of these are combined into a Cost utility. In an experiment we potentially learn about a number of quantities and, in their example, [8] represented this collection in four groups, each of which had a marginal utility based on the distance of our posterior expectation from the true value. These were combined into a Benefit utility. Finally the Cost and Benefit utilities were combined in an overall utility for the chosen design.

3 Imprecision in utility trade-offs

3.1 Mutually utility independent hierarchies

In order to introduce imprecision into the trade-offs between attributes, [8] proposed a general class of multi-attribute utility functions which uses the concept of mutual utility independence among sets of attributes in order to impose a structure on the utility function. Attributes $\underline{Y} = (Y_1, ..., Y_k)$ are *utility independent* of the attributes $\underline{Z} = (Z_1, ..., Z_r)$ if conditional preferences over lotteries with differing values of $\underline{\mathrm{Y}}$ but fixed values, $\underline{\mathrm{z}}$, of $\underline{\mathrm{Z}}$, do not depend on the particular choice of $\underline{\mathrm{z}}$. Attributes $\underline{X} = (X_1, ..., X_s)$ are *mutually utility independent* if every subset of $\underline{X}$ is utility independent of its complement. If attributes $\underline{X}$ are mutually utility independent, then [14] showed that the utility function for $\underline{\mathrm{X}}$ must be given by the *multiplicative form*

$$U(\underline{X}) = B^{-1} \left\{ \prod_{i=1}^{s} [1 + k a_i U_i(X_i)] - 1 \right\}, \quad (1)$$

where B does not depend on $U_1(X_1), \ldots, U_s(X_s)$, or the *additive form*

$$U(\underline{X}) = \sum_{i=1}^{s} a_i U_i(X_i), \quad (2)$$

where $\mathrm{U}_i(\mathrm{X}_i)$ is a conditional utility function for attribute X_i, namely an evaluation of the utility of X_i for fixed values of the other attributes. The coefficients in (1) and (2) are the *trade-off parameters*; the a_i reflect the relative importance of the attributes and k reflects the degree to which rewards may be regarded as complementary, if $k > 0$, or as substitutes, if $k < 0$.

The assumption of mutual utility independence is enough in itself to reduce the problem to one of considering a finite number of parameters.

The next step is to form a hierarchical structure, in which, at each node, several utilities are merged into a combined utility. This combined utility is merged with others at a node in the next level until, finally, one overall utility function is formed. If, at each node, we have mutual utility independence for the utilities combined at that node, then we term such a utility function a *Mutually Utility Independent Hierarchic (MUIH)* utility. Thus, in a MUIH utility, at each node we combine utilities using either (1) or (2).

This hierarchical structure allows us to relax the requirement for overall mutual utility independence by allowing the user to specify utility independence just at the nodes of the hierarchy and, of course, the user can choose this structure.

Nodes in the hierarchy, other than the marginal nodes, are termed *child nodes* and classified by [8] into the following three types:

1. an *additive node*, where utilities are combined as in (2) with $\sum_{i=1}^{s} a_i \equiv 1$ and $a_i > 0$ for $i = 1, \ldots, s$;
2. a *binary node*, where precisely two utilities are combined, where we rescale the combined utility as
$$U = a_1 U_1 + a_2 U_2 + h U_1 U_2 \quad (3)$$
where $0 < a_i < 1$ and $-a_i \leq h \leq 1 - a_i$, for $i = 1, 2$, and $a_1 + a_2 + h \equiv 1$. Note that (3) is derived by setting $s = 2$ and $h = k a_1 a_2$ in (1).
3. a *multiplicative node*, where more than two utilities are combined and the parameter k in (1) may be nonzero. We scale the utility using
$$B = \prod_{i=1}^{s} (1 + k a_i) - 1 \quad (4)$$
with $a_1 \equiv 1, k > -1$ and, for $i = 1, \ldots, s$, we have $a_i > 0$ and $k a_i > -1$. When $k = 0$ we obtain (2).

At each child node n, we have a collection $\underline{\phi}_n = (\phi_{n,1}, \ldots, \phi_{n,r_n})$ of trade-off parameters which determine how the parent utilities at node n are combined to give the value at the child node. If there are N child nodes, then we denote by $\underline{\theta} = (\underline{\phi}_1, \ldots, \underline{\phi}_N)$ the collection of all the trade-off parameters in the hierarchy. A hierarchy in which imprecision is allowed in some of the elements of $\underline{\theta}$ is called an *imprecise independence hierarchy (IIH).* If the hierarchy contains only additive and binary nodes, then the specification is a *simple imprecise independence hierarchy (SIIH)*

So that the interpretation of utility values does not depend on the choice of trade-off parameters, we place all utilities in the hierarchy on a *standard scale.* Each marginal utility is normed to lie between 0, the worst outcome that we shall consider for the problem, and 1, the best outcome. The relative weights of the marginal utilities are governed by the trade-off parameters at the nodes of the hierarchy and these are chosen to reflect this norming. Consider a child node n. Let C_n be an outcome such that all marginal predecessor nodes have utility 1, and c_n be an outcome such that all marginal predecessor nodes have utility 0. The scalings described above for additive, binary and multiplicative nodes ensure that, at n, the utilities of C_n and c_n are 1 and 0 respectively. Therefore, a utility value of u at node n may always be interpreted as the utility of a gamble giving C_n with probability u and c_n with probability $1 - u$, irrespective of the chain of trade-off parameters in the hierarchy.

3.2 Specification of imprecise utility trade-offs

In standard utility theory, the decision maker must make statements which define the preferences between all combinations of outcomes. In the case of imprecise utility, the decision maker may state preferences just for some, but not all, choices of outcome combinations. Imprecise utility is defined by obeying all of the constraints implied by the stated preferences. In [8, 9, 10] it was supposed that the decision maker could make preference statements over all outcomes of each individual attribute, and so could specify precise marginal utilities, but could only make preference statements for some, but not all, combinations of the various attributes. Each such preference statement imposed constraints on the tradeoff parameters which are used to combine the individual attributes into an imprecise multi-attribute utility. These constraints together specify a feasible region R for $\underline{\theta}$. Comments on the process of elicitation are made in [8, 9, 10].

In Section 4 below we will drop the assumption that the decision maker has to specify precise marginal utilities.

3.3 Analysis with imprecise utility trade-offs

In earlier work [8, 9, 10], methods have been developed which exploit the IIH structure to reduce the number of choices to be considered and select choices and to explore the sensitivity of choices. Our aim in Section 4 below will be to extend the structure to allow imprecision in the marginal utility functions while preserving the various results derived and retaining our ability to carry out these analyses. In this section we briefly summarise these results and methods.

Having obtained our imprecise specification for the parameters of our multi-attribute utility function we can reduce the number of possible choices, that is designs, by retaining only choices which are Pareto optimal (non-dominated) with respect to the range R of the parameters $\underline{\theta}$.

We have to choose from a set $\mathcal{D}$ of choices. We denote the utility of a particular choice $A \in \mathcal{D}$, evaluated with trade-off parameters $\underline{\theta}$ as $U_{A\underline{\theta}}$. This is evaluated as the expected value of U_θ, with respect to the probability distribution, induced by the choice A, over the marginal attributes involved in U. For two alternatives, A, B, let $d_{AB}(\underline{\theta}) = U_{A\underline{\theta}} - U_{B\underline{\theta}}$.

We write $A \succeq B$, if $U_{A\underline{\theta}} \geq U_{B\underline{\theta}} \, \forall \underline{\theta} \in R$. We say that A is preferred to B over R, written $A \succ B$, if $A \succeq B$ and $U_{A\underline{\theta}} > U_{B\underline{\theta}}$ for some $\underline{\theta} \in R$, and that A is equivalent to B, written $A \simeq B$, if $U_{A\underline{\theta}} = U_{B\underline{\theta}} \, \forall \underline{\theta} \in R$. We call alternative A *Pareto optimal* for R if there is no other allowable alternative B for which $B \succ A$ over R. We restrict attention to Pareto optimal alternatives. Furthermore, if we form equivalence classes of equivalent decisions $A_1 \simeq A_2 \simeq ... \simeq A_r$, then it is reasonable to restrict attention to only one representative member of each equivalence class.

To reduce the number of choices further, [9] introduced the concept of *ε-preference* as follows. Let $\varepsilon \geq 0$ be a value chosen to indicate a practical indifference between utility values. For two alternatives A and B, we say that A is almost-preferable with tolerance ε, or, more concisely, "*ε-preferable*" to B, written $A \succeq_\varepsilon B$, over the set R of parameter specifications if $\inf_R(d_{AB}(\underline{\theta})) \geq -\varepsilon$. Two alternatives A, B are said to be almost-equivalent with tolerance ε, or, more concisely, "*ε-equivalent*", written $A \simeq_\varepsilon B$, if both $A \succeq_\varepsilon B$ and $B \succeq_\varepsilon A$. Note that ε-preference does not define a complete ordering of the alternatives and nor does ε-equivalence define an equivalence relation. Alternative A is said to *ε-dominate* alternative B, written $A \succ_\varepsilon B$, if $A \succeq_\varepsilon B$ but $B \not\succeq_\varepsilon A$, where the negation of the relationship is indicated in the usual way. Setting $\varepsilon = 0$, an alternative which is not 0-dominated by any other is Pareto optimal. The notation is extended to collections of alternatives as follows. The collection $\mathcal{A}$ is ε-preferable to the collection $\mathcal{B}$ of alternatives, written $\mathcal{A} \succeq_\varepsilon \mathcal{B}$ if, for each $B \in \mathcal{B}$, there is at least one $A \in \mathcal{A}$ for which $A \succeq_\varepsilon B$.

In [9] a number of results are derived concerning the properties and uses of ε-preference in IIH utilities. In particular, an algorithm is presented for gradually reducing the number of choices by increasing ε from zero and eliminating choices while our retained list remains an ε-Pareto set. Eventually we are left with a single choice d^*. Notice that this choice is made without having to specify a value for ε in advance.

In [10] methods for exploring the sensitivity of choices are presented. In particular the *boundary linear utility*, which had been introduced in [8], is described and results concerning its properties and uses with IIH utilities are given. Let P be the set of vertices of R. In [8] it is shown that, for a SIIH utility, Pareto optimal alternatives for R are the same as Pareto optimal alternatives for P. This forms part of the motivation for the boundary linear utility

$$\bar{U}_\lambda = \sum_{i=1}^{s} \lambda_i U_i$$

where U_i is the utility function determined by the choice of trade-offs $\underline{\theta}_i \in P = \{\underline{\theta}_1, \ldots, \underline{\theta}_s\}$ and $\lambda_1, \ldots, \lambda_s$ are nonnegative constants with $\sum_{i=1}^{s} \lambda_i = 1$.

The results and methods which are developed, some of which may be extended to the case of general IIH utilities, allow us to exploit the idea that, by varying the λ weights, we can change the emphasis which is placed on different parts of the feasible region.

4 Imprecise risk aversion

4.1 Use of basis functions

Now we consider dropping the assumption that the decision maker can give a precise specification of each marginal utility function. Recall that two utility functions, U_A and U_B, are strategically equivalent if $U_B = c + dU_A$ where c and d are constants with $d > 0$. Therefore, without loss of generality we can rescale a marginal utility function to be on the standard scale, as in [8, 9, 10]. Without loss of generality we can also rescale a scalar attribute Z so that the "worst value" is $z = 0$ and the "best value" is $z = 1$. All that is left is to determine the shape of the utility curve between the points $(0, 0)$ and $(1, 1)$. The shape will typically reflect the degree of risk aversion, with a concave curve representing a risk-averse utility function and a convex curve representing a risk-seeking utility function, with respect to the (rescaled) attribute Z. See, for example, Section 4.4.1 of [14].

We could introduce imprecision into the shape of a marginal utility function $U(z)$ by introducing a collection of basis functions $U_1(z), \ldots, U_s(z)$ so that $U(z) = \sum_{i=1}^{s} b_i U_i(z)$ with $b_i \geq 0$ for all i and $\sum_{i=1}^{s} b_i = 1$. We would then elicit a feasible region for the weights $b_1, \ldots, b_s$. An important feature of this approach is that, in effect, we are simply adding an extra layer to the utility hierarchy by making each marginal utility an additive node and introducing the basis functions as new marginal quantities which are parents to the previously marginal nodes. Therefore all of the theory and methods developed previously for the case where imprecision applied only to the trade-offs extends to cover imprecision in the marginal utility functions as well.

A simple example of basis functions is given by quadratic functions. Consider $U_i(z) = c_0 + c_1 z + c_2 z^2$. The constraints $U(0) = 0$ and $U(1) = 1$ simplify this to $U(z) = cz + (1-c)z^2$. The constraints $U'(0) \geq 0$ and $U'(1) \geq 0$, where $U'(z) = dU(z)/dz$, imply $0 \leq c \leq 2$. With $c = 0$, we obtain $U_1(z) = z^2$ and, with $c = 2$, we obtain $U_2(z) = 2z - z^2$. Let $b = c/2$. Then

$$U(z) = (1 - b)U_1(z) + bU_2(z)$$

with $0 \leq b \leq 1$. If $b > 1/2$ we have a risk averse utility function, with $b = 1/2$ it is risk neutral and with $b < 1/2$ it is risk seeking. Curves with $b = 0,\ 0.25,\ 0.5,\ 0.75,\ 1$ are shown in Figure 2.

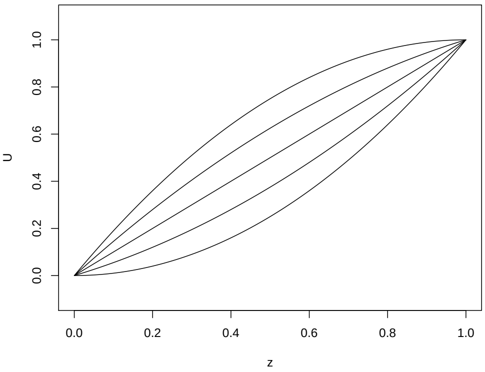


Figure 2: Quadratic utility curves with $b = 0.0, 0.25, 0.5, 0.75, 1$.

Note that we can rewrite the basis functions as $U_1(z) = z - h(z)$ and $U_2(z) = z + h(z)$ with, in this case, $h(z) = z - z^2$. It can be seen from Figure 2 that this offers a rather limited range of shapes. While restricting ourselves to monotonic functions, the greatest range that we can obtain in this form is with $h(z) = z$ $(0 \leq z \leq 1/2)$ and $h(z) = 1 - z$ $(1/2 < z \leq 1)$. Even this is somewhat restricted in range and certainly in shape. We can obtain greater flexibility with a more direct approach to eliciting the utility function.

While an elicitation procedure for use in practice might involve more refined questions, in principle we can use the probability-equivalent method. In its simplest form, to determine a range for $U(z^*)$ where $0 < z^* < 1$, we offer the decision maker a choice between d_A : the attribute value corresponding to $z = z^*$, with certainty, and d_B : with probability α, the attribute value corresponding to $z = 1$ and, with probability $1 - \alpha$, the attribute value corresponding to $z = 0$. For large α the decision maker will choose d_B, for small α the decision maker will choose d_A but for an intermediate range the decision maker may express no clear preference. The *lower utility* for z^*, $U_1(z^*)$ is the largest value of α at which the decision maker would choose d_A and the *upper utility* for z^*, $U_2(z^*)$ is the smallest value of α at which the decision maker would choose d_B. By repeating this process at a range of values z^* and using suitable interpolation, we obtain lower and upper utility functions, $U_1(z)$ and $U_2(z)$. These can then be our two basis functions. Linear interpolation may well be adequate.

With two basis functions, all allowable utility functions are weighted averages of these two. We could obtain more degrees of flexibility in the shape by adding additional basis functions, for example one which is closer to $U_1(z)$ for some of the range of z and otherwise closer to $U_2(z)$. This would, of course, require more sophisticated elicitation procedures.

4.2 Effect on trade-offs

While the standard scale ensures that all utilities are in $[0,1]$, where in that range they are likely to be will be different for the lower and upper utility functions. In itself this does not cause a problem. Of more concern is the fact that $U'(z)$ may be different between the lower and upper marginal utility functions. This could affect our consideration of the trade-off at the immediate successor node in the hierarchy. For example, suppose that our marginal utility is U_z and, at the child node, this is combined with another utility U_x to give $U_n = a_n U_z + (1-a_n)U_x$. Then, if $U_z = (1-b)U_1(z) + bU_2(z)$, the effect on U_n of a fixed change in z may depend on the choice of b. This may be acceptable. After all, the *average* gradient, given a uniform distribution for Z, will remain 1. However the decision maker, with the help of the analyst, needs to consider this consequence of allowing imprecision in the shape of $U_z(z)$. A possible solution would be to elicit a joint feasible region for a and b (or, more generally, for all of the parameters involved at the marginal and child nodes) so that the range of a can depend on the choice of b. If the child node is an additive node it can be extended straightforwardly to include all the basis functions at its parent (marginal) nodes as separate parents. If the child node is a binary node then it can similarly be extended although its new form will not imply mutual utility independence between all of its new parents.

5 Sample size example

To illustrate the method we consider a simple example. Suppose we wish to design a trial, for example a clinical trial, with two treatments and binary outcomes (*eg* cure/not cure). For $g = 1, 2$, we will give treatment g to n_g subjects and observe the number X_g of successes. Using these data, a choice will be made between these treatments for use with future cases.

Suppose that the unknown success rate with treatment g is θ_g. For simplicity assume that our terminal prior gives a Beta$(a_{t,g}, b_{t,g})$ distribution to θ_g with θ_1 and θ_2 independent and that our terminal utility is such that we will choose whichever treatment has the greater posterior probability of success. That is we choose treatment g if the posterior expectation of θ_g is greater than that of $\theta_{g'}$. We set $a_{t,1} = a_{t,2} = b_{t,1} = b_{t,2} = 1.5$.

In our design prior, θ_1 and θ_2 are not independent. A number of methods are available for constructing this joint distribution. For example we could use a bivariate normal distribution for the logits or probits of θ_1 and θ_2 or we could link beta marginal distributions using a copula. However, in this example, the prior is constructed using a mixture distribution. In each component, c, we give θ_1 and θ_2 independent beta distributions, Beta$(a_{c,g}, b_{c,g})$, $g = 1, 2$. The effect of the mixture is to induce correlation between θ_1 and θ_2. A three component mixture is used, with parameters as given in Table 1. The advantage of this form of prior distribution is that prior predictive distributions for the observations can be calculated analytically within each component leading to simple calculations of expected utilities. The results can then be averaged over components.

Component	Probability	Parameters			
c		$a_{c,1}$	$b_{c,1}$	$a_{c,2}$	$b_{c,2}$
1	0.25	7.5	3.0	4.5	4.5
2	0.50	4.5	3.0	3.0	4.5
3	0.25	4.5	6.0	3.0	6.0

Table 1: Parameters of design prior mixture distribution. Within each component $\theta_g \sim \text{Beta}(a_g, b_g)$.

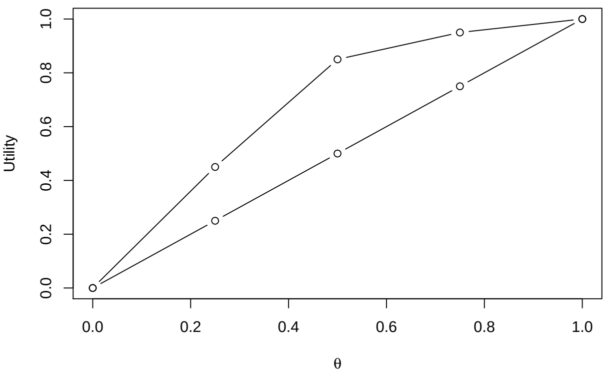


Figure 3: Lower and upper benefit utility functions.

For simplicity in this example we use a simple (precise) form for the marginal cost design utility. Let $n_{\max,1}$ and $n_{\max,2}$ be the largest sample sizes which we would consider. Let

$$Z_{C,g} = \begin{cases} 1 & (n_g = 0) \\ 1 - \frac{h_{0,g} + h_{1,g} n_g}{h_{0,g} + h_{1,g} n_{\max,g}} & (n_g > 0) \end{cases} . \qquad (5)$$

Then the marginal cost utility is $U_C = a_{c,1}Z_{C,1} + a_{c,2}Z_{C,2}$. We use $a_{c,1} = a_{c,2} = 0.5$, $h_{0,1} = h_{0,2} = 10$, $h_{1,1} = h_{1,2} = 1$, $n_{\max,1} = 100$ and $n_{\max,2} = 60$.

The overall design utility is $U = b_C U_C + b_B U_B$. We use $0.03 \leq b_C \leq 0.07$ and $b_B = 1 - b_C$.

Order	n_1	n_2	ε	Order	n_1	n_2	ε	Order	n_1	n_2	ε
	17	13		25	19	15	0.000084	12	20	15	0.000022
37	0	0	0.004334	24	16	12	0.000067	11	25	19	0.000018
36	19	16	0.000724	23	16	10	0.000048	10	25	16	0.000018
35	14	12	0.000571	22	15	11	0.000048	9	22	19	0.000013
34	18	15	0.000295	21	22	18	0.000048	8	21	17	0.000010
33	21	18	0.000271	20	18	14	0.000044	7	23	17	0.000009
32	13	10	0.000220	19	16	15	0.000043	6	16	16	0.000008
31	15	12	0.000134	18	18	16	0.000043	5	23	19	0.000008
30	21	16	0.000126	17	17	15	0.000040	4	13	13	0.000007
29	17	14	0.000114	16	16	11	0.000037	3	19	17	0.000002
28	13	11	0.000095	15	15	15	0.000033	2	24	18	0.000001
27	24	19	0.000092	14	15	13	0.000023	1	20	16	0.000001
26	16	13	0.000088	13	12	12	0.000022				

Table 2: Results of selection by ε-preference. The order of dropping is shown. The last-retained design is $n_1 = 17,\ n_2 = 13$.

The benefit utility depends on the outcomes for future patients. For a future patient i, let Z_i be 1 or 0 depending on the success or failure of the treatment. This suggests an attribute of the form $Z_B = \sum_{i=1}^{\infty} k_i Z_i$ with $\sum_{i=1}^{\infty} k_i = 1$. For example, we could use $k_i = (1-\lambda)\lambda^{i-1}$ with $0 < \lambda < 1$. Another possibility is $k_i = m^{-1}$ for $i = 1, \ldots, m$ and $k_i = 0$ for $i > n$. For simplicity in this example we adopt the second form and furthermore let $m \to \infty$ so that, given a value of θ, $Z_B \to \theta$.

Using the probability-equivalent method we elicit a lower and an upper utility function $U_{B,L}(\theta)$ and $U_{B,U}(\theta)$ with evaluations at a range of values of θ and linear interpolation. At $\theta = 0, 0.25, 0.5, 0.75, 1$, the lower values are chosen to be $U_{B,L}(\theta) = \theta$, giving risk neutrality. The upper values are $U_{B,L}(\theta) = 0.00, 0.45, 0.85, 0.95, 1.00$, giving risk aversion. These two functions are shown in Figure 3.

Let $\underline{\theta} = (\theta_1, \theta_2)^T$ and $\underline{x} = (x_1, x_2)^T$. We can write the joint probability density of component c, parameters θ_1, θ_2, observations X_1, X_2, and the benefit utility U_B, given sample sizes n_1, n_2, as

$$P = \Pr(c) f_{c,\theta,X}(\underline{\theta}, \underline{x} \mid c) f_U(U_B \mid \underline{x}, \underline{\theta}, c) \qquad (6)$$

where

$$\begin{aligned} f_{c,\theta,X}(\underline{\theta}, \underline{x} \mid c) &= \prod_{g=1}^{2} f_{c,g}(\theta_g \mid c) f_{X|\theta,n_1}(x_g \mid \theta_g) \\ &= \prod_{g=1}^{2} f_{X|n_g}(x_g \mid c) f_{c,g|x}(\theta_g \mid x_g, c) \end{aligned}$$

where $f_{X|n_g}(x_g \mid c)$ is the prior predictive probability function of X_g, given c, and $f_{c,g|x}(\theta_g \mid x_g, c)$ is the conditional posterior density, using the design prior, given c, of θ_g after observing the data $X_g = x_g$. The density of U_B depends on x_1 and x_2 both because we use the posterior density of θ_1 and θ_2 and because the choice of treatment (and hence θ_1 or θ_2) for future cases depends on the posterior distributions, given x_1 and x_2, using the terminal prior. From (6) we can see that we can evaluate conditional expectations within each component of the mixture straightforwardly and then average over the mixture components. The conditional posteriors are beta distributions and the conditional prior predictive distributions for X_g can be evaluated analytically.

With $0 \le n_1 \le 100$ and $0 \le n_2 \le 60$, there are 6161 potential designs. Of these, 38 are non-dominated. With the exception of $(0, 0)$, all of the non-dominated designs have $12 \le n_1 \le 25$, all have $0.6n_1 < n_2 \le n_1$ and all but three have $0.7n_1 < n_2 \le n_1$. Applying the ε-preference algorithm described in Section 5.2 of [9], we obtain the results shown in Table 2. Designs are eliminated one by one as we increase the value of the tolerance ε. Finally one design, $n_1 = 17,\ n_2 = 13$, is left. Interestingly, the last eliminated design is the null experiment, reflecting the fixed cost of any non-null experiment given in (5).

6 Concluding comments

Imprecision in the shape of the marginal utility functions is a natural extension of the earlier work on imprecision in utility trade-offs. In this paper this extension has been made in a way which preserves the results from the earlier work.

The remaining extension to give a fully imprecise analysis would be to allow imprecision in the probability distributions for outcomes given choices. In fact, if our utility hierarchy is fully additive then we can work directly in terms of previsions of marginal utilities and thus deal with this imprecision in the same way as we have done in this paper. When our multiattribute utility involves products of marginal utilities then incorporation of imprecision in our beliefs in this way would still be possible if we were prepared to regard all of the marginal utilities as uncorrelated. The generalisation to the case without this assumption awaits further work. See, for example, [4] for a different approach.

The simple example in this paper presented no serious computational difficulty. However more complicated experimental design problems will often present computational challenges, both because of the number of potential designs to be compared and, particularly in cases where computationally intensive methods would normally be used to evaluate posterior distributions, the difficulty of evaluating the expected utility for any proposed design. These difficulties apply even without the introduction of imprecision. One possible approach in such cases is to use a simulation-based method, as in [18]. Another possibility is to use a method which does not require such intensive computation, such as Bayes linear methods [8] or Bayes linear kinematics [11, 22] and such an approach, using Bayes linear kinematics is under investigation.

Acknowledgement

The author is grateful for the helpful comments of three reviewers on an earlier version of this paper.

References

[1] R. A. Bailey. Designs for two-colour microarray experiments. *Applied Statistics*, 56:365–394, 2007.

[2] J. Butler, J. Jia and J. Dyer. Simulation techniques for the sensitivity analysis of multi-criteria decision models. *European Journal of Operational Research*, 103:531–546.

[3] K. Chaloner and I. Verdinelli. Bayesian experimental design – a review. *Statistical Science*, 10:273–304, 1995.

[4] M. Danielson, L. Ekenberg and A. Larsson. Distribution of expected utility in decision trees. *International Journal of Approximate Reasoning*, 46:387–407, 2007.

[5] M. Danielson, L. Ekenberg and A. Riabacke. A prescriptive approach to elicitation of decision data. *Journal of Statistical Theory and Practice*, 3:157–168, 2009.

[6] M. Farrow. Optimal Experiment Design, Bayesian. In *Encyclopedia of Systems Biology* (W. Dubitzky, O. Wolkenhauer, K-H. Cho and H. Yokota, Eds). Springer, 2013.

[7] M. Farrow and M. Goldstein. Reconciling costs and benefits in experimental design. In *Bayesian Statistics 4* (J.M.Bernardo, J.O.Berger, A.P.Dawid and A.F.M.Smith, Eds). Oxford University Press, 607-615, 1992.

[8] M. Farrow and M. Goldstein. Trade-off sensitive experimental design: a multicriterion, decision theoretic, Bayes linear approach. *Journal of Statistical Planning and Inference*, 136:498–526, 2006.

[9] M. Farrow and M. Goldstein. Almost-Pareto decision sets in imprecise utility hierarchies. *Journal of Statistical Theory and Practice*, 3:137–155, 2009.

[10] M. Farrow and M. Goldstein. Sensitivity of decisions with imprecise utility trade-off parameters using boundary linear utility. *International Journal of Approximate Reasoning*, 51:1100-1113, 2010.

[11] M. Goldstein and S. Shaw. Bayes linear kinematics and Bayes linear Bayes graphical models. *Biometrika*, 91:425–446, 2004.

[12] G. B. Hazen. Partial information, dominance and potential optimality in multiattribute utility theory. *Operations Research*, 34:296-310.

[13] A. Jiménez, S. Ríos-Insua and A. Mateos. A decision support system for multiattribute utility evaluation based on imprecise assignments. *Decision Support Systems*, 36:65-79, 2003.

[14] R. L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-offs.* John Wiley & Sons, 1976.

[15] D. V. Lindley. The choice of sample size. *The Statistician*, 46:129–138, 1997.

[16] A. Mateos, A. Jiménez and S. Ríos-Insua. Modelling individual and global comparisons for multiattribute preferences. *Journal of Multi-Criteria Decision Analysis*, 12:177–190, 2003.

[17] A. Mateos, A. Jiménez and S. Ríos-Insua. Dominance, potential optimality and alternative ranking in imprecise multi-attribute decision making. *Journal of the Operational Research Society*, 58:326–336, 2007.

[18] P. Müller. Simulation-based optimal design. In *Bayesian Statistics 6* (J.M.Bernardo, J.O.Berger, A.P.Dawid and A.F.M.Sith Eds.) Oxford University Press, 459–474, 1999.

[19] D. J. Spiegelhalter, K. R. Abrams and J. P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* John Wiley & Sons, 2004.

[20] S. B. Tan and A. F. M. Smith. Exploratory thoughts on clinical trials with utilities. *Statistics in Medicine*, 17:2771–2791, 1998.

[21] F. Wang and A. E. Gelfand. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 17:193–208, 2002.

[22] K. J. Wilson and M. Farrow. Bayes linear kinematics in the analysis of failure rates and failure time distributions. *Journal of Risk and Reliability*, 224:309–321, 2010.

Computing with Confidence

Scott Ferson
Applied Biomathematics
100 North Country Road
Setauket, New York 11733 USA
scott@ramas.com, sandp8@gmail.com

Michael Balch
Applied Biomathematics
100 North Country Road
Setauket, New York 11733 USA
michael.balch@arctan-group.com

Kari Sentz
Los Alamos National Laboratory
P.O. Box 1663, MS F609
Los Alamos, New Mexico 87545 USA
ksentz@lanl.gov

Jack Siegrist
Applied Biomathematics
100 North Country Road
Setauket, New York 11733 USA
jack@ramas.com

Abstract

Traditional confidence intervals are useful in engineering because they offer a guarantee of statistical performance through repeated use. However, it is difficult to employ them consistently in analyses and assessments because it is not clear how to propagate them through mathematical calculations. Confidence structures (c-boxes) generalize confidence distributions and provide an interpretation by which confidence intervals at any confidence level can be specified for a parameter of interest. C-boxes can be used in calculations using the standard methods of probability bounds analysis and yield results that also admit the confidence interpretation. Thus analysts using them can now literally compute with confidence. We illustrate the calculation and use of c-boxes for some elementary inference problems and describe R functions to compute them and some Monte Carlo simulations demonstrating the coverage performance of the c-boxes and calculations based on them.

Keywords. confidence intervals, confidence structures, c-boxes, p-boxes, probability bounds analysis, binomial probability, imprecise beta model, *t*-distribution

1 Introduction

When frequentist confidence intervals are constructed across many separate data analyses based on different experiments, the proportion of such intervals that contain the true value of the parameter will match[1] the confidence level, which can be specified in advance to produce any statistical performance that may be desired. Such a guarantee is very attractive to engineers because it allows them to ensure that their conclusions based on confidence intervals will perform according to a specified standard. Bayesian methods in general lack such guarantees that could ensure statistical performance over the long run, and this fact may explain much of the reticence among engineers about adopting the Bayesian framework (Mayo 1996; cf. Vick 2002). On the other hand, Bayesian methodology allows convenient use of its posterior estimates in subsequent calculations, which is usually quite difficult with confidence intervals because it is not clear how knowledge of confidence intervals for parameters can be translated into a confidence interval for an arbitrary function of those parameters using traditional methods.

This paper introduces the notion of confidence structures, or c-boxes. These structures are defined by a traditional confidence interpretation yet admit computations that produce results that also have the confidence interpretation. The next section briefly reviews confidence distributions, which c-boxes generalize. The following sections informally describe c-boxes, give some numerical examples, and compare one of these examples with Walley's imprecise beta model. The paper includes a discussion of the prospects of using c-boxes to compute with confidence, both literally and figuratively, including how to project c-boxes characterizing parameters to estimate the distributions of observable random variates from distributions that depend on those parameters. We provide software functions to compute c-boxes for several important cases and simulate their coverage properties by Monte Carlo methods. Such simulations are useful to determine whether and how conservative the c-boxes are, and thus how useful they are likely to be in practice.

[1] That is, the average frequency of coverage will be at least the specified confidence level.

2 Confidence and Confidence Distributions

The notion of a confidence interval was introduced by Neyman (1937). A confidence interval for parameter θ with coverage α has the property that, among all confidence intervals computed by the same method, at least a proportion α will contain the true value of θ. A confidence interval can serve as an estimate of the parameter that is more sophisticated than any point estimate could be because it encodes not only the available data but also the sampling uncertainty they imply. Valid confidence intervals are more than merely subjective characterizations of uncertainty; they represent rigorous claims and their use establishes a standard of statistical performance that in principle can be checked empirically with Monte Carlo simulations. Credible intervals (sometimes called Bayesian confidence intervals in a usurpation of language) are often considered to be the Bayesian analogs of confidence intervals (Lee 1997), but credible intervals have no general accompanying guarantee like that of the frequentist notion.

Confidence distributions were introduced by Cox[2] (1958), but received little attention in the literature until a recent spike of interest (Efron 1998; Schweder and Hjort 2002; Singh et al. 2005; Xie et al. 2011; Xie and Singh 2012; inter alia). A confidence distribution is a *distributional estimate* for a parameter, in contrast with a point estimate like a sample mean or an interval estimate such as a confidence interval. It has the form of a distribution function on the space of possible parameter values that depends on a statistical sample in a way that encodes confidence intervals at all possible confidence levels. A confidence distribution for a parameter $\theta \in \Theta$ is a function C: $\Theta \rightarrow (0,1)$ such that, for every α in (0,1), $(-\infty, C^{-1}(\alpha)]$ is an exact lower-sided 100α% confidence interval for θ , where the inverse function $C^{-1}(\alpha) = C_n^{-1}(x_1,\ldots,x_n,\ \alpha)$ is increasing in α. This definition obviously also implies $[C^{-1}(\alpha),\ C^{-1}(\beta)]$ is a 100($\beta-\alpha$)% confidence for the parameter θ. Although related to many other ideas in statistical inference (Singh et al. 2005; Xie et al. 2011), a confidence distribution can be considered a purely frequentist concept (Schweder and Hjort 2002; Singh et al. 2005).

An important example of a confidence distribution is for the parametric mean of a normal distribution based on random sample data x_i, i = 1, 2, …, n. The confidence distribution in this case is

$$C_n(\mu) = F_{Tn-1}((\mu - \bar{x})\sqrt{n/s})$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, and F_{Tn-1} denotes the cumulative distribution function of Student's t-distribution with n–1 degrees of freedom. Confidence intervals for the normal's mean can be constructed directly from this confidence distribution as the inverse image of any subset of the confidence distribution's range that has measure equal to the intended confidence level. In particular,

$$[C_n^{-1}(\alpha),\ C_n^{-1}(\beta)] = \bar{x} + s\ [F_{Tn-1}^{-1}(\alpha),\ F_{Tn-1}^{-1}(\beta)] \ / \sqrt{n}$$

is a 100($\beta-\alpha$)% confidence interval on the mean. For the sake of clarity and convenience for readers, these formulas can be rendered as code for the R statistical computing language (R Development Core Team 2011):

```
pcnorm.mu = function(mu, x)
  pt(sqrt(length(x))*(mu-mean(x))/sd(x),length(x)-1)

cinorm.mu = function(x, c=0.95, alpha=(1-c)/2, beta=1-(1-c)/2)
  mean(x)+qt(c(alpha,beta),df=length(x)-1)*sd(x)/sqrt(length(x))
```

The function pxnorm.mu accepts random normal sample values in the x array and returns the value of the confidence distribution for every value in the mu array. The cinorm.mu function also takes the random samples in the x array, and returns a confidence interval for the mean of the normal distribution that generated those sample values at a confidence level set by the argument c, which defaults to 95%, or by alpha and beta if they are specified.

A Monte Carlo simulation can be implemented using the following R function to check that the confidence distribution indeed allows valid confidence intervals at any level to be constructed from it:

```
covnorm.mu = function(n,mu,sigma,many=1e4,lots=1e3, ... ) {
  ab = alphabeta(...)
  m = seq((mu-5*sigma),(mu+5*sigma),length.out=many)
  cov = 0
  for (i in 1:lots) {
    x = rnorm(n, mu, sigma)
    h = pcnorm.mu(m, x)
    ci = range(m[(ab[1]<=h) & (h<=ab[2])])
    if ((ci[1]<=mu)&(mu<=ci[2])) cov=cov+1 }
  cat(' Intended',diff(ab)*100,'%\n','Observed',100*cov/lots,'%\n')
  cov/lots }
alphabeta = function(c=0.95,a=(1-c)/2,b=1-(1-c)/2) sort(c(a, b))
```

This function can be exercised with a call like covnorm.mu(n, μ, σ), specifying just a positive integer n and the true mean and standard deviation to use in the simulation, which will return a value around 0.95, or a call like covnorm.mu(n, μ, σ, a=α, b=β) may also specify particular α and β levels.

Although a confidence distribution has the form of a probability distribution, it is usually not considered to be a probability distribution. It corresponds to no randomly varying quantity; the parameter it describes is presumed to be fixed and nonrandom. Some also emphasize that the value of the function C is not probability of θ, but

[2] Fraser (2011) argues that confidence distributions can be found in the work of Fisher (1930; 1935) under the name 'fiducial', and even in that of Bayes (1763) namelessly.

rather confidence[3] about θ (Cox 2006; cf. Lindley 1958). A confidence distribution is merely a ciphering device that encodes confidence intervals for each possible confidence level. Nevertheless, it might be reasonable and convenient to adopt a notation that only implicitly denotes the confidence distribution, so that, for instance, in the case of the normal mean, we can write

$$\mu \sim \bar{x} + s\, T_{n-1}/\sqrt{n}$$

where T_{n-1} denotes a random variable from Student's t-distribution (Student 1908) with $n-1$ degrees of freedom. This notation avoids the need to name the confidence distribution function. Note that this use of the tilde ~ extends conventional uses in statistics. We suggest that it can still be read as "has the distribution", or perhaps "has uncertainty like", but it obviously does not suggest that the left-hand side is a random variable. The left-hand side after all is a value that is fixed, though unknown. Instead, it says that the inferential uncertainty about the fixed parameter μ is characterized by the transformed t-distribution.

Despite their intimate connection with t-distributions, confidence distributions are not widely known in statistics, at least not under that name. Efron (1998) characterized bootstrap distributions as (approximate) confidence distributions, and so confidence distributions are widely used in modern statistics, albeit under the guise of bootstrap distributions.

The notion of confidence distributions is not without critics. Early association with fiducial inference has led to some confusion. Some readers seem to have difficulty accepting confidence distributions on their own terms. The arguments of Robert (2012) are paraphrased a bit more bluntly in his blog (http://xianblog.wordpress.com/2012/06/11/confidence-distributions/): "Either the confidence distribution corresponds to a genuine posterior distribution, in which case I think the only possible interpretation is a Bayesian one. Or the confidence distribution does not correspond to a genuine posterior distribution, because no prior can lead to this distribution, in which case there is a probabilistic impossibility in using this distribution." Of course confidence distributions are not trying to be Bayesian posterior distributions, so it should hardly be disquieting if they fail to be. The requisite interpretation of confidence distributions is of course Neyman confidence, which Bayesian posteriors do not generally have.

One potential practical disadvantage of confidence distributions is that they are not unique. Multiple functions may fill the bill, and there seems to be no general way to pick the best confidence distribution from among them. Of course, confidence intervals themselves are not unique either. There are usually lots of reasonable ways to construct a confidence interval for any parameter, even for fixed data and model. Neither form of non-uniqueness seems to impede the purpose of guaranteeing long-term statistical performance.

Another significant limitation on the use of confidence distributions is that not every important inferential problem has a solution. Confidence distributions are often constructed by inverting the upper limits of lower one-sided confidence intervals of all levels, but this is not possible for all important inferential problems. Notably, in particular, *there is no confidence distribution for the binomial probability*.

3 Confidence Structures (C-boxes)

Confidence distributions are special cases of more general confidence structures (Balch 2012), which we call 'confidence boxes' or 'c-boxes' because they may often be characterized by two bounding distributions like probability boxes (Ferson et al. 2003). A c-box represents inferential uncertainty about a parameter that characterizes some distribution from which limited or poor or discrete data have been randomly sampled. Like a confidence distribution, a c-box is defined by the property that it can be used to construct Neyman confidence intervals at any confidence level for that parameter. C-boxes generalize confidence distributions because both are estimators of unobservable parameters, but c-boxes can be applied to problems with discrete observations, interval-censored data, and even inference problems in which no assumption about the distribution shape can be made.

Methods for deriving c-boxes are varied (Balch 2012). Generally, wherever a meaningful and valid confidence interval can be defined, a c-box can also be defined. If a confidence interval is based on a pivot, that pivot can be used to directly define a c-box. Any defined confidence distribution can be generalized to a c-box when its data are encoded not as point values but as intervals to account for mensurational uncertainty from the inability to measure individual quantities with perfect precision (Nguyen et al. 2012; Ferson et al. 2007). When a confidence interval is based on a significance function, i.e., a function (of parameters and data) that produces p-values in a significance test, the significance function can be used to construct a consonant confidence structure, encoded as a Dempster–Shafer structure which can then be transformed, with some loss of information (Ferson et al. 2003), into a p-box (Balch 2012).

The formula and R function for this c-box of the normal mean can be generalized for the case of interval-censored data using a straightforward but non-trivial algorithm that

[3] Of course, confidence is a probability in a different domain; confidence is the probability realized by frequency that those defined intervals $(-\infty, C^{-1}(\alpha)]$ actually enclose the parameter over some in some future, perhaps hypothetical series of experiments.

extremizes $C_n(\mu)$ over possible configurations of point x-values within their respective interval ranges (Nguyen et al. 2012; Ferson et al. 2007). In case the intervals all overlap any value of μ, the result is vacuous (i.e., the interval [0,1]) for that value. For example, if interval-censored random samples from a normal distribution are {[8,11], [5.5,6.9], [−1.3,0.3], [3.5,7.5], [0.8,1], [2.8,4.2], [1.8,5.2], [2.2,5.2], [3.5,5.7], [5.3,6.1]}, a c-box for the normal mean is shown in Figure 1.

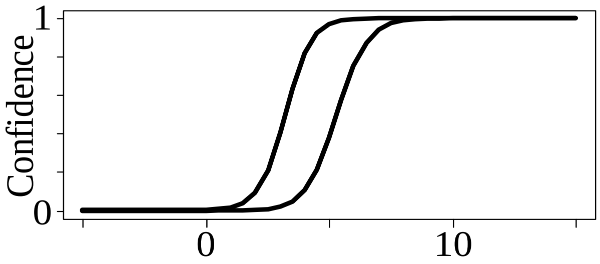


Figure 1: C-box for the normal mean from interval data.

To extract a confidence interval from a c-box, select values of α and β that imply a desired confidence level 100(β−α)%, and map these values from the confidence axis to the x-axis. The larger value β is mapped through the *right* bound, and the smaller value α is mapped through the *left* bound.

4 Computing with Confidence

Many authors (e.g., Grosof 1986) have suggested using ordinary confidence procedures to obtain interval inputs for use with interval analysis (Moore 1966) for bounding numerical results that depend on sample data. For example, EPA (2002) guidance instructs risk analysts to use the upper bound from the 95% confidence interval for a pollutant's mean concentration rather than the actual sample mean of observed concentration values in order to be protective of the public health in the face of sampling uncertainty arising from sometimes very small sample sizes. Although this may be a reasonable strategy when there is only a single variable for which sampling uncertainty is a major concern, it is not statistically defensible when such uncertainties for several variables must be combined together. Statistical confidence intervals are not rigorous intervals guaranteed to enclose the value they estimate, and therefore confidence intervals do not formally admit interval calculation in the sense of Moore (1966).

Some limited statements are possible using ad hoc application of Bonferroni or Šidák corrections or Boole or Fréchet inequalities (e.g., Ferson 1996). For example, if we combine, say by addition, two 95% confidence intervals using simple interval arithmetic, we might expect the result to be a ~90% confidence interval for the sum because the conjunction of the two probability statements would imply multiplying the two probability levels, at least assuming independence between them. If seven such confidence intervals were combined in some mathematical function, the implied probability level under independence would be less than 70%. Without the independence assumption, the level could fall as low as 65%. To achieve 95% confidence for the result, one would presumably have to use input confidence intervals with confidence level equal to the seventh root of 95%, which is greater than 99%. Because confidence intervals often get substantially wider as the confidence level rises, this approach is rarely workable in practice.

The alternative approach of computing with confidence distributions is also not practical just because (precise) confidence distributions often do not exist for important problems. This limitation may be alleviated by c-boxes because they generalize confidence distributions and more easily provide solutions. Although Cox (2006) counseled that analysts should not try to use confidence distributions in calculations as though they were true probability distributions, Balch (2012) proved that two or more independent c-boxes can be propagated through a function to yield a valid c-box. This is much more efficient than propagating individual confidence intervals because the combinations do not require application of the Bonferroni or Šidák corrections and they deliver results at all confidence levels all at once.

For example, suppose one were interested in computing a 95% confidence interval on the mean difference between two normal populations with both unknown mean μ and unknown standard deviation σ. Suppose we collect four random samples from each population, say, {2.71, 5.46, 5.45, 5.50}, and {1.88, 1.54, 1.15, 0.46}. One approach to obtaining the desired interval would be to take the interval-difference of the 97.468% confidence intervals on the two population means. The resulting estimate would be $\mu_2 - \mu_1$ = [0.37, 6.67] with 95% confidence. Alternatively, one could take the stochastic difference of the two c-boxes on the uncertain means which are (shifted and scaled) t-distributions. This yields a much tighter 95% central confidence interval on the difference, [1.10, 5.94], although it is somewhat more difficult to compute because it involves a subtractive convolution rather than merely an interval difference. Still, it can be calculated via Monte Carlo simulation in R using only three lines:

```
rcnorm.mu = function(m, z)
   mean(z)+sd(z)*rt(m, length(z)-1)/sqrt(length(z))
d = sort(rcnorm.mu(m, x) - rcnorm.mu(m, y))
range(d[round(c(0.025*m, (1-0.025)*m))])
```

where x and y are the vectors of sample values, m is the number of Monte Carlo simulations. In fact, this result is the same as the 95% credible interval that would be obtained using Bayesian inference with a Jeffreys prior. The convolution of the confidence distributions yields confidence intervals by a purely frequentist analysis that supports a traditional confidence interpretation in this and other cases generally. The following R function can be used to implement straightforward Monte Carlo simulations that demonstrate the confidence intervals produced by this approach have the prescribed coverage:

```
covnorm.mudiff=function(n,mu,sigma,many=1e4,lots=1e3,...){
 ab = alphabeta(...)
 truediff = mu[1] - mu[2]
 cov = 0
 for (i in 1:lots) {
  x = rnorm(n[1], mu[1], sigma[1])
  y = rnorm(n[2], mu[2], sigma[2])
  ci=range(sort(rcnorm.mu(many,x)-
            rcnorm.mu(many,y))[round(many*ab)])
  if ((ci[1] <= truediff) & (truediff <= ci[2])) cov = cov + 1 }
 cat(' Intended',diff(ab)*100,'%\n','Observed',100*cov/lots,'%\n')
 cov/lots }
```

This function can be called like covnorm.mudiff(n, μ, σ), where n, μ and σ are now each *pairs* describing the sample sizes and parameters for the two populations. For instance, covnorm.mudiff(c(10,20),c(5,1),c(2,3)) will return a value around 0.95.

5 C-box for the Binomial Probability

A Bernoulli random variable has only two possible values, perhaps designated {failure, success}, or more conveniently {0, 1}. A binomial random variable is a random variable whose value is a count of Bernoulli successes observed over $n > 0$ independent identical trials, each of which has the same probability p of success, which produces k successes from those n trials (where $0 \leq k \leq n$). A fundamental problem in risk analysis and statistics generally is to characterize what can be inferred about p from observing k successes out of n trials, under the assumption that the trials are independent and the binomial probability p is fixed across the trials.

In fact, the original problem in the famous paper of Bayes (1763) was about the estimation of the binomial probability. The paper begins "*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named" (Bayes 1763, page 376). The same page also says "By *chance* I mean the same as probability." We take this to be asking, given k successes and $n-k$ failures out of n trials where $k \sim$ binomial(n, p), what is $\Pr(p \in [p_1, p_2])$, for any values p_1 and p_2?

Balch (2012) offers a c-box solution to this problem:

$p \sim$ [beta(k, $n-k+1$), beta($k+1$, $n-k$)],

where p is the binomial parameter (which is a fixed but unknown value), and the two beta distributions are the left and right edges of the c-box that characterizes the inferential uncertainty about p. Note that we continue to use the ~ symbol even though the right-hand side has the form of a p-box. The ~ can be read as "has uncertainty like". We understand this to entail that the parameter on the left-hand side has inferential uncertainty characterized by a confidence distribution consistent with or inside the c-box, that is, a distribution that is bounded in the cumulative by the two edge distributions of the c-box.

Figure 2 depicts an example using $k = 2$ and $n = 10$ in a graph whose abscissa consists of the possible values of the parameter p and whose ordinate is confidence (probability).

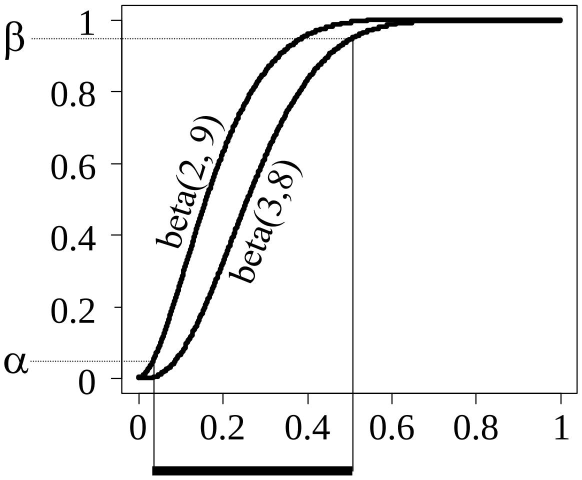


Figure 2: C-box and a 100(β–α)% confidence interval for probability from 2 successes in 10 trials.

The c-box in Figure 2 has a confidence interpretation, which means that one can generate from it true confidence intervals for the binomial probability p at any desired level of confidence. For example, the depicted interval is the symmetric 90% confidence interval [0.037, 0.507]. The confidence intervals obtained in this way are identical to the classical Clopper–Pearson (1934) confidence intervals on the binomial probability. One-sided confidence intervals can be obtained by setting α to zero or β to one. The c-box approach readily provides results for cases involving $k = 0$ and $k = n$, and even the no-data case where $n = 0$, without the overthinking required by a Bayesian analysis constrained to a single precise distribution (Winkler et al. 2002).

Of course the Bayesian and frequentist approaches are trying to do different things. In the c-box approach, p_1 and p_2 are sought to be functions of the data and probabilities are conditional on some hypothetical (but unknown) value of p. In contrast, Bayes explicitly conditions on the data, and asks about the probability of p as a latent variable. These approaches are asking a very different questions: c-boxes ask about coverage for a fixed value of *p*, whereas Bayes is asking about the probability of *p* as a latent random variable.

The c-box and arbitrary confidence intervals for the binomial probability given *k* successes out of *n* trials can be computed in R with the functions:

```
pcbinom.p = function(p, k, n)
  list(left=pbeta(p, k, n-k+1), right=pbeta(p, k+1,n-k))

cibinom.p = function(k, n, c=0.95, alpha=(1-c)/2, beta=1-(1-c)/2)
  qbeta(c(alpha,beta), c(k,k+1), c(n-k+1,n-k))
```

Straightforward Monte Carlo simulation can demonstrate the confidence intervals perform statistically.

Note that the c-box also answers Bayes' question about the chance *p* is in some range, but it gives an interval rather than a single precise probability. The c-box says $\Pr(p \in [p_1, p_2]) \in [\min(0, B_R(p_2) - B_L(p_1)),\ B_L(p_2) - B_R(p_1)]$, where B_L denotes the cumulative beta distribution with parameters *k* and *n–k*+1, and B_R is the cumulative beta with parameters *k*+1 and *n–k*. The lower bound can be called confidence, and the upper bound plausibility, and together they characterize the chance sought by Bayes.

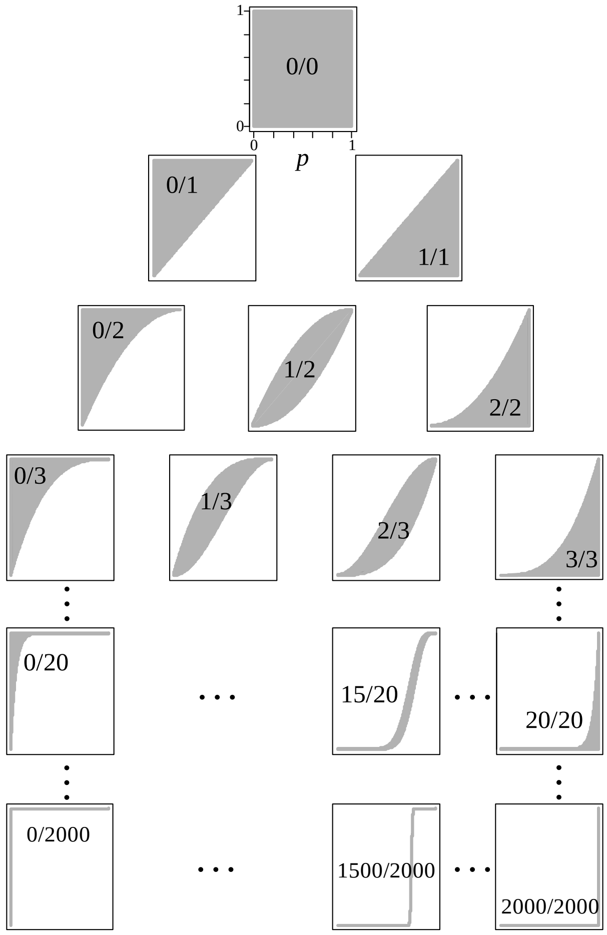


Figure 3: C-boxes for the binomial probability implied by *k*/*n* successes out of trials.

Figure 3 shows the first few c-boxes for sample sizes between zero and three. Notice that the c-box for the null case when *n* = 0 corresponds to the entire unit square. Thereafter, the possible c-boxes for any given sample size partition the unit square. As sample size increases, of course the c-box approaches a precisely specified beta distribution which becomes steeper and steeper and centered on the observed frequency *k*/*n*.

What determines whether the solution to an inference problem is a precise confidence distribution or a non-degenerate, imprecise c-box? For the normal mean the solution is precise unless the data are themselves imprecise from interval-censoring (as in Figure 1). For binomial probability, however, the solution is imprecise even for well identified data. The reason is what ecologists call "demographic" uncertainty (Akçakaya 1991), which is the variation that arises simply because of the constraint that data must come as integers. The discrete nature of binomial sampling means that evidence cannot reflect patterns as well as continuous data can. Demographic uncertainty is only important for small sample sizes, but it cannot be neglected in such cases.

5.1 Comparison with the Imprecise Beta Model

The c-box solution to the binomial probability estimation problem can be compared to the imprecise beta model (IBM) first suggested by Dempster (1966) but elaborated and championed by Walley (1991; 1996; Walley et al. 1996; Bernard 2005). The IBM employs a class of prior distributions beta(*st*, *s*(1–*t*)), $t \in [0,1]$, defined by a single, fixed value $s > 0$ that measures resistance (maybe stubbornness) of the model to new data. After observing *k* successes in *n* trials, the posterior is the class beta(*st*+*k*, *s*(1–*t*)+*n*–*k*). Extremizing *t* from 0 to 1 yields the posterior p-box [beta(*k*, *s*+*n*–*k*), beta(*s*+*k*, *n*–*k*)] whose expectation is the interval [*k*/(*s*+*n*), (*s*+*k*)/(*s*+*n*)]. As data become available and the model is updated, the left and right beta distributions incrementally converge in accordance with a rate defined by the parameter *s*. Figure 4 illustrates, for three different values of *s*, how the vacuous prior (top row) contracts to a posterior with the addition of each binary datum in the sequence {0, 0, 1, 0}. Each graph shows eleven beta distributions evenly distributed across the posterior class.

The IBM is an example of Bayesian sensitivity analysis or robust Bayes analysis (Berger 1985). It may be thought of as many simultaneous Bayesian analyses with many priors ranging between the limiting distributions beta(0,1) and beta(1,0), in which at least one posterior may be improper if *k* is equal to *n* or zero. Walley (1991) has demonstrated that robust Bayes analysis is part of a more general theory based on imprecise probabilities of very broad scope and flexibility, for which there is a firm theoretical foundation based on respecting consistency and coherence requirements but which avoids making unwarranted assumptions to obtain quantitative answers. The most important feature of the IBM is that it does not require the analyst to select some precise probability distribution as prior. The IBM instead intends to specify a reasonable class of priors. The idea is that no single distribution could be reasonable as a model of prior ignorance, but considered as a whole, the class of beta distributions with all possible means specified by IBM is arguably a reasonable model for ignorance.

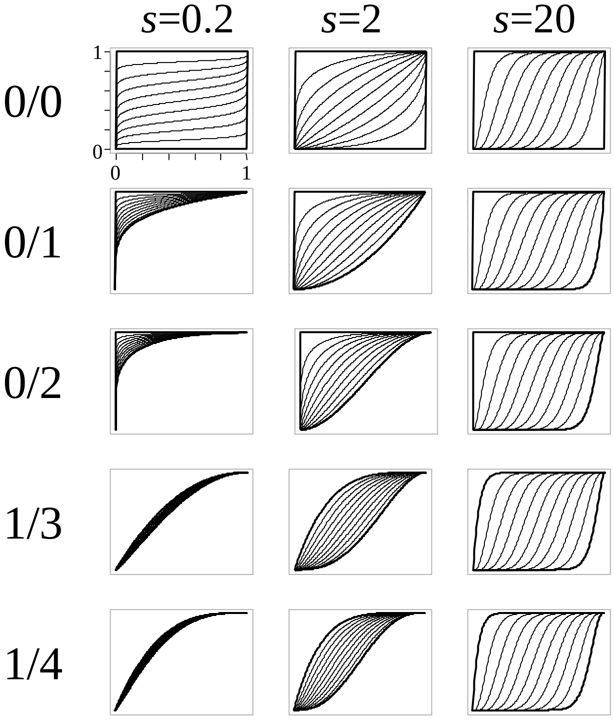


Figure 4: IBMs and their beta distributions for different values of s as data accumulates.

In the degenerate initial case, when the sample size is zero before any data are collected, the posterior is the same as the prior, and the IBM yields a vacuous posterior that effectively says the probability could be anywhere in the interval [0,1], which is arguably the only sensible inference when there are no data at all. When the sample size is very large, the posterior is a tight p-box that tends to the observed frequency, as all Bayesian analyses do. In the practical intermediate cases of small sample sizes, the posterior from the IBM is a credal set containing a range of beta distributions whose breadth reflects the uncertainty about the prior that a traditional Bayesian analysis ignores. Importantly, this breadth is not too wide to be useful, but yields answers whose imprecision is roughly what one might expect to see across a community of competent Bayesians (Walley 1991).

A user of the IBM must chose a value for the parameter s. This value determines the speed of convergence with which data cause the initially vacuous state of uncertainty to condense into the precise posterior approaching the observed frequency k/n. High values of s cause the IBM to converge slowly. For a given value of t, larger values of s cause the variance of the distribution beta(st, $s(1-t)$) to be smaller, so when the distribution is considered as an estimate of θ, larger s means there is more precision about the parameter. Walley (1996; Walley et al. 1996) recommended using $s = 1$ or $s = 2$, with preference for the larger value.

The c-box approach described in the previous section conforms with an IBM using $s = 1$, although the IBM and c-box have rather different interpretations. Walley (1996) noted the IBM's frequentist coverage characteristics, though he did not mention these coverage characteristics could be propagated through mathematical calculations based on the IBM. The most immediate difference between the IBM and the c-box approach might be that IBM users must select a value for s. Users of the c-box approach do not need to choose such a value, as the parameter is not used in the derivation of the approach.

There are also fundamental differences. The prior and posterior structures of the IBM are credal sets, but they are rather delicate credal sets in that they consist only of beta distributions with particular, constant values of s (as depicted in Figure 4). A c-box is a much coarser and fuller structure. It effectively includes all the beta distributions that are in the IBM plus infinitely many other distributions that might also be considered reasonable. The choice of the beta family is of course a result of the happenstance of mathematical conjugacy between the beta distribution and binomial sampling. One notable difference and possible conceptual advantage of the c-box approach is that it does not depend on the fiction that the appropriate prior actually or necessarily has some beta shape. Thus, in contrast with the imprecise *beta* model, one might consider the c-box solution to be an imprecise model for the binomial probability, or even *the imprecise model* for the binomial probability. Such presumptuousness in doing so might eventually be forgivable if it turns out that the c-box provides a slightly tidier solution to Bayes' original problem of estimating the binomial probability.

Perhaps more important than any tidiness or even the ability to propagate the confidence interpretation through mathematical functions is the fact that the solution strategy for the inference about binomial probability can now be contextualized as an instance of a general approach based on confidence that can be applied in many other inference problems. In contrast with the IBM, which seems to be a *sui generis* solution for one parameter of one particular sampling model, the c-box solution clearly generalizes to other problems. Balch (2012) discusses these prospects.

6 Predictive Distributions and P-boxes

If the first estimation problem given a sample of observable values $X_i \sim F(\theta)$ is to characterize the sampling or inferential uncertainty associated with a putatively fixed but unknown parameter θ governing the stochastic process that created those observable values, the second estimation problem, which is discussed in this section, is to characterize what can be inferred about a future observable value X_{n+1} that might be collected. In addition to the sampling uncertainty associated with the inference step that arises from not having measured every possible sample value, this characterization also has a component of pure aleatory uncertainty associated with the underlying stochastic process F.

The characterization is a predictive distribution, or more generally a predictive p-box. This output is analogous to a Bayesian posterior predictive distribution and related to prediction intervals common in frequentist analyses. Note that the output is a proper p-box because it is a collection of probability distributions constrained by a pair of bounding distributions. But this p-box is special in that it also inherits the confidence interpretation.

The predictive distribution or p-box can be understood to be, and evaluated as, the composition $F(C(\theta))$ of the distribution function F and the c-box C estimating the parameter θ. For example, the Bernoulli distribution can be composed with the c-box for the binomial probability to create the predictive p-box for the next randomly sampled Bernoulli deviate. For this case, the composition can be done analytically: Given a Bernoulli process generating zeros and ones where the probability of one is p which has a constant but unknown value, and n random observations of which k values are ones and $n-k$ values are zeros, the predictive p-box, i.e., the p-box estimate of the distribution for the next binary observation, is $[\mathrm{B}(k/(n+1)),\ \mathrm{B}((k+1)/(n+1))]$, where B denotes a Bernoulli distribution. Likewise, the predictive p-box for the next binomial deviate, that is, the number of ones in N Bernoulli trials, is $[\mathrm{BB}(k,\ n-k+1,\ N),\ \mathrm{BB}(k+1,\ n-k,\ N)]$ where BB denotes a beta-binomial distribution.

Straightforward Monte Carlo simulations can demonstrate that the interval $[\mathrm{BB}_1^{-1}(\alpha),\ \mathrm{BB}_2^{-1}(\beta)]$ will contain the next binomial deviate with coverage probability $\beta-\alpha$, where BB_1^{-1} and BB_2^{-1} are the quantile functions of the beta-binomial distributions $\mathrm{BB}(k,\ n-k+1,\ N)$ and $\mathrm{BB}(k+1,\ n-k,\ N)$ respectively.

When the c-box is described numerically rather than analytically, probability bounds analysis provides for numerical composition. For one-parameter distribution families, this involves discretizing the parameter's c-box $C = [C_1(\theta), C_2(\theta)]$ into to m+1 equal-confidence intervals $[C_1^{-1}(i/(m+1)),\ C_2^{-1}((i+1)/(m+1))]$, $i = 0, 1, ..., m$, where the superscripts denote appropriate inverse or quasi-inverse functions. Each of these intervals in turn define a p-box. Each of these p-boxes is the distribution function F with that interval for the parameter θ. All of the p-boxes are then aggregated using stochastic mixture which reverses the dissolution into many intervals. Equal weights are used for the mixture so long as the original discretization of the c-box was into intervals with equal partitions of confidence. (For details about this operation, see sections 2.3 and 3.2.1.6 of Ferson et al. 2003.)

7 Summary and Conclusions

This paper gives a brief introduction to a new class of estimators for a broad variety of inference problems called confidence boxes (c-boxes) that both embody a traditional confidence interpretation yet also support propagation of inferential uncertainty through mathematical operations. C-boxes can be thought of as the confluence of classical notions of confidence (Neyman 1937) embodied in confidence distributions (Cox 1958) with more recent ideas about imprecise probabilities (Walley 1991) expressed as probability boxes (Ferson et al. 2003). The paper omits the derivations of the c-box solutions described by Balch (2012), but emphasizes that their statistical performance can be checked via Monte Carlo simulations and provides R functions for this purpose.

C-boxes capture much of the flexibility of Bayesian posteriors. However, by consistently supporting a Neyman confidence interpretation, c-boxes also establish a clear connection to the underlying empirical reality, a connection which both Walley (1991) and Mayo (1998) have called for. This means that engineering and statistical calculations can be constructed using c-boxes that ensure a particular standard of performance. This approach should be useful for many applications in medical statistics, engineering in novel environments, market research, survey sampling, etc., whenever statistical performance is desired but sample data are in short supply.

In the inference for the binomial probability, the c-box is very similar to the imprecise beta model (IBM, Walley 1996). However, the c-box arises in a purely frequentist framework, and it does not refer to or depend on any priors. Its results include more than beta distributions. Unlike the IBM, the c-box approach for the binomial probability has clear connections to other inference problems such as those involving normal sampling models, and the pathway for extending these solutions to other problems is much more straightforward.

Because confidence boxes can be used in subsequent calculations involving compositions and convolutions using standard methods of probability bounds analysis, and the resulting structures also have the same Neyman confidence interpretation, analysts using c-boxes will be able, both figuratively and literally, to compute with confidence. For instance, a c-box for a parameter can be composed with the distribution function of a sample model to create a p-box that characterizes the distribution of the next sample. The result is a new type of p-box that also has the confidence interpretation. Convolutions of c-boxes yielding sums, differences or other mathematical results likewise preserve the confidence interpretation.

Point estimators ignore uncertainties altogether. Interval estimators such as confidence intervals can be unwieldy for several reasons. Even detail-rich distributional estimators like confidence distributions or Bayesian posteriors may give an incomplete characterization under demographic uncertainty when continuous parameters must be estimated from discrete data. C-boxes are more general than distributional, interval or point estimators. C-boxes can express inferential uncertainty arising from demographic uncertainty, as well as both sampling uncertainty from small sample sizes and mensurational uncertainty arising from the inability to measure quantities with infinite precision. The new estimators have the form of p-boxes, so that they may rightly be described as *p-box estimators* of parameters. C-boxes provide inferential tools to complement and support the theory of p-boxes and probability bounds analysis.

Acknowledgements

We thank Matthias Troffaes of Durham University and Jason O'Rawe, Dan Rozell and Masatoshi Sugeno of Applied Biomathematics for helpful discussions. Support was provided by the National Library of Medicine, a component of the National Institutes of Health within the United States Department of Health and Human Services, through a Small Business Innovation Research grant (RC3LM010794) to Applied Biomathematics funded under the American Recovery and Reinvestment Act. The views and opinions expressed herein are solely those of the authors and not those of the National Library of Medicine or the National Institutes of Health.

References

[1] H.R. Akçakaya (1991). A method for simulating demographic stochasticity. *Ecological Modelling* 54: 133–136.

[2] M S. Balch (2012). Mathematical foundations for a theory of confidence structures. *International Journal of Approximate Reasoning* 53: 1003–1019.

[3] T. Bayes [and R. Price] (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 53: 370–418. http://rstl.royalsocietypublishing.org/content/53/370.full.pdf. Reprinted (1958). *Biometrika* 45: 296–315.

[4] J.O. Berger (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag.

[5] J.-M. Bernard (2005). An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning* 39: 123–150.

[6] C. Clopper and E.S. Pearson (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26: 404–413.

[7] D.R. Cox (1958). Some problems with statistical inference. *The Annals of Mathematical Statistics* 29: 357–372.

[8] D.R. Cox (2006). *Principles of Statistical Inference*. Cambridge University Press.

[9] A.P. Dempster (1966). New methods for reasoning towards posterior distributions based on sample data. *The Annals of Mathematical Statistics* 37: 355–374. http://www.stat.purdue.edu/~chuanhai/projects/DS/docs/66Annals.pdf

[10] A.P. Dempster (1967). Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics* 38: 325–339.

[11] B. Efron (1998). R.A. Fisher in the 21st century *Statistical Science* 13: 95–122.

[12] EPA [U.S. Environmental Protection Agency] (2002). Calculating upper confidence limits for exposure point concentrations at hazardous waste sites. OSWER 9285.6-10, Office of Emergency and Remedial Response, Washington, DC. http://www.epa.gov/oswer/riskassessment/pdf/ucl.pdf

[13] Ferson, S. (1996). Reliable calculation in probabilistic logic: accounting for small sample size and model uncertainty. *Intelligent Systems: A Semiotic Perspective*, NIST, Gaithersburg, MD. Pp. 115–121.

[14] S. Ferson, V. Kreinovich, L. Ginzburg, K. Sentz and D.S. Myers (2003). *Constructing Probability Boxes and Dempster–Shafer Structures*. SAND2002-4015, Sandia National Laboratories, Albuquerque, New Mexico. http://www.ramas.com/unabridged.zip

[15] S. Ferson, V. Kreinovich, J. Hajagos, W.L. Oberkampf and L. Ginzburg (2007). *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*. SAND2007-0939, Sandia National Laboratories, Albuquerque, New Mexico. http://www.ramas.com/intstats.pdf

[16] R.A. Fisher (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society* 26: 528–535.

[17] R.A. Fisher (1935). The fiducial argument in statistical inference. *Annals of Eugenics B*: 391–398.

[18] B.N. Grosof (1986). An inequality paradigm for probabilistic knowledge: the logic of conditional probability intervals. *Uncertainty in Artificial Intelligence*, L.N. Kanal and J.F. Lemmer (eds.), Elsevier Science.

[19] P.M. Lee (1997). *Bayesian Statistics: An Introduction.* Arnold.

[20] D.V. Lindley (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society, Series B* 20: 102–107.

[21] D. Mayo (1996). *Error and the Growth of Experimental Knowledge*. Chicago University Press.

[22] R.E. Moore (1966). *Interval Analysis*. Prentice-Hall.

[23] J. Neyman (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society* A237: 333–380.

[24] H.T. Nguyen, V. Kreinovich, B. Wu and G. Xiang (2012). *Computing Statistics under Interval and Fuzzy Uncertainty*. Springer Verlag.

[25] R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/

[26] C.P. Robert (2012). Comments on "Confidence distribution, the frequentist distribution estimator of a parameter—a review" by Min-ge Xie and Kesar Singh. *International Statistical Review* [in press] http://arxiv.org /pdf/1206.1708.pdf. See http://xianblog.wordpress.com/ 2012/06/11/confidence-distributions/

[27] T. Schweder and N.L. Hjort (2002). Confidence and likelihood. *Scandinavian Journal of Statistics* 29: 309–332.

[28] K. Singh, M. Xie and W.E. Strawderman (2005). Combining information from independent sources through confidence distributions. *The Annals of Statistics* 33: 159–183.

[29] Student [W.S. Gosset] (1908). The probable error of a mean. *Biometrika* 6: 1–25. http://www.york.ac.uk/ depts/maths/histstat/student.pdf

[30] S.G. Vick (2002). *Degrees of Belief: Subjective Probability and Engineering Judgment*. ASCE Press, Reston, Virginia.

[31] P. Walley (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall,.

[32] P. Walley (1996). Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B* 58: 3–57.

[33] P. Walley, L. Gurrin and P. Barton (1996). Analysis of clinical data using imprecise prior probabilities. *The Statistician* 45: 457–485.

[34] R.L. Winkler, J.E. Smith and D.G. Fryback (2002). The role of informative priors in zero-numerator problems: being conservative versus being candid. *The American Statistician* 56: 1–4. See also Comments by Browne and Eddings and Reply. *The American Statistician* 56: 252–253.

[35] M. Xie and K. Singh (2012). Confidence distribution, the frequentist distribution estimator of a parameter—a review. *International Statistical Review* [in press].

[36] M. Xie, K. Singh and W.E. Strawderman (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association* 106(493): 320–333.

Entropy based classification trees

Paul Fink
Ludwig–Maximilians–Universität, Munich
paul.fink@stat.uni-muenchen.de

Richard J Crossman
University of Warwick
r.j.crossman@warwick.ac.uk

Abstract

One method for building classification trees is to choose split variables by maximising expected entropy. This can be extended through the application of imprecise probability by replacing instances of expected entropy with the maximum possible expected entropy over credal sets of probability distributions. Such methods may not take full advantage of the opportunities offered by imprecise probability theory. In this paper, we change focus from maximum possible expected entropy to the full range of expected entropy. We present an entropy minimisation algorithm using the non–parametric inference approach to multinomial data. We also present an interval comparison method based on two user–chosen parameters, which includes previously presented splitting criteria (maximum entropy and entropy interval dominance) as special cases. This method is then applied to 13 datasets, and the various possible values of the two user–chosen criteria are compared with regard to each other, and to the entropy maximisation criteria which our approach generalises.

Keywords. Imprecise probability, classification trees, nonparametric predictive inference

1 Introduction

The process of classification involves the splitting of a heterogeneous data space into homogeneous disjoint subspaces with respect to the nominal class(ification) variable C, with the aim of predicting future values of C. This is achieved by determining the splits through the values of feature/attribute variables $(X_1, \dots, X_n)$. Let C take values/categories in $\mathcal{C} = \{c_1, \dots, c_K\}$ and each X_i take values in the corresponding set $\mathcal{X}_i$, where for reasons of simplicity the feature variables are assumed to be on a nominal scale. The key consideration is how the homogeneous subspaces are to be constructed.

One method is a classification tree, which partitions the data space into orthotope shaped subspaces. The tree is grown from the root node, which corresponds to the complete data set, and ends in disjoint subsets known as leaves; this is done by recursively applying a splitting procedure. In this paper we consider only k–array splitting as in [4] which is based on Quinlan's ID3 [12] algorithm. In each step an optimal split variable with respect to an impurity criterion is evaluated, which is then assigned to the node; the data contained in the node are then split according to the values of this split variable. If no such optimal split variable may be found the node is declared as a leaf. A value of C is assigned to each leaf, this value is the most frequent category in its corresponding data subset (in the case of a tie, the most frequent category in the data subset of its parent node is used, and so on).

The optimality of a split candidate within a node is measured by the gain in a pre–specified information measure IM. Let N be the data relevant to the node. The information criterion for the node, $IM(N)$, and for each of the unassigned attribute variables X_i, $IM(N|X_i)$ (the information criterion evaluated following a split in X_i of N according to the values of X_i), are then calculated. A split is performed if $IM(N) < IM(N|X_i)$ for some X_i.

A reasonable measure is the *Information Gain*, based on Shannon's entropy [13]. Define $n^N = |N|$ and n_j^N the number of instances within N of class c_j, and furthermore denote the relative frequencies

$$p_j^N = \frac{n_j^N}{n^N}, \; p_j^{\hat{x}_i} = \frac{n_j^{\hat{x}_i}}{n^{\hat{x}_i}}, \tag{1.1}$$

with $\hat{x}_i = \{\boldsymbol{d} \in N | X_i = x_i\}$, then the information of N following a split in X_i is defined as

$$I(N, X_i) = \sum_{x_i \in \mathcal{X}_i} p(X_i = x_i) H(\boldsymbol{p}^{\hat{x}_i}), \tag{1.2}$$

where $p(X_i = x_i)$ is also estimated by relative frequencies and $H(\cdot)$ is the Shannon–Entropy defined as

$$H(\boldsymbol{p}) = -\sum_{j=1}^{K} p_j \ln(p_j), \tag{1.3}$$

for probability distribution $\boldsymbol{p}$. $H(\boldsymbol{p})$ attains its minimum (0) for some $p_j = 1$ and its maximum ($\ln(K)$) for the uniform distribution. While the probability distribution attaining the maximum is unique for fixed K, this does obviously not hold for the one attaining the minimum.

Finally, the Information Gain is defined as

$$IM(N, X_i) = H(\boldsymbol{p}^N) - I(N, X_i). \tag{1.4}$$

In determining the split variable only $I(N, X_i)$ in (1.4) is relevant. Maximising (1.4) implies minimising $I(N, X_i)$ which requires minimising entropy.

Up to this point the probabilities $p_j^{\cdot} = P(C = c_j|\cdot)$ were estimated by classical relative frequencies and thus too is the associated probability distribution. In [4] this single distribution is replaced by a credal set of probability distributions estimated by the Imprecise Dirichlet Model (IDM), giving intervals for $p_j^{\cdot}$ of

$$p_j^{\hat{x}_i} \in \left[\frac{n_j^{\hat{x}_i}}{n^{\hat{x}_i} + s}, \frac{n_j^{\hat{x}_i} + s}{n^{\hat{x}_i} + s} \right]. \tag{1.5}$$

Note that s influences the degree of imprecision; this parameter is commonly set to $s = 1$ or $s = 2$.

There are alternatives to the IDM; the Non-Parametric Predictive Inference (NPI) approach [6] is one. This is applied in [5] and [8] to replace the IDM with the multinomial NPI and ordinal NPI, respectively. A short introduction to this method follows.

The NPI approach is is designed to assume as little as is possible about a distribution from which observations are taken. Assume n observations $x_1, \ldots, x_n$ have been made. In the ordinal case, these are relabelled so that $x_1 < x_2 < \ldots x_n$. It is then assumed that observation x_{n+1} has probability $\frac{1}{n+1}$ of being smaller than x_1, the same probability of being larger than x_n, and the same probability of lying in any given data interval $I_{j+1} = [x_j, x_{j+1}]$ for $1 \leq j \leq n-1$ (we set $I_1 = (-\infty, x_1]$ and $I_{n+1} = [x_n, \infty)$). This is known as Hill's assumption, $A_{(n)}$ [11].

By using a latent variable approach, a category c_j in $\mathcal{C}$ can be considered as equivalent to some interval IC_j overlapping the data intervals. The interval IC_j itself is unknown (though IC_1 and IC_K have known bounds at negative and positive infinity, respectively), but its bounds must lie within data intervals which have an observation c_j as exactly one bound. Therefore each interval I_k can be said to be either entirely within IC_j, partially within it, or wholly outside it. The lower probability that $x_{n+1} \in c_j$ is then simply calculated by summing the probability mass of all intervals I_k which lie entirely within IC_j. The upper probability that $x_{n+1} \in c_j$ is calculated by summing the probability mass of all intervals I_k with a non-zero intersection with IC_j.

In the case of multinomial data, these intervals are represented as slices on a probability "wheel", with the observations that forming the interval boundaries representing the lines separating those slices. Observation x_{n+1} has equal chance $\frac{1}{n}$ of falling within any given slice on the wheel. This is referred to as the circular Hill assumption, or circular-$A_{(n)}$.

All observations of the same category are adjacent on the wheel, and any slices between those observations must be assigned to that category. Slices between two different observations can be assigned to either or both those observations, and/or to a previously unobserved condition (since slices for a given category are adjacent, a given unobserved category can be assigned to at most one such slice).

Therefore the lower probability of category j is equal to the probability mass of those slices with category j observations on either side. An exception is the case in which all observations come from a single category, one slice is left unassigned, resulting in a lower probability of $\frac{n-1}{n}$.

The upper probability is equal to the probability mass of of all those slices with category j observations on at least one side. An exception is the case in which c_j in unobserved; in this case the upper probability is equal to $\frac{1}{n}$, as only one slice can be assigned that category.

In the multinomial NPI case, then, the interval in (1.5) is replaced with

$$\left[\max\left(0, \frac{n_j^{\hat{x}_i} - 1}{n^{\hat{x}_i}}\right), \min\left(\frac{n_j^{\hat{x}_i} + 1}{n^{\hat{x}_i}}, 1\right) \right]. \tag{1.6}$$

In this paper trees are generated by the IDM and the multinomial NPI. The splitting criterion is based on an entropy interval comparison as in [8]. For the IDM, algorithms to obtain the minimum and maximum entropy already exist, as in [1] and [4]. For the multinomial NPI, a maximum entropy algorithm is given in [2], and we present a minimum algorithm in section 2. This algorithm will be employed in section 3 to define our splitting criterion. In section 4 the performance of our proposed splitting criterion is evaluated in a simulation study.

2 Minimum and maximum entropy distribution algorithm for multinomial NPI

The maximum entropy algorithm for the multinomial NPI model was already developed and discussed in [2]. Actually two versions to compute the maximum entropy are presented there. One algorithm computes the approximate maximum entropy, which is in structure and proof similar to its IDM counterpart as it assumes the obtained probabilities form a closed and convex set, whereas the other is an exact one, enforcing the restrictions of the probability wheel when assigning probability mass to unobserved categories. In the following only the exact algorithm will be applied.

We now describe an algorithm to calculate the minimum entropy distribution for the f–probability intervals, in the sense of Weichselberger [15]. The intervals for the multinomial NPI were proved to be f–probability intervals in [7].

We begin with a series of lemmas which demonstrate the algorithm's validity, and follow with a schematic outline of the algorithm itself. This algorithm has been adapted from the minimum entropy algorithm for ordinal NPI given in [8].

In what follows $\boldsymbol{L}$ is the vector of lower probabilities and $\boldsymbol{U}$ the vector of the upper probabilities for each category, and we choose elements of $\boldsymbol{L}$ to add mass to until we reach a probability distribution, $\boldsymbol{p}'$. The following four lemmas are required to prove our algorithm minimises entropy. In everything that follows in this section it is assumed that more than one category has been observed; minimising entropy in the case of only one observed category is trivial.

Lemma 1. *Let n_j denote the number of observations of category c_j. For two categories i and j such that n_i and n_j are strictly positive, $U_j - L_j = U_i - L_i = \frac{2}{n}$.*

Proof. Follows directly from the definition of the multinomial NPI model. □

Lemma 2. *Consider elements L_i and L_j, and mass $0 \leq m \leq \frac{2}{n}$. When assigning mass m to either or both of these elements, entropy is minimised by assigning m to c_i if and only if $L_i \geq L_j$, where i and j are interchangeable if $L_i = L_j$.*

Proof. The contributions of p'_i and p'_j to the entropy are $-p'_i \ln(p'_i)$ and $-p'_j \ln(p'_j)$. Note that $H_1(x,y) := -(x\ln(x) + y\ln(y))$ is a concave function in the domain $(x,y) \in [0,1]^2$. Therefore, for any $0 \leq c \leq m$

$$H_1(p_1 + m - c, p_2 + c) \geq H_1(p_1, p_2 + m),$$
$$H_1(p_1 + m - c, p_2 + c) \geq H_1(p_1 + m, p_2),$$

and hence to minimise H_1, all mass m should be fully assigned to either L_i or L_j. The fact that it should go to the larger of these values also follows from the concave nature of the function. When $L_i = L_j$, the mass must be fully assigned to either, but it makes no difference which is chosen. □

Lemma 3. *The probability distribution $\boldsymbol{p}'$ that minimises entropy is such that $L_i < p'_i < U_i$ holds for at most one i.*

Proof. Assume the contrary, that $L_i+\epsilon_i = p'_i = U_i-\delta_i$ and $L_j + \epsilon_j = p'_j = U_j - \delta_j$ both hold, where all constants in $S := \{\epsilon_i, \epsilon_j, \delta_i, \delta_j\}$ are strictly positive. Further assume $p'_i \leq p'_j$. By the nature of the concave function H_1

$$H_1(p'_i, p'_j) > H_1(p'_i - \min\{S\}, p'_j + \min\{S\})$$

hence minimum entropy has not been achieved. This holds true of any $i \neq j$, meaning at most only one p'_i can have this property. □

Lemma 4. *No mass is assigned to unobserved categories when minimising entropy.*

Proof. By the definition of the multinomial NPI model, $n_i = 0 \Leftrightarrow U_i - L_i = \frac{1}{n}$. We first prove that it is possible to avoid assigning mass to any unobserved category; this follows immediately in the non–trivial case (i.e. $n > 0$) from the definition of the multinomial NPI probability wheel.

It therefore follows that to assign mass to an unobserved category c_k, mass is being "denied" to two observed categories c_i and c_j (again, this follows from the probability wheel). Let $p'_k = m_1 + m_2$, $p'_i = U_i - m_1$, and $p'_j = U_j - m_2$, where $0 < m_1 + m_2 \leq \frac{1}{n} = U_k$. It immediately follows from Lemma 2 that entropy is minimised when $m_1 = 0$ and when $m_2 = 0$. □

Theorem 1. *Entropy is minimised in a structure defined by the multinomial NPI model by assigning the maximum possible mass to the largest element in $\boldsymbol{L}$, then the next largest, and so on until all mass is assigned. When two elements are equally large, choose one of those elements at random.*

Proof. From Lemmas 1 and 4 we will only assign mass to intervals of length $\frac{2}{n}$. Therefore we have that $p'_i \neq L_i \Rightarrow p'_i \in \{U_i - \frac{1}{n}, U_i\}$, where by Lemma 3 $p'_i = U_i - \frac{1}{n}$ holds for at most one i.

If no such i exists, then using Lemma 2 the minimisation algorithm works as follows: assign all $\frac{2m}{n}$ mass (with m an integer) to the m largest elements of L_i, choosing at random between equally large elements.

If one such i, denoted i^*, does exist, we assign $\frac{2m-1}{n}$ mass as above. It is immediately clear that i^* is such that $L_{i^*} = \max_{j\in M}\{L_j\}$ where M is the set of categories with no mass currently assigned to them. All that remains is to demonstrate that the entropy cannot be lowered further by swapping the mass assignment for category c_{i^*} with that of any category $c_k \in M^c$. However, this follows automatically by Lemma 2 for all c_k for which $L_k > L_j$. For any $L_k = L_{i^*}$, swapping as above does not change the entropy. □

Note that this algorithm does not produce the minimum entropy for a general structure. The algorithm can fail when $L_i > L_j > 0$ and $U_j > U_i$ both hold, as it is no longer the case that the stepwise assignment of mass to the largest lower bounds automatically produces the lowest entropy. It might instead be better to assign mass to smaller lower bounds in order to reach larger upper bounds than would otherwise be possible. The NPI multinomial model avoids this problem, as in that model $L_j \geq L_i \Rightarrow U_j \geq U_i$. It is worth noting that the distribution given by this algorithm is not necessarily a unique minimiser. However, the distribution will be unique up to rearranging the elements in ascending order.

Example 1. Consider the case of $K = 5$ classes with six observations $(1, 0, 2, 3, 0)$. From [5] we obtain that the minimum and maximum entropy distribution is contained within the set

$$\frac{1}{6}\left([0,2],[0,1],[1,3],[2,4],[0,1]\right).$$

Applying the exact maximum entropy algorithm as in [2] we obtain the distribution with maximum entropy already in the first step as $\frac{1}{6}(1, 1, 1, 2, 1)$.
The minimum entropy algorithm as described above obtains the following *working distributions* in each iteration step:

1. $\frac{1}{6}(0, 0, 1, 2, 0)$, 2. $\frac{1}{6}(0, 0, 1, 4, 0)$,
3. $\frac{1}{6}(0, 0, 2, 4, 0)$.

The entropy interval is then $[0.6365, 1.5607]$. Note that for a distribution over five classes the entropy must lie in the interval $[0, 1.6094]$.

3 Imprecise decision approach to classification trees

We begin by highlighting the differences between the approach in [8] and our approach here. In the former, an imprecise classification tree was defined as a set of classification trees. A decision in each node of the tree was made by comparing the obtained entropy intervals using interval dominance. A tree was then generated for each undominated split variable, hence creating an ensemble of classification trees. Therefore, the work in [8] can be seen as a generalisation of that in [3], which compares only the upper bounds of the entropy intervals, and also allows the generation of multiple trees, though only when considering potential root nodes.

Interval dominance is a strong condition, which means the method in [8] leads in general to a large ensemble of very small trees, as oppose to the smaller ensemble of larger trees created in general by the method in [3]. In particular, this means generating a single tree (and therefore generalising to Abellán and Moral's one–step classification tree algorithm [4]) will in general lead to an overly conservative classification model. In contrast, the Abellán and Moral method can allow splits based on very slight evidence, or even on contradictory evidence which the method ignores. It is not obvious, for example, that a variable with entropy range $[0.39, 0.4]$ should be considered a better choice to split upon than a variable with entropy range $[0, 0.41]$, but the splitting decision in the Abellán and Moral will do so, based just on the difference of 0.01 in the maximum entropy and ignoring the intervals' widths entirely.

Therefore, in this paper we explore whether, when constructing a single tree, there can be found an interval comparison method which is neither so strong as interval dominance, nor so weak as determining the lowest upper bound, and which generates an optimal tree. Our choice to limit consideration to single trees is for the sake of simplicity of comparison; the methods used here can easily be generalised to allow the construction of multiple trees. We refer to the trees generated for this paper as imprecise, as the splitting criterion compares entropy ranges derived from credal sets; note this is a different definition of imprecision to that given in [8]. The split criterion used in this paper is now described.

We note first that any simple comparison of intervals without additional properties is likely to involve one or more of three direct comparisons: comparing the upper bounds, comparing the lower bounds, and comparing the interval lengths. To some extent this third consideration is bound up in the first and second, since of course an interval's length is completely determined by its upper and lower bounds. It is possible that length cannot be completely dealt with by comparison of corresponding bounds, however, otherwise it would be equally easy to choose between intervals $[0.01, 0.95]$ and $[0, 1]$ as to choose between intervals $[0.11, 0.15]$ and $[0.1, 0.2]$, and this is not clearly true. On the other hand, comparing the lengths explicitly would lead to three separate comparisons, which is arguably overkill, and would require the use of three comparison functions where, for the sake of simplicity, we wish to only use two. We therefore implicitly compare interval length in the comparison of lower bounds shown below. This is done in the comparison of lower bounds rather than that of upper bounds in order to ensure our method is a generalisation of the one found in [2].

Our method of comparing entropy intervals requires two parameters set by the user, that of γ and T_0. We define

$$T = (1-\gamma)A_L + \gamma A_U, \tag{3.1}$$

where A_L and A_U reflect comparisons of the lower and upper bounds respectively (as in Definition 1 below), and $0 \leq \gamma \leq 1$. For each comparison, we choose to split only if $T < T_0$. Therefore the larger the value of T_0 chosen by the user, the less conservative the splitting criterion. Moreover, the greater the value of γ, the more weighting we place upon the comparison of the upper bounds. Therefore $\gamma = 1$ in the Abellán and Moral method, which considers only upper bounds. While in the methods in [8] and [3] the stopping rule is implicitly built-in, in our method we need one explicitly as T is a continuous function of the compared intervals. We now define A_L and A_U.

Definition 1. For the entropy interval $I = [a, b]$ over a data set, and an expected entropy interval $I_i = [a_i, b_i]$ following splitting on attribute variable X_i, we define

$$A_L = \frac{a_i - a}{b_i + |a - a_i|}, \tag{3.2}$$

and further

$$A_U = \frac{\ln(K) - b}{\ln(K) - b_i}. \tag{3.3}$$

Note that A_L is 0 when the lower bounds are equal, and grows larger (smaller) as the lower bound for I_i gets larger (smaller) in comparison to the lower bound for I. Hence a larger value of A_L represents a less desirable split, with respect to the lower bounds. Note also that A_U is equal to 1 when the upper bounds are equal, and gets smaller as the upper bound for I_i gets smaller in comparison to the upper bound for I. Hence a larger value of A_U represents a less desirable split, with respect to the upper bounds. Without any further restriction on when considering upper bound comparison A_U may take values larger than 1 for $b_i > b$, which is covered in what follows.

As noted, in Abellán and Moral's method the splitting is entirely based on the upper bounds comparison. This has the advantage that if there is a split, the maximum entropy is reduced. This property guarantees at least some subgroups which will be more homogeneous. Therefore we also only consider an attribute variable X_i as a split candidate if $b_i < b$. As T, defined by (3.1), does not satisfy this property of a decreasing expected maximum entropy in the split, we need to enforce more restrictions on our splitting criterion. Therefore we define T^* as follows, dealing with the above mentioned case and interval dominance.

Definition 2. For the entropy interval $I = [a, b]$ over a data set, and an expected entropy interval $I_i = [a_i, b_i]$ following splitting on attribute variable X_i, we define the combined splitting criterion

$$T_i^* = \begin{cases} 1 & \text{if} \quad b_i \geq b \\ T & \text{if} \quad b > b_i \geq a \\ T-3 & \text{if} \quad a > b_i \end{cases}. \tag{3.4}$$

This ensures that T and therefore A_U is only calculated in situation when $A_U < 1$. Thus in situations when T is actually evaluated it holds that $T \in [-1, 2)$. In the case $a > b_i$ we have I_i interval dominating I. Without the above definition, we would lack the ability to compare among interval dominating split candidates. As $T \in [-1, 2)$ for $b > b_i$ by subtracting three we obtain an always smaller value of T^* for interval dominating split candidates than for those situations where interval dominance does not occur, which allows us to consider both dominated intervals and undominated intervals via the same measure.

The fact that A_L and A_U, along with T, increase as the corresponding bound comparisons become less supportive of a split justifies the choice to split only when $T^* < T_0$. The variable X_{i^*} is chosen to split upon if it is the variable amongst the split candidates with the smallest value of T^*. With T_0 we are able to enforce a specific degree of support for a split. Note that for the Abellán and Moral method, A_L is ignored and A_U is required to be less than one, so the Abellán and Moral method is a special case of our method, with $(\gamma, T_0) = (1, 1)$. A splitting method requiring interval domination may be obtained by setting $T_0 = -1$. With our approach we are able to flexibly adapt the splitting criterion to situations where splits only in case of interval dominance or according to the Abellán and Moral method are favourable.

Although T_0 and γ were said to be chosen by the user in advance, when it is uncertain which actual splitting method to favour, they may be set data-driven, essentially functioning as so called tuning parameters.

4 Simulation

In order to evaluate the performances of the splitting criterion proposed in this paper, simulations were carried out on real–world data sets. The simulation was performed with two major questions in mind: Firstly, what is the general performance of the proposed splitting criterion and secondly, how does varying the tuning parameters T_0 and γ affect it.

For that purpose 13 different databases from the UCI repository of machine learning [10] were analysed. For each database one classification variable was predicted with the exception of the *Pittsburgh Bridges* database, where five classification variables were independently predicted[1]. Table 1 outlines the number of instances (N), number of continuous and nominal attribute variables (Num and Nom) and total missing values (NA), along with the ranges of the different states of the classification variable (K) and the predicting variables (R).

Database	N	Num	Nom	NA	K	R
abalone	4177	7	1	0	28	3-5
anneal	798	6	32	0	5	1-8
cmc	1473	2	7	0	3	2-5
credit	690	6	9	67	2	2-14
ecoli	336	7	0	0	8	2-5
hepatitis	155	6	13	167	2	2-5
lenses	24	0	4	0	3	2-3
monks1	432	0	6	0	2	2-4
bridges (deck type)	108	1	7	52	2	2-5
bridges (material)	108	1	7	48	3	2-5
bridges (span)	108	1	7	62	3	2-5
bridges (rel. span)	108	1	7	51	3	2-5
bridges (type)	108	1	7	48	7	2-5
po	90	0	8	3	4	2-4
soybean	683	1	34	2337	19	2-5
spect	267	0	22	0	2	2-2
zoo	101	0	16	0	7	2-6

Table 1: Database Overview

In a data pre–processing step any missing values were replaced by the mean or mode for continuous and nominal attributes respectively [2]. Discretisation was applied to the continuous variables by splitting them into five ideally equal frequency intervals, according to the quantiles [3]. Any variables with less than five unique values were not further discretised. Despite being commonly used in such situations, Fayyad and Irani's popular discretisation method [9] was rejected, as for some databases it returned for a notable proportion of predicting variables just one class, essentially removing those variables from the scope of predicting variables. In contrast to previously mentioned decision in the leaves, when there were ties in the most frequent categories, all of those most frequent categories were returned, thus allowing the classification tree to be *imprecise* in the prediction as well. The simulation was completely performed with the open–source statistical programming language R [14].

For each database different configurations of the splitting criterion were analysed: γ was varied over the range $[0, 1]$ and T_0 over $[-1, 1]$. As the configuration $(1, 1)$ corresponds to the maximum frequency criterion of Abellán and Moral, our criterion is implicitly compared to it. Furthermore the case of interval domination is included as T_0 is set to -1 in some configurations. For each setting 50 bootstrap samples were generated and the achieved accuracy on each was reported. On the training data two imprecise classification trees were grown. Both trees employ our proposed splitting criterion, but the underlying models to obtain the set of probability distributions differ: one employs the multinomial NPI and the other a local IDM. The accuracy of the trees was measured in terms of correct classification rate on the determinately predicted instances on the test set [4]. The correct classification rate was evaluated for each tree type on their determinate test data's observations.

To assess the first motivation of the simulation, for each database the optimal configuration of (γ, T_0) is chosen according to the average correct classification rate over the bootstrap sample. However, configuration $(1, 1)$ was not taken into account when evaluating the optimal configuration, because it serves as reference. According to the Wilcoxon signed rank there was a significant difference on a significance level of $\alpha = 0.05$ in the achieved accuracy in favour of our proposed splitting method when comparing it to the Abellán and Moral tress for both the NPI and the IDM approach.

As for the second aspect, there are differences present between the databases, even for the underlying estimation model. For all databases it was found that varying γ resulted in notable variation; only the dataset *po* demonstrated results independent of γ. In general, varying T_0 resulted in very little variation. Overall, the observed behaviour seems reasonable as a change in the weighting may change our splitting criterion drastically, while a change in T_0 only defines the cut point of the splitting criterion when we have non–interval dominating split candidates in a node. Overall, with our method we are not able to advocate a globally optimal γ as it appears database dependent. For the *Pittsburgh bridges (material)* database

[1] This means effectively splitting the database into 5 new databases.

[2] Following the data set description of the *annealing* database, the missing values were considered to be a category in themselves.

[3] Ideally in the sense that no overlapping of categories was permitted and so some categories attained larger/smaller frequencies.

[4] Whenever an observation leads to a prediction of a single class, this observation is said to be determinate, in all other cases, whether two or more classes, it is said to be indeterminate.

low values in γ led to higher accuracy, whereas for *anneal* and *hepatitis* the accuracy was greater for larger values of γ; these comparisons are with respect to the correct classification rate on the IDM–based trees, but similar examples may be found for those based on the NPI method.

Interestingly, there is also a substantial difference between the two tree types: for instance for the *ecoli* data set a high valued γ performs better for the IDM–based trees, but the opposite is true for the NPI–based trees. Moreover on this database for the IDM–based trees the accuracy is higher for $T_0 < 0$ as in comparison to $T_0 > 0$, but for the method based upon NPI the opposite holds.

To further outline the difference between the splitting methods, the performance of each configuration was compared to the one achieved by using the Abellán and Moral splitting. Therefore a Wilcoxon signed rank test was carried out. For most databases there was no significant difference between them for most configurations. However, on the *anneal* and *Pittsburgh Bridges (T or D)* datasets, most configurations achieve a significantly lower accuracy, whereas for the *cmc* and *Pittsburgh Bridges (material)* datasets, with some configurations we are able to significantly improve the accuracy with our splitting criterion.

Furthermore, if there were any significant differences present for a database those were all in the same direction, in the sense that accuracy was either non–increasing or non–decreasing with respect to γ and T_0, with the exception of just three occurrences (two in *soybean* and one in *hepatitis*).

As the previously mentioned difference between the tree–types with respect to changes in γ and T_0 may suggest, substantial differences also exist when comparing variations in those values with the fixed values used in the Abellán and Moral method. However, a significant difference in a certain configuration for the IDM–based tree does not necessarily imply one for the NPI–based and vice versa. On the other hand, for most databases, if there are significant differences present, they are in the same direction, i.e both are greater/less. Exceptions are the *spect* and *zoo* where on some configurations the accuracy is significantly improved using the IDM, but for the NPI on some (other) configurations we are predicting significantly worse.

In general, taking all databases into account, there is only a small difference between our splitting criterion and the Abellán and Moral one. On some databases we are able to improve the achieved accuracy with a certain database specific configuration of γ and T_0, while on others we are losing some accuracy for some settings. However, in most cases there is a significant difference between the Abellán and Moral splitting approach and our more general (and also more complicated) approach. The choice of the underlying probability model naturally influences the obtained results. Our results concur with [2] in that we find no significant difference between the NPI– and IDM–based trees when comparing them according to their best performance on each database. However, the NPI approach has a slightly poorer performance with our method in comparison to the Abellán and Moral splitting criterion. Generally, we are not able to identify an overall optimal configuration of (γ, T_0). This difficulty in predicting the effects of a change in parameter casts doubt on the ability of users to sensibly choose parameter values for the current model.

In our simulation we did not consider a comparison of our method to the underlying ID3 splitting mechanism. As [4] pointed out in their simulations, their splitting method has the ability to successfully compete against the even more advanced splitting algorithm C4.5.

5 Conclusions and further aspects

In this paper an approach to building classification trees using entropy range comparisons was outlined and tested. This process required the creation of an entropy minimisation algorithm, which was presented here for the multinomial NPI method. This algorithm was then used to compare trees built using the splitting criterion suggested in [4], which considers only the upper bounds of the entropy interval, and our method, which compares both upper and lower bounds of the entropy interval, with a user–defined weighting on these two comparisons determining which is the more important. A second user–defined criterion determines the amount of dissimilarity between entropy intervals necessary to justify a split; the ranges of these two user–defined criteria means our model includes both that described in [2] (which applies the model in [4] to the NPI case) and that described in [8] (in which interval dominance is required to allow splitting). These methods were compared over 13 datasets, and the resulting simulation bore interesting results. Whilst it is not the case that there exists a specific combination of user–defined criteria that improves upon consideration of the upper bounds alone, it is possible in many cases to find a combination that does improve upon that method for the specific dataset. Moreover, our results support the hypothesis that in situations in which comparison of upper bounds strongly support splitting it can make a noticeable difference to accuracy whether or not splits are allowed for variables with associated intervals which have higher lower bounds than the interval for the dataset.

Therefore it can be stated that our method has the potential to improve accuracy, but more work is required in determining under what circumstances this is the case. Related to this, further work is required in justifying this method or one similar to it through a decision theoretical foundation.

It also remains to be explored how our method performs in comparison with [3] when the former is used to generate ensemble trees. As mentioned in the text, reducing to the case of a single tree allows for quicker and more easily interpreted comparisons, but our method was created with ensemble trees in mind, and this should be considered further.

To allow for a comparison with precise classifiers future simulations will also include a precise classifier. Furthermore an investigation about the tree's actual length for the optimal configuration is worth carrying out. Larger trees, especially with ensembles in mind, induce a higher computational cost, even if it decreases in the future with more powerful hardware architecture.

Appendix

Algorithm 1 gives an outline of the minimum entropy algorithm as proposed in section 2.

Algorithm 1 Minimum Entropy Algorithm for NPI

```
Input:    F-probability intervals $[l_i, u_i]_1^n$
          as generated by the NPI
Output:   A probability distribution
          $\hat{p} = (\hat{p}_1, \hat{p}_2, \dots, \hat{p}_n)$

Helping functions:
 Sum(x):              returns the sum of
                      the elements of array x
 getMaxIndex(x, S):   returns the first index
                      of the maximum value
                      of the array x considering
                      only indices in S

Initialization: S ← $1, \dots, n$

$minEntropyNPI(l, u, \hat{p})$ {
    for ($i = 1$ to $n$) do {$\hat{p}_i \leftarrow l_i$}
    $mass \leftarrow 1 - Sum(\hat{p})$
    while ($mass > 0$) do {
        $index \leftarrow getMaxIndex(\hat{p}, S)$
        $d \leftarrow u_{index} - \hat{p}_{index}$
        if ($d \leq mass$) then {
            $\hat{p}_{index} \leftarrow u_{index}$
            $S \leftarrow S - \{index\}$
            $mass \leftarrow mass - d$
        } else {
            $\hat{p}_{index} \leftarrow \hat{p}_{index} + mass$
            $mass \leftarrow 0$
        }
    }
}
```

When considering its computational complexity, it mainly depends on the ordering of the $[l_i, u_i]_1^n$. The proposed algorithm requires generally the least steps, when $[l_i, u_i]_1^n$ is sorted according to decreasing l_i. Any of the popular sorting algorithms may be applied to obtain such a sorting, with complexity ranging from $O(n)$ to $O(n^2)$. The initialisation step means just copying l to p and generation of and index set. Due to the special ordering of l, $getMaxIndex$ in the **while** () **do**-loop finds the return value immediately as it is j when in the jth loop. Furthermore, because of the special representation of the multinomial NPI on a probability wheel, it is immediately clear that the **while** () **do**-loop has at maximum $\lceil \frac{n}{2} \rceil$ iterations. Therefore the algorithm without the sorting runs in linear time. Hence the computational most time intensive part is the chosen sorting algorithm.[5].

In the following the splitting procedure is outlined, considering the splitting process within a node N. Let $\mathcal{L}_N$ be the set of the attribute variables which are not used splitting variables on the path from the root node to N. Finding the optimal split requires three steps:

1. T_i^* is calculated for each $X_i \in \mathcal{L}_N$[6];
2. X_{i^*} is chosen as reasonable splitting candidate among the X_i in $\mathcal{L}_N$, where $T_{i^*}^* = \min_i (T_i^*)$;
3. A comparison of $T_{i^*}^*$ and T_0 is made. Only if $T_{i^*}^* < T_0$ is X_{i^*} chosen as the split variable, otherwise the node N is declared terminal.

Acknowledgement

We are grateful for the remarks of three anonymous reviewers and feel that some remarks need to be separately considered in further research. The work of R. Crossman has been partly supported by the Spanish “Consejería de Economía, Innovación y Ciencia de la Junta de Andalucía” under project TIC-6016.

References

[1] Joaquín Abellán. Uncertainty measures on probability intervals from the imprecise Dirichlet

[5] In the simulation the Shell-Sort algorithm was applied as it is implemented in R [14]

[6] The entropy interval I_i required to calculate T_i^* is obtained in the same way as in [8]

model. *International Journal of General Systems*, 35(5):509–528, 2006.

[2] Joaquín Abellán, Rebecca M. Baker, and Frank P.A. Coolen. Maximising entropy on the nonparametric predictive inference model for multinomial data. *European Journal of Operational Research*, 212(1):112–122, 2011.

[3] Joaquín Abellán and Andres Masegosa. An ensemble method of using credal decision trees. *European Journal of Operations Research*, 205(1):218–226, 2010.

[4] Joaquín Abellán and Serafín Moral. Building classification trees using the total uncertainty criterion. *International Journal of Intelligent Systems*, 18(12):1215–1225, 2003.

[5] Rebecca M. Baker. *Multinomial Nonparametric Predictive Inference: Selection, Classification and Subcategory Data*. PhD thesis, 2010. www.etheses.dur.ac.uk/257/.

[6] F.P.A. Coolen and T. Augustin. Learning from multinomial data: a nonparametric alternative to the imprecise dirichlet model. In *ISIPTA'05: Proceedings of the Fourth International Symposium on Imprecise Probability: Theories and Applications*, pages 125–134, Carnegie Mellon, 2005. SIPTA.

[7] Frank P. A. Coolen and Thomas Augustin. A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories. *International Journal of Approximate Reasoning*, 50(2):217–230, 2009.

[8] Richard J. Crossman, Joaquín Abellán, Thomas Augustin, and Frank P. A. Coolen. Building imprecise classification trees with entropy ranges. In F. Coolen, G. de Cooman, Th. Fetz, and M. Oberguggenberger, editors, *ISIPTA'11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, pages 129–138, Innsbruck, 2011. SIPTA.

[9] Usama M. Fayyad and Keki B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Thirteenth International Joint Conference on Articial Intelligence*, volume 2, pages 1022–1027. Morgan Kaufmann Publishers, 1993.

[10] Andrew Frank and Arthur Asuncion. UCI machine learning repository, 2010.

[11] B.M. Hill. Posterior distribution of percentile: Baye's theorem for sampling from a population. *Journal of the American Statistical Association*, 63:677–691, 1968.

[12] John R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1):81–106, 1986.

[13] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[14] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[15] Kurt Weichselberger. The theory of interval–probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24(2–3):149–170, 2000.

On Open Problems Connected with Application of the Iterative Proportional Fitting Procedure to Belief Functions

Radim Jiroušek
Faculty of Management, Univ. of Economics
Jindřichův Hradec, Czech Republic
radim@utia.cas.cz

Václav Kratochvíl
Inst. of Inform. Theory and Automation
Acad. of Sciences, Prague, Czech Republic
v.kratochvil@gmail.com

Abstract

In probability theory, Iterative Proportional Fitting Procedure can be used for construction of a joint probability measure from a system of its marginals. The present paper studies a possibility of application of an analogous procedure for belief functions, which was made possible by the fact that there exist operators of composition for belief functions.

In fact, two different procedures based on two different composition operators are introduced. The procedure based on the composition derived from the Dempster's rule of combination is of very high computational complexity and, from the theoretical point of view, practically nothing is known about its behavior. The other one, which uses the composition derived from the notion of factorization, is much more computationally efficient, and its convergence is guaranteed by a theorem proved in this paper.

Keywords. Marginal problem, belief functions, algorithm, multidimensional model, convergence.

1 Introduction

In probability theory, by a marginal problem we understand a task to find out whether there exists a joint probability measure having a given system of low-dimensional measures for its marginals, and/or the problem how to find such a joint probability measure. In statistics this problem appears, for example, as a subtask of multidimensional contingency tables analysis. In 1980s, the problem was often solved in connection with a design of probabilistic knowledge-based systems [1, 10, 13]. In these expert systems, marginal measures represent pieces of local knowledge and the looked for multidimensional measure represents a knowledge base.

For a solution of a discrete marginal problem famous Iterative Proportional Fitting Procedure (IPFP) was suggested by Deming and Stephan in 1940 [3]. Though this iterative procedure was applied to practical problems since that time, it was only in 1975 when Csiszár proved its convergence [2].

The goal of this paper is to show that an analogous iterative procedure can be, in principle, applied also for construction of a multidimensional belief function. However, as the title of the paper suggests, this application is connected with several open questions.

1.1 Notation

In this paper we use the notation from the ISIPTA 2011 paper [4]: $\mathbb{X}_N = \mathbb{X}_1 \times \mathbb{X}_2 \times \ldots \times \mathbb{X}_n$, denotes a finite multidimensional space, and its subspaces (for all $K \subseteq N$) are denoted by

$$\mathbb{X}_K = \bigtimes_{i \in K} \mathbb{X}_i.$$

For a point $x = (x_1, x_2, \ldots, x_n) \in \mathbb{X}_N$ its projection into subspace $\mathbb{X}_K$ is denoted $x^{\downarrow K} = (x_i)_{i \in K}$, and for $A \subseteq \mathbb{X}_N$

$$A^{\downarrow K} = \{y \in \mathbb{X}_K : \exists x \in A, x^{\downarrow K} = y\}.$$

By a *join* of two sets $A \subseteq \mathbb{X}_K$ and $B \subseteq \mathbb{X}_L$ we understand a set

$$A \bowtie B = \{x \in \mathbb{X}_{K \cup L} : x^{\downarrow K} \in A \;\;\&\;\; x^{\downarrow L} \in B\}.$$

Let us note that if K and L are disjoint, then $A \bowtie B = A \times B$, if $K = L$ then $A \bowtie B = A \cap B$, and, generally, for $C \subseteq \mathbb{X}_{K \cup L}$, C is a subset of $C^{\downarrow K} \bowtie C^{\downarrow L}$, which may be proper.

A *basic assignment* m on $\mathbb{X}_K$ ($K \subseteq N$) is a real valued function on $\mathcal{P}(\mathbb{X}_K)$, for which

$$\sum_{\emptyset \neq A \subseteq \mathbb{X}_K} m(A) = 1.$$

Notice that in agreement with Shenoy's papers (see e.g., [12]) we admit also negative values of a basic assignment. This is why we will call a basic assignment to be *proper* if all its values are nonnegative. If

$m(A) \neq 0$, then A is said to be a *focal element* of m. Considering two proper basic assignments m_1, m_2 on the same space $\mathbb{X}_K$, we say that m_1 is *dominated* by m_2, if for all $A \subseteq \mathbb{X}_K$: $m_1(A) > 0 \Longrightarrow m_2(A) > 0$.

Having a basic assignment m on $\mathbb{X}_K$ one can consider its *marginal assignments*. On $\mathbb{X}_L$ (for $L \subseteq K$) it is defined (for each $\emptyset \neq B \subseteq \mathbb{X}_L$):

$$m^{\downarrow L}(B) = \sum_{A \subseteq \mathbb{X}_K : A^{\downarrow L} = B} m(A).$$

Each basic assignment m on $\mathbb{X}_K$ can uniquely be represented by its *commonality function*, which is a set function $Q : \mathcal{P}(\mathbb{X}_K) \longrightarrow [0, +\infty)$ defined for each $A \subseteq \mathbb{X}_K$

$$Q(A) = \sum_{A \subseteq B \subseteq \mathbb{X}_K} m(B).$$

Recall the formula from [11] yielding for each commonality function the respective basic assignment:

$$m(A) = \sum_{A \subseteq B \subseteq \mathbb{X}_K} (-1)^{|B \setminus A|} Q(B)$$

for each $A \subseteq \mathbb{X}_K$.

1.2 Operators of composition

In this paper we will take advantage of the fact that the probabilistic IPFP can easily (and elegantly) be expressed with the help of the so called operator of composition [5] that was defined in ISIPTA paper [8] also for belief functions. In [6] (see also an extended version of this conference contribution, which is to appear in IJAR [7]) it was shown that the operator of composition can also be defined within the Shenoy's valuation based systems (VBS) [12] that, as a generic uncertainty calculus, covers not only probability theory but also some other uncertainty calculi like Spohns epistemic belief theory, Dempster-Shafer belief function theory, and others.

In VBS's the operator of composition is derived from the operation of *combination* $\oplus$ and its inverse operation called *removal* $\ominus$. For two basic assignments m_1, m_2 on $\mathbb{X}_K$, $\mathbb{X}_L$, respectively, the operator of composition is defined as

$$m_1 \triangleright m_2 = m_1 \oplus m_2 \ominus m_2^{\downarrow K \cap L}, \qquad (1)$$

from which one immediately sees its semantics: we combine knowledge contained in m_1 and m_2, and to prevent double counting of knowledge when double counting matters, we remove the knowledge contained in $m_2^{\downarrow K \cap L}$.

In Dempster-Shafer theory, the role of this general operator of composition $\oplus$ is played quite naturally by the Dempster's rule of combination $\oplus_D$. Thus, for m_1, m_2 on $\mathbb{X}_K$, $\mathbb{X}_L$, respectively, for each nonempty $A \subseteq \mathbb{X}_{K \cup L}$

$$\begin{aligned}(m_1 \oplus_D m_2)(A) \\ = \Gamma^{-1} \sum_{B \subseteq \mathbb{X}_K, C \subseteq \mathbb{X}_L : B \bowtie C = A} m_1(B) \cdot m_2(C),\end{aligned}$$

where Γ is the normalization factor

$$\Gamma = \sum_{B \subseteq \mathbb{X}_K, C \subseteq \mathbb{X}_L : B \bowtie C \neq \emptyset} m_1(B) \cdot m_2(C).$$

It is not an easy task to specify in terms of basic assignments the removal operator that should be an inverse to the Dempster's rule of combination. Therefore we take advantage of the fact famous from [11] saying that the commonality function $(Q_1 \oplus_D Q_2)$ corresponding to the basic assignment $(m_1 \oplus_D m_2)$ can easily be got as the pointwise product of commonality functions Q_1 and Q_2 corresponding to basic assignments m_1 and m_2, respectively. More precisely

$$(Q_1 \oplus_D Q_2)(A) = \Gamma^{-1} Q_1(A^{\downarrow K}) \cdot Q_2(A^{\downarrow L}),$$

where Γ is again a normalization constant, which is now computed

$$\Gamma = \sum_{A \subseteq \mathbb{X}_{K \cup L}} (-1)^{|A|+1} Q_1(A^{\downarrow K}) \cdot Q_2(A^{\downarrow L}).$$

From the definition of the combination operator for commonality functions, one can immediately see that the inverse removal operator must be defined for all $A \subseteq \mathbb{X}_{K \cup L}$

$$(Q_1 \ominus_D Q_2)(A) = \begin{cases} \Gamma^{-1} \frac{Q_1(A^{\downarrow K})}{Q_2(A^{\downarrow L})} & \text{if } Q_2(A^{\downarrow L}) > 0, \\ 0 & \text{otherwise,} \end{cases}$$

with

$$\Gamma = \sum_{A \subseteq \mathbb{X}_{K \cup L} : Q_2(A^{\downarrow L}) > 0} (-1)^{|A|+1} \frac{Q_1(A^{\downarrow K})}{Q_2(A^{\downarrow L})}.$$

So, following the results from [7], within D-S theory the proper operator of composition is defined

$$m_1 \triangleright_D m_2 = m_1 \oplus_D m_2 \ominus_D m_2^{\downarrow K \cap L}.$$

Its main disadvantage is its great computational complexity following, among others, from the fact that we do not know other way how to compute the composition $\triangleright_D$ of basic assignments than first transforming basic assignments $m_1, m_2, m_2^{\downarrow K \cap L}$ into the corresponding commonality functions, computing $Q_1 \triangleright_D$

$Q_2 = Q_1 \oplus_D Q_2 \ominus_D Q_2^{\downarrow K\cap L}$, and afterwards transforming the resulting composed commonality function back into the corresponding basic assignment.

One of the results from [7] says that the operator of composition $\triangleright_D$ is different from the one defined in [8], which we are going to introduce now. In what follows, the operator from [8] will be denoted $\triangleright_F$.

Consider two arbitrary basic assignments m_1 on $\mathbb{X}_K$ and m_2 on $\mathbb{X}_L$ ($K \neq \emptyset \neq L$) a *composition* $m_1 \triangleright_F m_2$ is defined for each $C \subseteq \mathbb{X}_{K\cup L}$ by one of the following expressions:

[a] if $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) > 0$ and $C = C^{\downarrow K} \bowtie C^{\downarrow L}$ then

$$(m_1 \triangleright_F m_2)(C) = \frac{m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L})};$$

[b] if $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) = 0$ and $C = C^{\downarrow K} \times \mathbb{X}_{L\setminus K}$ then

$$(m_1 \triangleright_F m_2)(C) = m_1(C^{\downarrow K});$$

[c] in all other cases $(m_1 \triangleright_F m_2)(C) = 0$.

Let us note that similarly to $\triangleright_D$, also the operator $\triangleright_F$ can be expressed in the form of formula (1) but, naturally, with a different operator of combination. We will not need it in this paper, nevertheless let us mention for the interested reader that the corresponding operator $\oplus_F$ for m_1, m_2 on $\mathbb{X}_K, \mathbb{X}_L$, respectively, is defined by the following formula (for each $A \in \mathbb{X}_{K\cup L}$)

$$(m_1 \oplus_F m_2)(A) = \begin{cases} \Gamma^{-1} m_1(A^{\downarrow K}) m_2(A^{\downarrow L}) & \text{if } A = A^{\downarrow K} \bowtie A^{\downarrow L}, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\Gamma = \sum_{A \subseteq \mathbb{X}_{K\cup L}: A = A^{\downarrow K} \bowtie A^{\downarrow L}} m_1(A^{\downarrow K}) \cdot m_2(A^{\downarrow L}).$$

Returning back to the main topic of this paper, let us summarize that in this section we have introduced two operators of composition $\triangleright_D$ and $\triangleright_F$. Though they differ from each other, as expressed in the following Proposition (for proofs see [8, 7]), both of them meet the basic properties required from an operator of composition.

Proposition 1 *Let m_1 and m_2 be basic assignments defined on $\mathbb{X}_K, \mathbb{X}_L$, respectively. Then both operators of composition $\triangleright_D$ and $\triangleright_F$ meet the following properties:*

1. *$m_1 \triangleright m_2$ is a basic assignment on $\mathbb{X}_{K\cup L}$;*
2. *$(m_1 \triangleright m_2)^{\downarrow K} = m_1$;*
3. *$m_1 \triangleright m_2 = m_2 \triangleright m_1 \iff m_1^{\downarrow K\cap L} = m_2^{\downarrow K\cap L}$;*
4. *For $M \subseteq K$, $m_1 = m_1^{\downarrow M} \triangleright m_1$.*

The reader probably noticed that Property 2 guarantees that if $L \subseteq K$ then $m_1 \triangleright_D m_2 = m_1 \triangleright_F m_2$. It is really an easy task to show that the same equality holds true also when $K \cap L = \emptyset$. Nevertheless, not too much is known about other situations. It is clear that the above conditions are not necessary. Namely, the same equality holds true when one composes Bayesian basic assignments (i.e. basic assignments whose all focal elements are singletons). This is why we can formulate the first open problem.

Open Problem 1 *Is it possible to specify necessary and sufficient conditions under which $m_1 \triangleright_D m_2 = m_1 \triangleright_F m_2$?*

2 IPFP

In this section we will describe the Iterative Proportional Fitting Procedure with the help of the operator of composition. It can be applied to a system of basic assignments using any of the two operators of composition introduced in the previous section. This is why we use just the symbol $\triangleright$. It is important to realize, that for this computational process we need an operator possessing all the properties from Proposition 1, and we do not know any other operator meeting these properties.

Assume there is a system of n low-dimensional basic assignments $m_1, m_2, \ldots, m_n$ defined on $\mathbb{X}_{K_1}, \mathbb{X}_{K_2}, \ldots, \mathbb{X}_{K_n}$, respectively. During the computational process, an infinite sequence of basic assignments $\mu_0, \mu_1, \mu_2, \mu_3, \ldots$ is computed, each of them defined on $\mathbb{X}_{K_1\cup\ldots\cup K_n}$. In case this sequence is convergent, its limit is the result of this process.

Algorithm IPFP Define the starting basic assignment μ_0 on $\mathbb{X}_{K_1\cup K_2\cup\ldots\cup K_n}$.
Then compute

$$\begin{aligned} \mu_1 &= m_1 \triangleright \mu_0 \\ \mu_2 &= m_2 \triangleright \mu_1 \\ \mu_3 &= m_3 \triangleright \mu_2 \\ &\vdots \\ \mu_n &= m_n \triangleright \mu_{n-1} \\ \mu_{n+1} &= m_1 \triangleright \mu_n \\ &\vdots \end{aligned}$$

focal elements	m
$\{a_1\bar{a}_2a_3, \bar{a}_1a_2a_3\}$	0.2
$\{a_1a_2\bar{a}_3, a_1\bar{a}_2a_3\}$	0.3
$\{a_1a_2\bar{a}_3, a_1\bar{a}_2\bar{a}_3, \bar{a}_1a_2\bar{a}_3, \bar{a}_1\bar{a}_2a_3\}$	0.5

Table 1: Three-dimensional assignment m

	focal elements	values
m_1	$\{a_1\bar{a}_2, \bar{a}_1a_2\}$	0.2
	$\{a_1a_2, a_1\bar{a}_2\}$	0.3
	$\{a_1a_2, a_1\bar{a}_2, \bar{a}_1a_2, \bar{a}_1\bar{a}_2\}$	0.5
m_2	$\{a_2a_3\}$	0.2
	$\{a_2\bar{a}_3, \bar{a}_2a_3\}$	0.3
	$\{a_2a_3, a_2\bar{a}_3, \bar{a}_2a_3, \bar{a}_2\bar{a}_3\}$	0.5
m_3	$\{a_1a_3, \bar{a}_1a_3\}$	0.2
	$\{a_1a_3, a_1\bar{a}_3\}$	0.3
	$\{a_1\bar{a}_3, \bar{a}_1, a_3, \bar{a}_2\bar{a}_3\}$	0.5

Table 2: Consistent assignments m_1, m_2, m_3

$$\begin{aligned}
\mu_{2n} &= m_n \triangleright \mu_{2n-1}\\
\mu_{2n+1} &= m_1 \triangleright \mu_{2n}\\
&\vdots
\end{aligned}$$

As said in Introduction, when this algorithm is applied to probability measures, it has some nice and useful properties, most of which were proved by Csiszár in his famous paper [2]. So it is not surprising that the general properties formulated and proved here for belief functions (including the presented proofs) are based on the Csiszár's results.

Theorem 1 *If the sequence $\mu_0, \mu_1, \mu_2, \mu_3, \ldots$ computed by the Algorithm IPFP converges then the basic assignment*

$$\mu^* = \lim_{i\to+\infty} \mu_i$$

is a common extension of all $m_1, m_2, \ldots, m_n$, i.e.,

$$(\mu^*)^{\downarrow K_j} = m_j$$

for all $j = 1, \ldots, n$.

Proof. Consider any $j \in \{1, 2, \ldots, n\}$. From Property 2. of Proposition we get that m_j is marginal of all the assignments $\mu_j, \mu_{n+j}, \mu_{2n+j}, \mu_{3n+j}, \ldots$, and therefore m_j is marginal also to the limit of this subsequence

$$(\lim_{k\to+\infty} \mu_{kn+j})^{\downarrow K_j} = m_j.$$

From the basic course on mathematical analysis we know that if a sequence converges, then all their subsequences converge, too, and the limits are the same. Therefore, $(\mu^*)^{\downarrow K_j} = m_j$. □

2.1 IPFP with $\triangleright_F$

Example 1 Let us first illustrate and comment the process on a simple example. Consider a three-dimensional space $\mathbb{X}_{\{1,2,3\}}$, with $\mathbb{X}_i = \{a_i, \bar{a}_i\}$. To be sure that the considered system of two-dimensional basic assignments is consistent, i.e., that there exists their common extension, consider the three-dimensional assignment on $\mathbb{X}_{\{1,2,3\}}$ with three focal elements from Table 1. Its two-dimensional marginal assignments $m_1 = m^{\downarrow\{1,2\}}, m_2 = m^{\downarrow\{2,3\}}$ and $m_3 = m^{\downarrow\{1,3\}}$ are in Table 2.

The computational process starting with $\mu_0(A) = 1/255$ for all nonempty $A \subseteq \mathbb{X}_{\{1,2,3\}}$ is depicted in Table 3. We do not present here assignments μ_1 and μ_2, because they have 99, and 15 focal elements, respectively. Starting with μ_3 all the remaining computations concern only six focal elements represented by six rows of Table 3. Looking at this table the reader perhaps believes that the process converges, and that the limit assignment has eventually only four focal elements.

The convergence of the procedure in the previous example is not surprising because for $\triangleright_F$ we can use the ideas from the Csiszár's proof [2] to get the following theorem.

Theorem 2 *Consider a system of proper basic assignments $m_1, m_2, \ldots, m_n$ defined on $\mathbb{X}_{K_1}, \mathbb{X}_{K_2}, \ldots, \mathbb{X}_{K_n}$ and a proper basic assignment μ_0 on $\mathbb{X}_{K_1\cup\ldots\cup K_n}$. If there exists a proper basic assignment ν on $\mathbb{X}_{K_1\cup\ldots\cup K_n}$ such that ν is dominated by μ_0, and ν is a common extension of all $m_1, m_2, \ldots, m_n$, then the sequence $\mu_0, \mu_1, \mu_2, \mu_3, \ldots$ computed by the Algorithm IPFP with $\triangleright_F$ converges.*

The proof is based on the following auxiliary assertion.

Lemma 1 *Consider two basic proper assignments μ, ν on $\mathbb{X}_L$, and let $K \subseteq L$. Denote*

$$D(\nu\|\mu) = \sum_{A\subseteq\mathbb{X}_L:\mu(A)>0} \mu(A)\log\frac{\mu(A)}{\nu(A)}.$$

If ν dominates μ (i.e., $\nu(A) = 0 \Rightarrow \mu(A) = 0$) then

$$D(\nu\|\mu) = D(\nu\|\mu^{\downarrow K} \triangleright_F \nu) + D(\mu^{\downarrow K} \triangleright_F \nu\|\mu).$$

focal elements	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_{100}	μ_{1000}
$\{a_1\bar{a}_2a_3, \bar{a}_1a_2a_3\}$	0.156	0.200	0.166	0.166	0.200	0.172	0.195	0.199
$\{a_1a_2a_3, a_1\bar{a}_2a_3, \bar{a}_1a_2a_3, \bar{a}_1\bar{a}_2a_3\}$	0.043	0.040	0.033	0.033	0.031	0.027	0.004	$4 \cdot 10^{-4}$
$\{a_1a_2\bar{a}_3, a_1\bar{a}_2a_3\}$	0.146	0.146	0.300	0.211	0.211	0.300	0.293	0.299
$\{a_1a_2\bar{a}_3, a_1\bar{a}_2a_3, a_1\bar{a}_2\bar{a}_3\}$	0.153	0.153	0.124	0.088	0.085	0.079	0.006	$7 \cdot 10^{-4}$
$\{a_1a_2\bar{a}_3, a_1\bar{a}_2\bar{a}_3, \bar{a}_1a_2\bar{a}_3, \bar{a}_1\bar{a}_2a_3\}$	0.250	0.230	0.187	0.250	0.234	0.210	0.250	0.250
$\{a_1a_2\bar{a}_3, a_1\bar{a}_2\bar{a}_3, \bar{a}_1a_2\bar{a}_3, \bar{a}_1\bar{a}_2a_3, \bar{a}_1\bar{a}_2\bar{a}_3\}$	0.250	0.230	0.187	0.250	0.234	0.210	0.250	0.250

Table 3: $m_1 \triangleright_F \mu_0$

Proof.

$$
\begin{aligned}
&D(\nu\|\mu)\\
&= \sum_{A\subseteq\mathbb{X}_L:\mu(A)>0} \mu(A)\log\left(\frac{\mu(A)}{\nu(A)}\cdot\frac{(\mu^{\downarrow K}\triangleright_F \nu)(A)}{(\mu^{\downarrow K}\triangleright_F \nu)(A)}\right)\\
&= \sum_{A\subseteq\mathbb{X}_L:\mu(A)>0} \mu(A)\log\frac{\mu(A)}{(\mu^{\downarrow K}\triangleright_F \nu)(A)}\\
&\quad + \sum_{A\subseteq\mathbb{X}_L:\mu(A)>0} \mu(A)\log\frac{(\mu^{\downarrow K}\triangleright_F \nu)(A)}{\nu(A)}\\
&= D(\mu^{\downarrow K}\triangleright_F \nu\|\mu)\\
&\quad + \sum_{A\subseteq\mathbb{X}_L:\mu(A)>0} \mu(A)\log\frac{(\mu^{\downarrow K}\triangleright_F \nu)(A)}{(\nu^{\downarrow K}\triangleright_F \nu)(A)}.
\end{aligned}
$$

The last modification is based on Property 4 of Proposition.

Realize, now, that the last summation is performed over those $A \subseteq \mathbb{X}_L$ for which $\mu(A) > 0$, and therefore, due to the assumed dominance, $\nu(A) > 0$, too. Therefore, both $(\mu^{\downarrow K} \triangleright_F \nu)(A)$ and $(\nu^{\downarrow K} \triangleright_F \nu)(A)$ are computed according to case [**a**] of the respective definition getting

$$
\frac{(\mu^{\downarrow K}\triangleright_F \nu)(A)}{(\nu^{\downarrow K}\triangleright_F \nu)(A)} = \frac{\frac{\mu^{\downarrow K}(A^{\downarrow K})\cdot\nu(A)}{\nu^{\downarrow K}(A^{\downarrow K})}}{\frac{\nu^{\downarrow K}(A^{\downarrow K})\cdot\nu(A)}{\nu^{\downarrow K}(A^{\downarrow K})}} = \frac{\mu^{\downarrow K}(A^{\downarrow K})}{\nu^{\downarrow K}(A^{\downarrow K})}.
$$

So, we can proceed further in computation of $D(\nu\|\mu)$:

$$
\begin{aligned}
&D(\nu\|\mu)\\
&= D(\mu^{\downarrow K}\triangleright_F \nu\|\mu)\\
&\quad + \sum_{A\subseteq\mathbb{X}_L:\nu(A)>0} \mu(A)\log\frac{\mu^{\downarrow K}(A^{\downarrow K})}{\nu^{\downarrow K}(A^{\downarrow K})}\\
&= D(\mu^{\downarrow K}\triangleright_F \nu\|\mu)\\
&\quad + \sum_{\substack{B\subseteq\mathbb{X}_K\\ \nu(B)>0}} \sum_{\substack{A\subseteq\mathbb{X}_L\,:\,\nu(A)>0\\ A^{\downarrow K}=B}} \mu(A)\log\frac{\mu^{\downarrow K}(A^{\downarrow K})}{\nu^{\downarrow K}(A^{\downarrow K})}\\
&= D(\mu^{\downarrow K}\triangleright_F \nu\|\mu)\\
&\quad + \sum_{\substack{B\subseteq\mathbb{X}_K\\ \nu(B)>0}} \log\frac{\mu^{\downarrow K}(B)}{\nu^{\downarrow K}(B)} \sum_{\substack{A\subseteq\mathbb{X}_L\,:\,\nu(A)>0\\ A^{\downarrow K}=B}} \mu(A)\\
&= D(\mu^{\downarrow K}\triangleright_F \nu\|\mu)\\
&\quad + \sum_{B\subseteq\mathbb{X}_K:\nu(B)>0} \mu(B)\log\frac{\mu^{\downarrow K}(B)}{\nu^{\downarrow K}(B)},
\end{aligned}
$$

where the last modification is based on the formula for marginalization.

Regarding the fact that using analogous computations

$$
\begin{aligned}
&D(\nu\|\mu^{\downarrow K}\triangleright_F \nu)\\
&= \sum_{\substack{A\subseteq\mathbb{X}_L\\ (\mu^{\downarrow K}\triangleright_F\nu)(A)>0}} (\mu^{\downarrow K}\triangleright_F \nu)(A)\log\frac{(\mu^{\downarrow K}\triangleright_F \nu)(A)}{(\nu^{\downarrow K}\triangleright_F \nu)(A)}\\
&= \sum_{\substack{A\subseteq\mathbb{X}_L\\ (\mu^{\downarrow K}\triangleright_F\nu)(A)>0}} (\mu^{\downarrow K}\triangleright_F \nu)(A)\log\frac{\mu^{\downarrow K}(A^{\downarrow K})}{\nu^{\downarrow K}(A^{\downarrow K})}\\
&= \sum_{B\subseteq\mathbb{X}_K:\nu(B)>0} \mu(B)\log\frac{\mu^{\downarrow K}(B)}{\nu^{\downarrow K}(B)},
\end{aligned}
$$

we have finished the proof. □

Proof of Theorem 2. First notice that the function $D(\nu\|\mu)$ introduced in the previous Lemma is in fact the famous Kullback-Leibler divergence between two probability measures (let us stress that we assume that all the involved basic assignments are proper, because $\triangleright_F$ composition of two proper assignments is obviously also proper) defined on $2^{\mathbb{X}_L}$, which is known to be nonnegative, equals 0 if and only if $\nu = \mu$, and is finite if ν dominates μ. Moreover, since ν is assumed to be a common extension of all $m_1, m_2, \ldots, m_n$, it means that $\nu^{\downarrow K_j} = m_j$ for all $j = 1, 2, \ldots, n$.

So, following the idea of Csiszár, we can apply

Lemma 1 getting

$$D(\mu_0\|\nu) = D(\mu_0\|m_1 \triangleright_F \mu_0) + D(m_1 \triangleright_F \mu_0\|\nu),$$

where $m_1 \triangleright_F \mu_0 = \mu_1$ computed by Algorithm IPFP. Analogously,

$$\begin{aligned} D(\mu_1\|\nu) &= D(\mu_1\|\mu_2) + D(\mu_2\|\nu), \\ D(\mu_2\|\nu) &= D(\mu_2\|\mu_3) + D(\mu_3\|\nu), \\ &\vdots \end{aligned}$$

and therefore

$$D(\mu_0\|\nu) \geq \sum_{j=1}^{\infty} D(\mu_{j-1}\|\mu_j).$$

Since we assume that μ_0 dominates ν, $D(\mu_0\|\nu)$ is finite, and therefore

$$\lim_{j\to\infty} D(\mu_{j-1}\|\mu_j) = 0.$$

The required convergence of $\mu_0, \mu_1, \mu_2, \mu_3, \ldots$ follows directly from the fact that the last equality guarantees also that (for more details see [2])

$$\lim_{j\to\infty} \sum_{A\subseteq \mathbb{X}_{K_1\cup\ldots\cup K_n}} |\mu_{j-1}(A) - \mu_j(A)| = 0. \qquad \square$$

Example 2 Let us conclude this section with an example illustrating behavior of the Algorithm IPFP in case of an inconsistent system of basic assignments. It is clear that IPFP does not converge in this case, because, due to Theorem 1, otherwise it would have converged to a joint extension of the given assignments, which does not exist. However, based on our experiments, there exist converging subsequences. This phenomenon is known also from the probabilistic IPFP [14].

Let us consider three basic assignments m_1, m_2, and m_3 defined on $\mathbb{X}_{\{1,2\}}$, $\mathbb{X}_{\{2,3\}}$, $\mathbb{X}_{\{1,3\}}$, respectively, where, again, $\mathbb{X}_i = \{a_i, \bar{a}_i\}$. The focal elements of these assignments as well as the respective values are in Table 4.

Now, let us perform the IPFP process with μ_0 that is the same as in Example 1: $\mu_0(A) = 1/255$ for all nonempty $A \subseteq \mathbb{X}_{\{1,2,3\}}$. A part of the computational process is depicted in Table 5.

In this situation, the beginning of the process is not interesting. But after a several cycles, we can see that the iteration process goes through cyclical changes. From this example we can see that there are three convergent subsequences, namely

$$\begin{aligned} &\mu_1, \mu_4, \mu_7, \ldots, \mu_{3k+1}, \cdots \\ &\mu_2, \mu_5, \mu_8, \ldots, \mu_{3k+2}, \cdots \\ &\mu_3, \mu_6, \mu_9, \ldots, \mu_{3k}, \cdots \end{aligned}$$

	focal elements	values
m_1	$\{\bar{a}_1 a_2\}$	0.55
	$\{a_1\bar{a}_2, \bar{a}_1 a_2\}$	0.40
	$\{a_1 a_2, \bar{a}_1 a_2, \bar{a}_1\bar{a}_2\}$	0.05
m_2	$\{a_2 a_3\}$	0.63
	$\{a_2 a_3, a_2\bar{a}_3, \bar{a}_2 a_3\}$	0.22
	$\{a_2 a_3, a_2\bar{a}_3, \bar{a}_2\bar{a}_3\}$	0.15
m_3	$\{\bar{a}_1 a_3\}$	0.65
	$\{a_1 a_3, \bar{a}_1 a_3, \bar{a}_1\bar{a}_3\}$	0.35

Table 4: Inconsistent assignments m_1, m_2, m_3

In all our computational experiments it appeared that the length of the cycle which the process goes through corresponds to the number of basic assignments entering the computational process, and that the respective subsequences converged.

2.2 IPFP with $\triangleright_D$

Let us say at the very beginning of this section that considering the operator $\triangleright_D$ leads to many open problems. One of the reasons is connected with the computational complexity of this operator. Namely, computational complexity of composition operators is, naturally, closely connected with the number of focal elements to be enumerated. As a rule, D-operator produces a higher number of focal elements in comparison with F-operator. Moreover, in case of F-operator the enumeration of a value of a basic assignment for each focal element is got as a product of the respective projections of the focal element (i.e. a product of only two numbers), for D-operator one needs to process all the supersets of the respective projections. Thus, we can apply the IPFP Algorithm with D-operator only to very simple examples and even for them we cannot compute too long sequences $\mu_0, \mu_1, \mu_2, \mu_3, \ldots$. Other difficulties connected with application of this operator of composition will be formulated as open problems. The first one is connected with the fact, that in contrast to $\triangleright_F$, composition $\triangleright_D$ of two proper basic assignments need not be proper - it can achieve negative values.

Open Problem 2 *What are the necessary and sufficient conditions guaranteeing that $\triangleright_D$ composition of two proper assignments is also proper?*

Example 3 Consider first the same system of three consistent basic assignments as in Example 1, and start the computational process again with $\mu_0(A) = 1/255$ for all nonempty $A \subseteq \mathbb{X}_{\{1,2,3\}}$. Assignments μ_1 and μ_2 have now 99, and 70 focal elements, respectively. Starting with μ_3 all the remaining computations concern 44 focal elements, and nearly half

focal elements	μ_{13}	μ_{14}	μ_{15}	μ_{16}	μ_{17}	μ_{18}	μ_{43}	μ_{44}	μ_{45}
$\left\{\begin{matrix} a_1a_2a_3, \bar{a}_1a_2a_3, \\ \bar{a}_1a_2\bar{a}_3, \bar{a}_1\bar{a}_2\bar{a}_3 \end{matrix}\right\}$	0.049	0.150	0.142	0.049	0.150	0.142	0.050	0.150	0.142
$\left\{\begin{matrix} a_1a_2a_3, \bar{a}_1a_2\bar{a}_3, \\ \bar{a}_1\bar{a}_2a_3 \end{matrix}\right\}$	0.001	$2 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	$7 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	10^{-10}	10^{-10}	10^{-10}
$\left\{\begin{matrix} a_1a_2a_3, \bar{a}_1a_2a_3, \\ \bar{a}_1a_2\bar{a}_3, \bar{a}_1\bar{a}_2a_3 \end{matrix}\right\}$	0.001	$2 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	$7 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	10^{-10}	10^{-10}	10^{-10}
$\left\{\begin{matrix} a_1\bar{a}_2a_3, \bar{a}_1a_2a_3, \\ \bar{a}_1a_2\bar{a}_3 \end{matrix}\right\}$	0.400	0.219	0.208	0.400	0.219	0.208	0.400	0.220	0.208
$\{\bar{a}_1a_2a_3\}$	0.550	0.630	0.650	0.550	0.630	0.650	0.550	0.630	0.650

Table 5: IPFP $\triangleright_F$: inconsistent marginals

focal elements	μ_3	μ_4	μ_5	μ_6	μ_7	μ_{100}	μ_{1000}
$\{a_1\bar{a}_2a_3, \bar{a}_1a_2a_3\}$	0.020	0.030	0.033	0.031	0.039	0.085	0.095
$\{a_1a_2a_3, a_1\bar{a}_2a_3, \bar{a}_1a_2a_3, \bar{a}_1\bar{a}_2a_3\}$	0.017	0.042	0.046	0.042	0.047	0.031	0.010
$\{a_1a_2\bar{a}_3, a_1\bar{a}_2a_3\}$	0.141	0.208	0.233	0.168	0.203	0.294	0.299
$\{a_1a_2\bar{a}_3, a_1\bar{a}_2a_3, a_1\bar{a}_2\bar{a}_3\}$	0.103	0.152	0.140	0.101	0.122	0.014	10^{-4}
$\{a_1a_2\bar{a}_3, a_1\bar{a}_2\bar{a}_3, \bar{a}_1a_2\bar{a}_3, \bar{a}_1\bar{a}_2a_3\}$	0.097	0.226	0.208	0.232	0.260	0.413	0.476
$\{a_1a_2\bar{a}_3, a_1\bar{a}_2\bar{a}_3, \bar{a}_1a_2\bar{a}_3, \bar{a}_1\bar{a}_2a_3, \bar{a}_1\bar{a}_2\bar{a}_3\}$	0.097	0.226	0.208	0.232	0.260	0.413	0.476
$\{a_1a_2\bar{a}_3, a_1\bar{a}_2\bar{a}_3, \bar{a}_1a_2\bar{a}_3\}$	−0.047	−0.034	−0.045	−0.090	−0.051	−0.190	−0.228
$\{\bar{a}_1a_2\bar{a}_3, \bar{a}_1\bar{a}_2\bar{a}_3\}$	−0.021	−0.020	−0.015	−0.001	0.004	0.012	0.001

Table 6: IPFP $\triangleright_D$: converging sequence for consistent marginals

of them have negative values. After a thousand of iterative steps the changes are so small that we can take μ_{1000} as a limit of the computational process. In agreement with Theorem 1 we can see that $m_1 \doteq m_{1000}^{\downarrow\{1,2\}}, m_2 \doteq m_{1000}^{\downarrow\{2,3\}}$ and $m_3 \doteq m_{1000}^{\downarrow\{1,3\}}$.

A part of the computational process is depicted in Table 6. We selected 8 focal elements, the first 6 of them correspond to those from Table 3, the other 2 are chosen to present examples of focal elements with negative values. Observe that the last focal element switched its value from a negative one to a positive one during the IPFP.

Example 4 It shows up that in contrast to the application of $\triangleright_F$, the Algorithm IPFP with $\triangleright_D$ need not converge for a consistent system of marginal basic assignments. As an example consider the 3-dimensional assignment m from Table 7 and its marginals $m_1 = m^{\downarrow\{1,2\}}, m_2 = m^{\downarrow\{2,3\}}$, $m_3 = m^{\downarrow\{1,3\}}$. With μ_0 as in the previous examples, the sequence $\mu_0, \mu_1, \mu_2, \mu_3, \ldots$ computed by the Algorithm IPFP does not converge - it stabilizes in a loop of length 6 after approximately 560 iterations (i.e. $\mu_{601} = \mu_{607}, \mu_{602} = \mu_{608}, \ldots$). The strange behavior of this process is visible from Table 8, where a selected part of focal elements are presented. There are two phenomena that are in a way surprising. First, it is the length of the cycle (6), and the fact that even focal elements may variate during the cycle.

focal elements	m
$\{\bar{a}_1\bar{a}_2\bar{a}_3\}$	0.225
$\{a_1\bar{a}_2\bar{a}_3, \bar{a}_1a_2a_3\}$	0.126
$\{\bar{a}_1a_2\bar{a}_3, \bar{a}_1\bar{a}_2\bar{a}_3\}$	0.594
$\{\bar{a}_1\bar{a}_2a_3, \bar{a}_1\bar{a}_2\bar{a}_3\}$	0.024
$\{a_1\bar{a}_2a_3, a_1\bar{a}_2\bar{a}_3, \bar{a}_1\bar{a}_2a_3, \bar{a}_1\bar{a}_2\bar{a}_3\}$	0.031

Table 7: Three-dimensional assignment m

Open Problem 3 *Under what conditions does the sequence $\mu_0, \mu_1, \mu_2, \mu_3, \ldots$ computed by the Algorithm IPFP with $\triangleright_D$ converge? When is the limit assignment proper?*

3 Summary and Conclusions

Using two different operators of composition for belief functions that were studied in [6, 7], we designed two versions of the iterative procedure presented as Algorithm IPFP. If they converge, both of these algorithms yield basic assignments that have the input low-dimensional assignments for their marginals. But this is perhaps the only property common to both of them. Even in case that both the algorithms converge, the results may be different. In fact, we conjecture that these algorithms yield the same results

focal elements	μ_{601}	μ_{602}	μ_{603}	μ_{604}	μ_{605}	μ_{606}	μ_{607}	μ_{608}	μ_{609}
$\{\bar{a}_1 a_2 a_3\}$	-0.008	0	-0.047	0.036	0	-0.052	-0.008	0	-0.047
$\{a_1 \bar{a}_2 a_3\}$	-0.023	0.008	0.047	0.008	-0.007	0.052	-0.023	0.008	0.047
$\{a_1 \bar{a}_2 \bar{a}_3\}$	0.008	-0.003	-0.123	-0.036	0.006	0.033	0.008	-0.003	-0.123
$\{\bar{a}_1 \bar{a}_2 a_3\}$	-0.076	0.012	0.348	-0.038	-0.051	0.279	-0.076	0.012	0.348
$\{\bar{a}_1 \bar{a}_2 \bar{a}_3\}$	0.227	0.205	0	0.232	0.242	0	0.227	0.205	0
$\{a_1 \bar{a}_2 \bar{a}_3, \bar{a}_1 a_2 a_3\}$	0.110	0.082	0.157	0.066	0.071	0.157	0.110	0.082	0.157
$\{\bar{a}_1 a_2 a_3, \bar{a}_1 \bar{a}_2 \bar{a}_3\}$	0	0.043	0	0	0.054	0	0	0.043	0
$\{a_1 \bar{a}_2 a_3, \bar{a}_1 \bar{a}_2 a_3\}$	0.058	-0.004	0.055	0.050	0.001	0.055	0.058	-0.004	0.055
$\{a_1 \bar{a}_2 a_3, a_1 \bar{a}_2 \bar{a}_3\}$	0.076	0.034	0	0.038	0.046	0	0.076	0.034	0
$\{a_1 \bar{a}_2 \bar{a}_3, \bar{a}_1 \bar{a}_2 a_3\}$	0.044	0.020	-0.031	0.007	0.008	-0.031	0.044	0.020	-0.031
$\{\bar{a}_1 \bar{a}_2 a_3, \bar{a}_1 \bar{a}_2 \bar{a}_3\}$	0.044	0.020	-0.031	0.007	0.008	-0.031	0.044	0.020	-0.031
$\{a_1 \bar{a}_2 \bar{a}_3, \bar{a}_1 a_2 \bar{a}_3\}$	0.015	0.016	0.140	0.059	0.058	0.022	0.015	0.016	0.140
$\{\bar{a}_1 a_2 \bar{a}_3, \bar{a}_1 \bar{a}_2 \bar{a}_3\}$	0.535	0.576	0.593	0.542	0.535	0.505	0.535	0.576	0.593
$\{a_1 \bar{a}_2 \bar{a}_3, \bar{a}_1 \bar{a}_2 \bar{a}_3\}$	0.031	0.007	-0.140	0.031	0.033	-0.022	0.031	0.007	-0.140
$\left\{\begin{matrix} a_1 \bar{a}_2 a_3, a_1 \bar{a}_2 \bar{a}_3, \\ \bar{a}_1 \bar{a}_2 a_3, \bar{a}_1 \bar{a}_2 \bar{a}_3 \end{matrix}\right\}$	-0.044	-0.020	0.031	-0.007	-0.008	0.031	-0.044	-0.020	0.031

Table 8: IPFP $\triangleright_D$: non-converging sequence for consistent marginals

only in degenerate situations. As a rule, application of $\triangleright_D$ yields basic assignments with greater number of focal elements (compare Examples 1 and 3).

The algorithm employing $\triangleright_F$ manifests some of the nice properties of the probabilistic IPFP: its convergence is guaranteed for consistent systems of low-dimensional assignments. Moreover, its significantly lower computational complexity predestinates this version of the algorithm to practical applications. Another its advantage follows from the fact that if the input assignments are proper then the resulting basic assignment is also proper, which is not true for the Algorithm based on $\triangleright_D$. For example, when we randomly generated three-dimensional basic assignments, and applied the Algorithm IPFP with $\triangleright_D$ to their two-dimensional marginals, only about every fifteens solution was proper.

As it was highlighted in one of the referee reports, the application of the IPFP procedure may be extended beyond probability theory to other topics as, for example, that described in [9]. In fact, as the title of the paper suggests, the authors see several ways how to prolong the research in the field.

Acknowledgements

This work was supported by GAČR: the first author under the grant no. 403/12/2175, the second one under the grant no. 13-20012S.

References

[1] P. Cheeseman. In defence of probability. In *Proc. 8th Int. Joint Conf. on AI (IJCAI'85).* Los Angeles, CA, pp 1002–1009, 1985.

[2] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* 3, pp 146–158, 1975.

[3] W. E. Deming and F. F. Stephan. On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* 11, 427–444, 1940.

[4] R. Jiroušek. A note on local computations in Dempster-Shafer theory of evidence. In *Proc. of the 7th Symp. on Imprecise Probabilities and Their Applications,* STUDIA Universitatsverlag, Innsbruck, pp. 219–227, 20011.

[5] R. Jiroušek. Foundations of compositional model theory. *Int. J. General Systems*, 40, 6, pp 623–678, 2011.

[6] R. Jiroušek and P. P. Shenoy. Compositional models in valuation-based systems. In *Belief Functions: Theory and Applications*, T. Denoeux and M.-H. Masson, eds., Advances in Intelligent and Soft Computing 164, pp 221–228, Heidelberg, Springer, 2012.

[7] R. Jiroušek and P. P. Shenoy. Compositional models in valuation-based systems. In print *Int. J. Approx. Reasoning*, doi: 10.1016/j.ijar.2013.02.002.

[8] R. Jiroušek, J. Vejnarová and M. Daniel. Compositional models of belief functions. In *Proc. of the 5th Symp. on Imprecise Probabilities and Their Applications,* G. de Cooman, J. Vejnarová, M. Zaffalon, Eds., Praha, pp. 243–252, 2007.

[9] E. Miranda, M. Zaffalon. Coherence graphs. *Artificial Intelligence,* 173(1), 104-144, 2009.

[10] R. D. Shachter. Intelligent probabilistic inference. In: *Uncertainty in AI.* L. N. Kanal, J. F. Lemmer, eds. North Holland, Amsterdam, 1986.

[11] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, New Jersey, 1976.

[12] P. P. Shenoy. Conditional independence in valuation-based systems. *Int. J. Approx. Reasoning,* 10, 3, pp. 203–234, 1994.

[13] D. J. Spiegelhater. Probabilistic reasoning in predictive expert systems. In: *Uncertainty in Artificial Intelligence (UAI-85).* L. N. Kanal, J. F. Lemmer, eds. North Holland, Amsterdam, pp. 47–67, 1985.

[14] J. Vomlel. Integrating inconsistent data in a probabilistic model. *J. of Applied Non-Classical Logics,* 14, 3, pp. 367–386, 2004.

Dynamic Credal Networks: Introduction and Use in Robustness Analysis

Matthieu Hourbracq[1]
matthieu.hourbracq@lip6.fr

Cédric Baudrit[2]
cbaudrit@grignon.inra.fr

Pierre-Henri Wuillemin[1]
pierre-henri.wuillemin@lip6.fr

Sébastien Destercke[3]
sebastien.destercke@hds.utc.fr

Abstract

Dynamic Bayesian networks (DBN) are handy tools to model complex dynamical systems learned from collected data and expert knowledge. However, expert knowledge may be incomplete, and data may be scarce (this is typically the case in Life Sciences). In such cases, using precise parameters to describe the network does not faithfully account for our lack of information. This is why we propose, in this paper, to extend the notion of DBN to convex sets of probabilities, introducing the notion of dynamic credal networks (DCN). We propose different extensions relying on different independence concepts, briefly discussing the difficulty of extending classical algorithms for each concept. We then apply DCN to perform a robustness analysis of DBN in a real-case study concerning the microbial population growth during a French cheese ripening process.

1 Introduction

Dynamic Bayesian networks (DBNs) [36] extend Bayesian networks (BNs) [37, 38] and form a convenient formalism to describe complex dynamical systems. They also extend the well-known Hidden Markov Models (HMMs) [40] by representing the hidden state and the observation in terms of several random variables. The probabilistic and graphical natures of DBNs make them attractive tools to integrate both expert knowledge and data in a single representation. The concept of DBNs makes possible to (i) combine different sources of knowledge; (ii) easily modify the model thanks to its modular nature and (iii) integrate uncertainties. However, one limitation of DBNs lies in the specification of parameters that requires a substantial knowledge that is seldom available. This is particularly the case when experimental data are costly, such as in Life Sciences.

One way to overcome this difficulty is to use *credal sets* [35, 44], i.e., convex sets of probabilities to model the lack of knowledge about the parameters. Applied to Bayesian networks, this idea corresponds to the concept of credal networks (CN) [17, 19], in which each node of the network is associated to a convex set of conditional probabilities (possibly degenerated to a single element). Other approaches such as possibilistic [3] or evidential networks [45] follow the same objective but cannot be interpreted as a proper extension of classical Bayesian Networks.

While the notion of credal network has received much attention in the past years, it is not the case for its dynamic extension. Indeed, the only works dealing with such extension consider specific models related to Markov Processes [22, 26], in which computations on the full dynamic network can be done separately for each time-step. Although such cases are of high interest and can benefit from efficient algorithms, there are many other cases where one will need to perform inferences on a complete network not reducible to a Markov model. This is especially the case in Life and Food Sciences [1, 39], where the modelling of non-linear, multi-scale dynamic processes (maturation processes, evolution of interacting physicochemical phenomena, . . .) is often based on qualitative expert knowledge and on limited experimental data. The use of Dynamic Credal networks (DCNs) extending DBNs seems a good way to integrate such heterogeneous and scarce knowledge.

The goal of this paper is two-fold: first to provide in Section 3 a first theoretical and practical discussion of the DBNs extension into DCNs, second to apply in Section 4 the DCNs framework to achieve a robustness analysis of

[1] Laboratoire d'Informatique de Paris VI (UPMC, CNRS UMR7606) 75004 Paris, France

[2] UMR782 Génie et Microbiologie des Procédés Alimentaires. INRA/AgroParisTech, 78850 Thiverval-Grignon, France

[3] Université de Technologie de Compiegne U.M.R. C.N.R.S. 7253 Heudiasyc Centre de recherches de Royallieu F-60205 Compiegne Cedex FRANCE

DBNs in a real-world case study involving the growth of yeast population during the Camembert-type cheese ripening. Preliminary notions are briefly recalled in Section 2.

2 Preliminary notions : DBN and CN

2.1 Dynamic Bayesian Networks

DBNs are classical Bayesian networks in which nodes $\{X_i(t), i = 1 \ldots n\}$, representing (discrete) random variables, are indexed by discrete time t. They provide a compact representation of the joint probability distribution P for a finite time interval $[\![1, \tau]\!]$ (we use $[\![i, j]\!]$ to denote the finite set of time indices $\{i, \ldots, j\}$) defined as follows:

$$P(\mathbf{X}(1)\ldots, \mathbf{X}(\tau)) = \prod_{i=1}^{n}\prod_{t=1}^{\tau} P\left(X_i\left(t\right) \mid \mathbf{U}_i\left(t\right)\right) \tag{1}$$

where $\mathbf{U}_i(.)$ denotes the set of parent nodes of a node $X_i(.)$ and $P\left(X_i\left(t\right) \mid \mathbf{U}_i\left(t\right)\right)$ denotes the conditional probability function associated with the random variable $X_i(t)$ given $\mathbf{U}_i(t)$. $\mathbf{X}(t) = \{X_1(t), \ldots, X_n(t)\}$, is called a "slice" and represents the set of all variables indexed by the same time t. This joint probability $P(\mathbf{X}(1), \ldots, \mathbf{X}(\tau))$ represents the beliefs about possible trajectories of the dynamic process $\mathbf{X}(t)$. DBNs assume the *first-order Markov property* which means that the parents of a variable in time slice t must occur in either slice $t-1$ or t :

$$\mathbf{U}_i(t) \subset \mathbf{X}(t-1) \cup \mathbf{X}(t) \backslash \{X_i(t)\} \tag{2}$$

Moreover, the conditional probabilities are time-invariant (*first-order homogeneous Markov property*):

$$P\left(X_i\left(t\right) \mid \mathbf{U}_i\left(t\right)\right) = P(X_i(2)|\mathbf{U}_i(2))) \text{ ,} \forall t \in [\![2, \tau]\!]. \tag{3}$$

To specify a DBN, we need to define the intra-slice topology (within a time slice), the inter-slice topology (between two time slices), as well as the parameters, *i.e* conditional probabilities in Equation (3) for the first two time slices.

In this paper, we consider that $\mathbf{X}_i(t)$ are all discrete variables. Faced with continuous variables X_i, these ones will be discretized. Let P^t_{ijk} be the probability that $X_i(t) = k$, given that its parents have instantiation j, *i.e.*

$$P^t_{ijk} = P\left(X_i(t) = k \mid \mathbf{U}_i(t) = j\right), \tag{4}$$

for $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, c_i\}$ where c_i is the number of distinct configurations of $\mathbf{U}_i(t)$ and $k \in \{1, \ldots, r_i\}$ where r_i is the number of values that node i can take.

2.1.1 Parameter learning and local elicitation

The techniques for learning DBNs are generally extensions of the techniques for learning BNs. Different methods exist to learn the structure or the parameters from substantial and/or incomplete data (for overviews, we refer readers to [5, 34]). In our case, we consider that the topology is given (e.g., learned from expert knowledge).

The most commonly used and simplest method which will be used in this paper is to estimate P^t_{ijk} by the occurrence rate of the event $(X_i(t) = k, \mathbf{U}_i(t) = j)$ in a training database :

$$P^t_{ijk} = N^t_{ijk}/\Sigma_k N^t_{ijk} \tag{5}$$

where N^t_{ijk} denotes the number of times where the event $(X_i(t) = k, \mathbf{U}_i(t) = j)$ occurs in database. As we assume the first-order homogeneous Markov property (3), P^t_{ijk} does not depend on time and we can rewrite

$$\forall t' \in [\![2, \tau]\!], P^{t'}_{ijk} = \frac{\sum_t N^t_{ijk}}{\sum_t \sum_k N^t_{ijk}} \tag{6}$$

In the case where $N^t_{ijk} = 0$ for all k, the uniform distribution is traditionally used as it maximizes the Shannon entropy and corresponds to the Laplace indifference principle.

A practical methodology able to incrementally build and update model parameters from heterogeneous information has been developed in [2] on the basis of Dirichlet model. From a given network structure, it consists in using a priori Dirichlet distributions which are then updated through Bayesian inference by expressing new pieces of information into a frequentist form. This method also integrates the confidence level on the different sources of information.

2.1.2 Knowledge propagation - inference

The use of DBNs consists in "query" expressed as conditional probabilities. The most common task we wish to solve is to estimate the marginal probabilities

$$P\left(\mathbf{X}_Q(t')|\{\mathbf{X}_E(t), \forall t \in [\![1, \tau]\!]\}\right), \forall t' \in [\![1, \tau']\!] \tag{7}$$

where $\mathbf{X}_Q$ is a set of query variables, and $\mathbf{X}_E$ is a set of evidence variables. Inference consists in computing the probability of each state of a variable when we know the state of other variables. In general, DBN inference is performed using recursive operators and Bayes' theorem that update the belief state of the DBN as new observations become available [36]. Due to the natural time ordering of the modelled process, $\mathbf{U}_i(t)$ will usually be observed before $X_i(t)$ and that may help with the sequential updating of the conditional probabilities as well as with the preservation of the original conditional independence structure.

2.2 Credal Networks and strong extension

A credal network (CN) [17, 19] is an extension of BNs where imprecision is introduced in probabilities by means of credal sets [35]. CNs specify a closed convex set $K(\mathbf{X})$ of multivariate probability mass functions over the whole set of variables $\mathbf{X}$. Under the *strong extension* [17] hypothesis, the joint credal set $K(\mathbf{X})$ over $\mathbf{X}$ may be formulated as:

$$K(\mathbf{X}) = CH\left\{P(\mathbf{X}) : P(\mathbf{X}) = \prod_{i=1}^{n} P_i, P_i \in K_i\right\} \tag{8}$$

where CH denotes the convex hull, $P_i = P(X_i \mid \mathbf{U}_i)$ and $K_i = K(X_i \mid \mathbf{U}_i)$ is the closed convex set of probability mass function for the random variable X_i given $\mathbf{U}_i$. In practice, it is sufficient to focus on the extreme points $ext[K(X_i \mid \mathbf{U}_i)]$ of $K(X_i \mid \mathbf{U}_i)$ in Eq. (8). In our experiments, we will limit ourselves to credal sets specified by means of probability intervals [25], that is for all $i = 1 \dots n$ and $j = 1 \dots c_i$:

$$K_{ij} = CH\left\{\begin{array}{l} P_{ij} : P_{ijk} \in [\underline{P}_{ijk}, \overline{P}_{ijk}] \subseteq [0,1],\ \forall k \\ \sum_k P_{ijk} = 1 \end{array}\right\} \tag{9}$$

This model has the advantage of creating a small number of extreme points provided additional constraints : $\forall k : \overline{P}_{ijk} - \underline{P}_{ijk} = \{0, \epsilon\}$ and $\epsilon = 1 - \sum_k \underline{P}_{ijk}$ is a constant. For such a linear-vacuous mixture, the number of vertices of $K(X_i \mid \mathbf{U}_i = j)$ is precisely the cardinality of X_i – assuming there is no modality k for which $\overline{P}_{ijk} = \underline{P}_{ijk}$ in which case $\mid X_i \mid$ is an upper bound – each vertex corresponding to the selection of a modality k for which $P(X_i = k \mid \mathbf{U}_i = j) = \overline{P}_{ijk}$ and therefor $\forall k' \neq k : P(X_i = k' \mid \mathbf{U}_i = j) = \underline{P}_{ijk'}$.

Inferences on a credal network comes down to assess lower and upper probabilities, that is search bounds of $P(\mathbf{X}_Q | \mathbf{X}_E)$ within $K(\mathbf{X})$ (under the strong extension hypothesis) for some values of $\mathbf{X}_Q$.

3 Dynamical Credal Networks (DCNs) : definitions and algorithms

This section introduces the notion of Dynamic Credal Networks (DCNs) and discusses their features.

3.1 Definition of Dynamic Credal Networks (DCNs)

A dynamic credal network is a DBN where conditional probabilities $P(X_i(t) \mid \mathbf{U}_i(t))$ (noted P_i^t) are replaced by credal sets $K(X_i(t) \mid \mathbf{U}_i(t))$ (noted K_i^t). We assume the same *first-order Markov property* (2) as in DBNs (parents only originate from same or previous time slice) and Eq. (3) becomes

$$K(X_i(t) \mid \mathbf{U}_i(t)) = K(X_i(2) \mid \mathbf{U}_i(2)),\ \forall t \in [\![2, \tau]\!]. \tag{10}$$

Therefore, specifying a DCN requires the same effort as a DBN but allows the user to provide conditional credal sets rather than probabilities if these latter cannot be reliably estimated (from data and/or experts).

3.2 Independence in DCN

When working with probability sets rather than precise probabilities, the notion of stochastic independence can be extended in several ways [15]. Within graphical models, the most commonly used extension is *strong independence*, that induces the strong extension defined in Eq. 8. It can be interpreted as a robust model of a precise yet ill-known BN.

This is in contrast with the notions of epistemic irrelevance and independence whose semantic as belief models is clearer. However, these notions encounters severe computational difficulties [18], limiting their practical interest. Recent results show that for particular models such as Hidden Markov ones, efficient algorithms can be used [22], however they remain intractable for the kind of models considered in this paper. This is why we focus on extending the notion of strong extension to dynamic schemes.

The most straightforward extension is to simply apply strong independence to the whole network, i.e.,

$$K(\mathbf{X})_{st} = CH\left\{P(\mathbf{X}) : P(\mathbf{X}) = \prod_{i=1}^{n}\prod_{t=1}^{\tau} P_i^t, P_i^t \in K_i^t\right\} \tag{11}$$

where $\mathbf{X} = (\mathbf{X}(1), \dots, \mathbf{X}(\tau))$. We call this extension the *dynamic strong extension* and it is worth noting $P_i^t \neq P_i^{t'}$ is valid, $t, t' \in [\![2, \tau]\!]$.

However, when stepping to dynamic models, Condition (10) allows us to use the notion of *repetitive independence*. This condition states that if two variables X, Y have the same set of possible outcomes, that is $\Omega_X = \Omega_Y$, and governed by the same probability distribution belonging to $K(X)$, then the joint credal set $K(X, Y)$ is :

$$K(X, Y) = CH\{P(X)P(X) \mid P(X) \in K(X)\}. \tag{12}$$

Adapting this notion of independence to DCN, so that probabilities of each time slice are assumed to be identical, leads to a second extension, i.e.,

$$K(\mathbf{X})_{rp} = CH\left\{\begin{array}{l} P(\mathbf{X}) : P(\mathbf{X}) = \prod_{i=1}^{n}\prod_{t=1}^{\tau} P_i^t, \\ P_i^2 \in K_i^2 \text{ and } P_i^t = P_i^2\ \forall t \in [\![2, \tau]\!] \end{array}\right\} \tag{13}$$

that we call the *dynamic repetitive extension*. We have $K(\mathbf{X})_{rp} \subseteq K(\mathbf{X})_{st}$, as $K(\mathbf{X})_{rp}$ is more constrained. In practice, the strong extension assumes that the dynamic network is ill-defined and that its behaviour can change between time slices, while the repetitive extension assumes that we seek a precise classical DBN who is partially known.

Next sections investigate the differences between these two extensions. In particular, we will see that some algorithms extend more easily to one extension than to another.

3.3 Inference algorithms in DCN

(D)CNs can be queried as (D)BNs were in Section 2.1.2 to get information about the state of a variable given evidence about other variables, with respect to the network *extension*. However, the use of credal sets makes the updating problem much harder, as it becomes an optimization problem. As such, the computation of the lower bound on $P(X_Q \mid \mathbf{X}_E)$ requires to minimize a quotient containing polynomials :

$$\underline{P}(X_Q \mid \mathbf{X}_E) = \min \left\{ \frac{\sum\limits_{X_i \in \mathbf{X} \setminus X_Q \cup \mathbf{X}_E} \prod\limits_{i=1}^{n} \prod\limits_{t=1}^{\tau} P_i^t}{\sum\limits_{X_i \in \mathbf{X} \setminus \mathbf{X}_E} \prod\limits_{i=1}^{n} \prod\limits_{t=1}^{\tau} P_i^t}, P \in K_\omega(\mathbf{X}) \right\} \tag{14}$$

with $P : P(\mathbf{X}) \in K_\omega(\mathbf{X})$ belonging to the *dynamic strong extension* ($\omega = st$) or *dynamic repetitive extension* ($\omega = rp$) of the network. An upper bound can be obtained by maximizing (14). It is known that such a minimum (or maximum) is obtained at a vertex of the *dynamic strong/repetitive extension*.

Depending on (1) the structure of network, (2) the number of modality of variables and (3) the chosen extension (strong/repetitive), the updating problem will be more or less complex to solve. Because inferences are already hard in static credal networks, little work has been done on DCNs (except for special cases already mentioned). By unrolling a two-time slice network over T time steps, the number of possible vertex combinations goes from $\prod\limits_{i,t=0} |ext[K_i^t]| \prod\limits_{i,t=1} |ext[K_i^t]|$ (with $|ext[K_i^t]|$ the number of vertices of K_i^t) in the case of repetitive independence, to $\prod\limits_{i,t=0} |ext[K_i^t]| \prod\limits_{i,t=1} |ext[K_i^t]|^{T-1}$ in the case of strong independence. Given the potential number of vertices, approximate algorithms seem more appropriate regarding DCNs.

Many algorithms, exact and approximate, have been proposed to deal with CN. Some are generalizations of well known (D)BNs algorithms. Among the approximate algorithms, there are those that compute inner bounds, i.e. bounds that are enclosed by the exact ones, outer bounds, which enclose the exact ones, and those that perform randomly.

3.3.1 Exact inference algorithms

The 2U algorithm [27] performs an exact rapid inference in the case of binary tree-shaped (D)CNs with the assumption of *strong independence*.

The CCM transformation [9] turns a (D)CN into a (D)BN by adding transparent nodes before performing an Maximum A Posteriori (MAP) estimation over the latter to find the best combination of vertices. It has the same complexity as credal network inference, that is $NP^{PP}Complete$, and performs poorly with separately specified credal networks such as the one we used during our trials (because of the sheer number of vertices).

Optimization techniques such as branch and bound over local vertices of credal sets [21, 7] are also well suited to medium-sized networks and can be stopped at any time to give an approximate answer.

Other algorithms are based on a variable elimination scheme from (D)BNs, such as Separable Variable Evaluation [20, 42] which keeps the separately specified credal sets as separated as possible during propagation, and can be mapped to an integer or a multi-linear program [24, 23].

3.3.2 Approximate inference algorithms

Regarding binary and DAG-shaped (DAG : Directed Acyclic Graph) credal networks, algorithm L2U (Loopy 2U) [32] (similar to LBP (Loopy Belief Propagation) [46]) produces either inner or outer approximations, and its efficiency is mainly due to the bounded cardinality of variables and in lesser extent to ignoring loops.

Another way to handle credal sets complexity is to represent them by simpler means. Variational methods [31, 30] choose a family of functions to approximate the exact combination of credal sets to decrease computational costs. Those functions are optimized according to some criteria until convergence and the inference is then realized in the network with the original credal sets replaced by the new found functions.

The A\R(+)(+) algorithm [21] uses interval probability arithmetic to approximate credal sets in a propagation scheme in tree-shaped networks (with the use of some additional constraints limiting the information loss in its enhanced version). The intervals produced are outer bounds of the real ones. Although those algorithms are fast in medium-sized network, they either produce too many approximations or are too complex to work with DCNs.

Another popular family of approximate algorithms producing inner bounds is based on Monte-Carlo sampling [29]. Several methods have been proposed to better guide the search (simulated annealing [6], genetic algorithms [8]) among the vertices of the (conditional) local credal sets, but they require some tuning for more accurate results, otherwise they can lead to poor approximations.

Although there exist several inference algorithms, none allows to do inference, in a realistic and practical way, on networks capable of representing global complex system of Life Sciences especially in Food Sciences. Indeed, networks is composed of a large number of interacting variables capable of describing the behaviour of microscopic scales (as micro organism) involving macroscopic view (as the evolution of sensory properties). In further inferences, we used a simple Monte-Carlo sampling algorithm [29] which has the advantage as point of reference, as it applies with the same easiness to *dynamic repetitive* and *strong extensions* (with a faster convergence for *dynamic repetitive extension*).

4 DCN for Robustness in DBN

In this section, we apply the concept of DCN to perform a robustness analysis of a learned precise DBN (both repetitive and strong independence concepts well correspond to this idea). We first recall some elements about robustness in classical BN before proceeding to our study.

4.1 Robustness in BN

Roughly speaking, a robustness analysis is the study of the behaviour of a model given small perturbations in its parameters. Robustness in Bayesian network is commonly addressed using sensitivity analysis where the main concern is to analyse the relationships between local network parameters and global conclusions drawn based on the BN. Sensitivity analysis has been largely studied by many researchers [10, 4, 14, 11]. We propose here a small survey of the main approaches.

The most common case of sensitivity analysis in BN is the study of single-parameter influence [14, 33]. In a BN, a parameter is a number in the CPT : $p(x_i|u)$ where x_i is a possible value for a random variable X and u is a possible instantiation of the parents of X in the BN. In this framework, a perturbation ϵ consists in modifying $p(X|u)$ into

$$p(x_j|u)[\epsilon] = \begin{cases} \epsilon & \text{where } i = j \\ \frac{p(x_j|u)\cdot(1-\epsilon)}{1-p(x_i|u)} & \text{otherwise.} \end{cases} \tag{15}$$

Under covariation conditions, inferred posterior distribution of any variable in the BN then takes the form of a quotient of two linear functions: $\frac{c_1\cdot\epsilon+c_2}{c_3\cdot\epsilon+c_4}$. Efficient algorithms have been proposed to assess the values of the c_i [43]. This kind of study can further be generalized to n-way sensitivity analyses where n is the number of parameters. It has been applied for DBNs in [13]. However, the results are often difficult to interpret [33].

Testing the sensitivity of the results of an inference can be more globally performed in a different manner. Soft evidence (i.e. uncertain evidence) is a way to disturb global behaviour of the BN using (local) belief revision [41, 12]. However, even if the specification of the perturbations is different, this methods still faces the same difficulty to interpret the results when multiple local changes are performed [11].

Sensitivity analysis in BN proposes tools to analytically follow the change in posterior distributions as a function of the parameters (or the beliefs) in local CPTs. As attractive as it might be, this is not exactly what it is asked in robustness analysis. Indeed, the effects of numerous small perturbations is not easy to be estimated with such analysis (using derivative of sensitivity expressions for instance). One would like to obtain a set of possible distributions for the posterior as a result. [16] describes such an approach but with a framework (epistemic independence) difficult to use in the context of large and complex systems such as DBNs. The next section extends and implements this approach by using DCN as a dedicated tool for specification of sets of complex distributions.

4.2 DCN as a robustness analysis tool

In this paper, we propose a robustness analysis that consists in perturbing the precise DBN by means of conditional credal sets $K^t_{ij|\epsilon}$=$K_\epsilon(X_i(t)|\mathbf{U}(t) = j, \epsilon)$ such that for all $i = 1 \dots n$, $j = 1 \dots c_i$ and $\epsilon \in [0, 1]$:

$$K^t_{ij|\epsilon} = \left\{ \begin{array}{l} P^t_{ijk} \in [(1-\epsilon)P^t_{ijk}, (1-\epsilon)P^t_{ijk} + \epsilon], \\ \sum_k P^t_{ijk} = 1 \end{array} \right\} \tag{16}$$

The parameter ϵ may be understood as a perturbation coefficient: the higher it is, the more imprecise $K^t_{ij|\epsilon}$ becomes.

4.2.1 Choosing ϵ

The perturbation should depend on the quantity of data used to learn the DBNs as well as on the strength of the intended perturbation. While the strength of the perturbation should be the same over all the network, the number of data used may differ significantly in different places. We propose, to pick the ϵ used for a given (conditional) probability, to use a function $\psi(n, \beta) : \mathbb{N} \times [0, 1] \to [0, 1]$ where n corresponds to the quantity of data for learning each P^t_{ij} (that is $n = N^t_{ij}$ in our case) and β the strength of the perturbation, and to take $\epsilon = \psi(N^t_{ij}, \beta)$ to perturb the conditional probabilities P^t_{ij} of the network. The mapping ψ should satisfy the following constraints:

- $\psi(n, 0) = 0$ and $\psi(n, 1) = 1$
- ψ is decreasing in n
- ψ is increasing in β

The first conditions ensure that no perturbation will keep P^t_{ij} unchanged, while a full perturbation will make the network completely imprecise (this condition may be relaxed into requiring only that $\psi(0, 1) = 1$). The two other conditions ensure that a higher perturbation will induce more imprecision (for a given data set), while more data will result in less imprecision (for a given perturbation). We may also require that $\psi(0, \beta) = 1$ for any $\beta > 0$, that is no data means full imprecision (unless no perturbation is applied), and that $\lim_{n\to\infty} \psi(n, \beta) = 0$ for any β , that is the perturbation tends to the null perturbation as data accumulates.

The following function satisfies the conditions:

$$\psi(n, \beta) = \beta^{f(n)} \tag{17}$$

where $f(n)$ is an increasing function of n. The natural logarithmic operator ln satisfies these properties and we use $f(n) = \ln(n+1)$.

4.2.2 Keeping the constraint

Note that if $P^t_{ijk} = 0$ because it corresponds to an hard constraint in the network, it should be kept to 0 even when perturbing the whole network by making it imprecise (only non constraint probabilities should be made imprecise). We will see in the next section that preserving such (physical) constraints indeed play a very important role to ensure the good behaviour of the prediction dynamics.

4.3 Experiments on real-life case study

To illustrate our approach on a real case, we have focused on a typical French product, namely the process of the Camembert-type soft mould cheese ripening that is still ill known and complicated to control [28]. During the ripening process, cheese represents an ecosystem and a bioreactor where relationships exist between microbiological, physicochemical and organoleptic changes which depend on environmental conditions. Despite the number of areas involved in cheese research, available knowledge of the cheese ripening process remains fragmented and pervaded with uncertainty. None of the approaches or investigations carried out up to now makes it possible to provide an explicit overview of the causal structure of associations between the underlying variables and an objective interpretation of the cheese ripening process. From operational and scientific knowledge, the structure of a dynamic Bayesian network providing a qualitative representation of the coupled dynamics of micro-organism behaviour with their substrate consumptions influenced by temperature and involving the sensory changes of cheese during ripening has been defined [1]. Figure 1 displays a sub-section of the DBN structure providing a representation of the coupled dynamics of a yeast behaviour (*Kluyveromyces marxianus* concentration (Km)) with their substrate consumptions (lactose concentration (lo) influenced by temperature (T). We attempt to estimate the lower and upper mean time evolution

$$\begin{array}{l} \underline{X}_{Q|E,\epsilon}(t) = \min_{P \in K_\epsilon(\mathbf{X})} E_P(X_Q(t)|X_E(t), \forall t) \\ \overline{X}_{Q|E,\epsilon}(t) = \max_{P \in K_\epsilon(\mathbf{X})} E_P(X_Q(t)|X_E(t), \forall t) \end{array} \tag{18}$$

(where $E_P(X_Q(t)|X_E(t))$ denotes the mean time evolution of X_Q given X_E) under some perturbation. The initial precise model has been learned by integrating (1) experimental trials; (2) simulated database stemming from existing partial mechanistic models; (3) expert rules based to the conservation laws of microbial activities.

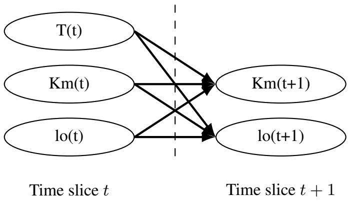


Figure 1: Dynamic Bayesian network representing the coupled dynamics Km growth versus lo consumptions influenced by temperature during the cheese ripening process.

In our experiment, a simple Monte-Carlo sampling algorithm over vertices is used to draw inference. The reasons for using such an algorithm are that (1) producing exact inference is too costly, even for small DCN with few time steps (here, 3 variables over 14 time steps), (2) it provides satisfactory bounds that are guaranteed to be inside exact ones and (3) it is sufficient in the present case, as our primary objective is not algorithmic efficiency.

4.3.1 Forward propagation

Forward propagation consists in trying to estimate

$$\begin{array}{c} Km(t)|\{Km(1), lo(1), T(1), \ldots, T(\tau)\} \\ lo(t)|\{Km(1), lo(1), T(1), \ldots, T(\tau)\} \end{array} \tag{19}$$

for all $t \in [1, \tau]$, using Eq. (18) to test the robustness of predictions. All temperatures are constant ($T(1) = \ldots = T(\tau) = 12^o$C) and τ corresponds to the day before the wrapping of cheeses, namely $\tau = 14$.

The Monte-Carlo sampling is stopped when lower and upper expectation bounds were not improved in the last 4000 samplings. In all our results about forward propagation, we have not observed differences between the dynamic strong and repetitive extension and we currently investigate whether it is always true in the case of forward propagation.

Figure 2 displays the upper and lower mean time evolutions of Km and lo for different perturbation levels where parameter learning have only been carried out from six experimental trials. We may observe that (1) the precise inferences of Km seem rather biased towards a rapid growth (line corresponding to $\beta = 0$ close to upper expectations); (2) Km may decrease (a physically impossible phenomena) even for relatively small perturbations ($\beta = 0.6$ and mean perturbation level $\epsilon = 0.124$) due to the absence of constraints based on conservation laws.

Figure 3 displays the upper and lower mean time evolutions of km and lo when constraints, based on conservation laws, are added. The effect of adding or preserving the constraints is obvious in the perturbed results. However, we may remark that the precise network is almost unchanged when constraints are added. This means that

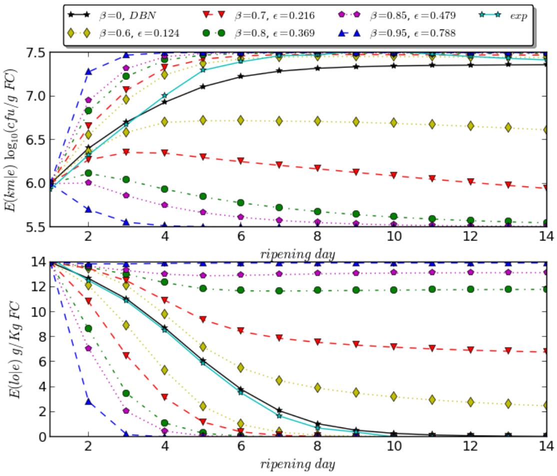


Figure 2: Upper and Lower mean evolutions of Km and lo according to different β values for forward propagation, without constraints. ϵ =mean contamination level

constraints play a secondary role when network parameters are well-estimated, however the comparison of Figures 2 and 3 shows that preserving them in case of bad estimation ensures more robustness in the inferences.

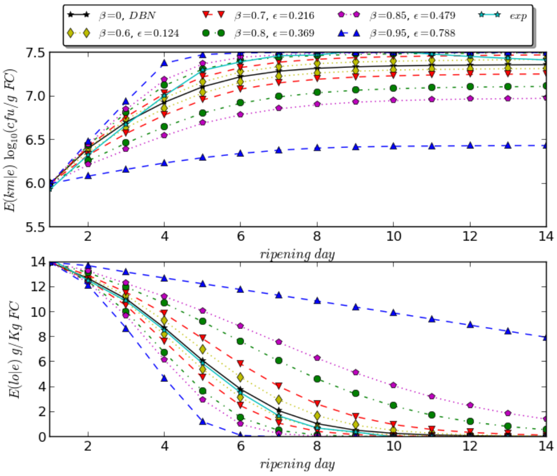


Figure 3: Upper and Lower mean evolutions of Km and lo according to different β values for forward propagation, with constraints. ϵ =mean contamination level

4.3.2 Forward-backward propagation

Forward-backward propagation consists in trying to estimate

$$\begin{gathered} Km(t)|\{Km(1), lo(1), Km(\tau), lo(\tau), T(1), \dots, T(\tau)\} \\ \text{and} \\ lo(t)|\{Km(1), lo(1), Km(\tau), lo(\tau), T(1), \dots, T(\tau)\} \end{gathered} \tag{20}$$

for all $t \in [1, \tau]$, using Eq. (18) in order to test the robustness of predictions. Monte-Carlo sampling was done as in the previous experiment.

Figure 4 displays the upper and lower mean time evolutions of km and lo for different perturbations with the *dynamic repetitive extension*, while Figure 5 displays the same results for the *dynamic strong extension* without the preservation of constraints.

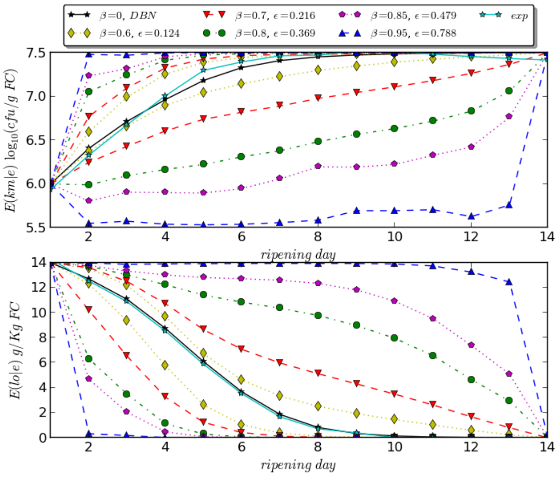


Figure 4: Upper and Lower mean evolutions of Km and lo according to different β values and *dynamic repetitive extension*.

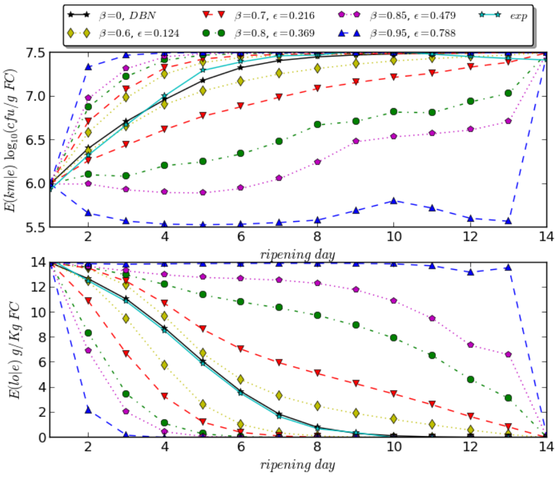


Figure 5: Upper and Lower mean evolutions of Km and lo according to different β values and *dynamic strong extension*.

In the case of forward-backward propagation, the results from the two extensions do not coincide in general. However, the bounds obtained with the *dynamic strong extension* are sometimes inside those obtained for the repetitive extension, meaning that the sampling algorithm has not reached optimal bounds (indeed, $K(\mathbf{X})_{rp} \subseteq K(\mathbf{X})_{st}$ by definition). We may also observe that the decreasing of Km is less severe than in forward propagation even for high β value because $Km(\tau)$ and $lo(\tau)$ are now evidences. Figure 6 displays the results of forward-backward inference with the *dynamic strong extension* when constraints are preserved. Again, we can see that preserving such constraints has a serious effect on the results precision.

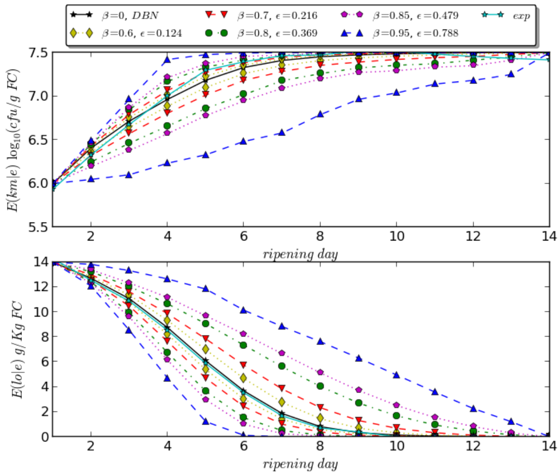


Figure 6: Upper and Lower mean evolutions of Km and lo according to different β values and *dynamic strong extension*, with constraints.

5 Conclusion

There are complex dynamical processes for which no deterministic model describing the complete process exists. In such cases, dynamic Bayesian networks are convenient models that allow to include expert knowledge, data and variable interaction in a single framework. However, they do not allow for a faithful representation of incomplete knowledge or of scarce data, features that are inherent to the complexity of bio-physicochemical phenomena occurring in Food and Life Sciences.

In this paper, we have discussed how DBNs can be extended to include credal sets and cope with such incompleteness and imprecision. We have introduced the concept of dynamic credal networks and have proposed the concepts of dynamic repetitive and strong extensions. While the latter can be seen as a straightforward extension of classical credal networks, the former considers repetitive independence to allow the model to preserve a temporal regularity. Inference algorithms of credal networks may extend better to one case than to the other, depending on their characteristics.

We have proposed to apply such DCN to the problem of robustness analysis, introducing an easy method to perturb a given precise network and performing some experiments on a real-case study concerning microbial population growth. These experiments have shown that including constraints (often provided by expert knowledge) in the network is essential in case of bad estimation of parameters, as they ensure more robustness, while such constraints seem unnecessary in case of good estimation.

We have also observed that in the case of forward propagation (evidences only on nodes without parents), inferences for the strong and repetitive extensions coincided. We are currently investigating under which conditions inferences of strong and repetitive extensions coincide.

In further works, DCNs should enable us to determine the contribution of imprecision and/or incompleteness on the outcomes of a model in order to know if an ambiguous answer is due to a lack of information or due to a random phenomenon. That is, we plan to develop refined sensitivity analysis techniques based on their use. They should thus determine key variables and/or key phenomena for which it will be necessary to acquire more information. Finally, we also plan to investigate their usefulness in determining optimal control commands.

References

[1] C. Baudrit, M. Sicard, PH Wuillemin, and N. Perrot. Towards a global modelling of the camembert-type cheese ripening process by coupling heterogeneous knowledge with dynamic bayesian networks. *Journal of Food Engineering*, 98(3):283–293, 2010.

[2] C. Baudrit, P.H. Wuillemin, and N. Perrot. Parameter elicitation in probabilistic graphical models for modelling multi-scale food complex systems. *Journal of Food Engineering*, 115(1):1 – 10, 2013.

[3] S. Benferhat and S. Smaoui. Hybrid possibilistic networks. *International journal of approximate reasoning*, 44(3):224–243, 2007.

[4] L.K. Blackmond. Sensitivity analysis for probability assessments in bayesian networks. In *Proceedings of the Ninth international conference on Uncertainty in artificial intelligence*, UAI'93, pages 136–142, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[5] W. Buntine. A guide to the literature on learning probabilistic networks from data. *Knowledge and Data Engineering, IEEE Transactions on*, 8(2):195–210, 1996.

[6] A. Cano, J. Cano, and S. Moral. Convex sets of probabilities propagation by simulated annealing on a tree of cliques. In *In: Proceedings of Fifth International Conference on Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU '94*, pages 4–8, 1994.

[7] A. Cano, M. Gómez, S. Moral, and J. Abellán. Hill-climbing and branch-and-bound algorithms for exact and approximate inference in credal networks. *Int. J. Approx. Reasoning*, 44(3):261–280, 2007.

[8] A. Cano and S. Moral. A genetic algorithm to approximate convex sets of probabilities. In *Proc. of the*

Int. Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pages 859–864, 1996.

[9] A. Cano and S. Moral. Using probability trees to compute marginals with imprecise probabilities. *International Journal of Approximate Reasoning*, 29(1):1 – 46, 2002.

[10] E. Castillo, R.M. Gutiérrez, and A.S. Hadi. Sensitivity analysis in discrete bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics*, 27:412–423, 1997.

[11] H. Chan and A. Darwiche. Sensitivity analysis in bayesian networks: From single to multiple parameters. In *Proceedings of the 20'th international Conference on Uncertainty in Artificial Intelligence*, UAI'04, 2004.

[12] H. Chan and A. Darwiche. On the revision of probabilistic beliefs using uncertain evidence. *Artif. Intell.*, 163(1):67–90, March 2005.

[13] T. Charitos and L.C. Van der Gaag. Sensitivity analysis for threshold decision making with dynamic networks. In *Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 72–79, Arlington, Virginia, 2006. AUAI Press.

[14] V.M.H Coupe and L.C. Van Der Gaag. Properties of sensitivity analysis of bayesian belief networks. In *Proceedings of the Joint Session of the 6th Prague Symposium of Asymptotic Statistics and the 13th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Union of Czech Mathematicians and Physicists*, pages 81–86, 1999.

[15] I. Couso, S. Moral, and P. Walley. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5(2):165–181, 2000.

[16] F.G. Cozman. Robustness analysis of bayesian networks with local convex sets of distributions. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, UAI'97, pages 108–115, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[17] F.G. Cozman. Credal networks. *Artificial Intelligence*, 120(2):199–233, 2000.

[18] F.G. Cozman. Separation properties of sets of probability measures. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 107–114, 2000.

[19] F.G. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39(2):167–184, 2005.

[20] J.C.F. da Rocha and F.G. Cozman. Inference with separately specified sets of probabilities in credal networks. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, UAI'02, pages 430–437, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[21] J.C.F. da Rocha, F.G. Cozmanl, and C.P. de Campos. Inference in polytrees with sets of probabilities. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, UAI'03, pages 217–224, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[22] J. De Bock and G. de Cooman. State sequence prediction in imprecise hidden markov models. In *Proceedings of the seventh International Symposium on Imprecise Probabilities: Theory and Applications*, pages 159–168, 2011.

[23] C.P. de Campos and F.G. Cozman. Inference in credal networks using multilinear programming. In *In Proceedings of the 2nd Starting AI Researchers' Symposium*, pages 50–61, 2004.

[24] C.P. de Campos and F.G. Cozman. Inference in credal networks through integer programming. In *In Proceedings of the 5th International Symposium on Imprecise Probability: Theories and Applications*, pages 145–154, 2007.

[25] L.M. De Campos, J.F. Huete, and S. Moral. Probability intervals: a tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2(2):167–196, 1994.

[26] K.V. Delgado, L.N. de Barros, F.G. Cozman, and R. Shirota. Representing and solving factored markov decision processes with imprecise probabilities. In *Proceedings ISIPTA, Durham, United Kingdom*, pages 169–178, 2009.

[27] E. Fagiuoli and M. Zaffalon. 2U: an exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106(1):77 – 107, 1998.

[28] Patrick F Fox. *Cheese - Chemistry, Physics and Microbiology; 3rd ed.* Elsevier, San Diego, CA, 2004.

[29] James E. Gentle. Monte carlo methods. *Encyclopedia of Statistical Sciences*, 2006.

[30] J.S. Ide and F.G. Cozman. Approximate Inference in Credal Networks by Variational Mean Field Methods. In *International Symposium on Imprecise Probabilities and Their Applications*, pages 203–212, 2005.

[31] J.S. Ide and F.G. Cozman. Approximate algorithms for credal networks with binary variables. *International Journal of Approximate Reasoning*, 48(1):275 – 296, 2008. Special Section: Perception Based Data Mining and Decision Support Systems.

[32] J.S. Ide and Cozman F.G. Ipe and l2u: Approximate algorithms for credal networks. In *Proceedings of the second starting AI Researcher Symposium*, pages 118–127. IOS Press, 2004.

[33] J. Kwisthout and L.C. Van der Gaag. The computational complexity of sensitivity analysis and parameter tuning. In David A. McAllester and Petri Myllymäki, editors, *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, July 9-12, 2008, Helsinki, Finland*, UAI'08, pages 349–356. AUAI Press, 2008.

[34] S.L. Lauritzen. The em algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, 19(2):191–201, 1995.

[35] I. Levi. *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probobility, and Chance*. MIT press, 1983.

[36] K.P. Murphy. *Dynamic Bayesian networks: representation, inference and learning*. PhD thesis, University of California, 2002.

[37] T.D. Nielsen and F.V. Jensen. *Bayesian networks and decision graphs*. Springer, 2007.

[38] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[39] N. Perrot, IC Trelea, C. Baudrit, G. Trystram, and P. Bourgine. Modelling and analysis of complex food systems: State of the art and new trends. *Trends in Food Science & Technology*, 22(6):304–314, 2011.

[40] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

[41] J.C. Richard. *The Logic of Decision*. University Of Chicago Press, 1983.

[42] J.C.F. Rocha and F.G. Cozman. Evidence propagation in credal networks: An exact algorithm based on separately specified sets of probability. In G. Bittencourt and G.L. Ramalho, editors, *Advances in Artificial Intelligence*, volume 2507, pages 376–385. 2002.

[43] L.C. Van der Gaag and U. Kjaerulff. Making sensitivity analysis computationally efficient. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, UAI'00, pages 317–325. Morgan Kaufmann Publishers, 2000.

[44] P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall London, 1991.

[45] H. Xu and P. Smets. Reasoning in evidential networks with conditional belief functions. *International Journal of Approximate Reasoning*, 14(2):155–185, 1996.

[46] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Generalized belief propagation. In *NIPS 13*, pages 689–695. MIT Press, 2000.

Second-Order Credal Combination of Evidence

Alexander Karlsson
Infofusion/Informatics Research Center
University of Skövde, Sweden
alexander.karlsson@his.se

David Sundgren
Department of Computer and Systems Sciences
Stockholm University, Sweden
dsn@dsv.su.se

Abstract

We utilize second-order probability distributions for modeling second-order information over imprecise evidence in the form of credal sets. We generalize the Dirichlet distribution to a shifted version, denoted the S-Dirichlet, which allows one to restrict the support of the distribution by lower bounds. Based on the S-Dirichlet distribution, we present a simple combination schema denoted as second-order credal combination (SOCC), which takes second-order probability into account. The combination schema is based on a set of particles, sampled from the operands, and a set of weights that are obtained through the S-Dirichlet distribution. We show by examples that the second-order probability distribution over the imprecise joint evidence can be remarkably concentrated and hence that the credal combination operator can significantly overestimate the imprecision.

Keywords. Second-order credal combination, imprecise probability, credal sets, second-order probability, combination, evidence

1 Introduction

One common way of representing belief imprecisely is by so called *credal sets* [19], i.e., a *closed and convex set of (discrete) probability distributions.* If one utilizes this belief structure in a *Bayesian context*, where a *prior* distribution is updated to a *posterior* by a *likelihood function*, one ends up with what is referred to as *credal set theory* [8, 18], similar to *robust Bayesian theory* [1, 2, 4, 14] but without the sensitivity interpretation. Credal set theory can be thought of as a straightforward generalization of Bayesian theory to imprecise probability since Bayes' theorem is applied point-wise on all elements in a prior credal set and a set of likelihood functions. In fact, Bayesian theory is the special case of credal set theory when all sets are singletons.

Karlsson et al. [18] have previously shown that credal set theory can yield posterior credal sets that are highly imprecise and that this extra imprecision can have a deteriorating effect on decision-making, even though there exists inherent imprecision in the decision situation. The main focus in this paper is to take one step further than only modeling imprecision by adding information in the form of *second-order probability* [9, 20, 28], thus qualifying the imprecision. This type of research was briefly initiated by Karlsson et al. [17] where the consequences of modeling second-order probability was explored by using the uniform distribution in the simple case where only two states were present in the state space. In that case, it was found that the second-order probability over posterior imprecision can be considerably skewed.

In order to model different second-order probability distributions, for any number of states, we will here generalize the *Dirichlet distribution* to allow for non-zero lower bounds, denoted the *shifted Dirichlet distribution* (S-Dirichlet). Compared to the Dirichlet distribution, the S-Dirichlet distribution has twice as many parameters since for each variable there is the usual Dirichlet parameter but also a lower bound that allows one to restrict the support of the distribution. The unique family of second-order distributions that factorizes into marginals, presented in Sundgren et al. [27], is a special case of the S-Dirichlet distribution with fixed Dirichlet parameters. In this case, the joint second-order distribution equals the normalized product of its own marginal distributions and do not model any dependencies between first-order probability other than the necessary requirement that the probabilities sum to one.

We will utilize the S-Dirichlet distribution for exploring the consequences of modeling second-order probability over *imprecise evidence*, in the form of credal sets, which we combine into a single imprecise joint evidence, also in the form of a credal set, while preserving second-order information. We introduce a simple particle based method for performing the computation and we denote such a schema by *second-order credal combination* (SOCC). The purpose is to utilize this schema in order to explore the concentration and placement of the resulting particle cloud. Our schema is based on considering likelihoods as first-order evidence and second-order probability in the form of the S-Dirichlet distribution, however, it should be noted that there exists previous approaches where likelihoods has been considered as a possibilistic second-order (hierarchical) model [5].

The paper is organized as follows: in Section 2, we formalize and give an overview of the problem of combining independent pieces of evidence from different sources. In Section 3, we generalize the Dirichlet distribution to the shifted version. In Section 4, we present a simple method for performing SOCC. Lastly, in Section 5, we summarize the paper and provide our conclusions.

2 Preliminaries

We here present some background material that the remaining paper relies on. We start by providing a general overview of the problem of combining independent pieces of evidence. Based on this overview, we then present how the combination problem can be tackled within the framework of credal sets [19], namely by using the *credal combination operator* also known as the *robust Bayesian combination operator* [1, 2].

2.1 Combination of Evidence

Combination of independent pieces of evidence from multiple sources, e.g., sensors, is a problem that has been extensively studied within many different variants of *evidence theory*, e.g., [10, 22, 24]. Common to all these theories is that evidence are represented imprecisely by so called *mass functions* which operate on the power set of some state space. Pieces of evidence are combined utilizing a so called *combination operator*, e.g., *Dempster's rule of combination* [22]. One important aspect to consider when performing such combination is that of independence. Loosely, this means that one piece of evidence should not be informative regarding the other piece (see further [23]).

The problem of combining evidence has not been equally well studied under a Bayesian perspective or within the framework of credal sets. However, Arnborg [1, 2] has explored the relationship between *robust Bayesian theory* [4, 14], which can be considered a sensitivity interpretation of credal sets, and evidence theory. He found that the results of these theories can even be disjoint. One key observation when considering the combination problem within a Bayesian or credal framework is that it is the *likelihoods* that constitute evidence. Let us further elaborate on this in the next section.

2.2 Credal Combination of Evidence

Since the *credal combination operator* [15, 18], introduced as the *robust Bayesian combination operator* by Arnborg [1, 2] (we deliberately avoid using this terminology since it imposes a sensitivity interpretation of the imprecision), is a direct generalization of its Bayesian counterpart, we start by elaborating on how the latter can be derived. The derivation is similar to Karlsson et al. [18], and is inspired by Arnborg [1, 2]. Assume that we have a random variable X for which we are uncertain about the true value. Let the state space for X be denoted by $\Omega_X \triangleq \bigcup_{i=1}^{n}\{x_i\}$ and that we can obtain observations $y_1 \in \Omega_{Y_1}$ and $y_2 \in \Omega_{Y_2}$. We can then use Bayes' theorem in order calculate the posterior distribution, or belief, regarding the true value of X given these observations:

$$p(X|y_1, y_2) = \frac{p(y_1, y_2|X)p(X)}{\sum\limits_{x \in \Omega_X} p(y_1, y_2|x)p(x)} \quad . \tag{1}$$

From the above equation, we see that the observations y_1 and y_2 only affect the posterior through the joint likelihood $p(y_1, y_2|X)$, which hence constitutes the evidence based on the observations. Now by assuming conditional independence, we obtain:

$$p(y_1, y_2|X) = p(y_1|X)p(y_2|X), \tag{2}$$

i.e., one observation is not informative about the other given that we know the true state of X. The above equation is essentially all that we need in order to combine two pieces of evidence in the form of likelihood functions into a single joint evidence. However, in order to avoid a monotonically decreasing joint evidence, it is convenient to normalize the joint evidence to a probability function. By also normalizing the likelihoods, we have constructed an operator where both the operands and result are evidences in the form of probability functions. Note that these normalizations do not affect the resulting posterior distribution since it is only the relative strengths of likelihoods that determines the posterior (see further Karlsson et al. [18] for more detail). Based on this line of reasoning, we are now ready to define *the Bayesian combination operator* [1, 2, 15, 18]:

Definition 1. *The Bayesian combination operator $\Phi_{\mathcal{B}}$ is defined as*

$$\Phi_{\mathcal{B}}(\hat{p}(y_1|X), \hat{p}(y_2|X))) \triangleq \frac{\hat{p}(y_1|X)\hat{p}(y_2|X)}{\sum\limits_{x\in\Omega_X} \hat{p}(y_1|x)\hat{p}(y_2|x)}, \quad (3)$$

where $\hat{p}(y_i|X)$, $i \in \{1,2\}$, are normalized likelihood functions satisfying conditional independence in the sense of Equation (2)*. The operator is undefined iff $\sum_{x\in\Omega_X} \hat{p}(y_1|x)\hat{p}(y_2|x) = 0$.*

Let us continue by elaborating on how *the credal combination operator* can be derived based on Def. 1. Since we now move into the domain of imprecise probability, we are allowed to utilize *a closed and convex set of probability distributions*, i.e., a *credal set* [19]. Convexity enables one to perform computation by the sets' *extreme points* (see further Karlsson et al. [18, Theorem 2]). Now, instead of evidence in the form of a single normalized likelihood function, we have credal sets of such functions, denoted by $\hat{\mathcal{P}}(y_1|X)$ and $\hat{\mathcal{P}}(y_2|X)$, which we want to combine into a single joint evidence $\hat{\mathcal{P}}(y_1, y_2|X)$. In order to regard the evidences as independent, the *extreme points* needs to factorize, denoted *strong independence* [7], i.e., for each extreme point $\hat{p}_e(y_1, y_2|X)$ we have:

$$\hat{p}_e(y_1, y_2|X) = \hat{p}(y_1|X)\hat{p}(y_2|X) \quad (4)$$

where $\hat{p}(y_i|X) \in \hat{\mathcal{P}}(y_i|X)$, $i \in \{1,2\}$. The combination is then performed by the credal combination operator, which simply applies the Bayesian combination operator point-wise on each combination of functions from the operand sets and as last step one applies the convex-hull operator [1, 2, 15, 18]:

Definition 2. *The credal combination operator $\Phi_{\mathcal{C}}$ is defined as*

$$\Phi_{\mathcal{C}}(\hat{\mathcal{P}}(y_1|X), \hat{\mathcal{P}}(y_2|X))) \triangleq \mathcal{CH}\left(\left\{\Phi_{\mathcal{B}}(\hat{p}(y_1|X), \hat{p}(y_2|X))) : \hat{p}(y_i|X) \in \hat{\mathcal{P}}(y_i|X), i \in \{1,2\}\right\}\right), \quad (5)$$

where $\hat{\mathcal{P}}(y_i|X)$, $i \in \{1,2\}$, are credal sets of normalized likelihood functions satisfying strong independence in the sense of Eq. (4)*, $\Phi_{\mathcal{B}}$ is the Bayesian combination operator, and $\mathcal{CH}$ denotes the convex hull. The operator is undefined iff there exist a pair $\hat{p}(y_1|X) \in \hat{\mathcal{P}}(y_1|X)$ and $\hat{p}(y_2|X) \in \hat{\mathcal{P}}(y_2|X)$ for which $\Phi_{\mathcal{B}}$ is undefined.*

Note that when only singleton sets are used as operands, the credal combination operator is equivalent to the Bayesian counterpart.

One important type of credal set that we will use throughout the article is the *probability simplex*, i.e., the set of all probability distributions over a given state space, formally defined as:

Definition 3. *The set of all probability distributions $\mathcal{P}^*(X)$, i.e., the probability simplex, over a given state space Ω_X is defined as*

$$\mathcal{P}^*(X) \triangleq \left\{p(X) : p(x) \geq 0, \sum_{x\in\Omega_X} p(x) = 1\right\}. \quad (6)$$

Another important concept with respect to imprecise probability is the *degree of imprecision* of a credal set. When we refer to "imprecision" in this article we perform averaging of the imprecision for single states [29, 16]:

Definition 4. *The degree of imprecision $\mathcal{I}(\hat{\mathcal{P}}(y|X))$ of a credal set of normalized likelihood functions $\hat{\mathcal{P}}(y|X)$ is defined as:*

$$\mathcal{I}(\hat{\mathcal{P}}(y|X)) \triangleq \frac{1}{|\Omega_X|}\sum_{x\in\Omega_X}\left(\max_{\hat{p}(y|X)\in\hat{\mathcal{P}}(y|X)} \hat{p}(y|x) - \min_{\hat{p}(y|X)\in\hat{\mathcal{P}}(y|X)} \hat{p}(y|x)\right) \quad (7)$$

Please note that we only include the above definition to unambiguously declare the term imprecision. It will not be utilized for any computation in the paper.

3 Shifted Dirichlet Distributions

Probability values can be considered as random variables themselves and the corresponding distributions over such variables is referred to as a *second-order probability distribution* [9, 20, 28]. Any probability distribution that has support on the probability simplex (Def. 3), e.g. a Dirichlet distribution, can be seen as a second-order probability distribution.

The Dirichlet family of distributions can be generalized to have support on subsets of the probability simplex by using lower bounds l_i on the random variables P_i corresponding to first-order probabilities. Just as with related models such as possibility measures [30], belief functions [22], Choquet capacities of order 2 [6] and coherent upper and lower probabilities [25], lower bounds l_i of probabilities determine upper bounds by $1 - \sum_{j\neq i} l_i$. There are other possibilities for lower and upper bounds for the support of second-order probability distribution, e.g., it is possible to give lower and upper bounds for all but one of the first-order probabilities as in, e.g., Sundgren et al. (2009) [26], but for simplicity we give lower bounds l_i to all

$n = |\Omega_X|$ random variables P_i such that $\sum_{i=1}^{n} l_i \leq 1$ and $\sum_{i=1}^{n} P_i = 1$, $l_i \leq P_i \leq 1 - \sum_{j \neq i} l_i$.

A probability distribution whose support has been shifted needs renormalization to remain a probability distribution. Let us then look at the probability density function of the *Shifted Dirichlet* family that allows for non-zero lower bounds. If $\sum_{i=1}^{n} l_i \leq 1$ and $\sum_{i=1}^{n} P_i = 1$, $l_i \leq P_i \leq 1 - \sum_{j \neq i} l_i$ then the function:

$$f(\{P_i\}_{i=1}^{n},\{\alpha_i\}_{i=1}^{n},\{l_i\}_{i=1}^{n}) = \frac{\Gamma\left(\sum_{i=1}^{n} \alpha_i\right) \prod_{i=1}^{n} (P_i - l_i)^{\alpha_i - 1}}{\left(1 - \sum_{i=1}^{n} l_i\right)^{\sum_{i=1}^{n} \alpha_i - 1} \prod_{i=1}^{n} \Gamma(\alpha_i)} , \tag{8}$$

is the probability density function of a probability distribution, where P_i are random variables, α_i are the parameters of a proper Dirichlet distribution and l_i are parameters that determine lower bounds of the variables P_i (see the Appendix for a proof). Note that f is a function of P_i only ($\{\alpha_i\}_{i=1}^{n}$ and $\{l_i\}_{i=1}^{n}$ are parameters).

4 Second-Order Credal Combination

Now assume that two *agents*, $i \in \{1, 2\}$, for two different types of sensors, have extracted features y_i and that the agents based on this express imprecise independent evidence through lower bounds on normalized likelihoods $\{l_i^j \leq \hat{p}(y_i|x_j)\}_{j=1}^{n}$ where $\sum_{j=1}^{n} l_i^j \leq 1$. These lower bounds can then be utilized in order to construct evidence in the form of credal sets of normalized likelihoods by:

$$\hat{\mathcal{P}}(y_i|X) \triangleq \left\{ \hat{p}(y_i|X) : l_i^j \leq \hat{p}(y_i|x_j), \sum_{j=1}^{n} \hat{p}(y_i|x_j) = 1 \right\} . \tag{9}$$

In addition, the agents also express theirs beliefs over these imprecise operands by specifying alpha-values for the S-Dirichlet distribution, i.e., $\{\alpha_j^i\}_{j=1}^{n}$. The goal then is to construct a schema for combination that do not only takes evidence in the form of credal sets into account but also *second-order probability* in the form of S-Dirichlet distributions. We will denote such schema by *second-order credal combination* (SOCC).

In order to achieve a computationally feasible schema for SOCC, we propose a simple method for approximating the second-order distribution over the joint evidence by simulation [12, 3]. Typically, these types of simulation utilize a set of so called *particles*[1], i.e., samples, and a set of corresponding *weights* of these particles. Such a representation has, as an example, previously been proposed as a model for *epistemic reliability* by Gärdenfors and Sahlin [11]. Now, we can obtain a set of particles, denoted $\{\hat{p}_j(y_i|X)\}_{j=1}^{m}$ where each $\hat{p}_j(y_i|X) \in \hat{\mathcal{P}}_j(y_i|X)$, and a set of weights, denoted $\{w_j^i\}_{j=1}^{m}$, by expanding a grid with a given precision over each operand. At each point in the grid we can compute the density value of the S-Dirichlet and then normalize with respect to all points in the grid [12, Chapter 11]. We can then use the grid as a basis for drawing m particles with replacement. Given these particles $\{\hat{p}_j(y_i|X)\}_{j=1}^{n}$, and the S-Dirichlet density, we can obtain the corresponding weights by:

$$w_j^i = \frac{f(\{\hat{p}_j(y_i|x_k)\}_{k=1}^{n}, \{\alpha_j^i\}_{j=1}^{n}, \{l_j^i\}_{j=1}^{n})}{\sum_{j=1}^{n} f(\{\hat{p}_j(y_i|x_k)\}_{k=1}^{n}, \{\alpha_j^i\}_{j=1}^{n}, \{l_j^i\}_{j=1}^{n})} . \tag{10}$$

where f is the S-Dirichlet density defined by Eq. (8). Since we now have particles and weights:

$$\Lambda_i \triangleq \{(\hat{p}_j(y_i|X), w_j^i)\}_{j=1}^{m} \tag{11}$$

from each operand $i \in \{1, 2\}$, we can compute an approximation of the second-order distribution over the joint evidence by combining pairs of particles:

$$\Lambda_{1,2} \triangleq \left\{ \left(\Phi_{\mathcal{B}}(\hat{p}_j(y_1|X), \hat{p}_j(y_2|X)), \frac{w_j^1 w_j^2}{\sum_{j=1}^{m} w_j^1 w_j^2} \right) \right\}_{j=1}^{m} . \tag{12}$$

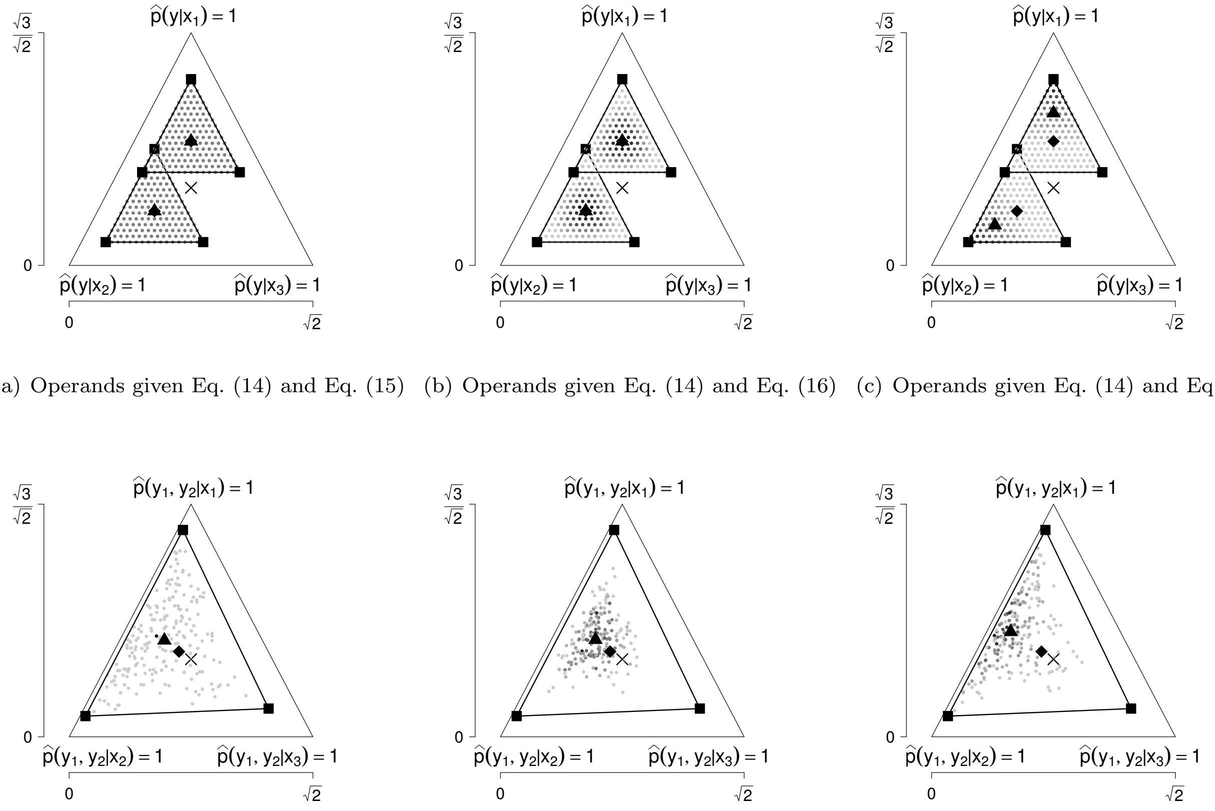
We can combine the above representation with a new operand by drawing a number of particles with replacement from $\Lambda_{1,2}$ where the sampling probability for each particle is given by its weight (similar to *resampling* in the case of particle filtering [3]). The difference is that since we already have a particle representation for $\Lambda_{1,2}$ we do not have to expand a grid for that operand.

In addition to the above described combination schema one can also perform credal combination, i.e:

$$\hat{\mathcal{P}}(y_1, y_2|X) \triangleq \Phi_{\mathcal{C}}(\hat{\mathcal{P}}(y_1|X), \hat{\mathcal{P}}(y_2|X))) . \tag{13}$$

By doing so, one preserve information about the extreme values, which could be valuable for a decision-maker in applications where there exists a strong risk component.

[1] The term “particle filtering” is frequently used in the tracking literature [3].



(a) Operands given Eq. (14) and Eq. (15) (b) Operands given Eq. (14) and Eq. (16) (c) Operands given Eq. (14) and Eq. (17)

(d) SOCC given Eq. (14) and Eq. (15) (e) SOCC given Eq. (14) and Eq. (16) (f) SOCC given Eq. (14) and Eq. (17)

Figure 1: The figures depict the probability simplex for three states where the upper figures show operands for Eq. (14) and Eqs. (15) – (17), and the lower figures show the result of performing SOCC. The intensity of grey shows the weights of the particle (identical particles have been merged by adding the weights), i.e., darker particles have more weights. Squares show extreme points of the credal sets, triangles show the expected value with respect to the particles, diamonds show the centroid of the credal sets, and the uniform distribution over Ω_X is indicated with a cross. The grid used for the operands has been obtained through probability vectors of rational numbers $(a, b, c)/40$, where $a+b+c = 40$, satisfying the lower bounds. In the figures, $m = 200$ particles have been sampled.

4.1 Examples

Let us now study SOCC through some examples where we use the S-Dirichlet distribution for expressing belief over imprecise operands and explore the appearance of the second-order distribution over the imprecise joint evidence. Assume that the two agents provide the following lower bounds on normalized likelihoods:

$$\begin{aligned} l_1 &= (0.1, 0.4, 0.1) \\ l_2 &= (0.4, 0.1, 0.1)\,, \end{aligned} \tag{14}$$

which then can be used in Eq. (9) for constructing $\hat{\mathcal{P}}(y_1|X)$ and $\hat{\mathcal{P}}(y_2|X)$, respectively. Given these lower bounds, we will explore the result of SOCC for three different, somewhat arbitrarily chosen although with some intuition, S-Dirichlet distributions:

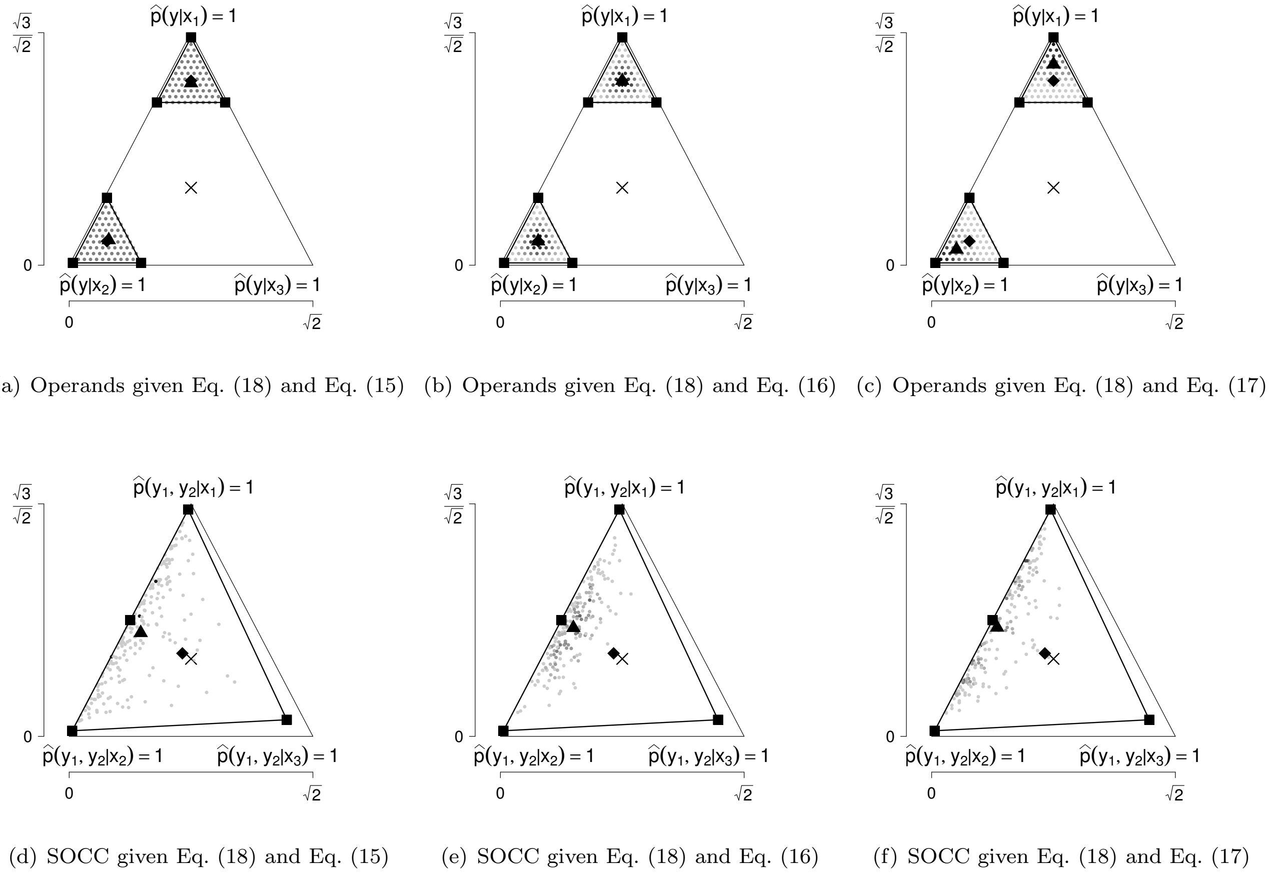
$$\alpha_1 = \alpha_2 = (1, 1, 1) \tag{15}$$

$$\alpha_1 = \alpha_2 = (3, 3, 3) \tag{16}$$

$$\alpha_1 = (1, 3, 1),\ \alpha_2 = (3, 1, 1)\,, \tag{17}$$

where Eq. (15) corresponds to the uniform (Bayes-Laplace) distribution; Eq. (16) is a case where the center of the credal sets is reinforced; and Eq. (17) is a case where the corner closest to some state has been reinforced. The result of applying SOCC on Eq. (14) and Eqs. (15) – (17) is shown in Fig. 1.

We see that the uniform distribution, i.e., Fig. 1(a) and 1(d), yields a particle cloud that is more scattered compared to the other S-Dirichlet distributions. Furthermore, we see that utilizing the S-Dirichlet



(a) Operands given Eq. (18) and Eq. (15) (b) Operands given Eq. (18) and Eq. (16) (c) Operands given Eq. (18) and Eq. (17)

(d) SOCC given Eq. (18) and Eq. (15) (e) SOCC given Eq. (18) and Eq. (16) (f) SOCC given Eq. (18) and Eq. (17)

Figure 2: The figures depict the probability simplex for three states where the upper figures show operands for Eq. (18) and Eqs. (15) – (17), and the lower figures shows the result of performing SOCC. The indicators and other settings are identical as in Fig. 1.

distribution that emphasizes the corners, defined by Eq. (17) and shown in Figs. 1(c) and 1(f), yields an expected value that has a lower probability for state x_3 than the others.

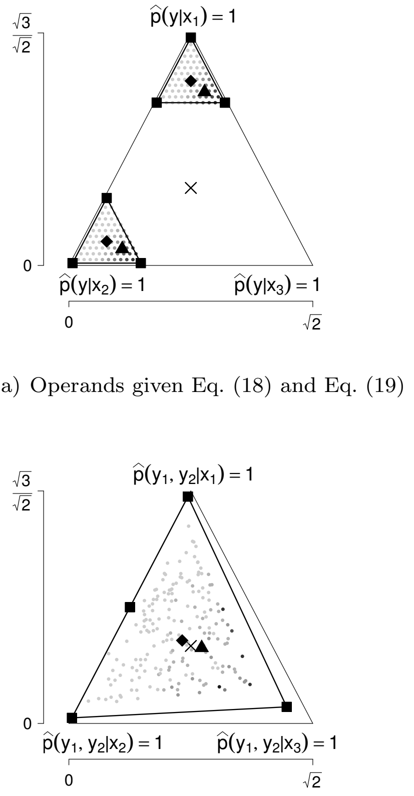
One key observation shown by Figs. 1(e) and 1(f), is that the particle cloud is fairly concentrated within the credal set, which in a sense means that the credal combination operator to some degree overestimates the imprecision. Such an overestimation is even more evident in the following example defined by:

$$\begin{aligned} l_1 &= (0.01, 0.7, 0.01) \\ l_2 &= (0.7, 0.01, 0.01)\,, \end{aligned} \tag{18}$$

and shown in Fig. 2. If one would have ignored the particle cloud in this example and only base a decision upon the posterior imprecision, it is quite likely that the true state could be x_3 since the lower right extreme point is quite close to the lower right extreme point of the simplex. Such results are somewhat counter-intuitive when interpreting what the evidence from the agents actually express, i.e., evidence for x_1 and x_2, and both pieces constitute counter evidence against x_3 since the operands are positioned far away from the corner corresponding to x_3. However, when combining the two lower right extreme points of the operands, the states x_1 and x_2 are more or less eliminated by the agents, since both of these extreme points are close to the boundary of the simplex where the probability of the these states is close to zero, in contrast to the probability of state x_3, which is not close to any boundary in relative terms. Therefore these lower right extreme points of the operand credal sets gets reinforced to the lower right extreme point of the joint evidence. This case bares close resemblance to Zadeh's counter example [31] against Dempster-Shafer theory [22]. In that example, when combining evidence in the form of *mass functions*, one ends up with a result where all mass lies on the single state that the pieces of evidence only weakly indicated (see further Karlsson et al. [18]). In contrast, from the particle clouds and expected values, we see a clear concentration around the left boundary of the joint evidence; hence in agreement with the intuition that the true state is not likely to be x_3. If the mass instead are concentrated around the lower extreme points of the operands, e.g., by using an S-Dirichlet with the following parameters instead:

$$\alpha_1 = (1, 1, 3),\ \alpha_2 = (1, 1, 3)\,, \tag{19}$$

we obtain the results seen in Fig. 3. In contrast to the former case, the mass is fairly uniformly distributed over the joint credal set.



(a) Operands given Eq. (18) and Eq. (19)

(b) SOCC given Eq. (18) and Eq. (19)

Figure 3: The figures depict the probability simplex for three states where the operands are defined by Eqs. (18) – (19). The indicators and other settings are identical as in Fig. 1.

The cases shown in Figs. 2 – 3 demonstrates that second-order information could be valuable to a decision-maker when imprecision in decision problems are modelled.

5 Summary and Conclusions

We have generalized the Dirichlet distribution to the S-Dirichlet distribution, where the Dirichlet parameters can be used to model different second-order probability distributions over a restricted region defined by lower bounds. Based on the S-Dirichlet distribution, we presented a simple combination schema, denoted as *second-order credal combination* (SOCC), which takes second-order probability into account. The combination schema is based on a set of particles, sampled from the operands, and a set of weights that are obtained through the S-Dirichlet distribution. We then gave some example of SOCC utilizing different types of S-Dirichlet distributions. By the examples, we showed that the particle cloud over the joint evidence can be remarkably concentrated in comparison to the credal set obtained by credal combination.

One new feature that is enabled through SOCC is that it provides a grounded way of selecting a single probability function from the credal set to base one's decision upon; simply use the expected value with respect to the particle cloud. Such a schema is useful when a single decision is necessary, something that is common in many application scenarios, and is similar to what Smets and Kennes [24] has proposed in the *transferable belief model*, i.e., as long as a single decision does not have to be implemented, use mass functions, otherwise transform the mass function to a single probability function and use that for deciding on a single state. Utilizing the expected value of the particle cloud should be put in contrast to utilizing the centroid distribution, i.e., the expected value with respect to a uniform second-order distribution over the joint evidence. Since uniformity is in general not the case, as is seen in Figs. 2(d) – 2(f) (see also Karlsson et al. [17]), there is in principle no reason to utilize the centroid. Another alternative is utilizing the maximum entropy distribution [1, 2], representing a cautious choice, however, in applications without a risk component, the maximum entropy distribution is likely to be too cautious.

Given the examples where the particle clouds seems to be quite concentrated in comparison to the resulting credal sets, one legitimate question is whether or not it is reasonable to utilize the credal combination operator solely, i.e., without modeling second-order probability. Could it be so that it is always preferable to model second-order probability when imprecision appears in a decision problems? Perhaps the credal combination operator can be appropriate to utilize when the imprecise operands are a consequence of small perturbations of some precise evidence as is done in sensitivity analysis (robust Bayesian theory) [4, 14]. In such a setting it seems reasonable to only model imprecision, and not second-order probability, due to that only low degrees of imprecision in the operands are considered. For these cases one is likely to infer the same conclusions irrespective of any second-order probability since the perturbation of the operands is performed so that every point in the perturbed set is a reasonable precise evidence. Consequently, every perturbed point in the resulting joint evidence is a reasonable joint evidence that a decision maker should be willing to act upon, irrespective of the amount of density such a point possesses.

When the imprecise operands are not a consequence of sensitivity analysis, i.e., when the degree of imprecision of the operands could be considerably higher, then, as our results suggest, second-order probability is likely to be an important modeling tool that cannot be neglected without consequences. In our future research, we will therefore continue by exploring how one can perform different modeling tasks using second-order probability, i.e., how SOCC can be applied in an application scenario.

Acknowledgements

This work was supported by the Information Fusion Research Program (University of Skövde, Sweden), in partnership with the Swedish Knowledge Foundation under grant 2010-0320 (URL: `http://www.infofusion.se`, UMIF project), and Formas. R [21] and the package rcdd [13] have been used for computations in this paper.

Appendix

We here prove Eq. (8). Let us use the following shorthand notation ($n = |\Omega_X|$):

$$\gamma \triangleq \frac{\Gamma\left(\sum_{i=1}^{n} \alpha_i\right)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \ . \tag{20}$$

We need to show that:

$$\int\limits_{\substack{\sum_{i=1}^{n} P_i = 1 \\ P_i \geq l_i}} \gamma \frac{\prod_{i=1}^{n} (P_i - l_i)^{\alpha_i - 1}}{\left(1 - \sum_{i=1}^{n} l_i\right)^{\sum_{i=1}^{n} \alpha_i - 1}} \, d\mathbf{P} = 1 \ . \tag{21}$$

Since a proper Dirichlet distribution has probability density function:

$$f(\{P_i\}_{i=1}^{n}, \{\alpha_i\}_{i=1}^{n}) = \gamma \prod_{i=1}^{n} P_i^{\alpha_i - 1} \,, \tag{22}$$

we know that:

$$\int\limits_{\substack{\sum_{i=1}^{n} P_i = 1 \\ P_i \geq 0}} \gamma \prod_{i=1}^{n} P_i^{\alpha_i - 1} \, d\mathbf{P} = 1 \ . \tag{23}$$

Let us replace P_i with $P_i - l_i$ and restrict the support from $\sum_{i=1}^{n} P_i = 1, P_i \geq 0$ to $\sum_{i=1}^{n} P_i = 1, l_i \leq P_i \leq 1 - \sum_{j \neq i} l_i$. Then, through the variable change:

$$Y_i = \frac{P_i - l_i}{1 - \sum_{i=1}^{n} l_i} \,, \tag{24}$$

where $i \in \{1, \dots, n\}$, we find that:

$$\begin{aligned}
&\int\limits_{\substack{\sum_{i=1}^{n} P_i = 1 \\ P_i \geq l_i}} \gamma \prod_{i=1}^{n} (P_i - l_i)^{\alpha_i - 1} \, d\mathbf{P} = \\
&\int\limits_{\substack{\sum_{i=1}^{n} Y_i = 1 \\ Y_i \geq 0}} \gamma \prod_{i=1}^{n} \left(Y_i \left(1 - \sum_{i=1}^{n} l_i\right)\right)^{\alpha_i - 1} \left|\frac{\partial \mathbf{P}}{\partial \mathbf{Y}}\right| d\mathbf{Y} = \\
&\int\limits_{\substack{\sum_{i=1}^{n} Y_i = 1 \\ Y_i \geq 0}} \gamma \prod_{i=1}^{n} \left(Y_i \left(1 - \sum_{i=1}^{n} l_i\right)\right)^{\alpha_i - 1} \\
&\qquad \left(1 - \sum_{i=1}^{n} l_i\right)^{n-1} d\mathbf{Y} = \\
&\int\limits_{\substack{\sum_{i=1}^{n} Y_i = 1 \\ Y_i \geq 0}} \gamma \prod_{i=1}^{n} Y_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{n} l_i\right)^{\sum_{i=1}^{n} \alpha_i - n} \\
&\qquad \left(1 - \sum_{i=1}^{n} l_i\right)^{n-1} d\mathbf{Y} = \\
&\left(1 - \sum_{i=1}^{n} l_i\right)^{\sum_{i=1}^{n} \alpha_i - 1} \int\limits_{\substack{\sum_{i=1}^{n} Y_i = 1 \\ Y_i \geq 0}} \gamma \prod_{i=1}^{n} Y_i^{\alpha_i - 1} \, d\mathbf{Y} = \\
&\left(1 - \sum_{i=1}^{n} l_i\right)^{\sum_{i=1}^{n} \alpha_i - 1} \ .
\end{aligned} \tag{25}$$

Therefore:

$$\frac{1}{\left(1 - \sum_{i=1}^{n} l_i\right)^{\sum_{i=1}^{n} \alpha_i - 1}} \tag{26}$$

is the normalization factor required for compensating the restricted support.

References

[1] Stefan Arnborg. Robust Bayesianism: Imprecise and paradoxical reasoning. In *Proceedings of the 7th International Conference on Information fusion*, pages 407–414, 2004.

[2] Stefan Arnborg. Robust Bayesianism: Relation to evidence theory. *Journal of Advances in Information Fusion*, 1(1):63–74, 2006.

[3] M. Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50:174–188, 2002.

[4] James O. Berger. An overview of robust Bayesian analysis. *Test*, 3:5–124, 1994.

[5] Marco Cattaneo. A generalization of credal networks. In Thomas Augustin, Frank P. A. Coolen, Serafín Moral, and Matthias C. M. Troffaes, editors, *ISIPTA '09, Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications*, pages 79–88. SIPTA, 2009.

[6] Gustave Choquet. Theory of capacities. *Annales de lInstitut Fourier*, 5:131–295, 1954.

[7] Inés Couso, Serafín Moral, and Peter Walley. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5:165–181, 2000.

[8] Fabio G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[9] Love Ekenberg and Johan Thorbiörnson. Second-order decision analysis. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 9, No 1*, 9(1):13–38, 2001.

[10] Dale Fixsen and Ronald P. S. Mahler. The modified Dempster-Shafer approach to classification. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 27:96–104, 1997.

[11] Peter Gärdenfors and Nils-Eric Sahlin. Unreliable probabilities, risk taking, and decision making. *Synthese*, 53:361–386, 1982.

[12] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 2004.

[13] Charles J. Geyer, Glen D. Meeden, and incorporates code from cddlib written by Komei Fukuda. *rcdd: rcdd (Computational Geometry)*, 2012. R package version 1.1-7.

[14] David Rios Insua and Fabrizio Ruggeri, editors. *Robust Bayesian Analysis*. Springer, 2000.

[15] Alexander Karlsson. *Evaluating Credal Set Theory as a Belief Framework in High-Level Information Fusion for Automated Decision-Making*. PhD thesis, Örebro University, School of Science and Technology, 2010.

[16] Alexander Karlsson, Ronnie Johansson, and Sten F. Andler. On the behavior of the robust Bayesian combination operator and the significance of discounting. In *6th International Symposium on Imprecise Probability: Theories and Applications*, 2009.

[17] Alexander Karlsson, Ronnie Johansson, and Sten F. Andler. An empirical comparison of Bayesian and credal set theory for discrete state estimation. In *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Communications in Computer and Information Science, Volume 80, ISSN 1865-0937*, 2010.

[18] Alexander Karlsson, Ronnie Johansson, and Sten F. Andler. Characterization and empirical evaluation of bayesian and credal combination operators. *Journal of Advances in Information Fusion*, 6:150–166, 2011.

[19] Isaac Levi. *The enterprise of knowledge*. The MIT press, 1983.

[20] Robert F. Nau. Uncertainty aversion with second-order utilities and probabilities. *Management Science*, 52(1):136–145, 2006.

[21] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[22] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[23] Philippe Smets. Analyzing the combination of conflicting belief functions. *Information Fusion*, 8:387–412, 2007.

[24] Philippe Smets and Robert Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

[25] Cedric A. B. Smith. Consistency in statistical inference and decision. *Journal of the Royal Statistical Society, Series B, xxiii*, pages 1–25, 1961.

[26] David Sundgren, Mats Danielson, and Love Ekenberg. Warp effects on calculating interval probabilities. *International Journal of Approximate Reasoning*, 50(9):1360–1368, 2009.

[27] David Sundgren, Love Ekenberg, and Mats Danielson. Shifted dirichlet distributions as second-order probability distributions that factors into marginals. In *Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications*, pages 405–410, 2009.

[28] Lev V. Utkin and Thomas Augustin. Decision making with imprecise second-order probabilities. In *Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications*, pages 547–561, 2003.

[29] Peter Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, 1991.

[30] Lotfi A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, pages 3–28, 1978.

[31] Lotfi A. Zadeh. Review of books: A mathematical theory of evidence. *AI Magazine*, 5:81–83, 1984.

Evaluation of Evidential Combination Operators

Alexander Karlsson
Infofusion/Informatics Research Center
University of Skövde, Sweden
alexander.karlsson@his.se

H. Joe Steinhauer
Infofusion/Informatics Research Center
University of Skövde, Sweden
joe.steinhauer@his.se

Abstract

We present an experiment for evaluating precise and imprecise evidential combination operators. The experiment design is based on the assumption that only limited statistical information is available in the form of multinomial observations. We evaluate three different evidential combination operators; one precise, the Bayesian combination operator, and two imprecise, the credal and Dempster's combination operator, for combining independent pieces of evidence regarding some discrete state space of interest. The evaluation is performed by using a score function that takes imprecision into account. The results show that the precise framework seems to perform equally well as the imprecise frameworks.

Keywords. Evidential combination, imprecise probability, credal sets.

1 Introduction

The problem of *combining independent pieces of evidence*, most often stemming from multiple sources of information, e.g., sensors, is common in many application scenarios [24]. Typically such applications involve one or several sensors where for each sensor a feature can be extracted and used for constructing an appropriate *evidence* with regard to the unknown state of interest. Even though the pieces of evidence might not be completely independent, in many application scenarios (e.g., [21]) where different sources are used, e.g., different type of sensors, it is reasonable to assume independence.

In order to investigate the question of how well different evidence combination operators, precise and imprecise, perform compared to each other, we design an experiment for an object recognition scenario. We restrict the comparison to three different evidential combination operators, Bayesian [1, 2, 20], credal [1, 2, 20], and Dempster's combination operator [26]. The latter is one of the most commonly used operator for combining pieces of evidence. The obvious difference between the operators is that the Bayesian one is precise, i.e., its operands are based on a single function, and the other two are imprecise, i.e., the operands are either a set of functions or can be cast to a set of functions.

Karlsson et al. [20] have previously empirically compared the performance of the Bayesian and credal combination operators. They found that the Bayesian combination operator performs better due to the fact that the credal counterpart could "overestimate" imprecision and in a sense become too "cautious". However, in that evaluation imprecision was inherent in the state estimation problem, i.e., imprecise operands were sampled directly, without any particular statistical information, and the Bayesian operator was applied on the centroids of these operators while the credal counterpart was applied directly on the operands.

In contrast to the work by Karlsson et al. [20], we here design an experiment specifically aimed at evaluating the performance of combination operators when only a limited statistical amount of information is available and used in the precise and imprecise statistical models, namely Dirichlet models. This type of situation, i.e., when only limited information is available, is often one of the main motivations for using *imprecise probability* (including credal sets) [31, 33]. In addition, we also include Dempster's combination operator, i.e., another imprecise operator, in our evaluation. Since the imprecise operators have exponential worst case complexity in comparison to the precise one, we are specifically interested in comparing these two classes of operators.

The paper is organized as follows: in Section 2, we describe the different operators considered in the experiment. In Section 3, we elaborate on the design, performance, and result of the empirical evaluation, and lastly, in Section 4, we summaries the work presented and discuss the results of the evaluation as well as include ideas for future research.

2 Preliminaries

In this section, we present the different *combination operators* that we later will use in the empirical evaluation. One important aspect to note for all of these combination operators is that the pieces of evidences must satisfy different types of independence requirements which will discuss preceding each formal definition of the operators. Due to this requirement, a joint evidence is most often stronger if both operands constitute strong evidence for a certain state; this state gets reinforced in the combination. This should be put in contrast to other operators that often goes under the name *aggregation operators* [30] which are typically more "consensus-inspired" which means that the joint result is an agreement between the operands, e.g., if both operands are identical the joint result is equivalent to the operands.

2.1 Bayesian Combination

A Bayesian approach to combining independent pieces of evidence can be derived by modeling evidence as *likelihood functions* [1, 2, 20]. To realize this, assume that we have a random variable X taking values $x \in \Omega_X$. Furthermore, assume that we can obtain observations y_1, y_2, from two different sources within the environment of interest and that these observations are informative about X in the sense that your *belief*, i.e., the *posterior probability* $p(X|y_1, y_2)$ could be affected. Now since:

$$p(X|y_1,y_2) = \frac{p(y_1,y_2|X)p(X)}{\sum_{x\in\Omega_X} p(y_1,y_2|x)p(x)}\,, \tag{1}$$

we see that the only way the observations can affect the belief $p(X|y_1, y_2)$ is through the *joint likelihood* $p(y_1, y_2|X)$. By assuming that the observations are *conditionally independent* given that we know the true state of X, we obtain:

$$p(y_1,y_2|X) = p(y_1|X)p(y_2|X)\ . \tag{2}$$

In terms of evidence, the above equation is a simple method for combining two independent pieces of evidence, i.e., likelihood functions, into a single *joint evidence*, i.e., a joint likelihood function. However, in order to avoid monotonically decreasing values of the joint evidence, we normalize after each combination to obtain a probability function. Such a normalization can be performed without loss of generality since it is the relative strength of the likelihoods that constitute the evidence structure and such relativeness is preserved under normalization (see further Karlsson et al. (2011) [20]).

Definition 1. *The Bayesian combination operator $\Phi_{\mathcal{B}}$ is defined as [1, 2, 20]:*

$$\Phi_{\mathcal{B}}(\hat{p}(y_1|X),\hat{p}(y_2|X)) \triangleq \frac{\hat{p}(y_1|X)\hat{p}(y_2|X)}{\sum_{x\in\Omega_X}\hat{p}(y_1|x)\hat{p}(y_2|x)}\,, \tag{3}$$

where $\hat{p}(y_i|X)$, $i \in \{1,2\}$, are normalized likelihood functions and where the joint evidence $\Phi_{\mathcal{B}}(\hat{p}(y_1|X),\hat{p}(y_2|X))$ satisfies the conditionally independence assumption in Eq. (2). The operator is undefined when $\sum_{x\in\Omega_X}\hat{p}(y_1|x)\hat{p}(y_2|x) = 0$.

Note that when the denominator is zero, the sources mutually exclude all possibilities within the state space which is a contradiction to the assumption that the truth exists within this space (given the closed world assumption). From this viewpoint it is quite natural that the operator is undefined for such cases. One way of handling such situation is to perform discounting [15, 20].

2.2 Credal Combination

Credal combination is a straightforward generalization of the Bayesian combination operator to *imprecise probability* [33]. It relies on the notion of *credal sets* [23, 10, 11], i.e., *closed convex sets of probability functions.* Such sets can be conveniently represented by *extreme points* and therefore one uses credal sets in the form of *polytopes* since such a structure guarantees a finite number of such points. The combination schema was introduced as the *robust Bayesian combination operator* by Arnborg [1, 2], and further studied by Karlsson et al. [20] as the *credal combination operator*[1].

The main reason for considering *imprecision* in the form of credal sets is that it allows one to model problems when only scarce information is available regarding the environment of interest [31]. In such cases it can be considered to be more realistic to express, e.g., probabilities in terms of intervals instead of single probability values. Credal sets can also be thought of as being a result of performing sensitivity analysis as in robust Bayesian theory [17, 4].

In order to generalize the Bayesian combination operator in Def. 1 to a credal counterpart, we start by modeling evidence by credal sets of normalized likelihoods functions, denoted $\hat{\mathcal{P}}(y_1|X)$ and $\hat{\mathcal{P}}(y_2|X)$, instead of a single normalized likelihood function, where

[1] We denote this operator as the credal combination operator for the simple reason that we do not want to impose any particular interpretation of the imprecision as "robust" imposes a sensitivity-analysis interpretation.

as previous X denotes a random variable for something unknown of interest in the environment and y_1 and y_2 are the observations. In order to model independent pieces of evidence we use a generalization of conditional independence denoted *strong independence* [9], which requires that all extreme points must factorize, i.e:

$$\hat{p}_e(y_1, y_2|X) = \hat{p}(y_1|X)\hat{p}(y_2|X)\,, \tag{4}$$

$\forall \hat{p}_e(y_1, y_2|X) \in \mathcal{E}(\hat{\mathcal{P}}(y_1, y_2|X))$ where $\mathcal{E}(\cdot)$ denotes the set of extreme points and where $\hat{p}(y_i|X) \in \hat{\mathcal{P}}(y_i|X)$, $i \in \{1, 2\}$. By using this independence assumption the credal combination operator can be defined in terms of applying the Bayesian combination operator point-wise on all combinations of functions in the operand credal sets and as a last step applying the convex-hull operator $\mathcal{CH}(\cdot)$ in order to fulfill convexity of the joint evidence.

Definition 2. *The credal combination operator $\Phi_{\mathcal{C}}$ is defined as [1, 2, 20]:*

$$\begin{aligned}\Phi_{\mathcal{C}}(\hat{\mathcal{P}}(y_1|X),&\hat{\mathcal{P}}(y_2|X)) \triangleq \\ &\mathcal{CH}\Bigg(\Big\{\Phi_{\mathcal{B}}(\hat{p}(y_1|X), \hat{p}(y_2|X)) : \\ &\hat{p}(y_i|X) \in \hat{\mathcal{P}}(y_i|X), i \in \{1, 2\}\Big\}\Bigg)\,,\end{aligned} \tag{5}$$

where $\hat{\mathcal{P}}(y_i|X)$, $i \in \{1, 2\}$, are credal sets of normalized likelihoods functions and where the joint evidence $\Phi_{\mathcal{C}}(\hat{\mathcal{P}}(y_1|X), \hat{\mathcal{P}}(y_2|X))$ satisfies the conditional independence assumption in Eq. (4). The operator is undefined if $\Phi_{\mathcal{B}}(\hat{p}(y_1|X), \hat{p}(y_2|X)))$ is undefined for any pair $\hat{p}(y_1|X) \in \hat{\mathcal{P}}(y_1|X)$, $\hat{p}(y_2|X) \in \hat{\mathcal{P}}(y_2|X)$.

Note that the credal combination operator inherits the property of being undefined for cases where the denominator is zero (see further the discussion after Def. 1) and that when only singleton sets are used the operator is equivalent to the Bayesian combination operator. For computation of $\Phi_{\mathcal{C}}$, one can restrict the application of the Bayesian combination operator to the extreme points of the operand credal sets [20, Theorem 2], i.e:

$$\begin{aligned}\Phi_{\mathcal{C}}(\hat{\mathcal{P}}(y_1|X),&\hat{\mathcal{P}}(y_2|X)) = \\ &\Phi_{\mathcal{C}}(\mathcal{E}(\hat{\mathcal{P}}(y_1|X)), \mathcal{E}(\hat{\mathcal{P}}(y_2|X)))\ .\end{aligned} \tag{6}$$

In order to measure the degree of imprecision of a credal set, we will utilize the following measure [20], which can be thought of as the average degree of imprecision for single events [31]:

$$\begin{aligned}\mathcal{I}(\mathcal{P}(X)) \triangleq \frac{1}{|\Omega_X|} \sum_{x \in \Omega_X} \Bigg(&\max_{p(X) \in \mathcal{P}(X)} p(x) - \\ &\min_{p(X) \in \mathcal{P}(X)} p(x)\Bigg)\end{aligned} \tag{7}$$

2.3 Dempster-Shafer Combination

Dempster-Shafer theory [12, 26], also known as *evidence theory*, is a variant of imprecise probability [33], where one models evidence imprecisely by so called *mass functions*. A mass function assigns mass to subsets $A \subseteq \Omega_X$. The idea is that this schema can be useful in cases where a source is only partly sure of the true value of X, e.g. for $\Omega_X = \{x_1, x_2, x_3\}$, a source might be able to exclude the alternative x_3 but not be able to specify more clearly whether the truth is x_1 or x_2.

Formally, a mass function is a mapping from the power set of the state space Ω_X, also known as the *frame of discernment*, to the interval $[0, 1]$:

$$m : 2^{\Omega_X} \to [0, 1] \tag{8}$$

$$m(\emptyset) = 0 \tag{9}$$

$$\sum_{A \subseteq \Omega_X} m(A) = 1 \tag{10}$$

Two additional functions that are often encountered when considering Dempster-Shafer theory are *belief* and *plausibility*, denoted $Bel(A)$ and $Pl(A)$, respectively, and defined by:

$$Bel(A) \triangleq \sum_{B \subseteq A} m(B) \tag{11}$$

$$Pl(A) \triangleq \sum_{B \cap A \neq \emptyset} m(B)\,, \tag{12}$$

where $Bel(A)$ can be interpreted as the sum of all evidence that supports A and $Pl(A)$ as the sum of all evidence that does not contradict A. Belief and plausibility can also be regarded as a *lower and upper bound* for the probability of A, i.e:

$$Bel(A) \leq p(A) \leq Pl(A)\ . \tag{13}$$

The concept of independent pieces of evidence in Dempster-Shafer theory, also known as *distinct evidences*, is a bit problematic [27]. However, when the mass functions only operate on singleton sets, independent pieces of evidence can be defined in the same way as for the Bayesian combination operator, i.e., by using an assumption of conditional independence [27]. In the other cases, this assumption does not work, however, according to Smets [27], independent pieces of evidence can "in practice" be defined as:

$$m_{1,2}(A) = \begin{cases} m_1(B)m_2(C) & if A = B \times C \\ 0 & \text{Otherwise} \end{cases}\,, \tag{14}$$

where $B \subseteq \Omega_{X_1}$ and $C \subseteq \Omega_{X_2}$, i.e., ordinary stochastic independence.

Given two independent pieces of evidence m_1 and m_2, e.g., in the sense of Eq. (14), we can combine them into a joint evidence $m_{1,2}$ utilizing *Dempster's combination operator* [12].

Definition 3. *Dempster's combination operator* $\Phi_{\mathcal{D}}$ *is defined as [12, 26]:*

$$\Phi_{\mathcal{D}}(A, m_1, m_2) \triangleq \frac{1}{1-k} \sum_{\substack{B \cap C = A \\ B,C \subseteq \Omega_X}} m_1(B) m_2(C) \,, \tag{15}$$

where k *is the conflict between evidence* m_1 *and* m_2*, defined by:*

$$k = \sum_{\substack{A \cap B = \emptyset \\ A,B \subseteq \Omega_X}} m_1(A) m_2(B) \ . \tag{16}$$

The operator is undefined when $k = 1$.

Dempster's combination operator is related to the Bayesian combination operator in the way that in case the mass is distributed only among singletons of Ω_X, the two operators produce the same results. Also note that similar to the Bayesian and credal combination operator, the operator is undefined in cases where the sources mutually exclude all possibilities of the state space.

Since a mass function imposes lower and upper bounds on a probability function, seen in Eq. (13), one can transform a mass function into a credal set. The question then arises if the mass function as a result of Dempster's combination operator applied on two operands yields a mass function that when transformed to a credal set is equivalent to the result of first transforming the same operands to credal sets and then use the credal combination operator? Arnborg [1, 2] has shown that this is not the case, in fact the resulting credal sets can even be disjoint, hence the credal and Dempster's combination operator are clearly different.

3 Empirical Evaluation

In this section we elaborate on the experiment design for evaluating the combination operators previously presented. We start by providing an overview of the application scenario where the combination takes place, and then move on to describe the design including assumptions, parameters, and score functions.

3.1 Overview

Assume that we want to implement an *object recognition* algorithm based on two different types of sensors: a camera and a microphone (we assume that the objects of interest produce some form of sound). Naturally, using both sensors for performing the recognition should yield a better result than only using one. Since we utilize different sensors, that observes different features of the object, it is fair to make the assumption that the sesnor readings yields independent pieces of evidence. As an example, if the object is positioned at an "unfamiliar" angle, yielding ambiguous output from an image analysis algorithm, this can be compensated for by the output from a pattern matching algorithm performed on the signal from the microphone. Also, if both sensor yields features that constitute evidence for one particular object, one would obtain an evidence that is reinforced towards that object.

Let the unknown object be denoted by X with a corresponding state space Ω_X. Assume that we use some technique to extract *discrete features* from each of the signals of the sensors. Let the features be denoted as y_1 and y_2 with corresponding feature spaces Ω_{Y_1} and Ω_{Y_2}. Furthermore, assume that we have performed a limited number of experiments where we have placed different objects at different positions in the range of the camera and microphone, and observed the extracted features. The goal then is to design an *agent*[2] $\mathcal{A}$ that uses this limited set of information in order to construct evidence based on the observed features y_1 and y_2 and combine these pieces of evidence for the purpose of predicting the true object. In the remainder of this section, we present three agents based on the combination operators described in Section 2.

3.2 The Bayesian Agent – $\mathcal{A}_{\mathcal{B}}$

We here describe how an agent based on the Bayesian combination operator in Def. 1 can be used in order to decide on an object $x \in \Omega_X$ based on features from the sensor readings and previous mentioned limited statistical information. Since the features are extracted from different types of signals, it is fair to assume that y_1 and y_2 are conditionally independent given object X. By using a uniform (Bayes-Laplace) *Dirichlet model* (used in many scenarios, e.g., [7]), which amounts to calculating the expected value of a posterior Dirichlet density [16, 6], we can construct non-normalized evidence by [18]:

$$p(y_i|X) \triangleq \frac{\alpha_{y_i|X} + 1}{\sum_{y_i \in \Omega_{Y_i}} \alpha_{y_i|X} + |\Omega_{Y_i}|}, \tag{17}$$

where $i \in \{1, 2\}$ and where $\alpha_{y_i|X}$ denotes the number of times a specific feature y_i has been extracted given

[2] The use of an agent paradigm for describing the empirical evaluation was inspired by Aughenbaugh and Paredis [3].

an object X. The evidence can then be normalized:

$$\hat{p}(y_i|X) = \frac{p(y_i|X)}{\sum_{x \in \Omega_X} p(y_i|x)}, \qquad (18)$$

and used as operands in the Bayesian combination operator in order to obtain a joint evidence $\hat{p}(y_1, y_2|X)$, i.e.:

$$\hat{p}(y_1, y_2|X) = \Phi_{\mathcal{B}}(\hat{p}(y_1|X), \hat{p}(y_2|X)) \ . \qquad (19)$$

Note that when we do not have any statistical information at all, $\hat{p}(y_1|X)$, $\hat{p}(y_2|X)$, and consequently $\hat{p}(y_1, y_2|X)$, are uniform. Finally, based on the joint evidence, the agent $\mathcal{A}_B$ can define the most probable object(s) by:

$$\mathcal{A}_B \triangleq \mathcal{O}(\hat{p}(y_1, y_2|X)) \,, \qquad (20)$$

where $\mathcal{O}(\cdot)$ is defined as:

$$\mathcal{O}(p(X)) \triangleq \Big\{ x \in \Omega_X : (\forall x' \in \Omega_X)\Big(p(x) \geq p(x')\Big) \Big\} \ . \qquad (21)$$

where $p(X)$ is a probability function (remember that $\hat{p}(y_1, y_2|X)$ is a normalized likelihood function, i.e., a probability function).

3.3 The Credal Agent – $\mathcal{A}_\mathcal{C}$

Consider the same setting but where one models the evidence by credal sets. Instead of utilizing the (precise) Dirichlet model, which was the case for the Bayesian agent, we utilize the corresponding imprecise model, denoted as the *imprecise Dirichlet model* [31, 32], where one calculates the expected value of a set of posterior Dirichlet densities. The difference between this model and the former is that one uses the imprecision, i.e., the "size" of a credal set, as a way of reflecting the amount of information that the evidence is based on. By utilizing the imprecise Dirichlet model, we can construct normalized evidence $\hat{\mathcal{P}}(y_i|X)$ by [18]:

$$\hat{\mathcal{P}}(y_i|X) \triangleq \Bigg\{ \frac{p(y_i|X)}{\sum_{x \in \Omega_X} p(y_i|X)} : (\forall x \in \Omega_X) \Bigg(\frac{\alpha_{y_i|x}}{\sum_{y_i \in \Omega_{Y_i}} \alpha_{y_i|x} + \beta} \leq p(y_i|x) \leq \frac{\alpha_{y_i|x} + \beta}{\sum_{y_i \in \Omega_{Y_i}} \alpha_{y_i|x} + \beta} \Bigg) \Bigg\}, \qquad (22)$$

where $i \in \{1, 2\}$ and the parameter β determines how the imprecision of the set $\hat{\mathcal{P}}(y_i|X)$ is affected by the sample size. Note that when the sample size increases, the imprecision $\mathcal{I}(\hat{\mathcal{P}}(y_i|X))$ decreases since the lower and upper bounds for each $p(y_i|x)$ in Eq. (22) converge [31, 32]:

$$\lim_{\left(\sum_{y_i \in \Omega_{Y_i}} \alpha_{y_i|x}\right) \to \infty} \left(\frac{\alpha_{y_i|x} + \beta}{\sum_{y_i \in \Omega_{y_i|x}} \alpha_{y_i|x} + \beta} - \frac{\alpha_{y_i|x}}{\sum_{y_i \in \Omega_{Y_i}} \alpha_{y_i|x} + \beta} \right) = 0 \,, \qquad (23)$$

i.e., imprecision is reflected by the sample size.

We can now utilize the credal combination operator in Def. 2 in order to obtain the joint evidence:

$$\hat{\mathcal{P}}(y_1, y_2|X) = \Phi_\mathcal{C}(\hat{\mathcal{P}}(y_1|X), \hat{\mathcal{P}}(y_2|X)) \ . \qquad (24)$$

Based on the joint evidence, agent $\mathcal{A}_\mathcal{C}$ can decide on the most probable object(s) in a similar way as in the Bayesian case:

$$\mathcal{A}_\mathcal{C} \triangleq \bigcup_{\hat{p}(y_1, y_2|X) \in \hat{\mathcal{P}}(y_1, y_2|X)} \mathcal{O}(\hat{p}(y_1, y_2|X)), \qquad (25)$$

where $\mathcal{O}(\cdot)$ is defined by Eq. (21). The intuition behind this set is that the agents includes all objects that are optimal for some probability function, i.e., there exists a probability function within the credal set that contains a probability that is highest for a given object within the set. In contrast to the Bayesian case, the above set is more likely to be non-singleton. This indicates that the agent does not possess enough information to distinguish between the objects within the set.

3.4 The Dempster-Shafer Agent – $\mathcal{A}_\mathcal{D}$

In order to define an agent based on Dempster-Shafer theory, we first need to elaborate on how mass functions could be constructed based on the credal set obtained from the imprecise Dirichlet model in Eq. (22). A credal set is a more general structure in comparison to a mass function [33, 2]. Hence, transforming a credal set to a mass function, e.g. by [26, Theorem 2.2]:

$$m(A) = \sum_{B \subseteq A} (-1)^{|A-B|} Bel(B) \,, \qquad (26)$$

cannot in general be performed without some approximation [1, 2, 8]. One way, demonstrated by Campos et al. [8], is to approximate the credal set by certain types of intervals and then apply the algorithm proposed by Lemmer and Kyburg [22]. Another way, suggested by Arnborg [2], is to utilize Eq. (26) and transfer mass upwards in the "set-lattice" to eliminate negative mass. Common to both these ways is that there is no unique minimal approximation that can be used for constructing the mass function. As a starting point for our evaluation, we here first consider a simple method for constructing a mass function based on the lower bounds of each single state $x \in \Omega_X$ of the credal set. Lower bounds on single states are often used in many models, e.g., [29], since they yield *simplices* which is the simplest type of imprecision.

By using lower bounds we can obtain the simplex:

$$\hat{\mathcal{P}}^*(y_i|X) \triangleq \left\{ \hat{p}^*(y_i|X) : \right. \\ \hat{p}^*(y_i|x) \geq \min_{\hat{p}(y_i|X) \in \hat{\mathcal{P}}(y_i|X)} \hat{p}(y_i|x), \\ \left. \sum_{x \in \Omega_X} \hat{p}^*(y_i|x) = 1,\ x \in \Omega_X \right\} . \tag{27}$$

The simplex $\hat{\mathcal{P}}^*(y_i|X)$ can then easily be transformed to a mass function using Eq. (26) on the belief function/lower probabilities defined by [1, 2]:

$$Bel(A) = \min_{\hat{p}^*(y_i|A) \in \hat{\mathcal{P}}^*(y_i|A)} \hat{p}^*(y_i|A)\,, \tag{28}$$

where $A \subseteq \Omega_X$, which results in a mass function of the following form [8]:

$$\begin{aligned} m_i(x) &= \min_{\hat{p}^*(y_i|X) \in \hat{\mathcal{P}}^*(y_i|X)} \hat{p}^*(y_i|x) \\ m_i(\Omega_X) &= 1 - \sum_{x \in \Omega_X} \min_{\hat{p}^*(y_i|X) \in \hat{\mathcal{P}}^*(y_i|X)} \hat{p}^*(y_i|x)\,, \end{aligned} \tag{29}$$

$\forall x \in \Omega_X$. Now, based on m_1 and m_2, obtained by Eq. (29), we can perform the combination:

$$m_{1,2}(A) = \Phi_{\mathcal{D}}(A, m_1, m_2)\ . \tag{30}$$

In order to be able to compare the results from the credal agents $\mathcal{A}_{\mathcal{C}}$ with the above mass function, we define a Dempster-Shafer agent which includes a transformation of $m_{1,2}$ back to a credal set by performing linear programming on the following set of constraints (cf. Eq. (13)):

$$\begin{aligned} \mathcal{P}_{1,2}(X) \triangleq \{p_{1,2}(X) : Bel_{1,2}(A) &\leq p_{1,2}(A) \\ &\leq Pl_{1,2}(A), A \subseteq \Omega_X\}\,, \end{aligned} \tag{31}$$

where $Bel_{1,2}$ is the belief, or lower probability, in Eq. (11) and $Pl_{1,2}$ is the plausibility, or upper probability, in Eq. (12), with respect to $m_{1,2}$. Now, based on the credal set $\mathcal{P}_{1,2}(X)$, the *Dempster-Shafer agent* $\mathcal{A}_{\mathcal{D}}$ can decide on objects according to:

$$\mathcal{A}_{\mathcal{D}} \triangleq \bigcup_{p_{1,2}(X) \in \mathcal{P}_{1,2}(X)} \mathcal{O}(p_{1,2}(X)), \tag{32}$$

where $\mathcal{O}(\cdot)$ is defined in Eq. (21).

3.5 Evaluation Schema

In order to evaluate the different agents, introduced in the previous sections, we consider a combination scenario where we have two sources $i \in \{1,2\}$ that report evidences, based on features $y_i \in \Omega_{Y_i}$ where $\Omega_{Y_i} \triangleq \{f_{i,1}, \ldots, f_{i,m}\}$, regarding a random variable $X \in \Omega_X$ where $\Omega_X = \{x_1, \ldots, x_m\}$ (we will instantiate the parameter m later when we describe the experiment in more detail). Note that $|\Omega_{Y_i}| = |\Omega_X| = m$. Let us now assume that the true state is x_1 and that each agent has a limited set of multinomial observations from the two sources to base evidence upon. We will simulate the limited information stemming from the sources by drawing n samples, where n is a small number that we will instantiate later, from a multinomial distribution, i.e., we sample a vector:

$$\vec{\alpha}^{n}_{y_i|x} \triangleq \left[\alpha_{f_{i,1}|x} \ldots \alpha_{f_{i,m}|x}\right] \tag{33}$$

$\forall x \in \Omega_X$ through:

$$\vec{\alpha}^{n}_{y_i|x} \sim \mathrm{Mu}(\vec{\alpha}^n | p(f_{i,1}|x), \ldots, p(f_{i,m}|x))\,, \tag{34}$$

where $\alpha_{f_{i,j}|x}$ denotes the number of times feature $f_{i,j}$ has been observed when the object is x and where $\mathrm{Mu}(\cdot|\cdot)$ denotes the multinomial distribution with parameters $p(f_{i,j}|x)$, $j \in \{1, \ldots, m\}$, i.e., the probability of observing a specific feature $f_{i,j}$ from source i given some object x. Note that

$$\sum_{j \in \{1,\ldots,m\}} \alpha_{f_{i,j}|x} = n \tag{35}$$

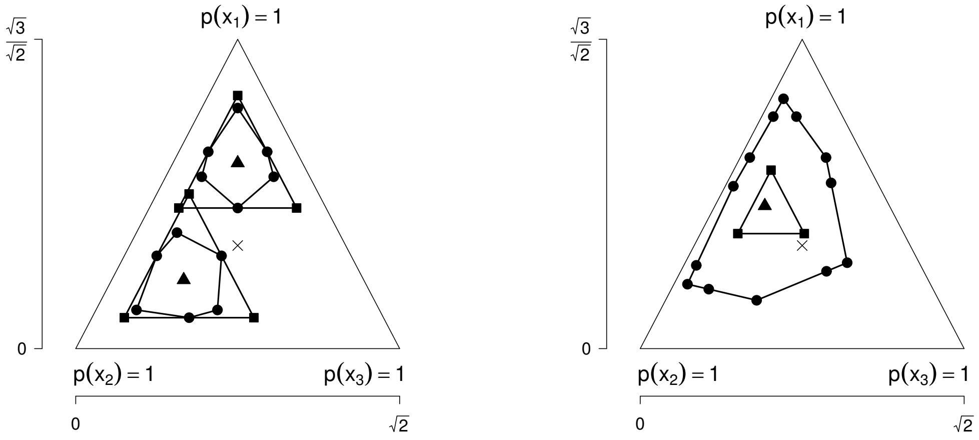
$\forall x \in \Omega_X$ and $i \in \{1,2\}$. The information contained in each sampled vector can then be used in the precise and imprecise Dirichlet models, Eqs. (17) – (18) and (22), by the agents in order to construct evidence. Since the imprecise agents are undefined in cases where the sources mutually excludes each other (see further the discussion following Def. 1 – 3), we simply omit such cases.

To give an example, assume that $m = 3$ and $n = 5$ and that we have sampled the following:

$$\begin{bmatrix} \vec{\alpha}^{5}_{y_1|x_1} \\ \vec{\alpha}^{5}_{y_1|x_2} \\ \vec{\alpha}^{5}_{y_1|x_3} \end{bmatrix} = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 3 & 1 \\ 1 & 0 & 4 \end{bmatrix} \tag{36}$$

$$\begin{bmatrix} \vec{\alpha}^{5}_{y_2|x_1} \\ \vec{\alpha}^{5}_{y_2|x_2} \\ \vec{\alpha}^{5}_{y_2|x_3} \end{bmatrix} = \begin{bmatrix} 3 & 1 & 1 \\ 0 & 5 & 0 \\ 0 & 1 & 4 \end{bmatrix}\ . \tag{37}$$

Further assume that an object $x \in \Omega_X$ have generated features $y_1 = f_{1,1}$ and $y_2 = f_{2,2}$. This would mean that we would utilize the first and second column of the matrices on the right hand side of Eqs. (36)



(a) Operands in Eqs. (36) and (37)

(b) Result of the agents

Figure 1: Precise and imprecise operands are shown in Fig. 1(a) and the result of applying agent $\mathcal{A}_\mathcal{B}$, $\mathcal{A}_\mathcal{C}$, and $\mathcal{A}_\mathcal{D}$ on the operands are shown in Fig. 1(b). The operands and results are denoted by triangles, circles, and squares for agent $\mathcal{A}_\mathcal{B}$, $\mathcal{A}_\mathcal{C}$, and $\mathcal{A}_\mathcal{D}$, respectively.

and (37), correspondingly, for contructing the evidences based on the precise and imprecise Dirichlet models in Eqs. (17) – (18) and (22). The operands and result of applying the different agents, i.e., $\mathcal{A}_\mathcal{B}$ in Eq. (20), $\mathcal{A}_\mathcal{C}$ in Eq. (25), and $\mathcal{A}_\mathcal{D}$ in Eq. (32), on this data is is shown in Fig. 1. Note in particular that the imprecision of the credal agent $\mathcal{A}_\mathcal{C}$ is considerably higher in comparison to the Dempster-Shafer agent $\mathcal{A}_\mathcal{D}$. From the figure we also see that the result from the credal agent $\mathcal{A}_\mathcal{C}$ contains extreme points that could be removed without changing the shape of the set significantly, however, we omit such removal in this experiment. In a real-world application such a removal would be performed to reduce computational complexity.

In order to compare the performance of the agents with each other, we will use a score function that takes imprecision into account [19, 20]:

$$\Upsilon(\mathcal{A}) \triangleq \begin{cases} \dfrac{1}{|\mathcal{A}|} & \text{if } (x_1 \in \mathcal{A}) \wedge (\mathcal{A} \neq \Omega_X) \\ 0 & \text{otherwise} \end{cases}, \quad (38)$$

i.e., if the agent manage to minimize the imprecision and is able to return the true state, the agent obtains the highest possible reward of one. If the set $\mathcal{A}$ contains two of the three states, where one of the states is the truth, i.e., x_1, the agent gets half of this reward since the set is still informative due to exclusion of one erroneous state. The other two cases, i.e., the truth is not contained in the set and all the states are reported, are considered to be non-informative; the latter due to that one already has modeled all possible states when the state space was designed.

Now, by simulating a large number of cases and apply the agents on each of these cases, we can obtain a good approximation of the expected score $\mathrm{E}[\Upsilon(\mathcal{A})]$ of each agent. The experiment, including simulation parameters, is then defined by the following step-wise description:

1. For each source $i \in \{1,2\}$ and $x \in \Omega_X$, draw γ according to:

$$\gamma \sim \text{Uniform}([0.7,\, 0.9])\,, \quad (39)$$

set:

$$p(f_{i,j}|x_k) \triangleq \begin{cases} \gamma & \text{when } k = j \\ \frac{1-\gamma}{m-1} & \text{otherwise} \end{cases} \quad (40)$$

and use these probabilities as multinomial parameters in Eq. (34) in order to draw vectors $\vec{\alpha}^{n}_{y_i|x}$. Note that given an object x_k, it is most likely that one observe the feature $f_{i,k}$ from source i.

2. Let us set $\beta = 2$ in Eq. (22) (this parameters is usually set to a value $1 \leq \beta \leq 2$, see further the discussion in [32, 5]) and sample new features to be used by the agents for predicting the true object by using the multinomial parameters in Eq. (40), i.e., sample $f_{i,j} \sim p(Y_i|x_1)$ (remember that x_1 is the true object) and apply the

agents $\mathcal{A}_{\mathcal{B}}$ in Eq. (20), $\mathcal{A}_{\mathcal{C}}$ in Eq. (25), and $\mathcal{A}_{\mathcal{D}}$ in Eq. (32), on the sampled vectors $\vec{\alpha}^{n}_{y_i|X}$ from Step 1.

3. Evaluate the agents by $\Upsilon(\mathcal{A})$ in Eq. (38), and store the score and repeat from Step 1, 10^3 times.
4. Approximate the expected score $\mathrm{E}\left[\Upsilon(\mathcal{A})\right]$ by using the stored scores from the previous step.

3.6 Results

The results are shown in Table 1, where the parameters m (dimension) and n (number of observations) have been varied. Taking the confidence interval into account, it seems that agent $\mathcal{A}_{\mathcal{B}}$ performs as well or better than agent $\mathcal{A}_{\mathcal{C}}$ and $\mathcal{A}_{\mathcal{D}}$. One reason for this is that agent $\mathcal{A}_{\mathcal{C}}$ and $\mathcal{A}_{\mathcal{D}}$ tends to be too cautious in many cases, as can be seen from the number of cases where the complete state space is reported. The difference in performance seems to increases when the state space m increases while n is constant.

One interesting and a bit surprising effect that one can observe is that the performance of agent $\mathcal{A}_{\mathcal{D}}$ deteriorates considerably when the size of the state space increases when maintaining the same number of observations (i.e., five). The low performance is due, as can be seen from the table, that the agent tends to increase the fraction of times it reports the complete state space, e.g., when $m = 7$ and $n = 5$, it reports the complete state space in 90.7% of the cases. An explanation for such behavior is that when m increases under a constant n, the sum of the upper constraints in the imprecise Dirichlet model in Eq. (22) increases and this means that the lower bounds of $\hat{\mathcal{P}}(y_i|X)$ in Eq. (22) decreases due to the normalization. Also, since it is not likely that we have observed a feature $f_{i,j}$ where $j \neq 1$ (x_1 is the truth and according to Eqs. (39) – (40) it is most likely to observe feature $f_{i,1}$), the sum of the lower bounds is also likely to have decreased and then more mass has been allocated to the complete state space in Eq. (29). This increased mass is then distributed among the states when transforming back to a credal set, which means that the number of cases where a non-singleton set is reported has increased. In other words, when m increases under a given n, the degree of imprecision based on the mass functions increases, as is also seen from the table. Also note that when we increase the number of observations to $n = 20$ when $m = 7$ the degree of imprecision decreases again.

The performance of the credal agent $\mathcal{A}_{\mathcal{C}}$ does not seem to be equally sensitive when increasing m. In fact, the performance increases slightly for both agent $\mathcal{A}_{\mathcal{B}}$ and $\mathcal{A}_{\mathcal{C}}$ and for the latter agent the average degree of imprecision decreases. One explanation for such a result could be that when the state space increases, the "noise probability" mass $1 - \gamma$ in Eq. (40) is distributed among more states, which could mean that it is less likely that a single erroneous state will be optimal for some probability function within the joint credal set due to noise.

It should also be noted that the credal combination operator can introduce substantial imprecision in the joint evidence, even though the operands are not that imprecise [20]. This can also be seen in Fig. 1, where both operands are less imprecise in comparison to the joint credal set. Such increasment in imprecsion of joint evidence mainly occur when the operands are in conflict with each other, i.e., when the operand credal sets are positioned at different positions within the simplex, especially when the operands are close to the boundary of the simplex.

We also observe that agent $\mathcal{A}_{\mathcal{B}}$ reports an erroneous set, i.e., when $\Upsilon(\mathcal{A}_{\mathcal{B}}) = 0$ and $|\mathcal{A}_{\mathcal{B}}| \neq \Omega_X$ in more cases compared to the other agents. This can be regarded as the usual trade-off between imprecision and precision, i.e., reducing erroneous output by increasing imprecision. It should be noted that the Bayesian agent $\mathcal{A}_{\mathcal{B}}$ has a quite crude way of reporting a decision set since the agent reports the most probable state also in cases where the difference of probability for this state in comparison to the other states is small. In a more refined Bayesian method one could use some form of thresholding (see further [20]). Nevertheless, even though such crude schema is utilized, $\mathcal{A}_{\mathcal{B}}$ performs well in comparison to the other agents.

4 Summary and Discussion

We have described an empirical experiment for evaluating and comparing the performance of different evidential combination operators. Besides comparing individual operators, our interest was also to compare precise and imprecise operators in general. The evaluation was restricted to the three operators, Bayesian combination (precise), credal combination (imprecise), and Dempster's combination (imprecise). For each combination operator we implemented a corresponding agent. The evaluation was based on the precise and imprecise Dirichlet models and a limited number of multinomial observations. To measure the agents performance we used a score function that, based on the informativeness of the outcome, assigned a reward to the agent.

The results showed that the Bayesian agent seems to perform at least equally well as its imprecise counterparts. Since the imprecise frameworks are often motivated by their suitability to situations where only scarce information is available, i.e. the case in the ex-

Parameters	Agent	$\mathrm{E}[\Upsilon(\cdot)]$	$\mathrm{E}[\mathcal{I}(\cdot)]$	$\Upsilon(\cdot) > 0$ (%)		$\Upsilon(\cdot) = 0$ (%)		
				$\lvert\cdot\rvert = 1$	$\lvert\cdot\rvert = 2$	$\lvert\cdot\rvert = 1$	$\lvert\cdot\rvert = 2$	$\lvert\cdot\rvert = m$
$m = 3, n = 5$	$\mathcal{A}_\mathcal{B}$	0.82 ± 0.03	0	80.5	2.3	17.0	0.1	0.1
	$\mathcal{A}_\mathcal{C}$	0.75 ± 0.02	0.43 ± 0.01	61.5	11.4	2.6	1.3	23.2
	$\mathcal{A}_\mathcal{D}$	0.80 ± 0.02	0.24 ± 0.00	69.3	16.5	6.0	1.1	7.1
$m = 5, n = 5$	$\mathcal{A}_\mathcal{B}$	0.87 ± 0.02	0	86.3	1.2	11.4	0.9	0.0
	$\mathcal{A}_\mathcal{C}$	0.81 ± 0.02	0.38 ± 0.01	73.7	5.1	0.7	0.1	16.0
	$\mathcal{A}_\mathcal{D}$	0.64 ± 0.02	0.41 ± 0.00	49.4	12.0	0.5	0.0	29.4
$m = 7, n = 5$	$\mathcal{A}_\mathcal{B}$	0.90 ± 0.02	0	89.1	1.8	8.9	0.1	0.0
	$\mathcal{A}_\mathcal{C}$	0.82 ± 0.02	0.36 ± 0.01	77.0	3.8	0.9	0.4	12.8
	$\mathcal{A}_\mathcal{D}$	0.17 ± 0.01	0.52 ± 0.00	1.1	2.8	0.0	0.0	90.7
$m = 7, n = 20$	$\mathcal{A}_\mathcal{B}$	0.83 ± 0.02	0	82.0	1.3	16.5	0.1	0.0
	$\mathcal{A}_\mathcal{C}$	0.79 ± 0.02	0.18 ± 0.01	66.4	21.3	1.2	2.1	7.1
	$\mathcal{A}_\mathcal{D}$	0.80 ± 0.02	0.18 ± 0.00	64.9	29.3	0.6	2.9	0.2

Table 1: Results of the empirical evaluation in terms of expected scores $\mathrm{E}[\Upsilon(\cdot)]$ and the degree of imprecision $\mathrm{E}[\mathcal{I}(\cdot)]$, all with 95% confidence intervals. In addition, we also see how the cardinality of the reported set of each agent is distributed among some sets that contains the truth and not.

periment (limited multinomial information), this outcome appears to be unexpected.

However, it might also be the case that the particular type of imprecision considered in this experiment does not do the imprecise operators justice. In particular we need to careful to judge the Dempster-Shafer agent based on this particular design of experiment, using the imprecise Dirichlet model and the given score function, since the agent was quite sensitive to the size of the limited statistical data in relation to the number of dimensions (which in principle could be reasonable). Also, in our experiment the mass function for the agent was derived by first approximating the credal set using a simplex and then transforming it into a mass function, which is a rather simple approximation. A more refined approximation might influence the performance of the Dempster-Shafer agent (see further the discussion in Section 3.4).

Note that the results are only valid in applications where the score function in Eq. (38) is accepted. It could still be useful in certain circumstances to report the entire state space, especially when you have the option to gather more information and thereby reduce the cardinality of the result. Depending on the application one needs to design the score function accordingly.

In our future research, we will explore different ways of obtaining the mass functions, and evaluate the imprecise operators found in this paper and also other variants, e.g., [13], with alternative forms of imprecision. One interesting way ahead is to simulate mass functions directly, which then can be transformed into a credal set, instead of constructing them from credal sets as was done in the experiment. In that case one could, e.g., use the *pignistic transformation* [28] on the mass functions in order to obtain operands for the Bayesian combination operator.

The overall conclusion that we infer from the results is that the Bayesian framework could still be suitable in applications where only limited statistical data is available. Taking into account that the Bayesian framework has considerably lower computationally complexity, this framework might even be the best choice for such type of applications.

Acknowledgements

We would like to thank the anonymous reviewers of this paper for their comments. This work was supported by the Information Fusion Research Program (University of Skövde, Sweden), in partnership with the Swedish Knowledge Foundation under grant 2010-0320 (URL: `http://www.infofusion.se`, UMIF project). R [25] and the package rcdd [14] have been used for the computations in this paper.

References

[1] Stefan Arnborg. Robust Bayesianism: Imprecise and paradoxical reasoning. In *Proceedings of the*

7th International Conference on Information fusion, pages 407–414, 2004.

[2] Stefan Arnborg. Robust Bayesianism: Relation to evidence theory. *Journal of Advances in Information Fusion*, 1(1):63–74, 2006.

[3] Jason Matthew Aughenbaugh and Christiaan J. J. Paredis. The value of using imprecise probabilities in engineering design. *Journal of Mechanical Design*, 128:969–979, 2006.

[4] James O. Berger. An overview of robust Bayesian analysis. *Test*, 3:5–124, 1994.

[5] Jean-Marc Bernard. An introduction to the imprecise dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39:123 – 150, 2005.

[6] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory.* John Wiley and Sons, 2000.

[7] Henrik Boström. Estimating class probabilities in random forests. In *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on*, pages 211–216, 2007.

[8] Luis M. De Campos, Juan F. Huete, and Serafin Moral. Probability intervals: A tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 02(02):167–196, 1994.

[9] Inés Couso, Serafín Moral, and Peter Walley. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*, 5:165–181, 2000.

[10] Fabio G. Cozman. *Decision Making Based on Convex Sets of Probability Distributions: Quasi-Bayesian Networks and Outdoor Visual Position Estimation.* PhD thesis, The Robotics Institute, Carnegie Mellon University, 1997.

[11] Fabio G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[12] Arthur P. Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society*, pages 205–247, 1969.

[13] Mihai Cristian Florea, Anne-Laure Jousselme, Éloi Bossé, and Dominic Grenier. Robust combination rules for evidence theory. *Information Fusion*, 10(2):183 – 197, 2009.

[14] Charles J. Geyer, Glen D. Meeden, and incorporates code from cddlib written by Komei Fukuda. *rcdd: rcdd (Computational Geometry)*, 2012. R package version 1.1-7.

[15] Rolf Haenni. Shedding new light on zadeh's criticism of dempster's rule of comb. In *7th International conference on Information Fusion (FUSION)*, pages 879–884, 2005.

[16] David Heckerman. Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1:79–119, 1997.

[17] David Rios Insua and Fabrizio Ruggeri, editors. *Robust Bayesian Analysis.* Springer, 2000.

[18] Alexander Karlsson. *Evaluating Credal Set Theory as a Belief Framework in High-Level Information Fusion for Automated Decision-Making.* PhD thesis, Örebro University, School of Science and Technology, 2010.

[19] Alexander Karlsson, Ronnie Johansson, and Sten F. Andler. An empirical comparison of Bayesian and credal set theory for discrete state estimation. In *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Communications in Computer and Information Science, Volume 80, ISSN 1865-0937*, 2010.

[20] Alexander Karlsson, Ronnie Johansson, and Sten F. Andler. Characterization and empirical evaluation of bayesian and credal combination operators. *Journal of Advances in Information Fusion*, 6:150–166, 2011.

[21] Max Krüger and Nane Kratzke. Monitoring of reliability in Bayesian identification. In *12th International Conference on Information Fusion*, 2009.

[22] John F. Lemmer and Henry E. Kyburg. Conditions for the existence of belief functions corresponding to intervals of belief. In *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence*, 1991.

[23] Isaac Levi. *The enterprise of knowledge.* The MIT press, 1983.

[24] Martin E. Liggins, David L. Hall, and James Llinas, editors. *Multisensor Data Fusion, Second Edition.* CRC Press, 2009.

[25] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[26] Glenn Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, 1976.

[27] Philippe Smets. Analyzing the combination of conflicting belief functions. *Information Fusion*, 8:387–412, 2007.

[28] Philippe Smets and Robert Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

[29] David Sundgren, Love Ekenberg, and Mats Danielson. Shifted dirichlet distributions as second-order probability distributions that factors into marginals. In *Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications*, pages 405–410, 2009.

[30] Vicenç Torra and Yasuo Narukawa. *Modeling Decisions.* Springer, 2007.

[31] Peter Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, 1991.

[32] Peter Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:3–57, 1996.

[33] Peter Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24:125–148, 2000.

Rationalizability under Uncertainty using Imprecise Probabilities*

Hailin Liu
Department of Philosophy
Carnegie Mellon University, PIttsburgh
hailinl@andrew.cmu.edu

Abstract

The notion of imprecise probability can be viewed as a generalization of the traditional notion of probability. Several theories and models of imprecise probability have been suggested in the literature as more appropriate representations of uncertainty in the context of single-agent decision making. In this paper I investigate the question of how such models can be incorporated into the traditional game-theoretic framework. In the spirit of *rationalizability*, I present two new solution concept called Γ-*maximin rationalizability* and E-*rationalizability*. They are intended to capture the idea that each player models the other players as decision makers who all employ Γ-maximin or E-admissibility as their decision rules. Some properties of these concept such as existence conditions and the relationships with rationalizability are studied.

Keywords. Normal form games, imprecise probabilities, rationalizability, Γ-maximin, E-admissibility.

1 Introduction

The theory of subjective expected utility (axiomatized by Savage [1954]) has become a widely-accepted normative theory for dealing with single-agent decision making under uncertainty. However, the assumption about the representation of uncertainty in this framework has often been criticized for being overly restrictive. In particular, Ellsberg [1961] has argued that uncertainty, as opposed to risk, cannot be adequately represented by a single personal probability distribution. Inspired by this challenge, various alternative theories of decision making under uncertainty have been developed in the literature, e.g., Gilboa and Schmeidler's multiple priors model [1989] and Schmeidler's Choquet expected utility model [1989]. In addition, there has been a vast amount of literature on alternative approaches to representing uncertainty in decision problems, such as upper and lower probabilities, sets of probability measures, belief functions, and so on. (See [Walley, 1991] for a detailed discussion of the models of imprecise probabilities)

The Ellsberg paradox arises in single-agent decision making situations where uncertainty regarding some exogenous event is involved. Nevertheless, one would expect that similar situations of uncertainty could arise in multi-agent, interactive scenarios, where the considerations underlying uncertainty for each player are the other players' strategy choices, rather than the state of nature. This naturally suggests a new line of research, which is to incorporate some model of uncertainty using imprecise probabilities into traditional game-theoretic frameworks. New conceptual issues arise in this approach to game under uncertainty, e.g., how should solution concepts be defined given the new decision theoretic foundations. In recent years, there has been a growing literature on applying the aforementioned theories of imprecise probabilities in the context of games, which can be divided roughly into two categories depending on the way of addressing these conceptual issues. On the one hand there are those that investigate the consequences of allowing players' beliefs to be represented by imprecise probabilities in the framework of Nash equilibrium or its refinements. Dow and Werlang [1994] introduce an equilibrium concept for two-player normal form games in which players' beliefs about the opponents' strategy choices are represented by non-additive probabilities and players are Choquet expected utility maximizers. Eichberger and Kelsey [2000] extend Dow and Werlang's equilibrium concept to normal form games with n-players and discuss some nice properties of this concept. By using the multiple priors model to represent players' uncertainty, Klibanoff [1994] and Lo [1996] provide two equilibrium-type solution concepts

*This paper is part of my master thesis, which extends the preliminary idea presented as a poster at the ISIPTA'11 conference. I have greatly benefited from many extensive discussions with Teddy Seidenfeld and Kevin Zollman. I also would like to thank the anonymous referees for pointing out several helpful references on this topic.

for normal form games with any finite number of players. Unlike these researchers, Liu [2011] presents a different solution concept called *robust equilibrium* by extending the framework of the so-called *linear tracing procedure* (Harsanyi and Selten [1988]), to accommodate games with uncertainty where players' initial beliefs are modeled by a set of probability measures rather than a common prior. This concept can be viewed as a refinement of Nash equilibrium.

On the other hand, several studies have attempted to generalize the concept of *rationalizability* (Bernheim [1984] and Pearce [1984]) in normal form games to accommodate notions of rationality other than subjective expected utility maximization. In addition to the idea of equilibrium with uncertainty aversion, another significant innovation introduced by Klibanoff [1994] is the characterization of common knowledge of rationality under uncertainty for normal form games where each player attempts to maximize the minimum expected utility. Epstein [1997] also considers normal form games and develops a general framework for discussing the implications of common knowledge of rationality in which the definition of rationality can accommodate different kinds of preference structures including the multiple priors model.

The approach to game theory with uncertainty I present in this paper is very much in the same spirit as Klibanoff's and Epstein's approaches, which embraces the essential idea of rationalizability, namely, to assume that each player models the opponents as the same kind of rational decision maker under uncertainty. As noted in previous literature, rationalizability captures the idea that each player attempts to deduce their opponents' rational behavior from the structure of the game by modeling her opponents as expected utility maximizers, where players' uncertainty about their opponents' strategy choices are fully described by a single probability distribution. This paper explores the possibility of adapting this standard assumption by using a set of probability distributions to model uncertainty in normal form games. However, even in single-agent decision theory, there is no generally accepted criterion for decision making under uncertainty when uncertainty is depicted by a set of probability distributions. In view of this, this paper develops a general theoretical framework to analyze the implications of rationality and common knowledge of rationality in the sense that each player employs the same decision rule to choose the best strategy with respect to a set of probability distributions. In particular, I consider here two familiar decision rules named *Γ-maximin* [Berger, 1985; Gilboa and Schmeidler, 1989] and *E-admissibility* [Levi, 1974]. According to the former rule, a decision maker should choose an option that maximizes the minimum expected utility with respect to a set of probability distributions, while the latter one constrains the decision maker's admissible choices to those options that maximizes expected utility for some probability in the set of probabilities. In analogy with rationalizability, I put forward two game-theoretic solution concepts under uncertainty, in which each player is required to model the other players as the same kind of decision makers who use either Γ-maximin or E-admissibility to make decisions. This gives rise to the two solution concepts that we shall call *Γ-maximin rationalizability* and *E-rationalizability* respectively. Just as Γ-maximin and E-admissibility are extensions of subjective expected utility theory, both Γ-maximin rationalizability and E-rationalizability turn out to be generalizations of rationalizability. Example 1 in Section 4 illustrates the distinction between these three solution concepts.

The main contribution of this paper is in providing a general game-theoretic framework which enables us to discuss how different decision rules can be incorporated into the framework of rationalizabiity in normal form games when uncertainty is depicted by a nontrivial set of probability distributions. This framework can be easily adapted to accommodate some other decision rules discussed in decision theory such as Maximality [Walley, 1991]. Although it turns out that the concept of Γ-maximin rationalizability coincides with Klibanoff's and Epstein's iterative definitions of rationalizability with uncertainty aversion (they used different terms for this concept), the current approach to rationalizability under uncertainty can be regarded as complementary work to their theories, since it provides an alternative way of characterizing the same solution concept. By applying this new definition, it is easier to check whether a strategy of a player is Γ-maximin rationalizable (or uncertainty aversion rationalizable). In a similar way, I define the concept of E-rationalizability, which, to my knowledge, has not been explored in any previous study.

The rest of this paper proceeds as follows: Section 2 presents a brief review of the solution concept rationalizability, and discusses some of its properties. Section 3 motivates the idea of using imprecise probabilities to represent uncertainty in games. I then propose two solution concepts called Γ-maximin rationalizability and E-rationalizability, which extend the framework of rationalizability to contexts where a set of probabilities is used to represent uncertainty. Section 4 studies some properties of these solution concepts, and also includes an example to illustrate their difference. Section 5 concludes the paper and suggests possible future work.

2 Rationalizability

In contrast with the concept of Nash equilibrium where each player's belief is required to coincide with her opponents' strategies, the concept of rationalizability, proposed independently by Bernheim [1984] and Pearce [1984], imposes a weaker requirement on players' beliefs. More precisely, it only demands players to obey the requirement of Bayesian rationality and common beliefs in Bayesian rationality. It attempts to account for rational behavior as the consequence of common knowledge of the game structure and the rationality of players, without imposing any further constraints on players' strategy choices.

Let us begin with some formal notations and definitions. Throughout this paper, we consider a finite normal or strategic form game $G \equiv \langle I, \{S_i\}, \{u_i\}\rangle_{i\in I}$, where $I = \{1, 2, \ldots, n\}$ is a finite set of players, S_i denotes a finite set of pure strategies (or actions) available to player i, and $u_i : S \to \mathbb{R}$ denotes player i's payoff function. We shall denote the set of player i's mixed strategies by Δ_i, which can be regarded as the set of all probability distribution over S_i. For each mixed strategy $\delta_i \in \Delta_i$, let $\delta_i(s_i)$ denote the probability assigned to s_i. Recall that a strategy profile is a *Nash equilibrium* if no player can benefit by merely changing her strategy while the other players keep theirs unchanged. More precisely, a mixed strategy profile $\delta^* \in \Delta$ is a (mixed strategy) Nash equilibrium if for each player i, $u_i(\delta^*_i, \delta^*_{-i}) \geqslant u_i(\delta_i, \delta^*_{-i})$ for every mixed strategy δ_i of player i. An alternative way to characterize the notion of Nash equilibrium is to define it in term of best response. We say that a strategy $\delta_i \in \Delta_i$ is a *best response* to δ_{-i} for player i if $u_i(\delta_i, \delta_{-i}) \geqslant u_i(\delta'_i, \delta_{-i})$ for all $\delta'_i \in \Delta_i$. Thus a strategy profile is a Nash equilibrium if each player's strategy is a best response to the other players' strategies. For an arbitrary set X of strategies, we denote by $\mathcal{H}(X)$ the convex hull of the set X, namely, the smallest closed convex set containing X.

It is well known that the concept of rationalizability attempts to characterize rational strategic behavior that are consistent with the assumption that both the structure of the game and the rationality of the players are common knowledge to them. To be more specific, rationalizability in normal form games is defined based on the following assumptions:

- **A1**: Each player employs a subjective personal probability to express her belief about the other players' strategy choice, which cannot conflict with any information available to her.
- **A2**: Each player attempts to maximize expected utility with respect to her subjective probability regarding her opponents' strategy choices.
- **A3**: The structure of the game, including the strategy space and payoff functions, and the fact that each player satisfies the above two assumptions are common knowledge.

Informally speaking, we can examine a player's rationality by checking whether the actions chosen by that player are "rational" or not. We say that an action of a player is rational if there exists some belief regulated by the assumptions given above such that it is a best response to that belief. Thus, a strategy δ_i of player i is *rationalizable* if she can justify her choice by explaining that (i) δ_i is rational, (ii) there exists some belief δ_{-i} such that δ_i maximizes her own expected utility with respect to δ_{-i}, and δ_{-i} assigns positive probability only to rational actions of her opponents, and (iii) there are beliefs of her opponents that make those actions rational and assign positive probability only to her rational actions, and so on. This suggests an intuitive way of defining rationalizability without invoking the iterative process originally suggested by Pearce (1984). In order to present this formal definition, we have to make the notion of a belief and what we mean by a strategy being rational explicitly.

Definition 1. *In a strategic form game G, a* belief *of player $i \in I$, denoted by μ_{-i}, about the other players' strategy choices is a probability distribution over the set of the other players' strategies $S_{-i} \equiv \prod_{j\neq i} S_j$.*

Here we should draw a clear distinction between the concepts of belief and mixed strategy. A belief about player i has the same mathematical form as a mixed strategy of player i, which is normally found in the literature. However, the interpretations of both concepts are different (see [Osborne and Rubinstein, 1994] for a comprehensive discussion on the interpretations of mixed strategies.). A mixed strategy of player i is usually viewed as an explicit randomization over her pure strategies in S_i. If player i chooses to play a mixed strategy, she commits herself to carry out the deliberate randomization. The main criticism of this interpretation of mixed strategy is that for each player there are usually infinitely many mixed strategies that yield her the same expected payoff as her mixed strategy equilibrium does, given her opponents' equilibrium behavior. But we are here concerned with a different solution concept called rationalizability. Thus, interpreting mixed strategies as objects of deliberate choice is appropriate within the current framework. On the other hand, a belief about player i is a probability distribution on the set of player i's mixed strategies, which represents another player's view about player i's strategy choice. It should not be confused with a randomization that is actually carried out by player i. In that sense, we can say that players' mixed strategies should be understood as the objects of the beliefs about players' strategy choices, and the probability distribution given by a belief about player i merely represents the likelihood that another player assigns to player i's mixed strategies.

Nevertheless, an essential feature of this formulation of belief is that it allows a player to believe that the other players choose their strategies according to certain correlated randomization devices, since a belief μ_{-i} of player i is a probability measure over S_{-i} and thus is an element of the set $\mathcal{H}(S_{-i})$. Note that a belief μ_{-i} of player i is not necessarily a product of independent probability distributions on each of the set S_j of actions for $j \in N\backslash\{i\}$. That is, a belief μ_{-i} of player i need not be identified as an element of the set of mixed strategies of her opponents S_{-i}. In addition, it is not difficult to see that the set S_{-i} is strictly smaller than the set $\mathcal{H}(S_{-i})$ in games with more than 2 players. Hence we have to use a different notation μ_{-i} for a belief in the current framework in order to distinguish it from a mixed strategy δ_{-i}.

It is assumed that each player always chooses an action to maximize her own expected payoff with respect to some belief about the opponents' strategies. A strategy being rational can then be defined precisely in terms of maximization of expected utility.

Definition 2. *A strategy δ_i of player i in a strategic form game G is a* rational *strategy if there exists a belief μ_{-i} of player i such that δ_i maximizes player i's expected utility, that is, $u_i(\delta_i, \mu_{-i}) \geqslant u_i(\delta_i', \mu_{-i})$ for all $\delta_i' \in \Delta_i$. In this case, we say that δ_i is a* best response *to the belief μ_{-i}.*

The key idea of the following characterization is to define an action (or pure strategy) to be rationalizable by considering each player's introspective process of justifying her own strategy choice, based on the analysis of her opponents' similar reasoning about their rational behavior.

Definition 3. *In a strategic form game G, an action $s_i \in S_i$ of player i is rationalizable if for each player $j \in I$, there exists a set $Z_j \subseteq S_j$ of actions such that: (i) $s_i \in Z_i$, and (ii) every action s_j in Z_j is a best response to some belief μ_{-j} of player j whose support is a subset of Z_{-j}.*

Whenever a new solution concept is put forward, a primary theoretical question is whether the proposed concept can give rise to at least one solution for games in general. Regarding the concept of rationalizability, the answer to this question is positive.

Proposition 2.1 (Pearce, 1984)**.** *For finite normal form games, the set of rationalizable strategies is always nonempty and contains at least one pure strategy for each player.*

We have considered above how to define the concept of rationalizability by using the notion of belief and the rationality of the players. As a matter of fact, the set of rationalizable actions can be further characterized for finite strategic games in terms of the familiar idea of dominance relations. As we shall see, this characterization for rationalizability gives rise to an operationalizable method for finding the set of rationalizable actions for finite games. Recall that the concept of rationalizability basically captures the idea that as a rational decision maker each player can only choose those strategies that are best responses to some beliefs regarding the other players' strategies. In other words, a rational player should not adopt a strategy that is not a best response to any belief about her opponents' strategy choices. In the game-theoretic terminology, such a strategy is called a never-best response strategy. Thus one can see that the concept of rationalizability is closely related to the notion of never-best response strategy as defined below.

Definition 4. *In a normal form game G, an action s_i of player i is a* never-best response *if it is not a best response to any belief of player i, that is, for every belief μ_{-i} of player i there exists a strategy $\delta_i \in \Delta_i$ such that $u_i(\delta_i, \mu_{-i}) > u_i(s_i, \mu_{-i})$.*

In other words, there is no belief μ_{-i} of player i about her opponents' strategies with respect to which a never-best response action s_i maximizes her own expected payoff. This coincides exactly with the central idea of rationalizability, namely that the players are rational in the sense of maximizing expected utility. As mentioned above, each player should rule out the actions that are not best response to any belief, namely, never-best response actions.

Let us now turn to the familiar notion of strict dominance which will play a crucial role in the characterization of rationalizable actions, as we shall see below.

Definition 5. *In a normal form game G, an action s_i of player i is* strictly dominated *if there exists a strategy $\delta_i \in \Delta_i$ such that $u_i(\delta_i, s_{-i}) > u_i(s_i, s_{-i})$ for all $s_{-i} \in S_{-i}$.*

In words, whatever the other players do, player i can benefit from playing some other strategy rather than a strictly dominated strategy. Clearly, a rational player would never use a strictly dominated strategy. Otherwise the player's choice violates the assumption of rationality in the sense of maximizing expected utility. At this point one may wonder whether the notion of never-best response is equivalent to the conception of strictly dominated action. It turns out that one can establish the equivalence between these two notions within the current framework.

Lemma 2.2 (Pearce, 1984)**.** *In a strategic form game G, an action s_i^* of player i is a never-best response if and only if s_i^* is strictly dominated.*

Suggested by the above lemma, we can show that the set of rationalizable actions can be obtained by iteratively deleting strictly dominated actions until we arrive at the stage where no more strictly dominated action can be further eliminated. Let us first formally define the process of iterated elimination of strictly dominated actions.

Definition 6. *Consider a normal form game G. Set $S_i^0 \equiv S_i$ for each $i \in I$. Then, for each $i \in I$ and for each $k \geqslant 1$, the set S_i^k is recursively defined as follows:*

$$S_i^k := \{s_i \in S_i^{k-1} \mid \nexists\, \delta_i \in \mathcal{H}(S_i^{k-1})$$
$$\text{such that } u_i(\delta_i, s_{-i}) > u_i(s_i, s_{-i}), \forall\, s_{-i} \in S_{-i}^{k-1}\}.$$

And define $S_i^\infty := \prod_{k=1}^{\infty} S_i^k$. The set S_i^∞ is the set of player i's actions that survives iterative elimination of strictly dominated actions.

Observe that after a finite numbers of steps the process of iterated elimination of strictly dominated actions will certainly halt in the sense that there is no action that can be further eliminated, since we restrict our attention to finite games. Moreover, one can show that the procedure of iterated elimination of strictly dominated actions does not depend on the order that we proceed the elimination, that is, it always yields the same surviving set of actions for each player.

With the aid of this procedure, we can thus easily identify the set of rationalizable actions for each player in finite games, which thus provides a nice algorithm for finding rationalizable actions.

Proposition 2.3 (Pearce, 1984)**.** *For any finite normal game G, the set of profiles of rationalizable actions coincides with the set of profiles that survives the process of iterated elimination of strictly dominated actions.*

3 Rationalizability with Imprecise Probabilities

Following the tradition of decision making under uncertainty, the concept of rationalizability assumes that each player's belief regarding the other players' strategies is represented by *a single personal probability measure.* However, there are many convincing arguments for supporting imprecision in beliefs - even in the context of single-agent decision problems (see [Ellsberg, 1961] and [Walley, 1999]). A number of alternative models to subjective expected utility theory have been proposed, which advocate the use of imprecise probabilities for dealing with uncertainty in decision problems (see, for instance, [Gilboa and Schmeidler, 1989] and [Levi, 1974]). It is thus natural to incorporate these ideas into the traditional game-theoretic framework. Based on the rules of Γ-maximin [Berger, 1985; Gilboa and Schmeidler 1989] and E-admissibility, we present here a generalized game-theoretic framework as an initial attempt to examine how modeling uncertainty with imprecise probabilities may provide insight into traditional game theory. In analogy with the concept of rationalizability, we propose two new game-theoretic solution concepts: the solution concept that we shall call *Γ-maximin rationalizability* attempts to capture the idea that each player models the other players as Γ-maximin decision makers, and the other one named *E-rationalizability* is meant to represent the idea that each player thinks of the other players as rational decision makers who respects the E-admissibility criterion.

An immediate question that is crucial to this investigation is: which model of imprecise probabilities should be assumed as representation of players' beliefs in strategic situations? There are a variety of mathematical models proposed in the literature to represent uncertainty in single-agent decision problems. For instance, lower previsions, upper and lower probabilities, sets of probabilities, non-additive probabilities, and belief functions (see [Walley, 1991]). Among these widely-discussed models of imprecise probabilities, a plausible method is to use *a convex set of probability distributions*, also called a credal set [Levi, 1980], to represent a decision maker's beliefs when confronted with uncertainty. A great advantage of this approach is that it allows us to deal with any state of insufficiencies in our information, including complete ignorance, in a unified way. Here we adopt this representation of uncertainty as the intended model for the players' beliefs regarding the other players' strategy choices. In order to distinguish it from the previous way of modeling beliefs, we will hereafter refer to a belief as a conjecture. Slightly modifying the formulation of belief in the framework of rationalizability, we define a conjecture of a player as follows:

Definition 7. *In a strategic form game G, a* conjecture *of player i, denoted by C_{-i}, about the other players' strategy choices is a (nonempty) convex set of probability measures over the opponents' actions S_{-i}.*

Note that this way of representing players' beliefs is a natural generalization of using a single probability distribution, as discussed earlier in the context of rationalizability. Moreover, this representation of beliefs admits the possibility of a correlated conjecture in the sense that, a player's conjecture may contain a probability distribution that cannot be obtained by independent mixtures over her opponents' strategies, for the elements of a conjecture are probability measures defined over S_{-i}.

One can interpret each member in a player's conjecture as the frequency of the strategy choices by her opponents, each of which is randomly drawn from a large population. More precisely, each player thinks that each of her opponents stands for a large set of players and has the same set of feasible choice. In this context, the probability distributions in player i's conjecture are viewed as the frequencies with which the members of the set S_{-i} are used by those large populations. In light of this, a probability distribution in a conjecture of a player has a completely different meaning from a mixed strategy, even though they may look the same from a mathematical point of view.

Under the preceding interpretation, it is reasonable to consider the cases where the set of strategies for some player is not convex, but players' conjectures are required to be convex. We understand that it is standard practice in game theory to consider the mixed extensions of games, that is, to include all the mixed strategies. Nevertheless, we may want to model circumstances where only the pure strategies are available to the players, which can be suitably described in the current framework with our interpretation.

In the context of single-agent decision making, several decision rules such as *Γ-maximin* [Berger, 1985; Gilboa and Schmeidler, 1989], *E-admissibility* [Levi, 1974], and *maximality* [Walley, 1999] have been discussed in the literature of imprecise probabilities (for a detailed comparison between these criteria see [Schervish et al., 2003], [Seidenfeld, 2004], and [Troffaes, 2007]). There is, however, no general agreement among decision theorists as to which is the right rule for judging rational decisions when uncertainty is expressed by a convex set of probability functions. Among these suggested criteria, the rule of Γ-maximin generalizes the principle of maximizing expected utility by simply taking the lower expected utility, thereby inducing a complete order on the decision set. More precisely, according to Γ-maximin, a rational decision maker should choose an option to maximize the minimum expected value with respect to a convex set of probabilities. This rule for decision making under uncertainty seems suitable for describing decision makers who are *uncertainty averse*, as it always takes the worst possible expected value as the base for maximization. Nevertheless, it has already been noted in [Seidenfeld, 2004] that the rule of Γ-maximin fails to distinguish between open and closed, convex and non-convex sets of probabilities, since choices based on this decision rule essentially reduces to binary comparisons which share the same supporting hyperplanes. It thus implies that the properties of closure and convexity concerning players' conjectures regarding their opponents' strategy choices are indistinguishable by Γ-maximin rationalizability.

The other decision criterion that we shall discuss below is often called E-admissibility, which was implicitly mentioned in [Savage, 1954] and extensively advocated by Issac Levi [1974]. According to this decision rule, an option is E-admissible if it maximizes expected utility relative to some probability distribution in the convex set of probabilities. In contrast with Γ-maximin, E-admissibility does not generate an order of options, but it does avoid the above-mentioned limitation, since it cannot be characterized by pairwise comparisons. As shown in the context of decision making, these two rules are not equivalent in the sense that they may recommend different sets of admissible options. Thus it is not surprising that the game-theoretic solution concepts defined based on these rules are not equivalent either, as illustrated by an example in the next section.

Under strategic situations, players are usually assumed to be uncertain about the other players' strategy behavior, and can only attempt to deduce their opponents' rational actions from the structure of the game and available information about their opponents' preferences. In most games, it is impossible for players to ascertain their opponents' actual behavior. Due to the insufficient information about preferences and irreducible strategic considerations, any level of uncertainty revealed by the imprecision in the set of probabilities may occur in situations of strategic interaction. Since Γ-maximin and E-admissibility have been often discussed in the literature of decision theory, it is therefore interesting to study the cases where all the players would use the rule Γ-maximin or E-admissibility to choose their strategies in games. By analogy to the framework of rationalizability, we need to be explicit about what we mean by a strategy being rational under uncertainty.

Definition 8. *In a strategic form game G, a strategy δ_i of player i is* Γ-rational under uncertainty *if there exists a conjecture C_{-i} of player i such that δ_i maximizes player i's minimum expected utility with respect to C_{-i}. In this case, we say that δ_i is a* Γ-maximin admissible *strategy relative to the conjecture C_{-i}.*

Likewise, we can define a notion called E-admissible strategy in a game where players are assumed to use E-admissibility as the criterion for strategy choices.

Definition 9. *In a strategic form game G, a strategy δ_i of player i is* E-rational under uncertainty *if there exists a conjecture C_{-i} of player i such that δ_i maximizes player i's expected utility for some probability* *in C_{-i}. In this case, we say that δ_i is an E-*admissible *strategy relative to the conjecture C_{-i}.*

Recall that the key idea of the concept of rationalizability is that each player regards the other other players as expected utility maximizers. It requires not only that players are rational in the sense of maximizing expected utility with respect some belief, but also that players' beliefs should be consistent with their opponents being rational in a similar way. The solution concept introduced below extends this idea to contexts, where each player is assumed to model the other players as decision makers who employ Γ-maximin or E-admissibility as the decision rule with respect to uncertainty. More specifically, we present a new solution concept that is meant to capture the idea that players are required to consider only those strategies that are rational under uncertainty, and that are supported by conjectures that do not contradict with their opponents being rational under uncertainty.

Now we need to specify the condition for a player's conjecture being consistent with her opponents' rationality in the senses of Definition 8 and Definition 9 rather than in a traditional decision-theoretic sense. A natural suggestion is to require that each element of the conjecture assigns positive probability only to those actions of her opponents that are rational under uncertainty. Putting these ideas together, we can formally define the new solution concept called *Γ-maximin rationalizability.*

Definition 10. *In a strategic form game G, an action $s_i \in S_i$ of player i is Γ-maximin rationalizable if for each player $j \in I$, there exists a set $A_j \subseteq S_j$ of actions such that: (i) $s_i \in A_i$, and (ii) every action s_j in A_j is Γ-maximin admissible relative to some conjecture C_{-j} of player j such that the support of each element of C_{-j} is a subset of A_{-j}.*

According to the above definition, one only needs to find a set of acts and a conjecture for each player in order to check whether a strategy is Γ-maximin rationalizable or not. Unlike the above formulation, Klibanoff [1996] has provided an alternative characterization of rationalizability with uncertainty aversion (see the definition before Theorem 4), which is defined as an iterative reduction process on the strategies. We shall see that his definition turns out to be equivalent to the concept of Γ-maximin rationalizability defined above. As noted in [Osborne, 2004], there are two distinct ways of defining rationalizability: one depends upon an iterated elimination procedure and the other does not. In the light of this, it seems fair to say that Klibanoff's characterization and the above formulation follow exactly the two different ways to generalize rationalizability in normal form games to accommodate uncertainty aversion, although they actually correspond to the same solution concept.

Analogously, the other solution concept that we call E-rationalizability can be formally defined as follows.

Definition 11. *In a strategic form game G, an action $s_i \in S_i$ of player i is E-rationalizable if for each player $j \in I$, there exists a set $A_j \subseteq S_j$ of actions such that: (i) $s_i \in A_i$, and (ii) every action s_j in A_j is E-admissible relative to some conjecture C_{-j} of player j such that the support of each element of C_{-j} is a subset of A_{-j}.*

4 Discussion of Properties

The aim of this section is to establish some properties of the solution concepts Γ-maximin rationalizability and E-rationalizability. Among other things, we will see that, Γ-maximin rationalizability can reasonably embrace a broader class of strategy profiles as outcomes under certain circumstances in comparison with rationalizability, whereas E-rationalizability can be distinguished from Γ-maximin rationalizability based on the ideas originated in decision theory. In addition, we will characterize the condition under which these three solution concepts coincide.

4.1 General Results

As we have noted, both of the decision rules, Γ-maximin and E-admissibility, can be regarded as simple extensions of the principle of maximizing expected utility to contexts where uncertainty is modeled by a set of probability measures. It is obvious that the former two rules lead to the same recommendations as the latter one when the set of probability measures is a singleton set. This enables us to show that the concepts of Γ-maximin rationalizability and E-admissibility generalize the notion of rationalizability to contexts where a set of probabilities is employed to represent uncertainty in games.

Proposition 4.1. *For any strategic form game G and each player i, if an action s_i^* of player i is rationalizable, then it is Γ-maximin rationalizable. This holds for E-rationalizability as well.*

Proof. Suppose that $s_i^* \in S_i$ is rationalizable. According to Definition 3, we have that there exists a set Z_j of actions for each player $j \in I$ such that both conditions specified in the definition are satisfied. Set $A_j \equiv Z_j$ for every player j. It immediately follows that $s_i^* \in A_i$. And it is clear that every action in A_j is both Γ-maximin admissible and E-admissible relative to some conjecture of player j by considering the set containing only one probability distribution over A_{-j}, as in this case both Γ-maximin and E-admissibility are equivalent to the principle of expected utility maximization. We can thus conclude that s_i^* is Γ-maximin rationalizable, and E-rationalizable as well. □

According to Proposition 2.1, the set of rationalizable actions of each player is nonempty for any finite normal form games. By applying this result, we can easily establish the existences of Γ-maximin rationalizable and E-rationalizable action in strategic games.

Corollary 4.2. *For any strategic form game, there always exists at least one* Γ*-maximin rationalizable action for each player* i*. This holds for* E*-rationalizable action as well.*

4.2 Comparisons

At this point, the reader may wonder whether the sets of Γ-maximin rationalizable and E-rationalizable actions are in fact identical to the set of rationalizable actions. It has already noted in [Epstein, 1997] that the concepts of Γ-maximin rationalizability and rationalizability are not equivalent when the analysis is restricted to only pure strategies. He also includes a generic game (see the game of Figure 1 in [Epstein, 1997]) that is designed to illustrate that difference. Yet he offers no explicit demonstration.

It has been pointed out in [Seidenfeld, 2004] that an option that is Γ-maximin admissible may not be Bayes admissible. Inspired by this result, I show by the following example that the concept of Γ-maximin may induce a larger set of solutions compared to rationalizability. It also serves the purpose of illustrating the definition of Γ-maximin rationalizability.

Example 1. Consider the 3 × 2 game shown in Figure 1. Unlike the usual setting which includes mixed strategies, we assume here that both players' feasible options are pure strategies only, that is, explicit randomization is excluded; no non-trivial mixed strategy is available to any player.

	L	R
U	$10, 1$	$0, 2$
M	$4, 10$	$4, 1$
D	$0, 1$	$10, 2$

Figure 1: A normal form game

It is easy to verify that only the pure strategies D and R are rationalizable for player 1 and 2 respectively. The previous argument basically relies on the fact that player 1's action M is strictly dominated when mixed strategies are taken into account. As a matter of fact, in this game the set of rationalizable action is the same, regardless of whether we allow explicit randomization or not. To see this, note that the action M is a never-best response, and thus does not belong to the support of any belief of her opponent. Therefore, the restriction imposed on the feasible options of the players does not alter the set of rationalizable actions for both players.

Nevertheless, I claim that all the actions of both player are Γ-maximin rationalizable in the sense of Definition 10. The crucial part for establishing the claim is to see that the action M of player 1 is actually Γ-maximin rationalizable, even though it is not rationalizable. This can be shown by considering the following construction: (i) let the sets of actions for both players be specified as follows: $A_1 = \{U, M\}$ and $A_2 = \{L, R\}$, and (ii) assume that player 1's and player 2's conjecture is depicted respectively by the following convex sets: $C_{-1} = \big\{\mathbb{P}_1(\cdot) : \{L, R\} \to [0, 1] \mid \mathbb{P}_1(\cdot) \text{ is a probability and } 0.2 \leqslant \mathbb{P}_1(R) \leqslant 0.6\big\}$ and $C_{-2} = \big\{\mathbb{P}_2(\cdot) : \{U, M, D\} \to [0, 1] \mid \mathbb{P}_2(\cdot) \text{ is a probability, } \mathbb{P}_2(D) = 0, \text{ and } 0.45 \leqslant \mathbb{P}_2(U) \leqslant 0.95\big\}$.

Under the specifications above, it is obvious that the first condition in Definition 10 is directly satisfied, since the action M belongs to the set A_1 specified for player 1. And it can be seen from Figure 2 and Figure 3 that the second condition is also satisfied, since player 1's lower expected payoff given by the actions U and M is the same with respect to the set C_{-1}, and the actions L and R also yield the same lower expected payoff to player 2 with respect to the set C_{-2}. We can thus say that every action in A_1 and A_2 is Γ-maximin admissible relative to the conjectures C_{-1} and C_{-2} respectively. In addition, note that every probability distribution in C_{-1} and C_{-2} assigns positive probability only to those action in A_2 and A_1 respectively. We can therefore conclude that the action M is Γ-maximin rationalizable. Once M can be Γ-maximin rationalized, it is then straightforward to verify that the other actions of both players are Γ-maximin rationalizable as well.

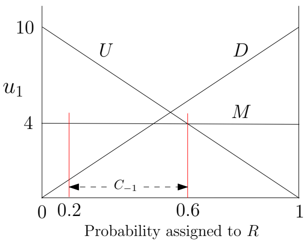


Figure 2: Expected utility to player 1

This example illustrates that the set of Γ-maximin rationalizable actions may differ from the set of rationalizable actions in some cases. In particular, the former

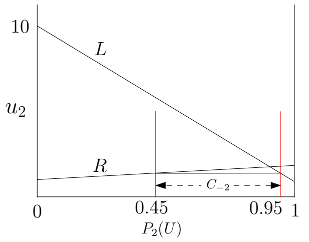


Figure 3: Expected utility to player 2

solution concept admits the action M as a candidate for the outcome of the game, which is ruled out by the concept of rationalizability. Intuitively, if player 1 is completely ignorant about player 2's strategy choices, it seems quite reasonable for player 1 to select M, as it has the highest security level. Thus one may say that the concept of Γ-maximin rationalizability does capture our intuition in some games.

The result suggested by the above example is not surprising, since the concept of Γ-maximin rationalizability in fact employs a richer representation of uncertainty than that assumed by rationalizability. More precisely, Γ-maximin rationalizability allows each player to model her opponents as Γ-maximin decision makers under uncertainty, which in fact includes the expected utility model considered by rationalizability as a special case. Hence, the concept of Γ-maximin rationalizability gives rise to a broader class of solutions under certain circumstances.

Nevertheless, it has been shown in [Seidenfeld, 2004] that the E-admissibility criterion differs from the rule of Γ-maximin in the context of individual decision making. It is therefore natural to expect that the solution concepts E-rationalizability and Γ-maximin rationalizability would not be equivalent in the game-theoretic context. In order to see this, consider again the game in Example 1. It is easy to see that player 1's option M is not E-admissible for any probability distribution over L, R. Based on this fact, we can then conclude that M is not E-rationalizable, which is Γ-maximin rationalizable as established above. Therefore, E-rationalizability and Γ-maximin rationalizability are not equivalent to each other in the sense that they may lead to different sets of admissible actions for players. It is worthwhile pointing out that E-rationalizability prescribes the same set of admissible actions as the one recommended by rationalizability in this example. It is not difficult to show that this holds for all finite normal form games. In this sense, the concept of E-rationalizability has a more intimate relationship with the traditional notion of rationalizability compared to Γ-maximin rationalizability.

Furthermore, there is another subtle difference between E-rationalizability and Γ-maximin rationalizability, which is based on some idea in decision theory. As mentioned before, in the context of individual decision making, Γ-maximin fails to distinguish among different convex sets of probabilities, while E-admissibility is capable of distinguishing between any two closed convex sets of probabilities. Putting this into a game-theoretic context, we can show that E-rationalizability and Γ-maximin rationalizability may lead to different sets of admissible options for a player given the same conjecture about opponents' strategy choices. In other words, even though the player holds the same belief model of the other players, E-rationalizability may recommend a different set of admissible options from the other suggested by Γ-maximin rationalizability. To see this, consider Example 1 again. Suppose that player 1's belief about player 2's strategy choice is represented by the conjecture $C_{-1} = \big\{\mathbb{P}_1(\cdot) : \{L, R\} \to [0,1] \mid \mathbb{P}_1(\cdot) \text{ is a probability and } 0.4 < \mathbb{P}_1(R) \leqslant 0.6\big\}$. Under this belief model, both M and D have the same infimum of expectation, and thus they are Γ-maximin admissible. However, only D is E-admissible, since D strictly dominates M with respect to C_{-1}. In this case, E-rationalizability and Γ-maximin rationalizability give rather different recommendations to player 1.

So far, we have shown how the notion of imprecise probabilities sheds light on the traditional game-theoretic framework, by illustrating the difference between Γ-maximin rationalizability and rationalizability, and further by examining the distinction between E-rationalizability and Γ-maximin rationalizability. However, it is also interesting to investigate when these solution concepts turn out to be equivalent. In other words, we want to give the conditions under which the decision rules Γ-maximin and E-admissibility reduce to the principle of expected utility maximization, including in cases where a convex set of probabilities is used to represent uncertainty.

Some basic notation and definitions are necessary for the following discussion. We are concerned here with finite decision problems where uncertainty is modeled by a closed convex set of probability functions. We let Ω denote a finite state space and let $\mathcal{O}$ denote a finite set of outcomes. An option (or act) f is a mapping from the state space Ω to the set of outcomes $\mathcal{O}$. Let $\mathcal{A}$ be a set of options available to the decision maker. As before, we will use the notation $\mathcal{H}(\mathcal{A})$ to denote the convex hull of $\mathcal{A}$. For sake of simplicity, we assume that the decision maker's values for outcomes are determinate and are represented by a cardinal utility function.

Definition 12. *Let $\mathcal{A}$ be a set of options and let $\mathcal{P}$ be a convex set of probability distributions on the un-*

derlying state space Ω. An option $f \in \mathcal{A}$ is Bayes admissible *with respect to $\mathcal{P}$ if there exists $\mathbb{P} \in \mathcal{P}$ such that f maximizes the expected utility under $\mathbb{P}$, that is, $\mathbb{E}_{\mathbb{P}}(f) \geqslant \mathbb{E}_{\mathbb{P}}(g)$ for all $g \in \mathcal{A}$.*

The above criterion recommends selecting those options in $\mathcal{A}$ that maximizes expected utility for at least one $\mathbb{P} \in \mathcal{P}$, which corresponds exactly to the idea of *E-admissibility.* We can now present the classic result (see Corollary 3.9.6 in [Walley, 1999] and Theorem 1 in [Schervish et al., 2003]) in decision theory, which plays a crucial role in establishing the central result of this section.

Proposition 4.3. *If the option set $\mathcal{A}$ is convex, then every option that is maximal admissible with respect to a closed convex set $\mathcal{P}$ of probability distributions is Bayes admissible with respect to $\mathcal{P}$. That is, if $f \in \mathcal{A}$ is not Bayes admissible, then there exists some $g \in \mathcal{A}$ different from f such that $\mathbb{E}_{\mathbb{P}}(g) > \mathbb{E}_{\mathbb{P}}(f)$ for all $\mathbb{P} \in \mathcal{P}$.*

We can now characterize the condition under which the concepts of Γ-maximin rationalizability and E-rationalizability are equivalent to rationalizability.

Proposition 4.4. *For any strategic form game G, if each player's choice set is convex and each player's conjecture regarding her opponents' choices is represented by a closed convex set of probabilities, then the set of Γ-maximin rationalizable actions is equal to the set of rationalizable actions. This holds for E-rationalizability as well.*

Proof. ($\Leftarrow$): It follows directly from Proposition 4.1.

($\Rightarrow$): Consider an arbitrary player $i \in I$. Suppose that s_i is not rationalizable. Then it follows from Proposition 2.3 that s_i is strictly dominated, which, by Lemma 2.2, implies that s_i is a never-best response. It thus follows that s_i is not a Bayes admissible action, since it is not a best response to any belief of player i. Note that each player's choice set is assumed to be convex. Hence, by Proposition 4.3, we have that s_i is not maximal admissible, that is, there exists some δ_i in player i's choice set such that player i's expected payoff to δ_i is strictly greater than her expected payoff to s_i with respect to any correlated belief regarding the other players' strategic behaviors. Accordingly, s_i is not Γ-maximin admissible relative to any conjecture, as any conjecture of player i is a subset of the set of correlated beliefs about her opponents' strategy choices. We can therefore conclude that the action s_i is not Γ-maximin rationalizable, as required.

The result concerning E-rationalizability can be established in a similar fashion. □

Klibanoff [1996] also establishes the equivalence between Γ-maximin (or uncertainty aversion) rationalizability and iterated strict dominance (see Theorem 4), whose proof depends heavily on the equivalence of the iterative definitions of uncertainty aversion rationalizability and rationalizability. By contrast, the proof I present here uses essentially Proposition 4.3, and thus has a decision-theoretic flavor. To some extent, the above proof makes explicit why such an equivalence holds by providing an alternative justification based on an important result in decision theory.

The above result implies that Γ-maximin rationalizability, E-admissibility and rationalizability suggest the same set of strategies for each player as rational decisions for games where players are allowed to consider the convex extensions of their choice sets. And it is quite standard in game theory to examine all the mixtures of the pure strategies. In view of this, we may say that the current framework provides a more general theoretical foundation for the concept of rationalizability. That is, the solutions suggested by rationalizability can be supported by a more general decision theory based on weaker assumptions. In that sense, rationalizability is a quite robust solution concept, which is implied merely by the assumption of common knowledge of players being Γ-maximin rational or E-rational.

5 Concluding Remarks

A variety of mathematical models have been discussed in the literature to deal with decision making under uncertainty in single-agent decision problems. In contrast with canonical Bayesian decision theory, which uses just one probability function to represent a decision maker's uncertainty, these models use imprecise probabilities, such as a nontrivial set of probability functions, to represent uncertainty. Based on this idea, I have developed in this paper a general theoretical framework for analyzing how different decision rules can be incorporated into the framework of normal-form rationalizability when uncertainty is represented by imprecise probabilities.

More precisely, I extended the notion of rationalizability to the case where players' conjectures about opponents' strategy choices are represented by a convex set of probability measures, instead of a unique probability function. In the spirit of rationalizability, I introduced a solution concept called Γ*-maximin rationalizability*, which captures the idea that each player models the other players as Γ-maximin decision makers with respect to sets of probabilities representing uncertainty; similarly, I also defined another solution concept named *E-rationalizability.* It is easy to see that both Γ-maximin rationalizability and E-*rationalizability* include the concept of rationalizability as a special case when the set of probability measures contains only a single probability function. In addition, I have shown by an example that these concepts are not equivalent. I have also identified the conditions under which these solution concepts coincide with each other.

Now I sketch some suggestions for future work along the current line of research. One natural project is to apply some other decision rules like maximality to interactive situations, in a way similar to the framework developed in this paper. And it also seems natural to extend the current framework to the context of extensive form games in which sequential decisions are involved. In this way, one can develop a general theory of games under uncertainty.

References

[1] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, 1985.

[2] D. Bernheim, Rationalizable Strategic Behavior. *Econometrica* 52: 1007-1028, 1984.

[3] J. Dow and S. Werlang, Nash Equilibrium under Knightian Uncertainty: Breaking Down Backward Induction. *J. Econ. Theory* 64: 305-324, 1994.

[4] J. Eichberger and D. Kelsey, Non-Additive Beliefs and Strategic Equilibria. *Games and Economic Behavior* 30: 183âĂŞ215, 2000.

[5] D. Ellsberg, Risk, Ambiguity and the Savage Axioms. *Quart. J. Econom.* 75: 643-669, 1961.

[6] L. Epstein, Preference, Rationalizability and Equilibrium, *J. Econom. Theory* 73: 1-29, 1997.

[7] I. Gilboa and D. Schmeidler, Maxmin Expected Utility with Non-Unique Prior. *J. Math. Econom.* 18: 141-153, 1989.

[8] J. C. Harsanyi and R. Selten, *A General Theory of Equilibrium Selection in Games.* MIT Press, Cambridge, MA, 1988.

[9] P. Klibanoff, Uncertainty, Decision and Normal Form Games, mimeo, 1994.

[10] K. C. Lo, Equilibrium in Beliefs under Uncertainty. J. Econom. Theory 71: 443-484, 1996.

[11] I. Levi, On Indeterminate Probabilities. *J. Phil.* 71: 391-418, 1974.

[12] I. Levi, *The Enterprise of Knowledge.* MIT Press, Cambridge, MA, 1980.

[13] H. Liu, Robust Equilibria under Linear Tracing Procedure. In *Proc. of 7th Int. Symp. on Imprecise Probabilities: Theories and Applications*, F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger (eds.), Innsbruck, Austria, 257-266, 2011.

[14] M. J. Osborne, *An Introduction to Game Theory.* Oxford University Press, New York, Oxford, 2004.

[15] M. J. Osborne and A. Rubinstein, *A Course in Game Theory.* MIT Press, Cambridge, MA, 1994.

[16] D. Pearce, Rationalizable Strategic Behavior and the Problem of Perfection. *Econometrica* 52: 1029-1050, 1984.

[17] L. J. Savage, *The Foundations of Statistics.* Wiley, New York, 1954.

[18] T. Seidenfeld, A contrast between two decision rules for use with (convex) sets of probabilities. *Synthese* 140: 69-88, 2004.

[19] M. J. Schervish, T. Seidenfeld, J. B. Kadane, and I. Levi, Extensions of Expected Utility Theory and Some Limitations of Pairwise Comparisons. In *Proc. of 3rd Int. Symp. on Imprecise Probabilities and Their Applications*, J. M. Bernard, T. Seidenfeld, M. Zaffalon (eds.), Lugano, Switzerland, Carleton Scientific, Waterloo, 496-510, 2003.

[20] D. Schmeidler, Subjective Probability and Expected Utility Without Additivity. *Econometrica* 57: 571-587, 1989.

[21] M. Troffaes, Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning* 45: 17-29, 2007.

[22] P. Walley, *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, New York, 1991.

Significance of a decision making problem under uncertainty

Kevin Loquin
LIRMM
161 rue Ada
34095 Montpellier Cedex 5 France
kevin.loquin@lirmm.fr

Abstract

In this paper, we work on the interval dominance based extension of the Savage Expected Utility Maximization (SEUM) approach. While usual probabilities only handle variability due uncertainty, imprecise probabilities additionally handle, in a unique framework, epistemic uncertainty. This side of uncertainty, often called imprecision, can generate incomparability between the acts of a decision problem. Incomparability is linked to information held by the imprecise probability model quantifying the outcomes uncertainty. Our proposal, in this paper, is that for a given decision problem, its significance is the quantity of information which makes the interval dominance based imprecise SEUM decision problem change from incomparable to decidable (and possibly still not comparable) or comparable (and possibly still not decidable). We discuss incomparability sources, a theoretical and a pragmatical definition of significance of a decision problem under uncertainty.

Keywords. Savage EUM, decision theory, imprecise probability, interval dominance, significance

1 Introduction

Decision making boils down to comparing the outcomes of many possible acts. Modeling a decision problem (DP) is as simple as ranking the set of possible acts according to a preference relation generally constructed from a quantification of the consequences of each act (by means of a utility function). The distinction between the notions of comparability and decidability is very important in our work. A DP is said to be comparable when its set of acts can be ranked according to a complete preference relation. A DP is said to be decidable when there is a unique act which is optimal according to its preference relation.

When no uncertainty pertains the problem, the preference relation is naturally complete: any act can be ranked according to this preference relation. Even if the DP is not decidable, optimal choice(s) can always be found.

Decision making under uncertainty stands for situations when an act does not lead to a unique outcome with certainty. Since the preference relation between the acts is constructed from their outcomes, it seems natural to admit incomparability when facing uncertainty. Nevertheless, what is important for most decision maker is to work with a comparable DP and if possible with a decidable one. Most of the last century advances in decision making under uncertainty aimed at making complete, the ranking between the acts, by means of axioms which are supposed to be consistent with a rational (subjective) behavior. Among them, the Savage axiomatic, from which is derived the Expected Utility Maximization criterion (SEUM), is the most popular one [14].

Our view is that it is an artificial and arbitrary task to force the comparability of a DP under uncertainty. Indeed, due to partial information, a DP is inherently incomparable. We propose to ground our definition of significance on this “informativist” view of decision under uncertainty. Thus, we propose to work with imprecise probability based decisions theories [11, 12, 17, 18] instead of usual probability based decision theories [14]. The main asset of imprecise probability theories [1, 16, 18] over the usual probability theory is that they handle partial information. Within these schemes, uncertainty is characterized by interval weights (probability or more generally utility) instead of point-valued weights due to partial information. An imprecise probability model is a convex family of precise probabilities.

As is done in SEUM, some of the imprecise probability based decision theories propose a complete ranking between acts (maximin, maximax or Hurwicz criteria [11]), but what is the main richness of the imprecise probability based decision models is that they can admit incomparability due to lack of knowledge in a rigorous way. For instance, the E-admissibility criterion of Isaac Levi [12], the maximality as proposed by Peter Walley [18] or the very simple interval dominance decision rule all admit incomplete preference relations between the acts.

We aim at proposing a notion of significance of a DP based on this informativity interpretation of uncertainty in decision theory. Let us take any DP under uncertainty (thus incomparable or undecidable in the general case), its significance is the smallest quantity of information required to make it comparable or decidable. This is a quite intuitive idea: one faces a decision problem which is incomparable, the amount of required information to disambiguate the problem naturally characterizes the significance of the original problem. In some sense, significance aims at measuring the missing information for making the DP complete or decidable.

This definition of significance of a DP under uncertainty is abstract and is not grounded on any decision or uncertainty theory. In order to derive more concrete definitions, we propose to define the significance of a DP under uncertainty from the interval dominance decision rule and imprecise probability assessments over the outcome space.

Section 2 is a reminder (or a presentation) about the lower prevision model which summarizes the materials required for a proper understanding of this paper. The notion of imprecise expectation is particularly stressed. Section 3 presents the usual generalizations to imprecise probabilities of the SEUM. Finally Section 4 discuss the notion of significance of a DP under uncertainty as we aim at presenting it. A general (unrealistic) definition of significance is proposed, followed by a pragmatic definition of a significance index of a DP. A toy example inspired from [8] is also proposed to illustrate the notion of significance.

2 Imprecise Probability Theory

It is generally obvious (and probably non discussable) for most readers that the uncertainty about the outcome of any experiment is modeled by a set of precise weights between 0 and 1 on all the possible outcomes of this experiment: the probability weights. The general idea behind most of the imprecise probability theories is that uncertainty should preferably be modeled by a set of probability weights in order to handle imprecision, partial information or lack of knowledge inherent to most systems. Such new uncertainty theories are more general and powerful models than probability because they jointly and consistently handle the distinct notions of uncertainty due to variability and uncertainty due to imprecision that is generally called epistemic uncertainty.

2.1 Imprecise Probability Models

This theory presentation (which can be bypassed by expert readers) will emphasize on lower previsions defined on *discrete* domains, i.e. on domains with finite cardinality.

Let X be an uncertain variable whose possible outcomes are on a (finite) space $\mathcal{X}$ containing N exclusive single elements. Let $\mathcal{L}(\mathcal{X})$ denote the set of bounded real-valued functions on $\mathcal{X}$. $\mathcal{L}(\mathcal{X})$ is called the set of gambles. Each element (gamble) $f \in \mathcal{L}(\mathcal{X})$ is interpreted as the function on $\mathcal{X}$ representing the rewards $f(x)$ associated to the occurence of any possible outcome $x \in \mathcal{X}$ of X. Since the outcome value $x \in \mathcal{X}$ is uncertain, $f(x)$ is also an uncertain reward and thus f is an uncertain gamble.

A lower prevision $\underline{E}$ on $\mathcal{L}(\mathcal{X})$ is defined as a mapping $\underline{E} : \mathcal{K} \subseteq \mathcal{L}(\mathcal{X}) \to \mathbb{R}$. Its behavioral interpretation advocated by Walley is as follows: $\underline{E}(f)$ is interpreted as the supremum buying price an agent would accept for the uncertain reward $f(x)$. In order to ease the understanding of this fundamental concept, a lower prevision $\underline{E}(f)$ can be seen as the lower bound of the expectations of the uncertain gamble f. To a lower prevision $\underline{E}$ is associated its dual upper prevision $\overline{E}$ (or upper expectation), defined as $\overline{E}(f) = -\underline{E}(-f)$.

A lower prevision is said to be **coherent** on its gamble domain $\mathcal{K} \subseteq \mathcal{L}(\mathcal{X})$ if it satisfies the following conditions:

- **(C1)** $\underline{E}(f) \geq \inf_{x\in\mathcal{X}} f(x)$ for all $f \in \mathcal{K}$ (accepting sure gain);
- **(C2)** $\underline{E}(\lambda f) = \lambda\underline{E}(f)$ for each $f \in \mathcal{K}$ and $\lambda \geq 0$ (positive homogeneity);
- **(C3)** $\underline{E}(f+g) \geq \underline{E}(f)+\underline{E}(g)$ for all $f, g \in \mathcal{K}$ (superadditivity).

A less restrictive class of lower previsions is the class of lower previsions avoiding sure loss. Let $\overline{G}(f)$ be the highest expected gain with a gamble $f \in \mathcal{K}$. It is naturally defined by $\overline{G}(f) = f - \underline{E}(f)$. We thus have a loss on f for the assessed prevision $\underline{E}$ when $\overline{G}(f) < 0$. Thus a lower prevision model is said **avoiding sure loss** when there is a set of gambles $(f_j)_{j=1,\ldots,n}$ of $\mathcal{K}$ fulfilling $\sum_{i=1}^{n} \overline{G}(f_j) \geq 0$, i.e. when there is at least a set of gambles whose combination avoids a sure loss.

A coherent lower prevision which is equal to its associated upper prevision is said to be linear. Therefore, a coherent linear prevision denoted by E, i.e. such that for all $f \in \mathcal{K}$, $\underline{E}(f) = \overline{E}(f)$ fulfills both super-additivity **(C3)** and sub-additivity[1] and thus the finite additivity axiom: $E(f+g) = E(f) + E(g)$. A linear prevision can be seen as a usual expectation operator.

A lower prevision can be associated to a convex set of linear previsions. The set of linear previsions dominating the coherent lower prevision $\underline{E}$, defined on $\mathcal{K}$, called the credal set, is defined by:

$$\mathcal{M}(\underline{E}) = \{E \in \mathcal{E}(\mathcal{X}) \mid (\forall f \in \mathcal{K})\,(\underline{E}(f) \leq E(f))\}\,, \tag{1}$$

where $\mathcal{E}(\mathcal{X})$ is the set of linear previsions on $\mathcal{X}$.

This object is particularly interesting since it links a lower expectation to its associated coherent set of dominating expectations.

An important particular case of coherent lower prevision is the lower probability. To any subset (or event) A of $\mathcal{X}$ can be associated its indicator function, which is a gamble, $\mathbb{1}_A \in \mathcal{L}(\mathcal{X})$. The lower probability of an event $A \subset \mathcal{X}$, denoted by $\underline{P}(A)$, is the lower prevision associated to this gamble $\mathbb{1}_A$. We denote by $\mathcal{B}(\mathcal{X})$, the set of indicator functions on $\mathcal{X}$, in order to remind the Borel algebra: $\mathcal{B}(\mathcal{X}) \subset \mathcal{L}(\mathcal{X})$ can be seen as the set of events on $\mathcal{X}$. To a lower probability is associated the dual notion of upper probability $\overline{P}(A) = 1 - \underline{P}(A^c)$, where A^c denotes the complement of A on $\mathcal{X}$.

Many other particular cases of the lower prevision model exist [3, 4, 19] that match the following inclusion: a necessity measure (dual of a possibility measure) is a particular case of belief function (whose pieces of evidence are consonant or nested [6]) ; a belief function is a particular case of convex Choquet Capacity ; a convex Choquet capacity is a particular case of lower probability ; a lower probability is a particular case of lower prevision.

2.2 Imprecise Expectation and natural extension

Facing a given quantity of information, a complete modeling of an uncertain variable X is done when a coherent lower prevision can be associated to any possible gamble f of $\mathcal{L}(\mathcal{X})$. However, in most real applications, information is limited to a lower prevision, denoted by $\underline{E}_{\mathcal{K}}$, defined on a subset of gambles $\mathcal{K} \subset \mathcal{L}(\mathcal{X})$. The natural extension procedure allows distributing (conveying) information held by $\underline{E}_{\mathcal{K}}$ to $\mathcal{L}(\mathcal{X})$ in the most conservative way. In other words, the natural extension is the most specific model, denoted by $\underline{E}$, that can be constructed on $\mathcal{L}(\mathcal{X})$ without any additional information incorporation, i.e. without reducing the model $\underline{E}_{\mathcal{K}}$.

Definition 2.1 (Natural Extension)
Suppose $\underline{E}_{\mathcal{K}}$ is a lower prevision on $\mathcal{K} \subset \mathcal{L}(\mathcal{X})$, then its natural extension $\underline{E}$ is defined, for any $f \in \mathcal{L}(\mathcal{X})$, by

$$\underline{E}(f) = \sup_{\mathbb{R}} \left\{ \begin{array}{c} \alpha : f - \alpha \geq \sum_{j=1}^{n} \lambda_j (f_j - \underline{E}_{\mathcal{K}}(f_j)), \\ \textit{for some } n \geq 0, f_j \in \mathcal{K}, \lambda_j \geq 0 \end{array} \right\}. \tag{2}$$

$\underline{E}(f)$ is the supremum buying price for the gamble f given that the linear combination of the highest gain $\overline{G}(f_j) = f_j - \underline{E}_{\mathcal{K}}(f_j)$ associated to any set of gambles f_j of $\mathcal{K}$ is still higher than the gain $\overline{G}(f)$ obtained on the gamble f for this price $\underline{E}(f)$.

When $\underline{E}_{\mathcal{K}}$ avoids sure loss $\underline{E}$ is the minimal coherent lower prevision which dominates $\underline{E}_{\mathcal{K}}$ on $\mathcal{K}$. This gives all its meaning to the expression “in the most conservative way”, which characterizes the way an uncertainty model $\underline{E}_{\mathcal{K}}$ on $\mathcal{K}$ is extended to $\underline{E}$ on $\mathcal{L}(\Omega)$. Note also that when $\underline{E}_{\mathcal{K}}$ is coherent, $\underline{E}_{\mathcal{K}}$ and $\underline{E}$ coincide on $\mathcal{K}$.

This tool is of prime importance since it tells us how to accomplish inference from the assessment of an imprecise prevision model on $\mathcal{K} \subset \mathcal{L}(\mathcal{X})$ to any gamble of $\mathcal{L}(\mathcal{X})$.

An interesting particular case is when $\mathcal{K} = \mathcal{B}(\mathcal{X})$. In that case, the natural extension procedure coincides with the computation of the lower expectation of any bounded function f: $\underline{E}(f)$ associated to the constraints provided by the lower probability model $\underline{P}$ defined on $\mathcal{B}(\mathcal{X})$. This is exactly what defines the imprecise expectation: this is the natural extension to $\mathcal{L}(\mathcal{X})$ of a lower prevision defined on $\mathcal{B}(\mathcal{X})$ (thus of a lower probability).

The imprecise expectation can only, in the general case of lower probability, be computed by using linear programming techniques. But, for a convex capacity (and any of its submodels: necessity, belief function,...), imprecise expectation can be computed by means of the Choquet integral [2].

3 Decision under uncertainty with Imprecise Probability

Uncertainty modeling has many available distinct theories generally associated to different interpretations. Decision modeling under uncertainty shares the same kind of diversity in its theories. In this section, we present the most encountered decision theories under uncertainty.

[1] Sub-additivity: axiom **(C3)** with the reverse inequality.

3.1 SEUM decision theory

In the SEUM decision theory, there is the set of possible acts, denoted by $\mathcal{A}$. Decision making under uncertainty stands for situations when an act does not lead, in general, to a unique outcome with certainty. Each act X of $\mathcal{A}$ is an uncertain variable with value in the finite outcome space, denoted by $\mathcal{X}$. This outcome space is a rather abstract space which can be numerical or not. For instance *patient healing* or *flood* are non numerical outcomes encountered in usual DP under uncertainty in the fields of medical decision or environmental risk assessment. In the SEUM approach, a utility function on the outcome space is used: $u : \mathcal{X} \to \mathbb{R}$ to quantify (and possibly rank) the acts (or their outcomes) on a utility scale.

Under the Savage axioms, the following complete preference relation $\succeq$ is constructed on $\mathcal{A}$ and defined, for X and Y in $\mathcal{A}$ by:

$$X \succeq Y \text{ iff } E_X(u) \geq E_Y(u). \tag{3}$$

In other words, an act X is preferred to another act Y when its associated expected utility is higher than the expected utility associated to Y. The optimal act(s) X^* is (are) such that

$$X^* \succeq Y, \ \forall Y \in \mathcal{A}.$$

At this point it is interesting to link some notations of the SEUM approach to notations of our imprecise probability presentation (IP) of Section 2. For instance, a gamble f in IP theory is similar to the utility function u of SEUM. Besides, the uncertain variable of IP and the uncertain outcome of SEUM, both denoted by X, are similar objects. We chose to incorporate the uncertain variable X to our IP presentation, which is generally not present in Walley theory and especially not in Walley's book [18], since it can easily be linked to the uncertain outcome of usual decision theories under uncertainty.

SEUM is a very elegant axiomatic construction [14] which entails a rational interpretation to preference structure (3). Many authors discussed and criticized the foundations of this approach by stressing too strong axioms [10]. Perhaps the most severe and constructive criticism is due to Ellsberg [7]. The SEUM is based on the idea that a decision maker behaves as if he possesses a complete and exhaustive knowledge of the possible states of the world, and moreover that, his assessment of the uncertainty about the outcomes may be represented as a unique finitely additive probability model. This idea has been termed as *probabilistic sophistication* [13]. Experimental evidence, as the Ellsberg paradox [7], has failed to support probabilistic sophistication as a good descriptive theory of behavior under uncertainty.

3.2 Imprecise SEUM generalizations and associated decision rules

Questioning the probabilistic sophistication principle of the SEUM approach has been done for many submodels of the lower prevision model: for possibility theory [5], for belief functions [10] or for capacities [15]. In such particular cases, the usual expectation operator based on the Lebesgue integral is replaced by a two-fold Choquet integral to compute the bounds of an imprecise utility expectation operator. The most general framework, i.e. obtained when uncertainty about the outcomes is modeled by a lower prevision $\underline{P}$, is computationally less tractable since it does not involve an explicit formulation of the imprecise expectation bounds but only linear optimization techniques.

Actually, most proposed generalizations of the SEUM to lower previsions were reduced to find meaningful ways to compare imprecise quantities: the imprecise expected utilities $[\underline{E}_X(u), \overline{E}_X(u)]$ instead of comparing precise quantities: the expected utilities $E_X(u)$. In other words, most approaches aim at finding a meaningful way to fulfill the first Savage axiom (which claims that *a preference relation is a complete ordering on the set of possible acts* $\mathcal{A}$) when the compared quantities are imprecise.

In order to expose some of the most encountered approaches, it is interesting to provide interpretations to $[\underline{E}_X(u), \overline{E}_X(u)]$. If u is a utility function on $\mathcal{X}$, $u(x)$ is uncertain due to the uncertainty on the outcomes of the act X, thus $\underline{E}_X(u)$ can be considered as the pessimistic expected utility associated to act X and $\overline{E}_X(u)$ can be considered as the optimistic expected utility associated to act X. In such framework, $u(x)$ represents a reward. It is therefore intuitive to term as optimistic the highest reward we can expect for uncertain outcomes of act X, i.e. $\overline{E}_X(u)$. Conversely, being pessimistic is to consider only the lowest reward we can expect with such model, i.e. $\underline{E}_X(u)$. Another remark is that when we are optimistic on a reward, we are pessimistic on a loss (and conversely) which is translated by relation $\overline{E}_X(u) = -\underline{E}_X(-u)$, since $-u$ is a loss when u is a reward.

This relevant interpretations of $\underline{E}_X(u)$ and $\overline{E}_X(u)$ as respectively the pessimistic and optimistic expected utility lead to propose a parametric optimal decision rule: the Hurwicz criterion, whose parameter r is a marker of the risk aversion of the decision maker. Actually, a decision maker has a high level of risk aversion when he considers for comparative quantities in its DP, the pessimistic expected utility. To favor the less risky problem posing and to be pessimistic are equivalent. Thus under a risk aversion (or pessimistic) attitude, the optimal decision rule is given by

$$X \succeq_P Y \text{ iff } \underline{E}_X(u) \geq \underline{E}_Y(u). \tag{4}$$

Optimism and risk are generally in accordance, thus the optimistic optimal decision rule is

$$X \succeq_O Y \text{ iff } \overline{E}_X(u) \geq \overline{E}_Y(u). \tag{5}$$

As a tradeoff between these rules stands the Hurwicz criterion. It is based on defining the expected utility for a risk aversion degree of r by :

$$E^r_X(u) = r\underline{E}_X(u) + (1-r)\overline{E}_X(u). \tag{6}$$

r is a sensible risk aversion index since $E^1_X(u) = \underline{E}_X(u)$ and $E^0_X(u) = \overline{E}_X(u)$. Thus the Hurwicz decision rule for a risk aversion degree r is

$$X \succeq^r_H Y \text{ iff } E^r_X(u) \geq E^r_Y(u). \tag{7}$$

Note that $\succeq_P$ is exactly $\succeq^1_H$ and $\succeq_O$ is exactly $\succeq^0_H$.

While the imprecise probability framework is supposed to model imprecision or epistemic uncertainty, to our view, the only consistent approaches, regarding this “informativist” view, are the approaches which allow incomparability between acts. At first sight admitting incomparability is problematic for providing optimal choices. However, this is a quite intuitive idea when facing epistemic uncertainty. In most cases where information is partial, admitting incomparability (and/or indecision) is safer than proposing a choice even if this choice is supposed to be obtained with a pessimistic rule. Let us consider an example of cancer diagnosis which illustrates a rational behavior under epistemic uncertainty: for most kind of cancers, abnormal blood tests results are not significant enough to diagnose cancer and an additional biopsy is generally required. Thus when information is partial (only the blood tests result), the physician admits incomparability and thus indecision. He will never claim that the patient has cancer and decide to start a heavy chemotherapy treatment only from these partial evidences.

Three decision rules admitting incomparability between acts are generally considered, the E-admissibility of Isaac Levi [12], the maximality as proposed by Peter Walley [18] or the very simple interval dominance decision rule. Interval dominance criterion is defined through the following incomplete preference relation:

$$X \succeq_{ID} Y \text{ iff } \underline{E}_X(u) \geq \overline{E}_Y(u). \tag{8}$$

This is certainly the most intuitive and simple decision rule admitting incomparability with imprecise probability. It says that an act X is preferred to an act Y if the imprecise expected utility of X completely (in terms of interval) dominates the imprecise expected utility of Y.

Actually this is the most cautious rule. Indeed a DP which is not comparable for the interval dominance criterion can be comparable for the E-admissibility and/or the maximality criteria. It implicitly means that available information is considered as insufficient for the interval dominance criterion while sufficient for the other criteria.

3.3 Sources of incomparability: a discussion

As already mentioned, a DP is said to be comparable when its set of acts can be ranked according to a complete preference relation and a DP is said to be decidable when there is a unique act which is optimal according to its preference relation. There is no inclusion relation between the decidability and the comparability of a DP. A decidable problem is not necessarily comparable. This is the case if there exists an optimal act for a partial preference ordering. Conversely, a comparable problem is not necessarily decidable. This is the case for any problem which results in more than one indifferent optimal acts for a complete preference ordering.

In this paper, we propose to use the non comparability of a DP under uncertainty to define its significance. Thus, it is interesting to discuss the incomparability sources of an imprecise SEUM problem. To our view, the sources of incomparability are twofold: 1/ epistemic (or reducible) uncertainty but also 2/ the problem construction itself. While they may not be the only sources of incomparability of an imprecise SEUM problem, they are certainly among these sources.

Indeed, 1/ the influence of the epistemic uncertainty on the comparability of a DP can easily be shown: let us take any incomparable imprecise SEUM problem, if uncertainty is reduced to a precise probability model then we recover a usual (i.e. precise) SEUM and thus a comparable DP.

And, 2/ the influence of the problem construction itself can be put forward: let us consider two different problems (i.e. two different utility functions) but with the same set of acts and associated uncertain outcomes and the same imprecise probability assessments for these parameters. We denote (P1) and (P2) these imprecise SEUM problems. We can find cases where (P1) provides a comparable decision framework, while (P2) is still incomparable.

Among the other possible sources of incomparability, we were wondering if the imprecise expectation operator which is used to pass from the uncertainty assessment step to the comparison step of an imprecise SEUM problem, has some impact on the comparability of the problem. Our answer is not clear yet but we showed some continuity results of the imprecise expectation operator in a working paper. These results tend to prove that the imprecise expectation operator does not impact the comparability of the problem. Indeed, continuity means that variations (measured with Hausdorff distances) between imprecise expectations are bounded by the variations between their generative imprecise probability models. Such stability is important in imprecise SEUM. It means that information rooting the uncertainty assessment of an imprecise SEUM problem is properly conveyed to the utility comparison step. More than this topological stability, it was already said that the natural extension is the most conservative extension of an imprecise probability model to the expectation of a utility function (or gamble).

4 Significance of a decision making problem under uncertainty

Now, let us reexamine an already considered situation: we are facing two different problems (i.e. with two different utility functions) with the same uncertain outcomes. Let us consider that both problems are non comparable and non decidable. If we progressively reduce the epistemic uncertainty associated to the uncertain outcomes of the problems, one problem, for instance (P1), should become comparable or decidable before the other problem (P2). It is thus natural to claim that problem (P1) is more significant than problem (P2) regarding the original pieces of information. Indeed, (P1) requires less artificial information addition than (P2) to become decidable or comparable.

The previous paragraph is the heart of this paper, since it explains the notion of significance as we hear it. We will say that a DP under uncertainty is fully significant if its associated ranking of the set of acts $\mathcal{A}$ is complete for the interval dominance or is decidable (even if non comparable). A DP under uncertainty is fully insignificant when the system must be reduced to a precise SEUM to become a comparable DP (decidable or not). Between these extreme cases, we will define the significance index of an incomparable and undecidable DP: it is the smallest quantity of information required to make it comparable or decidable.

In a sense, significance, as we aim at defining it, is a measure of "missing information" to make the problem comparable or decidable. Thus significance is a measure of meta-information: information about information. As for imprecise SEUM problems, information is modeled by lower previsions. It models information about a true underlying probability measure. Thus, meta-information can only be consistently quantified if we know the true underlying probability. In other words, it is impossible to judge information (i.e. to quantify meta-information) without knowing the truth. That is the reason why we ground our first definition of significance on the (unrealistic and unapplicable) assumption that we know the true underlying probability of an imprecise SEUM problem.

Note that all the involved lower probabilities in this definition of significance are consistant with the true underlying probability. It means that we only work with information which are not conflicting. Thus, we do not compete with formal decision frameworks which deal with ambiguity and conflict as separate types of uncertainty [9].

4.1 An unrealistic general definition of a significance index

The most general (but unrealistic) definition of a significance index that we will propose requires some preliminary definitions and notations.

Let $\underline{P_X}$ be a lower probability on the act X, which is an uncertain variable with values in the outcome space $\mathcal{X}$. Let P_0 be the true underlying probability modeling the uncertainty about X. We assume that $\underline{P_X}$ is consistant with P_0, i.e. $P_0 \geq \underline{P_X}$.

Let $\underline{\mathcal{P}}(\underline{P_X}) = \{\underline{P} : P_0 \geq \underline{P} \geq \underline{P_X}\}$ be the set of lower probabilities consistent with P_0 and dominating $\underline{P_X}$. It is the set of lower probability models more specific than $\underline{P_X}$, i.e. more informed, and still consistant with $\underline{P_X}$.

Let d be a distance between imprecise probabilities of $\underline{\mathcal{P}}(\underline{P_X})$ which respects the domination. We mean that, for three encapsulated (according to heir specificity) lower probabilities $\underline{P_1}$, $\underline{P_2}$ and $\underline{P_3}$, such that $\underline{P_1} \leq \underline{P_2} \leq \underline{P_3}$ then $d(\underline{P_1}, \underline{P_2}) \leq d(\underline{P_1}, \underline{P_3})$. This property is quite natural since it enables to use such distance for ranking the lower probabilities specificity-wise relative to a given lower probability. For instance, $d(\underline{P_1}, \underline{P_2}) \leq d(\underline{P_1}, \underline{P_3})$ means that, relatively to $\underline{P_1}$, we have that $\underline{P_2} \leq \underline{P_3}$, i.e. that $\underline{P_2}$ is more specific than $\underline{P_3}$. Note that the Hausdorff distance between sets of probabilities and thus between lower probabilities fulfills such natural property. It should be interesting to study other distances between lower probabilities respecting this property.

Let d_0 be this distance between any lower probability $\underline{P}$ of $\underline{\mathcal{P}}(\underline{P_X})$ and P_0: $d_0(\underline{P}) = d(P_0, \underline{P})$. We also define

$$d_{0X} = d_0(\underline{P_X}) = d(P_0, \underline{P_X})$$

as the distance between $\underline{P_X}$ and P_0.

Let α be the distance between any lower probability $\underline{P} \in \underline{\mathcal{P}}(\underline{P_X})$ and P_0 relative to the distance between $\underline{P_X}$ and P_0. α is defined by

$$\alpha(\underline{P}) = \frac{d_0(\underline{P})}{d_{0X}}.$$

This relative distance is such that $\alpha(\underline{P}) \in [0,1]$ for any lower probability $\underline{P} \in \underline{\mathcal{P}}(\underline{P_X})$ and $\alpha(P_0) = 0$ and $\alpha(\underline{P_X}) = 1$.

In other words, should we assume that P_0 exists and is known (which is not consistent with the Walley's behavioral imprecise probability framework), $\alpha(\underline{P})$ can be considered as a normalized index of non specificity (of imprecision) of $\underline{P}$.

Now, let us define, for a given imprecise SEUM problem (P), $\mathcal{C}$: the set of lower probabilities of $\underline{\mathcal{P}}(\underline{P_X})$ which make the problem comparable or decidable. Now we can propose a general unrealistic definition of the significance of an imprecise SEUM.

Definition 4.1 (Significance) *Let (P) be an imprecise SEUM problem: $\underline{P_X}$ is a lower probability on X defined on $\mathcal{X}$ and u is a utility function on $\mathcal{X}$.*

Let $\underline{P^*}$, be the least specific lower probability of $\underline{\mathcal{P}}(\underline{P_X})$, which makes (P) comparable or decidable. Then the significance of (P) is given by

$$\mathcal{S}_{(P)} = \alpha(\underline{P^*}). \tag{9}$$

An alternative definition can be proposed:

$$\mathcal{S}_{(P)} = \max_{\underline{P} \in \mathcal{C}} \alpha(\underline{P}). \tag{10}$$

The interpretation we can propose to this index is as follows. Significance is the maximal degree of imprecision (of epistemic uncertainty) which allows comparability. For a lower prevision model with an imprecision higher than $\mathcal{S}_{(P)}$, the problem is still incomparable, but for a lower prevision model with an imprecision lower than $\mathcal{S}_{(P)}$, the problem i scomparable or decidable.

Let us retake the example presented in the first paragraph of Section 4.1. We can rephrase it that way: the highest imprecision which makes the problem comparable or decidable is bigger for (P1) than for (P2) thus $\mathcal{S}_{(P1)} \geq \mathcal{S}_{(P2)}$.

Finally, if we are facing a problem (P) which is comparable regarding the provided information $\underline{P_X}$, then the significance of this problem should be the highest, i.e. should be equal to 1. With our definition, $\mathcal{S}_{(P)} = 1$, since $\mathcal{C} = \underline{\mathcal{P}}(\underline{P_X})$ and $\alpha(\underline{P_X}) = 1$. On the contrary, if we are facing a problem (P') which is comparable or decidable only when uncertainty is reduced to a linear probability, then the significance of this problem should be the lowest, i.e. should be equal to 0. With our definition, $\mathcal{S}_{(P')} = 0$, since $\mathcal{C} = \{P_0\}$ and $\alpha(P_0) = 0$.

4.2 Significance index : an applicable definition

Definition 4.1 of the significance is not applicable because P_0 is unknown (even if it exists). We propose in this section a pragmatic significance index for the imprecise SEUM approach with the interval dominance rule.

In Definition 4.1, the imprecision reduction is performed directly on the lower probability $\underline{P_X}$ modeling the uncertainty about X. In the applicable definition, we propose to perform this imprecision reduction directly on the interval utility expectations $[\underline{E}_X(u), \overline{E}_X(u)]$ associated to every act X.

This applicable definition is inspired from the Hurwicz risk aversion degree (6). In our case we define the relative imprecision index ρ of the imprecise expected utility as:

$$\begin{cases} \underline{E}^{\rho}_X(u) = (1-\rho)E_0(u) + \rho\underline{E}_X(u), \\ \overline{E}^{\rho}_X(u) = (1-\rho)E_0(u) + \rho\overline{E}_X(u), \end{cases} \tag{11}$$

where $E_0(u) = \frac{\underline{E}_X(u)+\overline{E}_X(u)}{2}$ is the middle of $[\underline{E}_X(u), \overline{E}_X(u)]$.

ρ is an index of imprecision relative to the imprecision of $\underline{E_X}$. We interpret $[\underline{E}^{\rho}_X(u), \overline{E}^{\rho}_X(u)]$ as the representation of $[\underline{E}_X(u), \overline{E}_X(u)]$ of relative imprecision ρ. Indeed, for a relative imprecision $\rho = 0$, $[\underline{E}^0_X(u), \overline{E}^0_X(u)] = \{E_0(u)\}$ and for a relative imprecision $\rho = 1$, $[\underline{E}^1_X(u), \overline{E}^1_X(u)] = [\underline{E}_X(u), \overline{E}_X(u)]$. In other words, $[\underline{E}^{\rho}_X(u), \overline{E}^{\rho}_X(u)]$ goes from $\{E_0(u)\}$ to $[\underline{E}_X(u), \overline{E}_X(u)]$ when ρ goes from 0 to 1.

We can thus define a new decision rule which is called the ρ-imprecise decision rule and which is the interval dominance decision applied to the ρ-imprecise interval $[\underline{E}^{\rho}_X(u), \overline{E}^{\rho}_X(u)]$:

$$X \succeq_{\rho} Y \text{ iff } \underline{E}^{\rho}_X(u) \geq \overline{E}^{\rho}_Y(u). \tag{12}$$

The proposed definition of the applicable significance is thus a direct application of Definition 4.1.

Definition 4.2 (Applicable Significance) *Let (P) be an imprecise SEUM problem: $\underline{P_X}$ is a lower probability on X defined on $\mathcal{X}$ and u is a utility* *function on $\mathcal{X}$. Let ρ^*, be the highest relative imprecision index, such that $\succeq_{\rho^*}$ becomes complete or makes (P) decidable. Then*

$$\mathcal{S}_{(P)} = \rho^*. \tag{13}$$

Compared to Definition 4.1, this solution, Definition 4.2 is feasible. Anyway, artificially increasing the informativity of an imprecise probability model is the only possible way to propose an applicable significance index. Indeed the informativity of any model can only be measured if we know the underlying true model, which is impossible or artificially possible.

Now let us illustrate this notion of significance on a toy example taken from [8].

Example

Assume that an individual with initial wealth ω is facing a risk of loss ℓ. There is uncertainty about the fact that this loss occurs or not. Each act X has two possible rewards: one if loss occurs, denoted by x_ℓ, and one if loss does not occur, denoted by $x_{\overline{\ell}}$.

One possible act for the individual would be not to buy any insurance. This can be represented by the act $X = (x_\ell, x_{\overline{\ell}}) = (\omega - \ell, \omega)$. Another act would be to buy full coverage at a premium π, yielding $Y = (y_\ell, y_{\overline{\ell}}) = (\omega - \pi, \omega - \pi)$. A third possible act would be to buy partial coverage at a premium π', yielding $Z = (z_\ell, z_{\overline{\ell}}) = (\omega - \ell + I - \pi', \omega - \pi')$ where I is the indemnity paid in case of damage.

We assume that the individual wealth is $\omega = \frac{3}{2}$, that its potential loss $\ell = \frac{1}{2}$, that the respective full and partial coverage are given by $\pi = \frac{1}{5}$ and $\pi' = \frac{1}{10}$ and that the indemnity is $I = \frac{1}{3}$. We also assume that the imprecise probability of loss is given by $\{(p, 1-p) : \text{ for } p \in [\frac{1}{3}, \frac{1}{2}]\}$. The utility function is $u(x) = x$ for $x \in \mathcal{X}$. Under such assumptions, the compared imprecise expectations are given by:

- $[\underline{E}_X(u), \overline{E}_X(u)] = [1.25, 1.33]$,
- $[\underline{E}_Y(u), \overline{E}_Y(u)] = \{1.3\}$,
- $[\underline{E}_Z(u), \overline{E}_Z(u)] = [1.288, 1.3166]$.

We can compute easily that the significance of this DP is 0.2 and that the associated optimal decision is Z. Indeed, for decreasing relative imprecision indices, Table 1 shows the evolution of the imprecise utility expectation when we artificially decrease imprecision.

We can see from Table 1 that the DP becomes decidable and completely ranked for $\rho = 0.2$ and that the associated optimal choice is Z. In other words, with a significance of 0.2 the individual should choose to buy the proposed partial coverage π'.

ρ	$[\underline{\overline{E}}_X(u)]$	$[\underline{\overline{E}}_Y(u)]$	$[\underline{\overline{E}}_Z(u)]$
0.3	[1.278, 1.302]	1.3	[1.2986 , 1.3069]
0.2	[1,282 , 1,298]	1.3	[1.3, 1.3056]
0.1	[1.286, 1.294]	1.3	[1.3014, 1.3042]

Table 1: Imprecise utility expectations for various relative imprecision

End of Example

It should be noted that the aim of our proposal is not to provide an optimal decision. Actually, with Definition 4.2, the optimal choice(s) is (are) always the optimal choice(s) for the center of the utility expectation intervals associated to the acts. The Hurwicz criterion with a risk aversion of $r = \frac{1}{2}$ gives the same result, i.e. the same optimal choice(s). However, our approach aims at providing a significance index which is not done with the Hurwicz criterion or any other decision rule. The proposed simplified and pragmatic definition is a simple way to explain and introduce the notions of interest in this paper. But more sensible and complex definitions of significance should be proposed in later works.

5 Conclusion

This article is a discussion paper. Its aim is mainly to define a new notion of significance of decision problem under uncertainty and to discuss its foundations. The idea is that if a decision problem is not comparable then the quantity of information which is required to make it comparable or decidable is directly linked to its significance. A theoretical definition of a significance index is proposed. This definition is constructed with the true underlying model of an imprecise probability and is thus unrealistic. A second artificial but pragmatical index is proposed. This index is very simple and inspired from the way the Hurwizc decision criterion is constructed.

The next step is to derive explicit formulations of other significance indices based on pragmatic constructions similar or different than the one found in Section 4.2 and obtained for different imprecise probability models. For instance with any submodel of the convex Choquet capacities, the imprecise expectation is explicitly computed with the Choquet integral. Thus explicit formulations of significance indices are possible. Experimental studies are now to be proposed.

Acknowledgment

This work has been supported by INSERM through Plan Cancer Program 2009-2013

References

[1] Dempster, A.: Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics 38, 325–339 (1967)

[2] Denneberg, D.: Non Additive Measure and Integral. Kluwer Academic Publishers, Dordrecht (1994)

[3] Destercke, S., Dubois, D., Chojnacki, E.: Unifying practical uncertainty representations: I. generalized p-boxes. International Journal of Approximate Reasoning 49(3), 649–663 (2008)

[4] Destercke, S., Dubois, D., Chojnacki, E.: Unifying practical uncertainty representations: Ii. clouds. International Journal of Approximate Reasoning 49(3), 664–677 (2008)

[5] Dubois, D., Fargier, H., Perny, P.: Qualitative decision theory with preference relations and comparative uncertainty: An axiomatic approach. Artificial Intelligence 148(1-2), 219 – 260 (2003)

[6] Dubois, D., Prade, H.: Consonant approximations of belief functions. International Journal of Approximate Reasoning 4(5-6), 419–449 (1990)

[7] Ellsberg, D.: Risk, ambiguity, and the savage axioms. The Quarterly Journal of Economics 75, 643–669 (1961)

[8] Etner, J., Jeleva, M., Tallon, J.M.: Decision theory under ambiguity. Journal of Economic Surveys 26(2), 234–270 (2012)

[9] Gajdos, T., Vergnaud, J.C.: Decisions with conflicting and imprecise information. Social Choice and Welfare pp. 1–26 (2009)

[10] Jaffray, J., Wakker, P.: Decision making with belief functions: Compatibility and incompatibility with the sure-thing principle. Journal of Risk and Uncertainty 7, 255–271 (1993)

[11] Jaffray, J.: Utility theory for belief functions. Operations Research Letters 8, 107–112 (1989)

[12] Levi, I.: The Enterprise of Knowledge. MIT Press, Cambridge MA (1980)

[13] Marinacci, M.: Probabilistic sophistication and multiple priors. Information Science 1, 1–3 (2001)

[14] Savage, L.: The foundations of statistics. John Wiley & Sons (1954)

[15] Schmeidler, D.: Subjective probability and expected utility without additivity. Econometrica 57, 571–587 (1989)

[16] Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)

[17] Troffaes, M.: Decision making under uncertainty using imprecise probabilities. International Journal of Approximate Reasoning 45(1), 17–29 (2007)

[18] Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, New York (1991)

[19] Walley, P.: Towards a unified theory of imprecise probability. International Journal of Approximate Reasoning 24(2-3), 125–148 (2000)

New Prior Near-ignorance Models on the Simplex

Francesca Mangili and Alessio Benavoli
IDSIA, Galleria 2, CH-6928 Manno (Lugano), Switzerland
email: francesca@idsia.ch, alessio@idsia.ch

Abstract

The aim of this paper is to derive new near-ignorance models on the probability simplex, which do not directly involve the Dirichlet distribution and, thus, that are alternative to the Imprecise Dirichlet Model. We focus our investigation to a particular class of distributions on the simplex which is known as the class of Normalized Infinitely Divisible distributions; it includes the Dirichlet distribution as a particular case. Starting from three members of this class, which admit a closed-form expression for the probability density function, we derive three new near-ignorance prior models on the simplex, we analyse their properties and compare them with the Imprecise Dirichlet Model.

Keywords. Prior near-ignorance, Normalized Infinitely Divisible distribution, Imprecise Dirichlet Model.

1 Introduction

The *Imprecise Dirichlet Model* (IDM) has been introduced by Walley [1] to draw inferences about the probability distribution of a categorical variable. Consider a variable Z taking values on a finite set $\mathscr{Z}$ of cardinality m and assume that we have a sample of size N of independent and identically distributed outcomes of Z. Our aim is to estimate the probabilities P_i for $i = 1, \ldots, m$, that is the probability that Z takes the i-th value. In other words, we want to estimate a vector on the m-dimensional simplex:

$$\Delta_m(p) = \left\{ (p_1, \ldots, p_m) : p_i \geq 0, \quad \sum_{j=1}^{m} p_j = 1 \right\}. \qquad (1)$$

A Bayesian approach consists in assuming a prior Dirichlet distribution for the vector of variables $(P_1, \ldots, P_m)$, and then taking the posterior expectation of P_i given the sample. The Dirichlet distribution depends on the parameters s, a positive real value, and $(t_1, \ldots, t_m)$, a vector of positive real numbers which satisfy $\sum_{i=1}^{m} t_i = 1$. In case of lack of prior information, an issue in Bayesian analysis is how to choose these parameters to reflect this condition of prior ignorance. To address this issue, Walley has proposed IDM, which considers the set of all possible Dirichlet distributions, with fixed value for s, in the simplex $\Delta_m(p)$:

$$\mathscr{M} = \left\{ \frac{\Gamma(s)}{\prod_{i=1}^{m} \Gamma(st_i)} \prod_{i=1}^{m} p_i^{st_i - 1} : \, t_i > 0, \, \sum_{i=1}^{m} t_i = 1 \right\}, \qquad (2)$$

where $\Gamma(\cdot)$ is the Gamma function and $s > 0$ is the prior strength. For a fixed value s, this is the set of all Dirichlet distributions obtained by letting $(t_1, \ldots, t_m)$ to freely vary in $\Delta_m(t)$. Walley has proven that IDM is a model of prior "near-ignorance" in the sense that it provides vacuous prior inferences for the probabilities $P(Z = z_i)$ for $i = 1, \ldots, m$. In fact, since $P(Z = z_i) = E[P_i] = t_i$, and t_i is free to vary in $\Delta_m(t)$, this means that $P(Z = z_i)$ is vacuous, which implies:

$$\underline{E}[P_i] = 0, \ \overline{E}[P_i] = 1, \qquad (3)$$

where $\underline{E}, \overline{E}$ denote the lower and respectively, upper expectations. This means that the prior mean of P_i is unknown, but this does not hold for all functions of $P_1, \ldots, P_m$, for example

$$\underline{E}[P_i P_j] = 0, \ \overline{E}[P_i P_j] = \frac{1}{4} \frac{s}{s+1}, \qquad (4)$$

while a prior ignorance model for $P_i P_j$ would have upper expectation equal to $1/4$. Walley has shown that prior ignorance can only be imposed on a subset of the possible functions of $P_1, \ldots, P_m$ otherwise it produces vacuous posterior inferences [2, Ch. 5], which means that we do not learn from data (for this reason the model is called near-ignorance). However, near-ignorance guarantees prior ignorance for many of the inferences of interest in statistical analysis and, at the same time, allows to learn from data and converges to the "truth" (be consistent in the terminology of Bayesian asymptotic analysis) at the increase of the number of observations.[1] Walley [3] has also proven that, besides near-ignorance, IDM satisfies several other desiderata for a model of prior ignorance.

Symmetry principle (SP): if we are ignorant a priori about P_i, then we have no reason to favour one possible outcome

[1] A full model of prior ignorance cannot learn from data [3].

of Z to another, and therefore our probability model on Z should be symmetric.
Embedding principle (EP): for each event $A \subseteq \mathscr{X}$, the probability assigned to A should not depend on the possibility space $\mathscr{X}$ in which A is embedded. In particular, the probability assigned a priori to the event A should be invariant w.r.t. refinements and coarsenings of $\mathscr{X}$.
Representation Invariance Principle (RIP): for each event $A \subseteq \mathscr{X}$, the posterior inferences of A should be invariant w.r.t. refinements and coarsenings of $\mathscr{X}$.
Learning/Convergence Principle (LCP): for each event $A \subseteq \mathscr{X}$, there exists $\overline{N}$ such that for $N \geq \overline{N}$ the posterior inferences about A should not be vacuous. Moreover, for $N \to \infty$, the posterior inferences should converge to $\lim_{N\to\infty} n_A/N$, where n_A is the number of occurrences of the event A in the N observations [4].[2]

Near-ignorance, SP and EP hold for any model on the simplex which satisfies $E[P_i] = t_i$ for $i = 1,\dots,m$ with $(t_1,\dots,t_m)$ are free to vary in $\Delta_m(t)$ [3],[3] while RIP holds if the lower and upper posterior expectations of the event A do not depend on the number of categories m [3]. Observe that IDM satisfies all the above principles and also the coherence (CP) and likelihood (LP) principles [1], [7]. Another important characteristic of the IDM is its computational tractability, which follows by the conjugacy between the categorical and Dirichlet distributions for i.i.d. observations. For instance the prior and posterior mean of P_i relative to a categorical-Dirichlet conjugate model are:

$$E[P_i] = t_i, \quad E[P_i|n_1,\dots,n_m] = \frac{n_i + st_i}{N+s}, \tag{5}$$

where n_i is the number of observations for the i-th category and, thus, $N = \sum_{i=1}^m n_i$. Hence, the lower and upper posterior mean derived from IDM can simply be obtained by

$$\begin{array}{rcccl} \frac{n_i+st_i}{N+s} & \overset{t_i\to 0}{=} & \frac{n_i}{N+s} & = & \underline{E}[P_i|n_1,\dots,n_m], \\ \frac{n_i+st_i}{N+s} & \overset{t_i\to 1}{=} & \frac{n_i+s}{N+s} & = & \overline{E}[P_i|n_1,\dots,n_m]. \end{array} \tag{6}$$

There are other models that involve the Dirichlet distribution which satisfy (some of) the above desiderata. For instance, a model which satisfies SP and RIP is defined by Walley in [1, Sec. 2.9] by further constraining the parameters $t_1,\dots,t_m$ of IDM.

The question we aim to address in this paper is to study if there are other models that satisfy the above desiderata, in particular near-ignorance, that are not directly derived from a Dirichlet distribution. We focus our investigation to a particular class of distributions on the simplex which is known as the class of *Normalized Infinitely Divisible* (NID) distributions [8]; it includes the Dirichlet distribution as a particular case. For this class, it is possible to derive general distributional properties and general moment formulae, briefly introduced in Section 2.1, which in some special cases, lead to explicit closed-form expressions [8]. In Sections 3 to 5, starting from three members of this class, which admit a closed-form expression for the prior density, we derive three new near-ignorance prior models on the simplex. We will show that all these new near-ignorance prior models satisfy EP, SP, LCP, CP and LP, and that, although they are not conjugate with the categorical distribution, the posterior inferences drawn from these models are still computationally tractable. In particular, we will show that for two of these models the lower and upper expectations of the P_i can be computed by means of simple algebraic expressions, while for one of these models, the lower and upper expectations can be computed efficiently by solving numerically one-dimensional integrals. Furthermore, we will show that one of this models also satisfies RIP and, given s, always provides inferences that are more conservative than those of IDM. On the other hand, the other two models, which do not satisfy RIP, have a posterior imprecision which increases linearly or almost linearly with the number of observed categories.

2 NID class

The aim of this section is to discuss some general properties that allow to characterize all infinitely divisible distributions. The most important of these properties follows from the Lévy-Khintchine representation theorem. Since the NID distributions studied in this paper admit a PDF, the use we will make of this general properties is limited to the derivation in eq. (8) of the moments of P_i; indeed, the reader that is not interested in a general description of the class of NID distributions can move on to Section 2.1.

Consider a collection of variables $X_1,\dots,X_m$ which are assumed to be independent and distributed according to a Gamma distribution with parameters $(\alpha_1,1),\dots,(\alpha_m,1)$, where $(\alpha_i,1)$ are respectively the shape and scale parameter of the Gamma distribution for the variable X_i. Define $W = X_1 + \cdots + X_m$ and $P_i = X_i/W$ for $i = 1,\dots,m$, then it can be shown that

$$(P_1,\dots,P_m) \sim Dir(\alpha_1,\dots,\alpha_m),$$

where $Dir(\alpha_1,\dots,\alpha_m)$ denotes the Dirichlet distribution with parameters $\alpha_1,\dots,\alpha_m$. In other terms, the Dirichlet distribution can be defined via normalization from a set of Gamma distributed independent variable divided by their sum. The Gamma distribution is infinitely divisible (ID), i.e for any $n \in \mathbb{N}$ and given variable X Gamma-distributed,

[2] We are assuming that the likelihood is categorical. For this reason, this is a weaker principle than the Strong Learning Principle proposed by Moral [5] which holds irrespectively from the type of the likelihood distribution. Unfortunately, the strong learning principle is not compatible with near-ignorance [5], [6].

[3] Since $P(Z = Z_i) = E[P_i] = t_i$, this implies that the lower and upper probabilities of the event A do not depend on $\mathscr{X}$.

there exists a collection of i.i.d. variables $Y_1, \dots, Y_n$ such that $X \overset{d}{=} Y_1 + \cdots + Y_n$ or, alternatively, the variable X can be separated into the sum of an arbitrary number of i.i.d. variables.

Consider then a collection of positive variables $X_1, \dots, X_m$ which are assumed to be independent and distributed according to, not necessarily coinciding, ID distributions [8]. According to the Lévy-Khintchine representation theorem [9, Ch. 16] for ID distributions, the moment generating function of X_i can be expressed by:

$$\psi_i(u) := E[e^{-uX_i}] = \exp\left(- \int_0^\infty (1 - e^{-ux}) \nu_i(dx) \right) \quad u \geq 0, \tag{7}$$

where E denotes the expectation w.r.t. the Lévy measure ν_i, which is any nonnegative Borel measure on $\mathbb{R}^+$ satisfying the condition $\int_0^\infty \min(1,x)\nu_i(dx) < \infty$, which completely characterizes the distribution of the random variable X_i, for each $i = 1, ..., m$.

Example 1. Consider the case where X is Gamma-distributed with parameters $(\alpha, 1)$, *in this case* $\nu(dx) = \alpha x^{-1} e^{-x} dx$, $E[e^{-uX_i}] = (u+1)^{-\alpha}$ *and, thus,*

$$E[X^n] = (-1)^n \frac{d^n}{du^n}(u+1)^{-\alpha}\Big|_{u=0},$$

which, for $n = 1, 2, \dots$ *gives the non-central moments of a Gamma distribution with parameters* $(\alpha, 1)$. *Thus,* $\nu(dx)$ *characterizes completely the distribution of* X. ■

Then, via the normalization approach $P_i = X_i/W$ for $i = 1, \dots, m$ with $W = X_1 + \cdots + X_m$, we can define a wide class of distributions for the vector $(P_1, \dots, P_m)$. In particular, as it holds for the distribution of X_i, each of these distributions for $(P_1, \dots, P_m)$ is completely characterized by the corresponding collection of Lévy measures $\nu_1, \dots, \nu_m$. This class of distributions is termed the normalized ID (NID) distributions. For this class, it is possible to derive general distributional properties and general moment formulae, which in some special cases, lead to explicit closed-form expressions. For instance, the mean of P_i can be determined:

$$E[P_i] = \int_0^\infty \left(\frac{d}{du} \psi_j(u) \right) e^{-\sum_{i=1}^m \psi_j(u)} du; \tag{8}$$

the proof can be found in [8, Prop. 2]. The class of NID distributions represents a natural extension of the Dirichlet distribution, which can be recovered as special case of NID distributions by assuming the collection of Lévy measures to be $\nu_i(dx) = \alpha x^{-1} e^{-x} dx$ for $i = 1, \dots, m$. The computations simplify in case X_i admits a probability density function (PDF) with respect to the Lebesgue measure on $\mathbb{R}^+$.

2.1 NID with a PDF

Assume that the PDF of X_i, denoted by f_i, admits a closed-form expression for every $i = 1, \dots, m$ and define $W = X_1 + \cdots + X_m$, $P_i = X_i/W$ for $i = 1, \dots, m$. Then, the PDF of the (NID) vector $(P_1, \dots, P_m)$ is:

$$g(p_1, \dots, p_{m-1}) = \int_0^\infty \prod_{i=1}^{m-1} f_i(p_i w) f_m \left(w - \sum_{i=1}^{m-1} p_i w \right) w^{m-1} dw. \tag{9}$$

where we have exploited the relationship $p_m = 1 - \sum_{i=1}^{m-1} p_i$. This can be proven by applying the change of variable theorem for PDFs.

Example 2. Consider again the case in which X_i *is Gamma-distributed with parameters* $(\alpha_i, 1)$, *then* $f(x_i) \propto x_i^{\alpha_i - 1} \exp(-x_i)$, *and, thus, from (9), neglecting the normalization constant, one derives:*

$$\begin{aligned} &\int_0^\infty \prod_{i=1}^{m-1} (p_i w)^{\alpha_i - 1} \exp(-p_i w) \\ &\cdot (w - w \textstyle\sum p_i)^{\alpha_m - 1} \exp(-(w - w \sum_{i=1}^{m-1} p_i)) w^{m-1} dw \\ &\propto p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} \cdots (1 - \textstyle\sum_{i=1}^{m-1} p_i)^{\alpha_m - 1}. \end{aligned} \tag{10}$$

■

Besides the Dirichlet distribution, further examples of NID distributions, which admits a PDF are the normalized inverse-Gaussian distribution [10], the normalized 1/2-stable [11, 8] and a NID distribution based on two degrees of freedom (2dof) Gamma variables [8, Sec. 3.5]. In the next section, we derive new prior near-ignorance models based on these three NID distributions and analyse their properties.

3 NID distribution based on 2dof Gammas

Consider the case in which $X_1, \dots, X_m$ have distribution $X_i \sim Ga(\alpha_i; \beta_i)$ (Gamma distributed) for $i = 1, \dots, m$ [8, Sec. 3.5]. The PDF of the NID vector $(P_1, \dots, P_m)$ is easily obtained by applying (9) leading to

$$\begin{aligned} &g(p_1, \dots, p_{m-1}) = \\ &\Gamma(s) \prod_{i=1}^m \frac{\beta_i}{\Gamma(a_i)} \prod_{i=1}^{m-1} p_i^{\alpha_i - 1} \left(1 - \sum_{j=1}^{m-1} p_j \right)^{\alpha_m - 1} \\ &\cdot \left(\sum_{i=1}^{m-1} \beta_i p_i + \beta_m \left(1 - \sum_{j=1}^{m-1} p_j \right) \right)^{-s} \end{aligned} \tag{11}$$

where $s = \sum_{i=1}^m \alpha_i$. Note that for $\beta = \beta_i$ for $i = 1, \dots, m$ we are back to the Dirichlet distribution. The r-th non-central moment of (11) is given by [8, Sec. 3.5]:

$$E[P_j^r] = \frac{\Gamma(\alpha_j+r)\prod_{i=1}^m \beta_i^{\alpha_i}}{\Gamma(\alpha_j)\Gamma(r)} \int_0^\infty \frac{u^{r-1}}{(\beta_j+u)^r \prod\limits_{i=1}^m (\beta_i+u)^{\alpha_i}} du. \tag{12}$$

To model prior near-ignorance, we consider the set of PDFs in (11) obtained by taking

$$\begin{array}{l} \alpha_i = st_i, \ \ \beta_i = t_i' \ \text{ for } i=1,\dots,m \text{ with} \\ (t_1,\dots,t_m) \in \Delta_m, \ \ (t_1',\dots,t_m') \in \Delta_m; \end{array} \tag{13}$$

we call this model *Normalized 2dof Gamma* (N2dG).[4]

Proposition 1. *N2dG model satisfies:*

$$\begin{array}{rclrcl} \underline{E}[P_i^r] & = & 0, & \overline{E}[P_i^r] & = & 1 \\ \underline{E}[P_iP_j] & = & 0, & \overline{E}[P_iP_j] & \geq & \frac{1}{4}\frac{s}{s+1}. \end{array} \tag{14}$$

for any i,j *and* $r = 1,2,\dots$. ■

The lower and upper expectations in (14) can be derived by noticing that for $t_i' = 1/m$ for $i=1,\dots,m$ the set of priors defined by (11) and (13) reduces to IDM. Thus, (14) follows by (3)–(4). We have not be able to compute the exact value of $\overline{E}[P_iP_j]$, our conjecture is that $\frac{1}{4} > \overline{E}[P_iP_j] > \frac{1}{4}\frac{s}{s+1}$. Consider now the set of posteriors obtained by combining via Bayes' rule the likelihood relative to the sequence of counts $(n_1,\dots,n_m)$, i.e.,

$$L(n_1,\dots,n_m|p_1,\dots,p_{m-1}) = p_1^{n_1}p_2^{n_2}\cdots\left(1-\sum_{i=1}^{m-1}p_i\right)^{n_m}, \tag{15}$$

and the set of priors defined by (11) and (13). From this set of posteriors, we can compute lower and upper posterior expectations of P_i for $i=1,\dots,m$.

Theorem 1. *The lower and upper posterior expectations of* P_i *are:*

$$\begin{array}{rcl} \underline{E}[P_i|n_1,\dots,n_m] & = & \max\left(0, \frac{n_i-s}{N}\right), \\ \overline{E}[P_i|n_1,\dots,n_m] & = & \min\left(1, \frac{n_i+s}{N}\right), \end{array} \tag{16}$$

for any $i=1,\dots,m$. ■

The proof can be found in Appendix. Observe that N2dG model satisfies near-ignorance, SP and EP; this follows by the first row in (14) by using the same arguments as for IDM. It also satisfies LP and CP; coherence follows by [2, Th. 7.8.1]. Notice that the prior lower and upper expectations do not depend on the number of categories m and, thus, N2dG model satisfies also RIP. Moreover, since $\underline{E}[P_i|n_1,\dots,n_m], \overline{E}[P_i|n_1,\dots,n_m] \to \frac{n_i}{N}$ for $N\to\infty$, it also satisfies LCP.

Corollary 1. *The lower and upper posterior expectations of* $\sum_{i\in J}P_i$*, where* J *denotes a subset of* $\{1,\dots,m\}$*, are:*

$$\begin{array}{rcl} \underline{E}[\sum_{i\in J}P_i|n_1,\dots,n_m] & = & \max\left(0, \frac{\sum_{i\in J}n_i-s}{N}\right), \\ \overline{E}[\sum_{i\in J}P_i|n_1,\dots,n_m] & = & \min\left(1, \frac{\sum_{i\in J}n_i+s}{N}\right). \end{array} \tag{17}$$

for any $i=1,\dots,m$. ■

The proof can be found in Appendix. By looking at (16)–(17), we can highlight the following difference w.r.t. IDM. The IDM lower probability for the second observation to be equal to the first, is $1/(1+s)$, i.e., $1/2$ for $s=1$. For N2dG with $s=1$, this lower probability is zero. Walley has shown that, in case $m=2$, IDM with $s=2$ encompasses all the Bayesian inferences derived from the Jeffreys ($s=1, t=0.5$), uniform ($s=2, t=0.5$) and Haldane ($s=0$) priors [3]. For N2dG, this is already true for $s=1$. Another difference with IDM, is that the lower and upper expectations derived in (16) are symmetric w.r.t. the sample mean n_i/N whenever $n_i-s\geq 0$ and $n_i+s\leq N$. Furthermore, the denominator in (16) depends only on N and not on s. Thus, for $n_i-s\geq 0$ and $n_i+s\leq N$, the imprecision $2s/N$ should not be interpreted as additional counts that are added to the observations but as a swing scenario in which s counts among the N are moved from a category to another. It should be pointed out that the lower and upper expectations in (16)–(17) coincide with those derived in [12, Sec. 5.2] for a near-ignorance model based on finitely additive priors obtained as limits of truncated exponential priors. Moreover, the inferences drawn from N2dG with $s=1$ coincide with those of the *Nonparametric Predictive Inference* model [13] in case all the categories have been observed at least once.

4 The normalized 1/2-stable distribution

Consider now the case where the ID variables $X_1,\dots,X_m$ have positive stable distribution $X_i{\sim}St(\gamma,\beta,\alpha_i,\mu)$ with characteristic exponent $\gamma>0$, skewness parameter $\beta=1$, scale parameter $\alpha_i>0$, and a location parameter $\mu=0$ [14]. Although, in general, the PDF of a stable distribution does not admit a closed-form expression, for this choice of parameters and $\gamma=1/2$, the PDF, hereafter referred to as 1/2-stable distribution, is given by:

$$f(x_i|\alpha_i) = \frac{\alpha_i}{(2\pi)^{1/2}}x_i^{-3/2}\exp\left(\frac{\alpha_i^2}{2x_i}\right), \quad x_i\in\mathbb{R}^+. \tag{18}$$

From (9) it follows that the NID vector $(P_1,\dots,P_m)$ arising from the normalization of the m 1/2-stable distributed variables $X_1,\dots,X_m$ has the *Normalized* $1/2-$*Stable distribution* (N1/2S) with PDF [15]:

$$g(p_1,\dots,p_{m-1}) = \frac{\Gamma(\frac{m}{2})\prod\limits_{i=1}^m\alpha_i}{\pi^{\frac{m}{2}}}\frac{\prod\limits_{i=1}^{m-1}p_i^{-\frac{3}{2}}(1-\sum\limits_{i=1}^{m-1}p_i)^{-\frac{3}{2}}}{[\mathscr{A}(p_1,\dots,p_{m-1})]^{\frac{m}{2}}}, \tag{19}$$

where $\mathscr{A}(p_1,\dots,p_{m-1}) = \sum_{i=1}^{m-1}\frac{\alpha_j^2}{p_i} + \frac{\alpha_m^2}{1-\sum_{i=1}^{m-1}p_i}$.

Although there is not a closed form expression for the normalized γ-Stable distribution (with $\gamma\neq 1/2$) we can compute its first moment for any γ by using (8) (a similar expression can be used to derive the mixed second moment

[4] From (11) it can be noticed that the constant $\sum_{i=1}^m\beta_i$ simplifies a-posteriori, and thus w.l.o.g. we can take $\sum_{i=1}^m\beta_i = 1$.

[15]),

$$E[P_i] = \tfrac{\alpha_i}{s}, \quad E[P_iP_j] = \tfrac{\alpha_i\alpha_j}{s^2}(1-\gamma), \tag{20}$$

where $s = \sum_{j=1}^{m} \alpha_j$.
Starting from a normalized γ-Stable distribution, we can thus obtain a prior near-ignorance model by considering the set of distributions obtained by taking:

$$\begin{array}{c} \alpha_i = st_i, \text{ for } i = 1,2,\ldots,m \text{ with} \\ s > 0 \text{ and } (t_1,\ldots,t_m) \in \triangle_m. \end{array} \tag{21}$$

Proposition 2. *For the set of priors defined from a γ-Stable distribution with parameters varying as in (21), it holds:*

$$\begin{array}{rclrcl} \underline{E}[P_i^r] & = & 0, & \overline{E}[P_i^r] & = & 1 \\ \underline{E}[P_iP_j] & = & 0, & \overline{E}[P_iP_j] & = & \tfrac{1}{4}(1-\gamma). \end{array} \tag{22}$$

for any i, j and $r = 1,2,\ldots$. ■

This can simply be obtained by first observing that $\alpha_i/s = t_i$ and, thus, by minimizing and maximizing w.r.t. $t_1,\ldots,t_m$ the expectations in (20). In the following, we only focus on the case $\gamma = 1/2$ where a closed form for the PDF of the γ-Stable distribution exists.[5] For this case, it can be noticed that the value of the parameter s is irrelevant. In fact, from the expression of the N1/2S PDF in (19), it is evident that the parameter s simplifies a-posteriori, and thus it does not affect the inferences produced by the N1/2S priors.

By considering the likelihood model (15) and the set of N1/2S priors defined by (19)–(21), a-posteriori we can derive the following.

Theorem 2. *Given the sequence of counts $(n_1,\ldots,n_m)$, the lower and upper posterior expectation obtained from the N1/2S set of priors are:*

$$\begin{array}{rcl} \underline{E}[P_i|n_1,\ldots,n_m] & = & \max\left(0, \frac{n_i-1/2}{N}\right), \\ \overline{E}[P_i|n_1,\ldots,n_m] & = & \min\left(1, \frac{n_i+\hat{m}/2}{N}\right), \end{array} \tag{23}$$

for any $i = 1,\ldots,m$, where $\hat{m}$ is the number of categories $j \neq i$ such that $n_j > 0$. ■

The proof can be found in Appendix. Note that, as for the N2dG, the denominators in (23) do not depend on s. N1/2G model satisfies near-ignorance, SP and EP; this follows by the first row in (22) by using the same arguments as for IDM. It also satisfies LP and CP; coherence follows by [2, Th. 7.8.1]. Moreover, since $\underline{E}[P_i|n_1,\ldots,n_m], \overline{E}[P_i|n_1,\ldots,n_m] \to \frac{n_i}{N}$ for $N \to \infty$, it also satisfies LCP. Since the upper expectation in (23) depends on $\hat{m}$, the RIP principle is not satisfied by N1/2S . As a consequence, the uncertainty about the expected value of P_j increases with the number of observed categories.

5 The normalized inverse-Gaussian distribution

Consider now m ID variables $X_1,\ldots,X_m$ having inverse-Gaussian distribution $X_i \sim IG(\alpha_i,\gamma)$ with shape parameter $\alpha_i > 0$ and scale parameter $\gamma = 1$. Their PDF is given by:

$$f(x_i|\alpha_i) = \tfrac{\alpha_i}{(2\pi)^{1/2}} \exp\left[-\tfrac{1}{2}\left(\tfrac{\alpha_i^2}{x_i} + x_i\right) + \alpha_i\right], \quad x_i \in \mathbb{R}^+. \tag{24}$$

From (9), it follows that the NID vector $(P_1,\ldots,P_m)$ arising from the normalization of the variables $X_1,\ldots,X_m$ has the normalized inverse Gaussian distribution (NIG) with PDF [10]:

$$\begin{array}{l} g(p_1,\ldots,p_{m-1}) = \dfrac{\exp\left(\sum\limits_{i=1}^{m} \alpha_i\right) \prod\limits_{i=1}^{m} \alpha_i}{2^{m/2-1}\pi^{m/2}} \times \\ \times \dfrac{\mathrm{K}_{-m/2}[\mathscr{A}(p_1,\ldots,p_{m-1})^{1/2}]}{\prod\limits_{i=1}^{m-1} p_i^{3/2}(1-\sum\limits_{i=1}^{m-1} p_i)^{3/2}[\mathscr{A}(p_1,\ldots,p_{m-1})]^{m/4}}. \end{array} \tag{25}$$

where $\mathrm{K}_{-m/2}$ is the modified Bessel function of the second kind of order $-m/2$. The first and mixed second moments of the NIG distribution are:

$$E[P_i] = \tfrac{\alpha_i}{s}, \quad E[P_iP_j] = \tfrac{\alpha_i\alpha_j}{s^2}(1 - s^2\mathrm{e}^s\Gamma(-2,s)), \tag{26}$$

where $s = \sum_{i=1}^{m} \alpha_i$ and $\Gamma(a,x) = \int_x^\infty t^{a-1}\exp(-t)dt$ denotes the incomplete gamma function.[6] To model prior near-ignorance, let us consider the set of NIG distributions obtained by taking:

$$\begin{array}{c} \alpha_i = st_i, \text{ for } i = 1,2,\ldots,m \text{ with} \\ s > 0 \text{ and } (t_1,\ldots,t_m) \in \triangle_m. \end{array} \tag{27}$$

Proposition 3. *For the set of priors defined by (25) and (27), it holds:*

$$\begin{array}{ll} \underline{E}[P_i] = 0, & \overline{E}[P_i] = 1, \\ \underline{E}[P_iP_j] = 0, & \overline{E}[P_iP_j] = \tfrac{1}{4}(1 - s^2e^s\Gamma(-2,s)). \end{array} \tag{28}$$

for any i, j. ■

These properties follow from the same arguments used for Proposition 2. A-posteriori, given the observed likelihood (15), it is not possible to provide a closed form expression of the lower and upper posterior expectation of P_i, but it is possible to indicate for which values of $t_1,\ldots,t_m$ the upper and lower can be found and provide a simplified integral expression for them, in the case where all counts are positive.

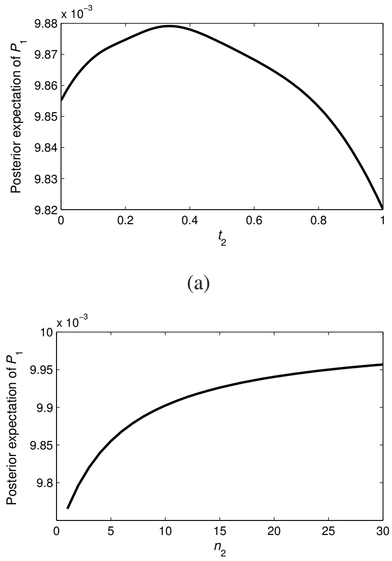
Conjecture 1. *Consider the set of priors defined by (25) and (27). Given the set of counts $(n_1,\ldots,n_m)$, the lower*

[5] For $\gamma = 1/2$, one has $\overline{E}[P_iP_j] = 1/8$ which coincided with the result obtained by IDM for $s = 1$.

[6] For $s = 1$, $\overline{E}[P_iP_j] = 0.175$ which is bigger than the result obtained by IDM for $s = 1$.

posterior expectation of P_i is found for $t_k = 1$, with $k = \arg\min_{j \neq i}(n_j)$, and the upper posterior expectation is found for $t_j = 1$. ■

Conjecture 1 is based on the experimental verification in several cases in which $n_j > 0$ holds for all $j \neq i$. However, we have not still been able neither to prove this conjecture nor to extend it to the cases in which $n_j = 0$ for some category $j \neq i$. As a verification of Conjecture 1 consider for instance Figure 1. Here we are computing the lower and upper posterior expectation of P_1. Figure 1.(a) shows that by taking only two parameters t_2 and t_3 different from 0, the minimum of $\underline{E}[P_i|n_1,\dots,n_m]$ is found for $t_2 = 1$ ($j = 2$ is in fact the category $j \neq 1$ with smaller number of observations). Figure 1.(b) shows that by taking $t_2 = 1$ the lower posterior expectation of P_i increases with n_2 (this means that the parameter t_k to be taken equal to 1 is the one corresponding to the category k with minimum number of counts n_k).



(a)

(b)

Figure 1: Posterior expectation of P_1 when $m = 5$, $n_1 = 1$, $N = 50$, $s = 1$ and (a) $n_2 = 3$ and $n_3 = 5$, $t_3 = 1 - t_2$ and t_2 spans the interval $[0,1]$ or (b) $t_2 = 1$, and n_2 ranges from 1 to 30.

Theorem 3. *Given the NIG set of priors defined by (25) and (27) and the set of counts $(n_1,\dots,n_m)$, with $n_j > 0$ for $j = 1,\dots,m$, the lower and upper posterior expectations of P_i for $t_k = 1$, with $k = \arg\min_{j \neq i}(n_j)$, and for $t_i = 1$ are, respectively,*

$$\underline{E}[P_i|n_1,\dots,n_m] = \frac{n_i - \frac{1}{2}}{N - n_k - \frac{1}{2}(m-1)} \times \\ \times \frac{\int_0^1 p_k^{n_k + \frac{m-6}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_k}}\right)(1-p_k)^{N-n_k-\frac{m-1}{2}}}{\int_0^1 p_k^{n_k + \frac{m-6}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_k}}\right)(1-p_k)^{N-n_k-\frac{m+1}{2}}}, \tag{29}$$

$$\overline{E}[P_i|n_1,\dots,n_m] = \\ = \frac{\int_0^1 p_i^{n_i + \frac{m-2}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_i}}\right)(1-p_i)^{N-n_i-\frac{m+1}{2}}}{\int_0^1 p_i^{n_i + \frac{m-6}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_i}}\right)(1-p_i)^{N-n_i-\frac{m+1}{2}}}.$$

■

The proof can be found in Appendix. Note that the lower posterior expectation in (29) depends on the minimum number of counts n_k observed for a value $z_k \neq z_i$ of Z. However, from Figure 1.(b), it appears that this dependence is weak and that it diminishes at the increasing of n_k. The NIG model satisfies near-ignorance, SP and EP; this follows by the first row in (28) by using the same arguments as for IDM. It also satisfies LP and CP, coherence follows by [2, Th. 7.8.1]. From Conjecture 1 and Theorem 3 it follows that, if there is at least one count for each value of Z considered, the lower and upper posterior expectations of P_i are not vacuous. Furthermore, the lower and upper posterior expectations of P_i converge to $\lim_{N \to \infty} \frac{n_i}{N}$; this follows from (29) by noticing that for large N and $n_j > 0$ for $j = 1,\dots,m$ the lower and the upper concentrate on $\frac{n_i}{N}$. Thus LCP is also satisfied. Yet, since both the lower and upper posterior expectations in (29) increase with m, the RIP principle is not respected by this set of priors. Figure 2 shows the upper and lower expectation for set of IDM, N2dG, N1/2S and NIG prior distributions for different values of $\hat{m}$ and a given case study. For the NIG set of priors, it can be noticed that the variation of the lower posterior expectation with $\hat{m}$ is negligible. Furthermore, it can also be noticed that the upper of the N1/2S and NIG set of priors are quite similar and both increase as $\frac{\hat{m}}{2N}$.

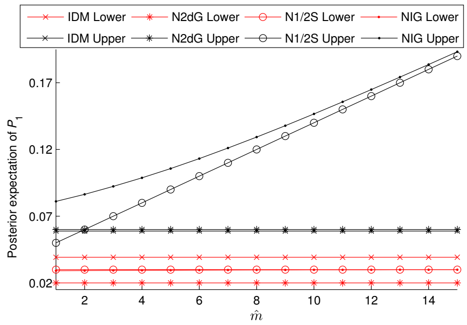


Figure 2: Posterior expectation of P_1 for $n_1 = 2$, $n_k = 3$ with $k = \arg_{j \neq 1} \min(n_j)$, $N = 50$, $s = 1$, and $\hat{m}$ ranging from 1 to 15.

	$\underline{P}(Z = red\|n_1,\dots,n_m)$			$\overline{P}(Z = red\|n_1,\dots,n_m)$		
	$\mathscr{Z}_1$	$\mathscr{Z}_2$	$\mathscr{Z}_3$	$\mathscr{Z}_1$	$\mathscr{Z}_2$	$\mathscr{Z}_3$
IDM	0.222	0.222	0.222	0.333	0.333	0.333
N2dG	0.125	0.125	0.125	0.375	0.375	0.375
N1/2S	0.188	0.188	0.125	0.438	0.313	0.313
NIG	0.176	0.178	0.120	0.489	0.371	0.368

Table 1: Upper and lower probabilities of drawing a red marble for different choices of $\mathscr{Z}$ and sets of priors ($s = 1$).

	$\mathscr{Z}_1$	$\mathscr{Z}_2$	$\mathscr{Z}_3$
IDM	[0.032; 0.681]	[0.032; 0.681]	[0.032; 0.681]
N2dG	[0.004; 0.710]	[0.004; 0.710]	[0.004; 0.710]
N1/2S	[0.016; 0.766]	[0.016; 0.648]	[0.004; 0.648]
NIG	[0.015; 0.778]	[0.015; 0.685]	[0.004; 0.683]

Table 2: 95% credible intervals for P_1.

6 Examples of inferences about a bag of marbles

To illustrate the difference between the three sets of priors proposed in this work and to compare them with the IDM, let us consider a bag of marbles containing coloured marbles of an unknown number of different colours [1]. Each colour represents a category z_i. Suppose we draw a sequence of $N = 8$ marbles 3 of which are blue, 1 green, 2 yellow, 1 light red, and 1 dark red. We consider three different possibility spaces $\mathscr{Z}_1 = \{red, blue, green, yellow\}$, $\mathscr{Z}_2 = \{red, all\ other\ colors\}$, $\mathscr{Z}_3 = \{light\ red, dark\ red, all\ other\ colors\}$. Tables 1 and 2 show, respectively, the upper and lower probabilities, $\underline{P}(Z = red|n_1,\dots,n_m)$ and $\overline{P}(Z = red|n_1,\dots,n_m)$, of drawing a red marble at the next trial and its 95% credible interval for the different possibility spaces $\mathscr{Z}$, and sets of priors.

Notice that the uncertainty of the estimates provided by the three sets of priors proposed in this paper is always larger than that of the IDM with $s = 1$. As expected, since the RIP principle is not respected by the N1/2S and the NIG sets of priors, the estimates provided by them depends on the possibility space $\mathscr{Z}$ adopted: their uncertainty increases with the number of categories in $\mathscr{Z}$. This dependence could appear unjustified in this example, since the definition of the categories is rather arbitrary, so that it is desirable that inference do not depend on them. However, in a situation where the categories could be objectively defined, the fact that uncertainty increases with the number of category, can reflect the idea that the knowledge of a system after a number of trials N is as more precise as simpler is the system, i.e., in this case, as smaller is the number of categories. To show an example where this property may be valuable, consider the following situation: assume to draw N marbles from a closed marble bag and to ask yourself what is the probability of drawing from the bag a marble of a new colour, not yet observed in any of the N trials. Said $\hat{m}$ the number of different colours observed after N trials, this corresponds to finding the probability that the event of observing a marble in the category $z_{\hat{m}+1}$ occurs at the $(N+1)$-th trial. The lower posterior expectation of $P_{\hat{m}+1}$ is zero, since, by hypothesis $n_{\hat{m}+1} = 0$. The upper posterior expectation is shown in Table 3, in the limiting cases where the number of values observed in $N = 100$ and $N = 1000$ trials is $\hat{m} = 1$ (only one category has been observed) or $\hat{m} = N$ (a different category has been observed in each drawn). In the first case, one obtains upper probabilities of observing a marble in a new category which goes to zero for large N; this same result is obtained if $\hat{m} = N$ for the IDM and N2dG sets of priors, whereas for the N1/2S priors the upper probability remains constant regardless of the number of trials N and for the NIG priors it converges for large N to a value close to 0.6. The result provided by the N1/2S and NIG sets of priors in this second case seems more appropriate than that provided by IDM, according to which the probability of observing a new category at the $N+1$-th trials goes to zero, although a new category has been, indeed, observed at each drawn of the N trials. This means that for predictive models the dependency of the lower and upper posterior expectations to the number of observed categories can lead to more intuitive inferences than the one derived by IDM or its predictive form [16].

	IDM	N2dG	N1/2S	NIG
$\hat{m} = 1, N = 100$	0.0099	0.0100	0.0050	0.0321
$\hat{m} = N, N = 100$	0.0099	0.0100	0.5000	0.6008
$\hat{m} = 1, N = 1000$	0.0010	0.0010	0.0005	0.0066
$\hat{m} = N, N = 1000$	0.0010	0.0010	0.5000	0.6000

Table 3: Upper probability of observing a marble in the new category $z_{\hat{m}+1}$.

7 Differences with IDM

In this Section, we briefly summarize the differences between the new prior "near-ignorance" models proposed in this paper and IDM. A characteristic of IDM, which has been criticized, is that the lower probability for the second observation to be equal to the first is equal to $1/(1+s)$. The values $s = 1$ or $s = 2$ lead to high values for this lower probability. However, it seems reasonable to assume that the lower probability of observing twice the same category is significantly large than 0 only if we have a strong prior belief that the number of categories is low. Instead, under complete prior ignorance, we may not want to bet on a category after we have seen it only once, but we would preferably wait until we see it for the second time before starting betting on it. If, for example, the process were a random generator, the probability of observing twice the same outcome would be 0 (see also [1, pages 43-44] and [13] for further discussion on this point). For the N2dG model, we have already seen that if $s \geq 1$ this lower probability is equal to 0. More generally, the lower probability of observing a specific category after N trials is equal to 0 until we observe at least $s+1$ realizations in that category. In this view, the parameter s can be interpreted as a threshold on the number of observations in a given category below which we would never bet on it, regardless of the reward. Thus, the N2dG model satisfies RIP but is also able to account for our prior ignorance about the number of categories. For the N1/2S model, the lower probability for the second observation to be equal to the first is $1/2$, so equal to that of IDM for $s = 1$.[7]

Another weak point of IDM is that, after N observations, the upper probability of observing a new category goes to zero as $s/(s+N)$. This upper probability does not depend on how much variety there has been in the previous observations, i.e., the upper probability in case we have observed the same category in all the N previous observations or N different categories is the same. However, if N different categories have been observed in N trials we may not want to bet against seeing a new category at the next trial, regardless of the reward. In this case, we would like the upper probability of observing a new category to be equal to 1. This weak point is also discussed by Walley in [1, page 50]. The N2dG model gives in practice the same upper probability of IDM. For the N1/2S and the NIG sets of priors, we have seen that the upper probability of observing a new category depends on how much variety there has been in the previous observations. Consider for instance N1/2S, as it has been shown in Section 6, if we observe N different categories in all the previous observations this upper probability is equal to $1/2$, while if we observe the same category in all the N previous observations, this upper probability is $1/2N$. This difference between IDM, N2dG and N1/2S, NIG seems in this case be due to the RIP property. IDM and N2dG satisfy RIP, while N1/2S and NIG do not. It has already been argued in [13, Sec. 5] that the RIP principle is not always a desirable property. In this paper, the authors stress that from the perspective of interval probability theory, the difference between lower and upper probabilities should depend on the amount of information available and the data representation. We think that this is especially true for predictive models in which we have no prior evidence about the number of categories and the inferences should depend on the number of observed category. Notice that none of the three models proposed in this paper meet at the same time both the desiderata here addressed: a lower probability for the second observation to be equal to the first equal to 0 and an upper probability of observing a new category having observed N different categories in N previous trials equal to 1. In this view, it could be interesting to extend the N1/2S model by considering a stable prior distribution with values of the γ parameter different from $1/2$. This way, the upper and lower probabilities predicted by the model would depend on γ, so that it might be possible to find a value of it (probably $\gamma = 1$) for which both desiderata can be met at the same time. Clearly, this would require working with the moment generating function since the PDF of the stable distribution does no admit a closed-form expression. On the other hand, the RIP property seems to be desirable for a prior ignorance model. In objective Bayesian analysis, a common practice is to impose invariance principles to derive non-informative priors. In this respect, the fact that IDM and N2dG satisfy EP, SP and RIP, while the commonly used precise non-informative priors do not, is valuable. In [17], the authors show that IDM can be derived starting from general invariance principle, in particular exchangeability and representation insensitivity (which is similar to RIP). This result reinforces the importance of IDM as a model of prior ignorance. In [17], the authors conclude the papers listing several open questions about representation insensitivity for predictive systems. One of this question was if there exist other models which satisfy RIP besides IDM. With the N2dG model derived in this paper, we have shown that this is the case.[8]

8 Conclusions

In this paper, we have derived new near-ignorance models for three members of the class of Normalized Infinitely Divisible distributions. We have shown that all these new near-ignorance prior models satisfy the embedding, symmetry, likelihood, learning and coherence principles, which are desirable properties for a model of prior ignorance. Furthermore, we have shown that one of these models satisfies the representation invariance principle while, for the other two models, the posterior imprecision depends linearly or almost linearly on the number of observed categories. As future work, we aim to complete the analysis of these three new near-ignorance models by proving the conjecture that we have discussed in the paper. Furthermore, we aim to extend our analysis to other members of the Normalized Infinitely Divisible distributions by working directly on the domain of the Infinitely Divisible distributions, that is before normalization. For a practical side, we plan to apply our models to solve classification and prediction problems and compare the results with the ones obtained by precise models and by the Imprecise Dirichlet Model.

[7] For the NIG prior we are not able to compute the lower probability in this case, since Theorem 3 is valid only if at least 1 observation has been collected for each category.

[8] The paper [17] discusses IDM as a predictive model. We plan to extend the N2dG model to predictive inferences and, thus, to verify if it satisfies the other properties listed in [17].

A Appendix: Proofs

A.1 Proof of Theorem 1

Without loss of generality, we assume that $i = 1$. For $n_1 - s \geq 0$ the lower can be derived by taking $\beta_1 = 1$ and applying the formula of IDM with α_1 replaced by $\alpha_1 - s$. For $n_1 + s \leq N$, let us consider the integral:

$$\int_0^1 dp_1 p_1^{n_1+\alpha_1-1} \int_0^{1-p_1} \cdots \int_0^{1-p_1-\cdots-p_{m-1}} \frac{p_2^{n_2+\alpha_2-1} p_3^{n_3+\alpha_3-1} (1-p_1-\cdots-p_{m-1})^{n_m+\alpha_m-1}}{\left(\sum\limits_{i=1}^{m-1} \beta_i p_i + \beta_m (1-p_1-\cdots-p_{m-1})\right)^s} dp_2 \cdots dp_{m-1} \tag{30}$$

Set $\beta_1 = 0$ and introduce the change of variables $p_i' = p_i/(1-p_1)$ for $i = 2,\ldots,m-1$ then, neglecting the normalization constant, the previous integral reduces to:

$$\textstyle\int_0^1 p_1^{n_1+\alpha_1-1}(1-p_1)^{N-n_1-\alpha_1-1} dp_1. \tag{31}$$

Therefore, the posterior expectation of P_1 for $\beta_1 = 0$ is

$$E[P_1|n_1,\ldots,n_m] = \frac{\int_0^1 p_1 p_1^{n_1+\alpha_1-1}(1-p_1)^{N-n_1-\alpha_1-1} dp_1}{\int_0^1 p_1^{n_1+\alpha_1-1}(1-p_1)^{N-n_1-\alpha_1-1} dp_1} = \frac{n_1+\alpha_1}{N},$$

where the last equality follows from the property of the Beta distribution. Hence, the upper posterior expectation of P_1 is $\overline{E}[P_1|n_1,\ldots,n_m] = (n_1+s)/N$. Consider now the case $n_1 + s > N$. For (30), we introduce the short notation: $\int_0^1 dp_1 p_1^{n_1+\alpha_1-1}(\ldots)$, where $(\ldots)$ denotes the multidimensional inner integration in (30), then for a chosen $\varepsilon \in (0,1)$ one has:

$$\begin{aligned} & E[P_1|n_1,\ldots,n_m] \\ & = \frac{\int_0^{1-\varepsilon} dp_1 p_1^{n_1+\alpha_1+1-1}(\ldots) + \int_{1-\varepsilon}^1 dp_1 p_1^{n_1+\alpha_1+1-1}(\ldots)}{\int_0^{1-\varepsilon} dp_1 p_1^{n_1+\alpha_1-1}(\ldots) + \int_{1-\varepsilon}^1 dp_1 p_1^{n_1+\alpha_1-1}(\ldots)} \\ & \geq \frac{\int_0^{1-\varepsilon} dp_1 p_1^{n_1+\alpha_1+1-1}(\ldots)}{\int_0^{1-\varepsilon} dp_1 p_1^{n_1+\alpha_1-1}(\ldots) + \int_{1-\varepsilon}^1 dp_1 p_1^{n_1+\alpha_1-1}(\ldots)} \\ & + \frac{(1-\varepsilon)\int_{1-\varepsilon}^1 dp_1 p_1^{n_1+\alpha_1-1}(\ldots)}{\int_0^{1-\varepsilon} dp_1 p_1^{n_1+\alpha_1-1}(\ldots) + \int_{1-\varepsilon}^1 dp_1 p_1^{n_1+\alpha_1-1}(\ldots)} \end{aligned} \tag{32}$$

Now, since for $\beta_1 \to 0$ it results that $\int_{1-\varepsilon}^1 dp_1 p_1^{n_1+\alpha_1-1}(\ldots) \to \infty$ (this can be derived from (30) by noticing that for $n_1 + s > N$ the argument of the integral goes to infinity at $p_1 = 1$ faster than $1/(1-p_1)$), while $\int_0^{1-\varepsilon} dp_1 p_1^{n_1+\alpha_1-1}(\ldots)$ and $\int_0^{1-\varepsilon} dp_1 p_1^{n_1+\alpha_1+1-1}(\ldots)$ are finite, then this implies that the right hand side of (32), is lower bounded by $1-\varepsilon$ which goes to 1 for $\varepsilon \to 0$. This shows that for $n_1 + s > N$, the upper posterior expectation is $\overline{E}[P_1|n_1,\ldots,n_m] = 1$. A similar approach can be used to prove that $\underline{E}[P_1|n_1,\ldots,n_m] = 0$ for $n_1 - s < 0$.

A.2 Proof of Corollary 1

The proof is similar to that of Theorem 1.

A.3 Proof of Theorem 2

Without loss of generality, we assume that $i = 1$. For $n_1 - 1/2 > 0$, i.e., $n_1 > 0$, the lower posterior expectation can be derived by taking $t_1 = 0$. Then, neglecting the normalization constant, the integral expression for the posterior expectation $E[P_1|n_1,\ldots,n_m]$ becomes:

$$\int_0^1 dp_1 p_1^{n_1-3/2+1} \int_0^{1-p_1} \cdots \int_0^{1-p_1-\ldots-p_{m-1}} \frac{\prod\limits_{i=2}^{m-1} p_i^{n_i-3/2+m/2} (1-\sum\limits_{i=1}^{m-1} p_i)^{n_m-3/2+m/2}}{\left[(1-\sum\limits_{i=1}^{m-1} p_i)\sum\limits_{i=2}^{m-1}\left(t_i^2 \prod\limits_{i\neq j=2}^{m-1} p_j\right) + t_m^2 \prod\limits_{i=2}^{m-1} p_i\right]^{m/2}} dp_2 \cdots dp_{m-1} \tag{33}$$

By introducing the change of variable $p_i' = p_i/(1-p_1)$, $i = 2,\ldots,m-1$ the previous integral and its normalization constant reduces to:

$$E[P_1|n_1,\ldots,n_m] = \frac{\int\limits_0^1 p_1^{n_1-\frac{3}{2}+1}(1-p_1)^{N-n_1-\frac{1}{2}}}{\int\limits_0^1 p_1^{n_1-\frac{3}{2}}(1-p_1)^{N-n_1-\frac{1}{2}}} = \frac{n_1-\frac{1}{2}}{N} \tag{34}$$

where the last equality follows from the property of the Beta distribution with $\alpha_1 = -1/2 + n_1 > 0$, $\alpha_2 = N - n_1 + 1/2 > 0$. A similar approach than that used in the proof of theorem 1 can be used to prove that $\underline{E}[P_1|n_1,\ldots,n_m] = 0$ if $n_1 = 0$.

For $n_1 + \hat{m}/2 < N$, i.e., $n_1 < N$, the upper expectation can be computed from $\overline{E}[P_1|n_1,\ldots,n_m] = 1 - \sum_{i=2}^m \underline{E}[P_i|n_1,\ldots,n_m]$. In the first part of this proof, we have shown that the lower expectation of P_i is $(n_i - 1/2)/N$ only for the $\hat{m}$ possible values $z_i \neq z_1$ of Z for which $n_i > 0$, whereas for the remaining $m - \hat{m} - 1$ values of Z for which $n_i = 0$ the lower expectation is zero. Then, $\sum_{i=2}^m \underline{E}[P_i|n_1,\ldots,n_m] = \frac{N-n_1-\hat{m}/2}{N}$, and one obtains the expression in (23). An approach similar to that used in the proof of Theorem 1 can be used to prove that $\overline{E}[P_j|n_1,\ldots,n_m] = 1$ if $n_i = N$.

A.4 Proof of Theorem 3

Without loss of generality, we assume that $k = 1$ and $i = 2$. Taking $t_1 = 1$ and $t_j = 0$, $j = 2,\ldots,m$, if $n_i > 0$, $i = 1,..,m$, the integral expression for the posterior expectation $E[P_1|n_1,\ldots,n_m]$, neglecting the normalization constant, can be written as:

$$\begin{aligned} & \int_0^1 dp_2 p_2^{n_2-\frac{3}{2}+\frac{m}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_2}}\right) \\ & \int_0^{1-p_1} p_2^{n_2-\frac{3}{2}+1} dp_1 \int_0^{1-p_1-p_2} \cdots \int_0^{1-p_1-\cdots-p_{m-1}} \\ & \prod_{i=3}^{m-1} p_i^{n_i-\frac{3}{2}} (1-\sum_{i=1}^{m-1} p_i)^{n_m-\frac{3}{2}} dp_3 \cdots dp_{m-1} \end{aligned} \tag{35}$$

By introducing the change of variable $p_i' = p_i/(1-p_1)$, for $i = 2,\ldots,m-1$, and $p_i'' = p_i'/(1-p_2)$, for $i = 3,\ldots,m-1$, the previous integral and its normalization constant reduces to:

$$\begin{aligned} & \frac{\int_0^1 p_1^{n_1+\frac{m-6}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_1}}\right)(1-p_1)^{N-n_1-\frac{m-1}{2}}}{\int_0^1 p_1^{n_1+\frac{m-6}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_1}}\right)(1-p_1)^{N-n_1-\frac{m+1}{2}}} \times \\ & \frac{\int_0^1 p'^{n_2-\frac{3}{2}+1}_2 (1-p'_2)^{N-n_1-n_2-\frac{m}{2}} dp_2'}{\int_0^1 p'^{n_2-\frac{3}{2}}_2 (1-p'_2)^{N-n_1-n_2-\frac{m}{2}} dp_2'}, \end{aligned} \tag{36}$$

where the second term of the product is equal to $\frac{n_i-1/2}{N-n_k-(m-1)/2}$ from the property of the Beta distribution with $\alpha_1 = -1/2 + n_1 > 0$, $\alpha_2 = N - n_1 - n_2 - m/2 + 1 > 0$. A similar approach can be used to prove the result for $t_2 = 1$.

Acknowledgements

This work has been partially supported by the Swiss NSF grant n. 200020-137680/1.

References

[1] P. Walley, “Inferences from multinomial data: learning about a bag of marbles,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 3–57, 1996.

[2] P. Walley, *Statistical Reasoning with Imprecise Probabilities*. New York: Chapman and Hall, 1991.

[3] P. Walley, “Measures of uncertainty in expert systems,” *Artificial Intelligence*, vol. 83, no. 1, pp. 1–58, 1996.

[4] A. Benavoli and M. Zaffalon, “A model of prior ignorance for inferences in the one-parameter exponential family,” *Journal of Statistical Planning and Inference*, vol. 142, no. 7, pp. 1960 – 1979, 2012.

[5] S. Moral, “Imprecise probabilities for representing ignorance about a parameter,” *Int. Journal of Approximate Reasoning*, vol. 53, no. 3, pp. 347 – 362, 2012.

[6] A. Piatti, M. Zaffalon, F. Trojani, and M. Hutter, “Limits of learning about a categorical latent variable under prior near-ignorance,” *Int. Journal of Approximate Reasoning*, vol. 50, no. 4, pp. 597–611, 2009.

[7] J. Bernard, “An introduction to the imprecise Dirichlet model for multinomial data,” *Int. Journal of Approximate Reasoning*, pp. 123–150, 2005.

[8] S. Favaro, G. Hadjicharalambous, and I. Prnster, “On a class of distributions on the simplex,” *Journal of Statistical Planning and Inference*, vol. 141, no. 9, pp. 2987 – 3004, 2011.

[9] B. Fristedt and L. Gray, *A modern approach to probability theory*. Birkhäuser Boston, 1996.

[10] A. Lijoi, R. H. Mena, and I. Prnster, “Hierarchical mixture modeling with normalized inverse-gaussian priors,” *Journal of the American Statistical Association*, vol. 100, no. 472, pp. 1278–1291, 2005.

[11] M. A. Carlton, “A Family of Densities Derived from the Three-Parameter Dirichlet Process,” *Journal of Applied Probability*, vol. 39, no. 4, pp. pp. 764–774, 2002.

[12] A. Benavoli and M. Zaffalon, “Prior near-ignorance for inferences in the k-parameter exponential family.” Available at http://www.idsia.ch/∼alessio/TR2011.pdf.

[13] F. P. A. Coolen and T. Augustin, “A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories,” *International Journal of Approximate Reasoning*, vol. 50, no. 2, pp. 217–230, 2009.

[14] J. Nolan, “An introduction to stable distributions.” Available at http://academic2.american.edu/ jpnolan.

[15] S. Favaro, G. Hadjicharalambous, and I. Prnster, “On a class of distributions on the simplex,” *Journal of Statistical Planning and Inference*, vol. 141, no. 9, p. 29873004, 2011.

[16] P. Walley and J.-M. Bernard, “Imprecise probabilistic prediction for categorical data.” Technical Report CAF-9901, Laboratoire Cognition et Activités finalisées, Université Paris 8, Saint- Denis, France (1999).

[17] G. De Cooman, E. Miranda, and E. Quaeghebeur, “Immediate prediction under exchangeability and representation insensitivity,” in *ISIPTA ’07 : proceedings of the fifth international symposium on imprecise probability: theories and applications* (G. De Cooman, J. Vejnarová, and M. Zaffalon, eds.), pp. 107–116, Society for Imprecise Probability: Theories and Applications (SIPTA), 2007.

A New Framework for Learning Generalized Credal Networks

Andrés R. Masegosa, Serafín Moral
Dpto. Ciencias de la Computación e Inteligencia Artificial
Universidad de Granada, 18071, Granada, Spain
{andrew,smc}@decsai.ugr.es

Abstract

In this paper we consider the problem of learning credal networks from observations when the prior information is given by a set of prior densities instead of a single one. We shall concentrate in the case in which the prior information is given by a family of Dirichlet distributions with uniform weights and different values of the equivalent sample size parameter. Also the use of imprecise probability models to specify the prior probability regarding the different graphs will be considered. The novelty is twice: first we assume an additional information on the set of possible values of the equivalent sample size parameter; and second we give a formalization of the problem which includes also as particular case different Bayesian approaches. Additionally, approximate and exact algorithms based on the A* search procedure are provided to compute the set of undominated decisions. Some preliminary experiments are reported.

Keywords. Credal networks, learning, imprecise sample size Dirichlet model, search algorithms.

1 Introduction

Learning Bayesian networks from a dataset of observations [15, 8, 12] is an important research challenge that despite the important effort made in past years, is far from having resulted in a satisfactory solution in all situations. In particular, in some cases it is assumed that providing a single network can be insufficient, especially if the sample size is small, as there can be several graphical structures with high posterior probability given the data. Model averaging approaches [7] provide a set of alternative networks each one with its own posterior probability. Inferences about any structural feature (for example the presence of a link) are carried out by averaging in the solution set. In this paper, we will show that even if our decision consists in selecting one graph, it makes sense to consider a set of alternative solutions if we use imprecise probability models to specify the prior probability regarding graphs and parameters [18]. In this paper we present a new framework which is based on imprecision in the prior probability about the graph associated to the network and in the value of the equivalent sample size of a BDeu score [2]. The result of the learning problem will be a set of undominated graphs (a *multigraph credal network* [5, 10]). The approach is based on our previous work [10] but assumes more information on the equivalent sample size (ϵ-contaminated model of a uniform distribution) and on the space of graphs. The main contributions of this paper are the following: we provide a new formulation of the learning problem depending on the prior information and the set of possible decisions; we transform the problems of computing undominated solutions in simpler problems involving the optimization of specific scores; and we propose different algorithms both approximate and exact for the associated optimization problems. Corani and Zaffalon [4] also consider multigraph credal networks in a classification problem. They are based on an imprecise prior information on the space of graphs but their approach is different of the one proposed in this paper and they concentrate in the classification problem and not in the computation of the graphs.

The paper is structured as follows: Section 2 reviews the problem of learning Bayesian networks; Section 3 is devoted to credal networks; Section 4 introduces our framework for learning multigraph credal networks [10]; Section 5 proposes approximate and exact algorithms to compute the set of undominated decisions; Section 6 reports some very preliminary experiments; while Section 7 is devoted to the conclusions.

2 Learning Bayesian Networks

Assume that we have a vector $\mathbf{X} = \{X_1, \ldots, X_m\}$ of variables. A *Bayesian network* about variables $\mathbf{X}$ is a directed acyclic graph G with a node for each variable X_i and a list of conditional probability distributions $(p(X_1|\Pi_1),\ldots,p(X_m|\Pi_m))$, where Π_i is the set of parents of variable X_i in graph G [13] and $p(X_i|\Pi_i)$ is a conditional distribution of variable X_i given Π. Given the independence relationships represented by G, the list of conditional distributions determines a joint probability distribution about $\mathbf{X}$ as a product of the conditional distributions (p *factorizes* with respect to G):

$$p(X_1,\ldots,X_m) = p(X_1|\Pi_1)\cdot\ldots\cdot p(X_m|\Pi_m) \qquad (1)$$

The number of different values of variables in Π_i is denoted by q_i, the set of possible values of Π_i by $\{\pi^i_1,\ldots,\pi^i_{q_i}\}$, the number of possible values of X_i is denoted by k_i, and the set of possible values of X_i by $\{w^i_1,\ldots,w^i_{k_i}\}$.

If we have a database of n observations D of variables $\{X_1,\ldots,X_m\}$, *learning* a Bayesian network consists first in estimating the graph G, and then estimating the conditional probability distributions associated to the graph [12]. In both cases, the most common approach is based on assuming a Dirichlet prior probability distribution for the values of each one of the conditional probability distributions $p(X_i|\Pi_i=\pi^i_j)$ with parameters $D(\alpha^i,\ldots,\alpha^i)$ (the weights are the same for the different values π^i_j of the parents, but they can be different for different variables). The value $s^i = k_i\alpha^i$ is called the *equivalent sample size*.

To estimate G, it is also assumed that there is a prior probability about the different graph structures (usually the uniform) and that the prior distributions for the different conditional distributions are independent. Under these conditions, it is possible to compute $p(G|D)$ which is a value proportional to (under the uniform prior for the graphs):

$$p(G|D) \propto P(D|G) = \prod_{i=1}^{m}\prod_{j=1}^{q_i}\frac{\Gamma(k_i\alpha^i)}{\Gamma(n_{ij}+k_i\alpha^i)}\prod_{l=1}^{k_i}\frac{\Gamma(\alpha^i+n_{ijl})}{\Gamma(\alpha^i)} \qquad (2)$$

where n_{ij} is the number of cases in D in which $\Pi_i=\pi^i_j$ and n_{ijl} the number of cases in D in which $\Pi_i=\pi_j$, $X_i=w^i_l$, and $\Gamma(.)$ is the gamma function. This value is called the score of the graph given the data and will be denoted as $Score(G|D)$.

The problem is then to find the graph with maximum score, and usually a greedy search algorithm is employed: given an initial graph, the set of graphs obtained by adding, removing, and inverting a link is computed, and for each one of them the score is found. Then the current graph is changed to the graph of the set with highest score, and the process is repeated while a score greater than the score of the current graph can be found.

Once a graph with highest score has been found, G, the conditional probabilities can be estimated (*parameter learning*). If for each conditional probability $p(X_i|\Pi_i=\pi^i_j)$ we have a Dirichlet prior with parameters $D(\alpha^i,\ldots,\alpha^i)$ and these prior distributions are independent, then the estimation can be done independently for each conditional probability. The estimated values are [12]:

$$p(X_i=w^i_l|\Pi_i=\pi^i_j,D) = \frac{n_{ijl}+\alpha^i}{n_{ij}+k_i\alpha^i} \qquad (3)$$

In both problems (learning the structure and the parameters), we have to specify how the weights, α^i, are computed. Usually the so called *Bayesian Dirichlet equivalent metric* or BDeu [2] is used in which a parameter s is fixed (the *global equivalent sample size*) and then the weights for each variable are computed as $\alpha^i = \frac{s}{q_ik_i}$. Though for the conditional probabilities it is also common to consider the Laplace correction which is equivalent to consider $\alpha^i=1$. In this paper, we will always consider $\alpha^i=\frac{s}{q_ik_i}$.

3 Credal Networks

A *credal network* [5] is a generalized Bayesian network where the probabilities can be imprecise. More concretely, a *locally defined credal network* [5] is a directed acyclic graph G and a list $(\mathcal{P}_1,\ldots,\mathcal{P}_m)$ where each $\mathcal{P}_i$ is a set of conditional distributions for variable X_i given its parents in G. The joint credal set associated to a credal network, is the set of probability distributions p that can be obtained as a product $p(X_1,\ldots,X_m)=p_1(X_1|\Pi_1)\cdot\ldots\cdot p_m(X_m|\Pi_m)$, where $p_i\in\mathcal{P}_i$. A *credal network* is a directed acyclic graph G and a credal set of joint probability distributions $\mathcal{P}$ such that any extreme probability $p\in\mathcal{P}$ factorizes with respect to G, i.e. it can be expressed as $p(\mathbf{X})=p_1(X_1|\Pi_1)\cdot\ldots\cdot p(X_m|\Pi_m)$, where $p(X_i|\Pi_i)$ is a conditional distribution of X_i given its parents in G. Not every credal network is locally defined.

A credal network can be obtained by determining a single graphical structure (either elicited from experts or learned with a Bayesian procedure) and a set of decomposable probability distributions learned with an imprecise probability procedure, for example using the Imprecise Dirichlet Model (IDM) [19]. The most simple application is the separable estimation of the conditional probabilities, being the result a locally defined credal network where $\mathcal{P}_i$ is the set of all the conditional probability distributions such that $p(X_i=w^i_l|\Pi_i=\pi^i_j,D)\in[\frac{n_{ijl}}{n_{ij}+s},\frac{n_{ijl}+s}{n_{ij}+s}]$, where s is a global parameter (usually $s=1$ or $s=2$). This procedure has a tendency to produce overly imprecise intervals when computing conditional imprecise probabilities given observations in these credal networks [21]. There is another alternative application of the IDM to learn the parameters: the global application [21] which produces more precise results, but it is more difficult to compute with it. There is a solution procedure for the naive credal classifier [21], but there is no efficient algorithm available for computing at the general case.

Sometimes the most natural result of learning is a family of graphs instead of a single one [4, 10]. To encompass this case, we define a *multigraph credal network* as a finite set a credal networks, i.e. a set of graphs and with each graph having a set of probability distributions that factorize according to the graph. In our approach, usually each graph will have a single probability distribution associated to it.

A multigraph credal network will be said to be *locally defined* when the variables can be sorted in such a way that for each variable X_i we have a family of possible sets of parents for this variable $\{\Pi_i^1, \dots, \Pi_i^{l_i}\}$ were Π_i^j is always included in the set of variables preceding X_i in the given order, and for each set of parents Π_i^j we have a set of conditional probability distributions for X_i given Π_i^j. A locally defined multigraph credal network will have an associated multigraph credal network: we only have to consider all the credal networks obtained by selecting a possible set of parents Π_i^j and its corresponding set of conditional probability distributions for each variable X_i. For a locally defined multigraph we have to give the family of possible parents and for each parent the set of possible conditional distributions. However, if l_i is the number of possible parents for X_i, the number of credal networks that can be obtained by selecting a possible set of parents for each variable is $\prod_{i=1}^m l_i$, which can be a very large number. Then it is clear that the local definition can be much more compact than the multigraph definition. Furthermore, it is possible to directly make inferences with the local definition [10].

To locally define a multigraph credal network, we need to specify an order of the variables in such way that the set of parents of a variable are selected from preceding variables. If we did not specify this order then, there could be two variables X_i and X_j, a possible set of parents of X_i, Π_i, containing X_j and a possible set of parents of X_j, Π_j, containing X_i. The two set of parents Π_i and Π_j are not compatible as they give rise to a cycle. More complex cycles could be created by circular relationships. So modularity would be lost, as we could not locally specify a family of parent sets for each variable without considering additional global restrictions for them.

4 Learning Multigraph Credal Networks

A directed acyclic graph G about variables $\mathbf{X}$ will be represented by a finite list $G = (\Pi_1, \dots, \Pi_m)$ of the set of parents of the different variables. The set of all the acyclic directed graphs will be denoted as $\mathcal{G}$.

To learn a multigraph credal network, we will follow a variant of the Imprecise Sample Size Dirichlet Model (ISSDM) introduced in [10]. As in the case of learning precise Bayesian networks, our model is based on assuming prior distributions $D(\alpha^i, \dots, \alpha^i)$ for the conditional probability distributions $p(X_i|\Pi_i = \pi_j^i)$, where $\alpha_i = \frac{s}{q_i k_i}$ and $s > 0$ is the global equivalent sample size. But now instead of an unique global equivalent sample size $s > 0$, it will be assumed that there is a finite set S of possible equivalent sample sizes[1].

In [10] two basic applications of the Imprecise Sample Size Dirichlet Model (ISSDM) have been considered:

- *The global approach.*- Given a graph G, the set of prior distributions for the conditional probabilities $p(X_i|\Pi_i = \pi_j^i)$ is the set of Dirichlet distributions $D(\alpha^i, \dots, \alpha^i)$ that are obtained by considering a value $s \in S$ and computing $\alpha^i = \frac{s}{q_i k_i}$.
- *The local approach.*- Given a graph G, the set of prior distributions for the conditional probabilities $p(X_i|\Pi_i = \pi_j^i)$ is the set of Dirichlet distributions $D(\alpha^i, \dots, \alpha^i)$ that are obtained by considering a value $s_i \in S$ for each variable X_i and computing $\alpha^i = \frac{s_i}{q_i k_i}$.

The difference between the local and the global approach is that in the global we have to select the same $s \in S$ to compute the weights of the prior distribution for each variable X_i, and in the local approach, we can select a different value $s_i \in S$ for each variable. The number of different prior distributions for the parameters is higher in the local approach than in the global approach. In the global approach it is $|S|$ (the cardinal of S) but in the local approach is $|S|^m$.

In this paper we will follow the local approach. So, given a graph G, we will consider that there is a prior distribution about the values of the conditional probabilities for each $\mathbf{s} = (s_1, \dots, s_m) \in S^m$, where the prior distribution for the conditional probabilities of variable X_i is $D(\alpha^i, \dots, \alpha^i)$ where $\alpha^i = \frac{s_i}{q_i k_i}$. There are reasons for assuming the possibility of a different $s_i \in S$ for each variable instead of the same $s \in S$ for all the variables. The value $s_i \in S$ determines

[1] In [10] it was assumed that S was an interval, but finally for computational reasons it was approximated by a finite set.

the prior probabilities for the conditional probability distribution of X_i: with small values of s_i the prior Dirichlet distribution is concentrated in the extremes (close to 0 and 1) and with high values of s_i the prior distribution is concentrated around the uniform distribution (all the probabilities close to $\frac{1}{k_i}$). If all the values are the same for all the variables, we are assuming that if prior density of X_i is concentrated in the extreme values (small s) so is the prior probability about the conditional probabilities for X_j. Assuming a different values of s_i for different variables, we are expressing that the probabilities for one variable can be extreme, while for other variable they can be close to the uniform distribution. In [3], we elaborate on these arguments when using precise Bayesian methods.

Given $\mathbf{s} \in S^m$, we can compute the score of a graph using (2) and to estimate the conditional probabilities associated to a graph with expression (3), with $\alpha^i = \frac{s_i}{q_i k_i}$. To emphasize the dependence upon $\mathbf{s}$, the score will be denoted by $Score(G|D,\mathbf{s})$ and the estimated conditional probability by $p_{\mathbf{s}}(X_i|\Pi_i, D)$.

Given an arbitrary set H, an ϵ-contaminated imprecise probability model of the uniform distribution in H where $\epsilon > 0$ is given by the set of all the probability distributions p defined on H and satisfying $p(h) \geq \frac{1-\epsilon}{k}$, where k is the number of elements in H. Equivalently this model can be characterized by the inequalities $\frac{p(h)}{p(h')} \geq \beta_\epsilon$, where $\beta_\epsilon = \frac{1-\epsilon}{1-\epsilon+k\epsilon}$. This is a convex set of probabilities and there is an extreme probability, p_h^ϵ, for each $h \in H$, given by $p_h^\epsilon(h) = \frac{1-\epsilon}{k} + \epsilon$ and $p_h(h') = \frac{1-\epsilon}{k}$ if $h' \neq h$. This probability can be expressed as a convex combination: $p_h^\epsilon = \epsilon p_h + (1-\epsilon)p_u$, where p_h is the probability degenerated on h (assigning probability 1 to h) and p_u is the uniform distribution. p_h^ϵ will be called the probability that *concentrates mass* ϵ on h and it is the probability for which the probability of h is maximized.

In [10] it was considered that the information on S^m was vacuous, but in this paper additional assumptions will be made.

- There will be an imprecise prior probability for the graph and the equivalent sample size vector, i.e. in $\mathcal{G} \times S^m$. The following options will be considered:
  - An ϵ-contaminated model of the uniform distribution in $\mathcal{G} \times S^m$, where $\epsilon > 0$. The associated set of probability distributions will be denoted as $\mathcal{P}_1$.
  - An ϵ-contaminated model of the uniform distribution in the space of directed acyclic graphs $\mathcal{G}$ and for each graph $G \in \mathcal{G}$ we have an uniform distribution in S^m conditioned to G. It will be denoted as $\mathcal{P}_2$.
- The set of possible decisions, $\mathcal{D}$, can be one of the following options:
  - The set of graphs $\mathcal{D} = \mathcal{G}$.
  - The set of graphs and equivalent sample sizes: $\mathcal{D} = \mathcal{G} \times S^m$.
- The utility function when $\mathcal{D} = \mathcal{G}$ is $U : \mathcal{G} \times \mathcal{D} \to [0,1]$, given by

$$U(G,G') = \begin{cases} 1 & \text{if } G = G' \\ 0 & \text{otherwise} \end{cases},$$

  where G is the true graph and G' our decision.

  If $\mathcal{D} = \mathcal{G} \times S^m$, the utility function is $U : \mathcal{G} \times S^m \times \mathcal{D} \to [0,1]$ given by:

$$U(G,\mathbf{s},G',\mathbf{s}') = \begin{cases} 1 & \text{if } G = G', \mathbf{s} = \mathbf{s}' \\ 0 & \text{otherwise} \end{cases}$$

  where G is the true graph and $\mathbf{s}$ the true equivalent sample size and $(G',\mathbf{s}')$ our decision.

A decision $d \in \mathcal{D}$ is said to be *dominated* by another decision $d' \in \mathcal{D}$ if and only if for any possible probability distribution, p, associated to the problem we have that $p(U(.,d)|D) < p(U(.,d')|D)$, where $U(.,d)$ is the function that assigns to each graph G (or pair $(G,\mathbf{s})$) the value $U(G,d)$ (or $U(G,\mathbf{s},d)$) and $p(U(.,d)|D)$ stands for the prevision or mathematical expectation of this function with respect to p conditioned to the dataset D.

In these conditions, *learning* is defined as the computation of all the undominated decisions. Assume that a probability p has been fixed in $\mathcal{G} \times S^m$. If $\mathcal{D} = \mathcal{G}$, then if $d = G$, we have that $p(U(.,d)|D) = p(G|D) = p(d|D)$, as $U(.,d)$ is a function in $\mathcal{G}$ that is equal to 1 in G and 0 otherwise. Analogously, in the case of $\mathcal{D} = \mathcal{G} \times S^m$ and $d = G \times \mathbf{s}$, we also obtain $p(U(.,d)|D) = p(G \times \mathbf{s}|D) = p(d|D)$.

Depending on the different options, we have the following situations.

4.1 $\mathcal{D} = \mathcal{G}$ and $\mathcal{P}_2$ as prior probability

In this case, we have an ϵ-contaminated model in $\mathcal{G}$ and for each graph an uniform distribution in S^m.

Given that $p(U(.,G)|D) = p(G|D)$, we have to compute all the graphs G such that for any graph G' there is a probability $p \in \mathcal{P}_2$ such that $p(G|D) \geq p(G'|D)$. Given that we have the ϵ-contaminated model in $\mathcal{G}$, then this probability exists, if and only if this inequality is satisfied for the probability p^ϵ_G in $\mathcal{G}$ that maximizes the probability of G (concentrates the ϵ mass on G).

So we have to compute all the graphs G such that $p^\epsilon_G(G|D) \geq p^\epsilon_G(G'|D), \forall G' \in \mathcal{G}$, for the probability $p^\epsilon_G \in \mathcal{P}_2$.

For any probability $p \in \mathcal{P}_2$, we have that $p(G|D) \propto p(G)p(D|G)$. In this expression, $p(D|G) = \sum_{\mathbf{s}\in S^m} p(D|G,\mathbf{s}) = \frac{1}{|S|^m}\sum_{\mathbf{s}\in S^m} Score(G|D,\mathbf{s})$, as given a graph, we have the uniform distribution in S^m. This probability is fixed and does not depend on the probability $p \in \mathcal{P}_2$ and will be denoted as $AScore(G|D) = \frac{1}{|S|^m}\sum_{\mathbf{s}\in S^m} Score(G|D,\mathbf{s})$.

As for any graph we have that $\frac{p^\epsilon_G(G')}{p^\epsilon_G(G)} = \beta_\epsilon$, then a graph G is undominated if and only if $AScore(G|D) \geq \beta_\epsilon AScore(G'|D)$ for any graph G'.

If G is the graph maximizing $AScore(G|D)$, then this graph is undominated (β_ϵ is always lower than 1), and another graph G' is undominated if and only if $AScore(G'|D) \geq \beta_\epsilon AScore(G|D)$.

$AScore(G|D)$ is the locally averaged score of a Bayesian network as defined by Cano et al. [3]. It is immediately clear that it is not necessary to average an exponential number of scores as,

$$AScore(G|D) = \frac{1}{|S|^m}\sum_{\mathbf{s}\in S^m} Score(G|D,\mathbf{s}) =$$

$$\prod_{i=1}^{m}\left(\frac{1}{|S|}\sum_{s_i\in S}\prod_{j=1}^{q_i}\frac{\Gamma(k_i.\alpha^i)}{\Gamma(n_{ij}+k_i.\alpha^i)}\prod_{l=1}^{k_i}\frac{\Gamma(\alpha^i+n_{ijl})}{\Gamma(\alpha^i)}\right).$$

In short, in this case the problem is to compute the set of graphs with an averaged score greater or equal to $\beta_\epsilon.MAXAVG$, where $MAXAVG$ is the maximum averaged score of a graph. If $\epsilon = 0$, we have the problem of computing the graph G optimizing the locally average score as considered in [3].

4.2 $\mathcal{D} = \mathcal{G} \times S^m$ and $\mathcal{P}_2$ as prior probability

In this case, we have to select a graph G and a vector of equivalent sample sizes $\mathbf{s} = (s_1, \ldots, s_m)$ such that there is not another pair $(G', \mathbf{s}')$ such that for any probability p we have that $p(G,\mathbf{s}|D) < p(G',\mathbf{s}'|D)$. Given that for any $p \in \mathcal{P}_2$ the prior probability in S^m given G is uniform, we have that $p(G,\mathbf{s}|D) \propto p(G)Score(G|D,\mathbf{s})$. So, a pair $(G,\mathbf{s})$ is undominated if and only if for any pair $(G',\mathbf{s}')$ there is a probability in $p \in \mathcal{P}_2$ for which $p(G)Score(G|D,\mathbf{s}) \geq p(G')Score(G'|D,\mathbf{s}')$. As $Score(G|D,\mathbf{s})$ does not depend on $p \in \mathcal{P}_2$, if this inequality is true for a probability $p \in \mathcal{P}_2$, it will also hold for the probability maximizing the probability of G, p^ϵ_G. So this is equivalent that for any $(G',\mathbf{s}')$ we have that $p^\epsilon_G(G)Score(G|D,\mathbf{s}) \geq p^\epsilon_G(G')Score(G'|D,\mathbf{s}')$. Taking into account that $\frac{p^\epsilon_G(G')}{p^\epsilon_G(G)} = \beta_\epsilon$ (if $G \neq G'$), this inequality is equivalent to $Score(G|D,\mathbf{s}) \geq \beta_\epsilon Score(G'|D,\mathbf{s}')$ if $G \neq G'$ and to $Score(G|D,\mathbf{s}) \geq Score(G'|D,\mathbf{s}')$ if $G = G'$.

If G is fixed, then $(G,\mathbf{s}')$ dominates $(G,\mathbf{s})$ if and only if $Score(G|D,\mathbf{s}') > Score(G|D,\mathbf{s})$. Then for a pair $(G,\mathbf{s})$ to be undominated, it is necessary that $\mathbf{s} = \arg_{\mathbf{s}'}\max Score(G|D,\mathbf{s}')$.

Let us define $MScore(G|D) = \max_{\mathbf{s}\in S^m} Score(G|D,\mathbf{s})$ and assume that $MAXMAX$ is the maximum of this score in the space of all the graphs and G^* the graph for which this score is obtained. We can prove the following result.

Proposition 1 *A pair $(G,\mathbf{s})$ is undominated if and only if $\mathbf{s} = \arg_{\mathbf{s}'}\max Score(G|D,\mathbf{s}')$ and $MScore(G|D) \geq \beta_\epsilon MAXMAX$.*

Proof: If the pair $(G,\mathbf{s})$ is undominated we know that $\mathbf{s} = \arg_{\mathbf{s}'}\max Score(G|D,\mathbf{s}')$. Also this pair can not be dominated by $(G^*,\mathbf{s}^*)$ where $s^* = arg_{\mathbf{s}'}\max Score(G^*|D,\mathbf{s}')$ and therefore $MScore(G|D) \geq \beta_\epsilon Score(G^*|D,\mathbf{s}^*) = \beta_\epsilon MAXMAX$.

On the other hand, assume $\mathbf{s} = \arg_{\mathbf{s}'}\max Score(G|D,\mathbf{s}')$ and $MScore(G|D) \geq \beta_\epsilon MAXMAX$. If $\mathbf{s} = \arg_{\mathbf{s}'}\max Score(G|D,\mathbf{s}')$ then the pair $(G,\mathbf{s})$ is not dominated by any pair $(G,\mathbf{s}')$ (a pair with the same graph and different vector of equivalent sample sizes). Also $MAXMAX \geq Score(G'|D,\mathbf{s}')$ and therefore for any pair $(G',\mathbf{s}')$ $Score(G|D,\mathbf{s}) \geq \beta_\epsilon MAXMAX \geq \beta_\epsilon Score(G'|D,\mathbf{s}')$, and the pair $(G,\mathbf{s})$ is not dominated by any pair $(G',\mathbf{s}')$ with $G' \neq G$ either. So, the pair $(G,\mathbf{s})$ is undominated. □

It is important to remark that $MScore(G|D)$ can be locally computed as

$$MScore(G|D) = \max_{\mathbf{s}\in S^m} Score(G|D,\mathbf{s}) =$$

$$\max_{\mathbf{s}\in S^m}\prod_{i=1}^{m}\left(\prod_{j=1}^{q_i}\frac{\Gamma(k_i.\alpha^i)}{\Gamma(n_{ij}+k_i.\alpha^i)}\prod_{l=1}^{k_i}\frac{\Gamma(\alpha^i+n_{ijl})}{\Gamma(\alpha^i)}\right) =$$

$$\prod_{i=1}^{m}\left(\max_{s_i\in S}\prod_{j=1}^{q_i}\frac{\Gamma(k_i.\alpha^i)}{\Gamma(n_{ij}+k_i.\alpha^i)}\prod_{l=1}^{k_i}\frac{\Gamma(\alpha^i+n_{ijl})}{\Gamma(\alpha^i)}\right).$$

In short, in this case the problem is to compute the set of graphs with a maximum score greater or equal than $\beta_\epsilon.MAXMAX$, where $MAXMAX$ is the optimal value of the maximum score of a graph and for each one of these graphs we have to determine the vector $\mathbf{s}$ maximizing the score. If $\epsilon = 0$, we have the problem of computing the graph $(G, \mathbf{s})$ optimizing the score $MScore(G|D, \mathbf{s}')$. This approach has been considered by Steck [16] to minimize the effect in the learned structure of a Bayesian network of the equivalent sample size parameter. However, in that paper a continuous set of possible parameters S is considered and a global approach is applied: the same parameter s must be applied to the conditional probability of each variable X_i.

4.3 $\mathcal{D} = \mathcal{G} \times S^m$ and $\mathcal{P}_1$ as prior probability

This case is similar to the one considered in Subsection 4.2. Now, we have that $p(G, \mathbf{s}|D) \propto p(G, \mathbf{s})Score(G|D, \mathbf{s})$ and we have an ϵ-contaminated model in $\mathcal{G} \times S^m$. If a pair $(G, \mathbf{s})$ is dominated by another pair $(G', \mathbf{s}')$, then this is equivalent to $p(G, \mathbf{s})Score(G|D, \mathbf{s}) < p(G', \mathbf{s}')Score(G'|D, \mathbf{s}')$ for any probability $p \in \mathcal{P}$, which given the structure of $\mathcal{P}_1$ is equivalent to $p^\epsilon_{G,\mathbf{s}}(G, \mathbf{s})Score(G|D, \mathbf{s}) < p^\epsilon_{G,\mathbf{s}}(G', \mathbf{s}')Score(G'|D, \mathbf{s}')$ where $p^\epsilon_{G,\mathbf{s}}$ is the probability in $\mathcal{P}_1$ maximizing the probability of $(G, \mathbf{s})$. And this is equivalent to $Score(G|D, \mathbf{s}) < \beta_\epsilon Score(G'|D, \mathbf{s}')$.

It is immediately clear that for a graph G there is a pair $(G, \mathbf{s})$ that is undominated if and only if $MScore(G|D) \geq \beta_\epsilon MAXMAX$. The difference with the computations in Subsection 4.2, is that for a given undominated graph G, now there can be several vectors of parameters $\mathbf{s} \in S^m$ such that $(G, \mathbf{s})$ is undominated: all the pairs for which $Score(G|D, \mathbf{s}) \geq \beta_\epsilon.MAXMAX$.

We can proceed as follows: we can compute the set of graphs with a $MScore(G|D)$ greater or equal than $\beta_\epsilon.MAXMAX$ as in the previous case. Then for each graph G we compute the set S' of parameters $\mathbf{s}$ such that $Score(G|D, \mathbf{s}) \geq \beta_\epsilon MAXMAX$. This computation can be difficult as the number of elements in $\mathbf{s} \in S^m$ is exponential and the problem can not be decomposed by computing a set of components, S_i, for each variable X_i and then computing $S' = S_1 \times \cdots \times S_m$. Whether a component s_i belongs to an undominated vector $\mathbf{s}$ depends on the other components in the vector and can not be separately computed for each variable.

4.4 $\mathcal{D} = \mathcal{G}$ and $\mathcal{P}_1$ as prior probability

This case poses an additional difficulty compared to above situations. A graph G is undominated if and only if for each graph G' there is a probability p such that $p(G|D) \geq p(G'|D)$. The difference is that now the probability p can depend on the graph G'. In previous cases, the problem could be simplified as it could be shown that if this happened we could select the same probability for any graph: the probability p^ϵ_G maximizing the probability of G. But this simplification is not possible in this case. A possible solution is to concentrate in the set *e-admissible* decisions [9]: a graph G is e-admissible if it maximizes $p(G|D)$ for a possible probability p. All the e-admissible decisions are undominated but not the reverse.

In the following we shall concentrate in computing e-admissible solutions. If G' maximizes the conditional probability $p(.|D)$ with p in a convex set $\mathcal{P}_1$, then G' will also be optimal for one extreme probability $p_{(G,\mathbf{s})} \in \mathcal{P}_1$. So we shall concentrate in finding the graphs that optimize the posterior probability for extreme probabilities.

Let us consider $p_{(G,\mathbf{s})}(G'|D)$ the posterior probability of graph G' when the prior probability in $\mathcal{G} \times S^m$ is $p_{(G,\mathbf{s})}$. We have the following situations:

- If $G \neq G'$, then

$$p_{(G,\mathbf{s})}(G'|D) \propto \sum_{\mathbf{s}' \in S^m} p_{(G,\mathbf{s})}(G', s')p_{(G,\mathbf{s})}(D|G', s') =$$

$$\sum_{\mathbf{s}' \in S^m} \frac{1-\epsilon}{k} Score(G'|D, \mathbf{s}') = \frac{1-\epsilon}{k'} AScore(G'|D),$$

  where $k' = \frac{k}{|S|^m}$. In the above equalities, we have that $p_{(G,\mathbf{s})}(D|G', s')$ is equal to $Score(G'|D, \mathbf{s}')$ as this conditional probability does not depend of the prior probability in $\mathcal{G} \times S^m$.

- If $G = G'$, then

$$p_{(G,\mathbf{s})}(G|D) \propto \sum_{\mathbf{s}' \in S^m} p_{(G,\mathbf{s})}(G', s')p_{(G,\mathbf{s})}(D|G, s') =$$

$$\sum_{\mathbf{s}' \in S^m} \frac{1-\epsilon}{k} Score(G|D, \mathbf{s}') + \epsilon Score(G|D, \mathbf{s}) =$$

$$= \frac{1-\epsilon}{k'} AScore(G|D) + \epsilon Score(G|D, \mathbf{s})$$

  where $k' = \frac{k}{|S|^m}$

Given above expressions, it can immediately be seen that if G' maximizes $p_{(G,\mathbf{s})}(G'|D)$ for one extreme probability, then $G' = G$. In that case, we have that $p_{(G',\mathbf{s})}(G'|D) = \frac{1-\epsilon}{k'} AScore(G'|D) + \epsilon Score(G'|D, \mathbf{s})$ and $p_{(G',\mathbf{s})}(G|D) = \frac{1-\epsilon}{k'} AScore(G|D)$, for $G \neq G'$. Then, if G' maximizes $p_{(G,\mathbf{s})}(G'|D)$, it will also do it when $\mathbf{s} = \arg_{\mathbf{s}'} \max Score(G'|D, \mathbf{s}')$, and in this case, $p_{(G',\mathbf{s})}(G'|D) = \frac{1-\epsilon}{k'} AScore(G'|D) +$

$\epsilon MScore(G'|D)$. So, we can express the condition for G' e-admissible: $\frac{1-\epsilon}{k'} AScore(G'|D) + \epsilon MScore(G'|D) \geq \frac{1-\epsilon}{k'} AScore(G|D)$, $\forall G \in \mathcal{G}$.

To compute the set of e-admissible graphs, we can start by computing the graph G^* maximizing $AScore(G|D)$. It is clear that this graph is e-amissible and that a graph G' is non dominated if and only if $\frac{1-\epsilon}{k'} AScore(G'|D) + \epsilon MScore(G'|D) \geq \frac{1-\epsilon}{k'} AScore(G^*|D)$, which is equivalent to $AScore(G'|D) + \frac{\epsilon k'}{1-\epsilon} MScore(G'|D) \geq AScore(G^*|D)$. If we call $MAMScore_\epsilon(G'|D)$ to $AScore(G'|D) + \frac{\epsilon k'}{1-\epsilon} MScore(G'|D)$. The computational problem is similar to the previous ones: to compute a family of graphs with a score above a threshold.

5 Algorithms

In this section we discuss some procedures to compute the set of undominated graphs or undominated graphs and equivalent sample sizes. In all the situations the procedure involves the computation of one graph maximizing a specific score, $Score1$ (in some cases the averaged and in others the maximum score). Then we have to compute all the graphs with a score ($Score2$) above B, where B is a value depending of the previous computed optimum. If A is the optimum of $Score1$, then this value will be denoted as $B = f(A)$. The scores of the first and second stages are not necessarily the same as in the case of $\mathcal{D} = \mathcal{G}$ and $\mathcal{P}_1$ as prior probability.

To carry out this task, we shall propose approximate and exact algorithms. Approximate algorithms can be based on algorithms that try to visit a significant set of networks with a high score as in Bayesian model averaging procedures [7]. In this paper we have considered a modification of the algorithms presented in Masegosa, Moral [11]. The procedure has two stages:

- First, it computes a graph G^* with a high value $Score1$ using the Max-Min hill climbing algorithms by Tsamardinos et al. [17], a state of the art algorithm to learn Bayesian networks. Compute $A = Score1(G^*|D)$ and $B = f(A)$.

- In a second step, a Markov Chain Monte-Carlo procedure is employed to compute the family of undominated graphs. It starts with a family $\mathcal{H}$ of graphs containing only G^*, the current graph G is initially equal to G^*. Then it randomly generates a graph G' from the neighboring graphs of the current graph G (the graphs obtained by adding, deleting, or reversing a link of G) and computes $Score2(G'|D)$. If $Score2(G'|D) \geq B$, then the current graph is set to G' and is added to $\mathcal{H}$.

  It also computes $Score1(G'|D)$ (there is no additional computation if $Score1 = Score2$) and if $Score1(G'|D) > A$, then we make $B = f(Score1(G'|D))$ and remove from $\mathcal{H}$ all the graphs G with $Score2(G|D) < B$. This step is done because the first stage is an approximate algorithm, and in this step we are visiting graphs with a high $Score2$. As the different scores are strongly correlated, there is a possibility that the computed optimum is improved by one of the graphs visited at this stage. This is specially convenient when $Score1 = Score2$ and there is not necessity in doing additional computations.

Recently, exact algorithms for computing the Bayesian networks maximizing a decomposable score have been presented [14, 6]. In this paper, we will concentrate on the A* algorithm proposed by Yuan et al. [20] and we will indicate how it can be generalized to compute the full family of undominated graphs in the case of a decomposable score such that $Score1 = Score2$. $Score$ is decomposable if and only if we can express $Score(G|D) = \prod_{i=1}^m Score_i(\Pi_i|D)$, i.e. it can be expressed as a product of scores associated to each variable and its sets of parents in the graph. This covers all the situations except the last one ($\mathcal{D} = \mathcal{G}$ and $\mathcal{P}_1$ as prior probability), as the score $MAMScore_\epsilon(G|D) = AScore(G|D) + \frac{\epsilon k'}{1-\epsilon} MScore(G|D)$ can not be expressed as a product of scores for each variable[2].

In the following $Score(G|D)$ is any decomposable score function and $LScore(G|D)$ is the logarithm of this score. We have that $LScore(G|D) = \sum_{i=1}^m LScore_i(\Pi_i|D)$, where $LScore_i(\Pi_i|D)$ is the logarithm of $Score_i(\Pi_i|D)$. First, we describe the A* algorithm for precise Bayesian networks (to compute the graph maximizing the score). It is assumed that we have a function $BestScore(X_i, R_i)$ that computes the optimal value of $LScore(\Pi_i|D)$ for $\Pi_i \subseteq R_i$ (see [20] for efficient procedures for this task) where R_i is a subset of $\mathbf{X}$. The learning problem is formulated as a search of the best path between two states. The set of states is the family of all the possible subsets $\mathbf{Y} \subseteq \mathbf{X}$. The initial state is the empty set and the final state is the full set $\mathbf{X}$. The set of children of a state $\mathbf{Y}$ is the set of states $\mathbf{Y} \cup \{X_i\}$, where $X_i \notin \mathbf{Y}$. The utility of going from one state $\mathbf{Y}$ to one of its children $\mathbf{Y} \cup \{X_i\}$ is $BestScore(X_i, \mathbf{Y})$. The utility of a path is the sum of the utilities of each one of its arcs, and the problem is to compute the path maximizing the utility of going from the initial state to the final

[2] As it is a linear combination of decomposable scores this does not pose any problem for the local computation under local changes (we locally compute $AScore(G|D)$ and $MScore(G|D)$ which are decomposable).

one. For that the A* algorithm is used with heuristic function $h(\mathbf{Y}) = \sum_{X_i \notin \mathbf{Y}} BestScore(X_i, \mathbf{X} \backslash \{X_i\})$. It can be proved that this heuristic is admissible [20] as it is an optimistic evaluation of the utility of the rest of the path. In these conditions a search procedure that expands the node with maximum value of $g(\mathbf{Y}) = u(\mathbf{Y}) + h(\mathbf{Y})$, where $u(\mathbf{Y})$ is the utility of the best path arriving to $\mathbf{Y}$, is guaranteed to find the optimal path the first time it chooses the final state $\mathbf{X}$ to be expanded.

A path from the initial state to the goal represents an ordering of the variables (if we go from $\mathbf{Y}$ to $\mathbf{Y} \cup \{X_i\}$, then all the variables from $\mathbf{Y}$ are predecessors of $\{X_i\}$). The utility of this path is the logarithm of the score of the best network that can be obtained restricted to this order (a variable can not be a parent of one of its predecessors). Knowing this path, we obtain the order with best score. The optimal graph can be found by considering for each variable X_i the first time this variable appears in the path from node $\mathbf{Y}$ to node $\mathbf{Y} \cup \{X_i\}$. The set of parents of X_i is the subset of $\mathbf{Y}$ for which the optimal value $BestScore(X_i, \mathbf{Y})$ is obtained.

In order to adapt this algorithm to our problem we have to compute all the graphs with an score greater than or equal to $\log(f(A))$, where A is the graph with best score. A first approximation can be to continue after the final node $\mathbf{X}$ has been expanded, and expands the nodes while $g(\mathbf{Y}) \geq \log(f(A))$ (the value of A is known after the first time the full node is expanded). In this way we obtain a set of paths, and for each path an undominated graph with the same procedure used in the optimal path. With this modification, it is necessary that if the same state is obtained with two different paths we keep the two copies of the state one for each path as they can lead to different solutions[3]. However, in this procedure we do not obtain all the undominated graphs, but the set of orders of the variables such that there is an undominated graph compatible with this order. However, only one undominated graph is obtained for each one these orders. Computing all the undominated graphs given an order can be difficult and perhaps a solution could be to decompose the problem and once an order is considered, to compute a set of different parents for each variable, for example for variable X_i we compute all the set of parents Π_i selected from the predecessors variables $\mathbf{Y}$ (the parent state), such that changing $BestScore(X_i, \mathbf{Y})$ as utility of the arc arriving to X_i by the value $LScore_i(\Pi_i|D)$ the cost of the path is greater than or equal to the threshold ($\log(f(A))$). This procedure has the advantage of decomposing the problem in local problems for each variable, and that we obtain a locally defined credal network. But not all the compatible networks are undominated: it is possible that changing the best parent for X_i or changing the best parent for X_j we obtain undominated graphs, but changing both of them the obtained graph is dominated.

A modification of the A* algorithm can be done in order to compute all the undominated graphs. For that, we change the set of states to the set of pairs $(\mathbf{Y}, \mathbf{T})$, where $\mathbf{T} \subseteq \mathbf{Y}$. $\mathbf{Y}$ will have the same interpretation as above, and $\mathbf{T}$ will be the set of parents of the last variable introduced in $\mathbf{Y}$. The problem starts with $(\emptyset, \emptyset)$ and the final states are $(\mathbf{X}, \mathbf{T})$ where $\mathbf{X}$ is the full set of variables. The children of an state $(\mathbf{Y}, \mathbf{T})$ are all the states $(\mathbf{Y} \cup \{X_i\}, \mathbf{T}')$, where $X_i \notin \mathbf{Y}$ and $\mathbf{T}' \subseteq \mathbf{Y}$. The utility of the arc going from $(\mathbf{Y}, \mathbf{T})$ to $(\mathbf{Y} \cup \{X_i\}, \mathbf{T}')$ is the logarithm of $LScore_i(\mathbf{T}'|D)$. The heuristic function on one state is computed as above $h(\mathbf{Y}, \mathbf{T}) = h(\mathbf{Y}) = \sum_{X_i \notin \mathbf{Y}} BestScore(X_i, \mathbf{Y} \setminus \{X_i\})$ The algorithm first computes the optimum value A of the score. For that, it works as the A^*, but not expanding all the nodes. $(\mathbf{Y}, \mathbf{T})$ is only expanded to nodes $(\mathbf{Y} \cup \{X_i\}, \mathbf{T})$, where $X_i \notin \mathbf{Y}$ and $\mathbf{T}$ is the subset of $\mathbf{Y} \setminus \{X_i\}$ maximizing $LScore_i(\mathbf{T}|D)$, i.e. the set of parents for which $BestScore(X_i, \mathbf{Y})$ is obtained. In this way, the behavior is very similar to the simple A* algorithm, with the only difference that here we make explicit the best set of parents for each variable. Afterwards, we compute all the undominated graphs, i.e. those with a score greater than or equal to the threshold, $\log(f(A))$. For that, for a node $(\mathbf{Y}, \mathbf{T})$ we expand all the children $(\mathbf{Y} \cup \{X_i\}, \mathbf{T}')$, such that $u(\mathbf{Y}, \mathbf{T}) + LScore_i(\mathbf{T}'|D) + h(\mathbf{Y} \cup \{X_i\}) \geq \log(f(A))$ (the cost of the path to arrive to the state plus the cost of the heuristic is above the threshold). This implies to compute all the set of parents $\mathbf{T}' \subseteq \mathbf{Y}$ with a score greater that or equal to $\log(f(A)) - u(\mathbf{Y}, \mathbf{T}) - h(\mathbf{Y} \cup \{X_i\})$ for a given variable X_i. Existing algorithms to compute $BestScore(X_i, \mathbf{Y})$ can be adapted to this task, as they make almost an exhaustive search in the set of all possible parents of X_i in $\mathbf{Y}$.

The algorithm expands a state with a new variable and a set of parents if the associated partial network could obtain a score above the threshold, by assuming an optimistic evaluation for the score of the rest of the variables $(\mathbf{X} \setminus (\mathbf{Y} \cup \{X_i\}))$. At the end, all the paths arriving to final states $(\mathbf{X}, \mathbf{T})$ represent undominated Bayesian networks (their utility is the log of the score, being the heuristic function equal to 0, so we have an exact value of the utility). In each path the graph is obtained by assigning to variable X_i the set of parents

[3] alternatively, we could maintain a unique state with a set of utilities, one of each path arriving to it. A utility value is active while the value plus the heuristic value is greater than or equal to the threshold. But for simplicity in the exposition, we shall assume that we repeat the states.

T_i in the node $(\mathbf{Y}, T_i)$ for which $X_i \in \mathbf{Y}$ for the first time (starting in the root node).

In the case of $\mathcal{D} = \mathcal{G} \times S^m$ and $\mathcal{P}_2$, with this algorithm we can compute all the graphs for which there is a vector $\mathbf{s}$ of equivalent sample sizes for which $(G, \mathbf{s})$ is undominated. If we want to compute all the undominated pairs $(G, \mathbf{s})$, we can proceed with a further modification of the A* algorithm. In this case, the set of states is the set of all triples $(\mathbf{Y}, \mathbf{T}, s)$ where now s represents the equivalent sample size with which the score of the set of parents $\mathbf{T}$ has been computed. The procedure is completely analogous to the above one (first only expanding the states with s maximizing the score) and then all the states with a cost plus heuristic above the threshold. In that case, in a path to a final state, we record the equivalent sample size of each variable and we have the pair $(G, \mathbf{s})$.

6 Experiments

We have done some experiments to illustrate the behavior of our approach. The experiments are done with the approximate algorithm for the case of $\mathcal{P}_2$ as prior probability. In both cases, we have to compute the set of graphs in which the score is greater or equal than a given threshold: $Score(G|D) \geq \beta_\epsilon$, where $Score(G|D)$ is the average score if $\mathcal{D} = \mathcal{G}$ and the maximum score if $\mathcal{D} = \mathcal{G} \times S^m$. In the last case, the decision involves the vector $\mathbf{s}$ for which the maximum score is obtained, but these data are not reported. We have considered a very well known network, the *Alarm* network [1]. This network has 37 nodes and 46 arcs. We have considered different data samples (50, 100, 500, 1000, 5000) and two different values of β_ϵ (0.8 and 0.9). For each one of them, we have computed the following values: NM number of different learned models (graphs); PI percentage of imprecise links (links appearing in some models and not in others without taking into account the direction) in relation with the total number of possible links (37*36/2); PE percentage of sure extra links (links appearing in all the learned models but not present in the original graph) with respect to the missing links in the original graph (37*36/2 - 46); PM percentage of sure missing links (links appearing in the original graph but missing in all learned networks) with respect to the number of links of the graph (46). Results are reported in Table 1.

We can observe that the imprecision decreases with the sample size (with $N = 5000$ we have almost no imprecision); it is greater if β_ϵ decreases (ϵ increases and introduce more imprecision in the ϵ-contaminated prior); and it is higher when using the *MScore* (we decide about the graph and about the equivalent sample sizes) than when using the *AScore* (we only decide about the graph).

Table 1: Results of the experimental evaluation.

Samples	50	100	500	1000	5000
AScore-0.8					
NM	67.00	22.50	10.40	4.40	3.40
PI	3.95	1.44	0.45	0.21	0.05
PE	8.50	5.82	2.18	1.58	0.76
PM	43.04	31.96	12.61	8.26	6.09
AScore-0.9					
NM	34.20	9.20	2.00	2.20	1.80
PI	1.98	0.66	0.06	0.09	0.00
PE	9.35	6.29	2.34	1.53	0.79
PM	51.22	45.78	31.65	27.26	19.30
MScore-0.8					
NM	126.20	36.70	7.10	3.30	2.10
PI	4.95	1.59	0.39	0.26	0.12
PE	6.97	5.98	2.92	2.24	0.98
PM	38.91	30.87	10.65	8.48	5.22
MScore-0.9					
NM	58.60	22.40	2.30	2.30	1.30
PI	2.37	0.87	0.11	0.15	0.05
PE	7.98	6.18	3.10	2.29	0.95
PM	42.61	32.17	11.09	7.83	5.22

On the other hand, it can be seen that the structural errors, PE and PM, strongly improve with the sample size. When β_ϵ decreases there are less missing links but more extra links (we have more models and, in consequence, more links are considered); and, at least for this BN, $MScore$ seems to obtain less structural errors than $AScore$ specially for $\beta_\epsilon = 0.9$ where PM is much higher (although more extensive experiments are needed to evaluate if this trend persists).

7 Conclusions

We have presented a general methodology for learning multigraphs credal networks. Our approach justifies the use of different scores that can be found in the literature and the fact that several networks are the result of the learning task. Even if our set of decisions consists in determining a single graph, it makes sense that the output of the learning task is a set of undominated decisions (graphs) if we have imprecise prior information. Though the problem is computationally more difficult than learning a single network, algorithms have been proposed, both for exact and approximate computation. In the future, we plan to make a more extensive experimentation including the exact algorithms and to compare with the results of learning a single network.

Acknowledgments

This research was jointly supported by the Spanish Ministry of Education and Science under projects TIN2010-20900-C04-01,02, the European Regional Development Fund (FEDER), and the Andalusian Research Program under project P08-TIC-03717. We are very grateful to the Isipta reviewers for their useful comments and suggestions.

References

[1] I.A. Beinlich, H.J. Suermondt, R.M. Chavez, and G.F. Cooper. The Alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the Second European Conference on Artificial Intelligence in Medicine*, pages 247–256. Springer-Verlag, 1989.

[2] W. Buntine. Theory refinement in Bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pages 52–60. Morgan Kaufmann, San Francisco, CA, 1991.

[3] A. Cano, M. Gómez-Olmedo, A. Masegosa, and S. Moral. Locally averaged Bayesian Dirichlet metrics for learning the structure and the parameters of Bayesian networks. *International Journal of Approximate Reasoning*, pages 526–540, 2013.

[4] G. Corani and M. Zaffalon. Credal model averaging: an extension of Bayesian model averaging to imprecise probabilities. In *ECML PKDD 2008: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 257–271. Springer, 2008.

[5] F.G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.

[6] C.P. de Campos and Q. Ji. Efficient structure learning of Bayesian networks using constraints. *Journal of Machine Learning Research*, 12:663–689, 2011.

[7] N. Friedman and D. Koller. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–125, 2003.

[8] D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[9] I. Levi. *The Enterprise of Knowledge.* MIT Press, London, 1980.

[10] A. Masegosa and S. Moral. Imprecise probability models for learning multinomial chances from data. Applications to learning credal networks. *International Journal of Approximate Reasoning*, Submitted, 2013.

[11] A.R. Masegosa and S. Moral. New skeleton-based approaches for Bayesian structure learning of Bayesian networks. *Applied Soft Computing*, 13:1110–1120, 2013.

[12] R.E. Neapolitan. *Learnig Bayesian Networks.* Prentice Hall, Upper Saddle River, 2004.

[13] J. Pearl. *Probabilistic Reasoning with Intelligent Systems.* Morgan & Kaufman, San Mateo, 1988.

[14] T. Silander and P. Myllymaki. A simple approach for finding the globally optimal Bayesian network structure. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 445–452. AUAI Press, 2006.

[15] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search.* Springer Verlag, Berlin, 1993.

[16] H. Steck. Learning the Bayesian network structure: Dirichlet prior versus data. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI2008)*, pages 511–518. AUAI Press, 2008.

[17] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 2006.

[18] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.

[19] P. Walley. Inferences from multinomial data: learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996.

[20] C. Yuan, B. Malone, and X. Wu. Learning optimal Bayesian networks using A* search. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11)*, pages 2186–2191, 2011.

[21] M. Zaffalon. Statistical inference of the naive credal classifier. In *Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, pages 384–393. Shaker Publishing, 2001.

Credal model averaging of logistic regression for modeling the distribution of marmot burrows

G. Corani
IDSIA (Switzerland)
giorgio@idsia.ch

A. Mignatti
Politecnico di Milano (Italy)
mignatti@elet.polimi.it

Abstract

Bayesian model averaging (BMA) weights the inferences produced by a set of competing models, using as weights the models posterior probabilities. An open problem of BMA is how to set the prior probability of the models. Credal model averaging (CMA) is a credal ensemble of Bayesian models, which generalizes BMA by substituting the single prior over the models by a set of priors. The base models of the ensemble are learned in a Bayesian fashion. We use CMA to ensemble base classifiers which are Bayesian logistic regressors, characterized by different sets of covariates. CMA returns indeterminate classifications when the classification is prior-dependent, namely when the most probable class depends on the prior probability assigned to the different models. We apply CMA for modelling the presence and absence of marmot burrows in an Alpine valley in Italy and show that it compares favorably to BMA.

Keywords. Bayesian model averaging, credal model averaging, logistic regression, classification, ecological modeling.

1 Introduction

Over the last years, classifiers based on imprecise probabilities have been mostly developed by extending probabilistic graphical models (see [30] for a pioneering work and [7] for a recent review) or decision trees (see [1] and the references therein). Alternatively, extension of the k nearest neighbors have been also proposed [11].

In this paper we consider the idea of credal model averaging (CMA) [8, 6], which can be described as a credal ensemble of Bayesian classifiers. In other words, the parameters of the base models are learned in a Bayesian way. The ensemble of the base models is instead carried out in an imprecise way, modelling a condition of ignorance about the prior probability of the different models.

Model uncertainty is the problem of many models being consistent with the available data. In this condition, there is substantial uncertainty about which model should be chosen for drawing inferences or computing predictions. Choosing a single model and then ignoring the substantial uncertainty of the model selection leads to overconfident inferences [3]. Bayesian model averaging (BMA) is a sound approach to deal with model uncertainty, based on the key idea of averaging the inferences produced by a set of different models, using the models' posterior probabilities as weights.

However, BMA requires to specify the prior probability of each model. This is a critical issue, as it is recognized in the BMA literature [5]. To tackle this issue, some authors repeat the BMA analysis assigning different prior probabilities to the models [21, 28]. From the viewpoint of the credal classification, it is well-known the relying on a single prior implies unavoidable arbitrariness, which entails the risk of drawing prior-dependent classifications.

CMA generalizes BMA, overcoming the problem of the prior specification by adopting a *set* of prior over the models. As a result, the posterior probability of the models lies within an interval rather being a punctual value. Moreover, CMA automatically detects the instances which are prior-dependent, namely whose most probable class varies depending on the prior probability assigned to the different models. On such instances, CMA suspends the judgment by returning more than one class, thus automating the sensitivity analysis. So far, CMA has been proven effective in ensembling probabilistic graphical models [8, 6].

We develop CMA for ensembling logistic regression models characterized by different feature sets. Indeed, BMA of logistic regressors was already used to model presence or absence of ecological populations [21, 28]; we then compare BMA and CMA on the case study of predicting the presence of marmot burrows in an Alpine valley.

2 Bayesian model averaging (BMA)

Let us consider a logistic regression model for predicting the value of a binary class C, with classes c_0 and c_1. The set of *covariates* (or *features*) is $\mathcal{X} = \{X_1, X_2, \dots X_k\}$; in a generic instance, the value of the covariates is $\mathbf{x} = x_1, \dots, x_k$. We denote $\pi_0 = P(C = c_0|\mathbf{x})$ and $\pi_1 = P(C = c_1|\mathbf{x})$. The logistic regression model is

$$y = \text{logit}(\pi_0) = ln\frac{\pi_0}{1-\pi_0} = ln\frac{\pi_0}{\pi_1} = \\ = \beta_0 + \sum_{j=1}^{j=k} \beta_j x_j \qquad (1)$$

where x_j is the observation of X_j.

Given k covariates, the model space $\mathcal{M}$ is composed of 2^k possible model *structures.* Each model structure includes a specific set of covariates. We denote by m_i the i-*th* structure. The model size is defined as the number of covariates included in the structure.

Feature selection is the problem of identifying the supposedly best set of covariates for the model. The traditional feature selection approach is to assess the significance of each covariate through hypothesis tests [10]. More modern approaches for feature selection are instead based on the so-called Information Criteria [3], such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC)[1]. Information Criteria have been recognized to be more effective than repeated hypothesis tests for the purpose of feature selection [3]. Yet, even adopting Information Criteria one could face the problem of *model uncertainty.* If, for instance, different models obtain a similar value of BIC, a substantial uncertainty underlies the choice of a single model. The subsequent inferences are hence overconfident if this uncertainty is disregarded.

BMA addresses model uncertainty by combining the inferences of multiple models, using as weights the posterior probability of the models. We denote by D the available dataset, by $P(m_i|D)$ the posterior probability of model m_i and by $P(Y|D)$ the entire posterior distribution of Y given D, from which posterior probabilities $P(y|D)$ of a specific value y can be obtained. The posterior of Y under BMA is [5]:

$$P(Y|D) = \sum_{m_i \in \mathcal{M}} P(Y|m_i, D)P(m_i|D) \qquad (2)$$

where:

$$P(m_i|D) = \frac{P(m_i)P(D|m_i)}{\sum_{m_k \in \mathcal{M}} P(m_k)P(D|m_k)}$$

$$P(D|m_i) = \int P(D|\boldsymbol{\beta_i}, m_i)P(\boldsymbol{\beta_i}|m_i)d\boldsymbol{\beta_i},$$

having denoted by $P(m_i)$ the prior probability of model m_i, $\boldsymbol{\beta_i}$ the vector of its parameters and $P(D|m_i)$ its marginal likelihood, which in the linear case can be exactly computed [25]. Equation (2) requires an extensive summation over 2^k models, which is usually carried out by sampling the model space . Only for small k it is possible to exhaustively treat the model space.

As a result of averaging across different models, $P(Y|D)$ is given by a sum of distributions and thus has a multi-modal shape. Inferences about other quantities of interest such as the parameter of the models can be obtained by averaging over the models as in Eq(2).

BMA requires to set a precise prior over the parameters and over the models. As a prior distribution on the parameters $P(\boldsymbol{\beta_i}|m_i)$ we adopt Zellner's g-prior [13], setting g equal to the number of observations. As for the prior over the

[1] The BIC provides a simple but effective approximation of the posterior probability of a given model [24].

models, we adopt the binomial prior [25, 13]; namely, every covariate has the same prior probability θ of being included in the model; moreover, the probability of inclusion of each covariate is *independent*. Thus, the prior probability of model m_i, which includes a number k_i of covariates, is:

$$P(m_i) = \theta^{k_i}(1-\theta)^{k-k_i}. \tag{3}$$

Once the prior probability of each possible model is specified according to Eq.3, it can be analyzed the prior distribution of the random variable constituted by the *model size*, namely the number of covariates included in the model. The model size follows a binomial distribution with mean θk and variance $\theta(1-\theta)k$ [19], where k is the total number of available covariates. An easy way to elicit the prior distribution over the models is to ask the expert his beliefs about the model size.

3 Credal Model Averaging(CMA)

CMA generalizes BMA by substituting the *single* binomial prior over the models by a *set* of binomial priors: thus, the prior probability of inclusion of each covariate varies within the range $[\underline{\theta}, \overline{\theta}]$; thus, the mean model size a priori varies within the range $[\underline{\theta}k, \overline{\theta}k]$. Thus, CMA allows eliciting from the expert an *upper* and a *lower* model size. If no expert is available, one can model a situation of ignorance a priori, by setting $\underline{\theta} = \epsilon$ and $\overline{\theta} = 1-\epsilon$. In our experiments we adopt this approach, setting ϵ=0.05.

Each model of the ensemble is learned in a Bayesian fashion, using a *precise* prior over the parameters. Instead, the prior probability of the models is imprecisely modelled. Hence CMA is a *credal ensemble of Bayesian models*. Because of imprecision, CMA computes for the logit the interval $[\underline{y}, \overline{y}]$ rather than a point value as in traditional logistic regression. The length of such interval varies instance by instance, showing the sensitivity of the prediction on the priors which has been set over the models, namely how much the BMA prediction would vary as a consequence of θ varying between $\underline{\theta}$ and $\overline{\theta}$. No coverage probability can be assigned to the CMA intervals. To compute $\overline{y}$ and $\underline{y}$, CMA solves a maximization and a minimization problem on each instance. Since the prior probability of inclusion θ is equal for all covariates, the optimization problem involves only a single variable.

Let us focus on the minimization case. We denote by a hat the estimated values. Given the data set D and the observation $\mathbf{x} = x_1, \ldots, x_k$ of the covariates, we denote the prediction of model m_i as $\hat{y}_i = \beta_0^i + \sum_{j=1}^{j=k_i} \beta_j^i x_j$, where β_0^i and β_j^i denote the parameters of m_i (the previous formula assumes, with no loss of generality, that for model m_i the covariates have been re-ordered, so that the first k_i covariates are those included in the model). For simplicity of notation we do not indicate the dependence of $\hat{y}_i$ on D and $\mathbf{x}$. The lower bound $\underline{\hat{y}}$ of the CMA interval is computed as:

$$\underline{\hat{y}} = \min_{\theta \in [\underline{\theta}, \overline{\theta}]} \sum_{m_i \in \mathcal{M}} \hat{y}_i P(m_i|D) =$$

$$\min_{\theta \in [\underline{\theta}, \overline{\theta}]} \sum_{m_i \in \mathcal{M}} \hat{y}_i \frac{P(D|m_i)P(m_i)}{\sum_{m_j \in \mathcal{M}} P(D|m_j)P(m_j)} =$$

$$= \min_{\theta \in [\underline{\theta}, \overline{\theta}]} \frac{\sum_{m_i \in \mathcal{M}} \hat{y}_i P(D|m_i)\theta^{k_i}(1-\theta)^{k-k_i}}{\sum_{m_j \in \mathcal{M}} P(D|m_j)\theta^{k_j}(1-\theta)^{k-k_j}}$$

$$:= \min_{\theta \in [\underline{\theta}, \overline{\theta}]} h(\theta)$$

Let us define the k sets $\mathcal{M}_1 \ldots \mathcal{M}_k$ which include all the models containing respectively $\{1, 2, \ldots, k\}$ covariates. For instance, $\mathcal{M}_2$ contains all the models which include two covariates. To address the optimization problem it is useful noting that all the models contained in the set $\mathcal{M}_j$ have the same prior probability $\theta^j(1-\theta)^{k-j}$. We introduce $Z_j = \sum_{m_v \in \mathcal{M}_j} \hat{y}_v P(D|m_v)$ and $L_j = \sum_{v \in \mathcal{M}_j} P(D|m_v)$ and then rewrite function $h(\theta)$ as:

$$h(\theta) = \frac{\sum_{j=0}^{k} \theta^j (1-\theta)^{k-j} Z_j}{\sum_{j=0}^{k} \theta^j (1-\theta)^{k-j} L_j} \tag{4}$$

In the interval $[\underline{\theta}, \overline{\theta}]$, the maximum and minimum of $h(\theta)$ should lie either in the boundary points $\theta = \overline{\theta}$ and $\theta = \underline{\theta}$, or in an internal point of the interval in which the first derivative of $h(\theta)$ is 0. Let us introduce $f(\theta) = \sum_{j=0}^{k} \theta^j(1-\theta)^{k-j} Z_j$ and $g(\theta) = \sum_{j=0}^{k} \theta^j(1-\theta)^{k-j} L_j$. The first derivative $h'(\theta)$ is:

$$h'(\theta) = \frac{f'(\theta)g(\theta) - f(\theta)g'(\theta)}{g(\theta)^2}, \tag{5}$$

where $g(\theta)$ is strictly positive because L_j is a sum of marginal likelihoods. We can therefore search the solutions looking only at the numerator $f'(\theta)g(\theta)-f(\theta)g'(\theta)$, which is a polynomial of degree $k(k-1)$ and thus has $k(k-1)$ solutions in the complex plain. We are interested only in the *real* solutions that lie in the interval $(\underline{\theta}, \overline{\theta})$. Such solutions, together with the boundary solutions $\theta = \overline{\theta}$ and $\theta = \underline{\theta}$, constitute the set of *candidate solutions*. To find the minimum and the maximum $h(\theta)$, we evaluate $h(\theta)$ in each candidate solution point, and eventually we retain the minimum and maximum among such values.

Having determined the upper and lower logit values $\underline{y}$ and $\overline{y}$, we obtain the upper and lower posterior probabilities of the two classes by inverting Eq.(1):

$$\begin{aligned}\overline{\pi}_0 &= \frac{\exp(\overline{y})}{1+\exp(\overline{y})}\\ \underline{\pi}_0 &= \frac{\exp(\underline{y})}{1+\exp(\underline{y})}\\ \overline{\pi}_1 &= 1-\underline{\pi}_0\\ \underline{\pi}_1 &= 1-\overline{\pi}_0\end{aligned}$$

CMA adopts the criterion of *interval-dominance* [27] to take decisions: class c_1 is returned if $\underline{\pi}_1 > \overline{\pi}_0$, namely if $\underline{\pi}_1 > 1/2$. Conversely, class c_0 is returned if $\underline{\pi}_0 > 1/2$. In these cases the instance is *safe* because the rank between the two classes is the same regardless the prior probabilities assigned to the competing models. If instead the intervals of the posterior probability of the two classes overlap, the judgment is suspended. The instance is *prior-dependent*, since the rank among the classes changes when different prior probabilities are assigned to the competing models.

A final note regards the relation between the logit computed by BMA and CMA. If the value of θ used to induce BMA is included in the interval $[\underline{\theta}, \overline{\theta}]$ used to induce CMA, the logit computed by BMA is included within the the logit interval computed by CMA. Thus when CMA returns a single class, this is the same class returned by BMA.

4 Case study

The study area is located in the Italian Alps, near the Stelvio National Park. The valley has an altitude comprised between 2100 and 3100 m above sea level. The field surveys identified the position of the Alpine marmot burrows and the characteristics of their surrounding territory. The censuses were carried out in the summers 2010 and 2011; three different areas of the valley were investigated. To develop the species distribution model we divide the area into cells of 100m^2; the censused area is overall of about 95 ha (9500 cells). Presence of burrows has been detected in about 4.5% of the cells. Each cell is then labelled as presence or absence.

The considered covariates are altitude, slope, aspect (the direction in which the slope faces) topographic ruggedness index (TRI) [26], hillshade, curvature, soil temperature and soil cover. For the aspect, we did not directly use the angle from North, but we divided the information into two sub-variables that we called *northitude* and *eastitude*. The *northitude* is calculated as the cosine of the angle from North, while the *eastitude* is calculated as the sine of the same quantity. While the former represents the attitude of the marmot to select sunny slopes, the latter represents the preference to have a sunny territory during the sunrise and the morning rather than during the sunset and the evening. To build the soil temperature map we relied on five different meteorological stations (altitude comprised between 1800 and 2600 m a.s.l.) located in the surroundings, which provide the data of air temperature and snow depth. The soil temperature is a mean yearly value and was calculated starting from the DEM (digital elevation model) and the data of air temperature and snow depth, using the model developed by [14]. Finally, we express the soil cover as the percentage of cells with debris and outcrops cover in the buffer area (see later for an explanation of the buffer area).

As a pre-processing step we removed some highly cross-correlated ($|\rho| > .8$) covariates: more precisely the soil temperature (anti-correlated with the altitude), the TRI (anti-correlated with the slope) and the hillshade, correlated with the *northitude*.

The Alpine marmot is a mobile species, which uses a huge territory for its activities. Thus, we supposed that the decision to dig a burrow in a given cell does depend also on the environmental conditions of surrounding cells. For this reason, the value associated to each cell (for each environmental variable) is calculated as the mean of the values of the variable in a surrounding of the same cell. We refer to this area with the term *buffer area*, and, in our case, it has a pseudo-circular shape, since we considered the cells within a circular area built around the given cell. The home range of the Alpine marmot ranges between 1 and 3 ha [23, 17]. We considered buffer areas of size 1 ha, 2 ha and 3 ha. Since the results are quite consistent when different buffer areas are considered, in the following we present results referring only to a buffer area of 2 ha.

5 Results

To gain understanding of the data and to investigate the role of the different covariates, we develop a BMA model using the entire dataset. The prior probability of inclusion of the covariates is set to 0.5, corresponding to a uniform prior probability over the models.

Under BMA the posterior probability of inclusion of a covariate is calculated as the sum of the posterior probability of the models in which the covariate is included. In Table 1 we report the posterior probability of inclusion of the covariates, the expected values and the standard deviations of the parameters of the models, obtained using the standardized values of the variables. The expected values and the standard deviations of the coefficients are calculated averaging over the models which do include the covariate.

Variable		2ha	
	p.inc.	EV	SD
altitude	1	-1.050	0.158
slope	1	0.491	0.067
curvature	0.02	0.001	0.011
northitude	1	-1.381	0.010
eastitude	1	-0.553	0.056
% of outcrops and debris cover	0.97	-0.399	0.122

Table 1: Posterior probability of inclusion of the covariates (p.inc), expected values (EV) and standard deviations (SD) of the model parameters.

The most important variables are the altitude, the slope, the *eastitude* and the *northitude*. The signs of the parameters confirm, for most of the variable, what is reported in literature. The coefficient of the altitude has a negative value, and the valley altitude ranges from ca. 2200 m a.s.l. and 3000 m a.s.l.. The suitable altitude for the marmot is approximately between 1650 m a.s.l. and 1950 m a.s.l. [4, 2] with maximum altitudes around 3000 m a.s.l.. Since the valley is above the optimal altitude range of the marmot, the fact that the suitability of the valley decreases with the altitude confirm the past results. The slope positively influences the presence of burrows. In this case, we have conflicting results reported in literature, with an optimal slope that varies from 0 to 60°[22]. The *northitude* negatively influences the presence of burrows, so that the marmot preference is for southerly exposed slopes, as previously reported in several studies [2]. The *eastitude* negatively influences the presence of burrows, contrary to what is reported in literature [2], with a preference for the westerly exposed slopes in the valley. This preference is probably due to the fact that, in the valley, the areas located at a higher elevation and with a low suitability, are mainly westerly exposed. This result seems therefore to be mainly due to the valley shape. A high percentage of outcrops and debris cover negatively influences the presence of marmot burrows, showing the importance of the alpine meadows for the species, as reported by [2, 22].

5.1 Comparing BMA and CMA

We compare BMA and CMA using training data sets of varying sample size. For comparing BMA and CMA, we downsample the original data set, generating training sets of size $n \in \{30, 60, 90 \ldots, 300\}$. For each sample size, we build 30 different training sets. The instances not contained in the training set constitute the test set. The training sets contain the same prevalence (fraction of presence data) of the entire dataset, namely 4.6%. For CMA we assume a situation of substantial ignorance a priori, setting $\overline{\theta} = 0.95$ and $\underline{\theta} = 0.05$.

CMA can be seen as dividing the instances into two groups: the *safe* ones, for which a single class is returned, and the *prior-dependent* ones, for which instead the judgment is suspended and both classes are returned. For the prior-dependent instances, presence or absence is more probable depending on the prior probability of the competing models.

The most common measure of performance in classification is the *accuracy*, defined as the proportion of instances correctly classified. To evaluate the effectiveness of CMA, we assess the accuracy of BMA on the safe and on the prior-dependent instances. As can be seen in Fig.1, BMA undergoes a sharp drop of accuracy on the instances indeterminately classified by CMA.

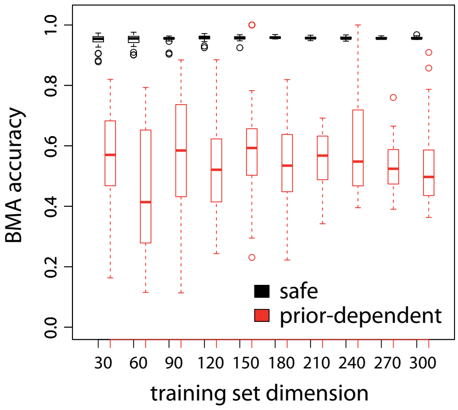


Figure 1: The accuracy of BMA drops on the prior-dependent instances. For each sample size, the boxplot refers to 30 experiments.

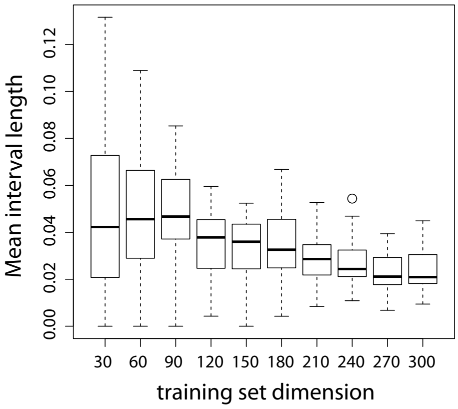


Figure 2: The length of the CMA interval $(\underline{\pi}_0, \overline{\pi}_0)$ decreases with the sample size. For each sample size, the boxplot refers to 30 experiments.

The length of the logit interval $[\underline{y}, \overline{y}]$ of CMA decreases with the dimension of the training set as shown in Figure 2: the larger the sample size, the less influential the prior probability of the models.

5.2 Credal Classification and Reject Option

Traditional classifiers can be equipped with a *reject option* [16], thus refusing to classify an instance if the posterior probability of the most probable class is below a certain threshold. To adopt the reject option, it is necessary setting the *rejection cost* which is incurred into when rejecting an instance. When classifying an instance, the *expected cost* [12] associated to decision of returning each class is computed. The instance is rejected if the expected classification cost of each class is higher than the rejection cost. This corresponds to rejecting all the instances in which the posterior probability p^* of the most probable class is below a threshold t [16].

However, the behavior induced by the reject option is quite different from that of a credal classifier. On a *large* data set the posterior probability of the classes is *not* sensitive on the choice of the prior; a credal classifier will generally return a single class. On the other hand, the determinate classifier could reject even a considerable number of instances, if the rejection cost is small. To fairly compare a traditional classifier equipped with rejection option against a credal classifier, it would be necessary making the credal classifier aware of the rejection cost. This point we leave for future research.

However, applying a *rejection option* to BMA does in general yield a behavior which is quite different from that of CMA. The point is that on the prior-dependent instances the BMA predictions are *not tightly* distributed around a 50% posterior probability; instead, there are many prior-dependent instances in which BMA estimates a posterior probability larger than 60-70% for the most probable class: see for an example Figure 3. Thus, BMA equipped with rejection option would reject only part of the prior-dependent instances. Conversely, it will instead reject some instances which are not prior-dependent.

5.3 Utility-discounted accuracy

To further compare CMA and BMA we adopt the utility-discounted accuracy introduced in [29]. We briefly summarize here the idea underlying this approach. The starting point is the *discounted accuracy*, which rewards a prediction

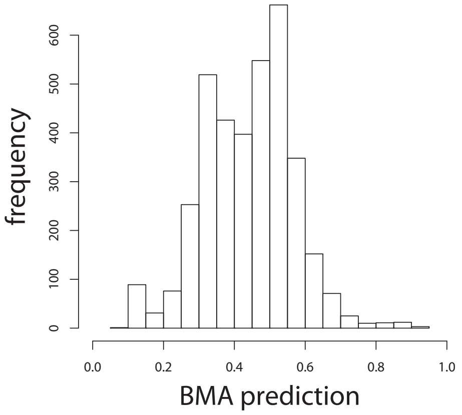


Figure 3: Distribution of the posterior probability associated by BMA to the most probable class in the *prior-dependent* instances. The figure refers to a training set of dimension n=210.

containing m classes with $1/m$ if it contains the true class, and with 0 otherwise. Within a betting framework based on fairly general assumptions, discounted-accuracy is the only score which satisfies some fundamental properties for assessing both determinate and indeterminate classifications; thus, the discounted accuracy of a credal classifier can be compared to the accuracy achieved by a determinate classifier. Yet discounted-accuracy has severe shortcomings. Consider two medical doctors, doctor *random* and doctor *vacuous*, who should diagnose whether a patient is *healthy* or *diseased*. Doctor *random* issues uniformly random diagnosis; doctor *vacuous* instead always returns both categories, thus admitting his/her ignorance. Let us assume that the hospital profits a quantity of money proportional to the discounted-accuracy achieved by its doctors at each visit. Both doctors have the same *expected* discounted-accuracy for each visit, namely 1/2. For the hospital, both doctors provide the same *expected* profit from each visit, but with a substantial difference: the profit of doctor vacuous has no variance. Any risk-averse hospital manager should thus prefer doctor vacuous over doctor random: under risk-aversion, the expected utility increases with expectation of the rewards and decreases with their variance [18]. To model this fact, it is necessary to apply a utility function to the discounted-accuracy score assigned to each instance. The utility function is designed as follows in [29]: the utility of a correct and determinate classification (discounted-accuracy 1) is 1; the utility of a wrong classification (discounted-accuracy 0) is 0. Therefore, the utility of a traditional determinate classifier corresponds to its accuracy. The utility of an accurate but indeterminate classification consisting of two classes (discounted-accuracy 0.5) is assumed to lie between 0.65 and 0.8. Two quadratic utility functions are then derived corresponding to these boundary values, and passing respectively through $\{u(0) = 0, u(0.5) = 0.65, u(1) = 1\}$ and $\{u(0) = 0, u(0.5) = 0.8, u(1) = 1\}$, denoted as u_{65} and u_{80} respectively. Since $u(1) = 1$, utility and accuracy coincide for determinate classifiers; therefore, utility of credal classifiers and accuracy of determinate classifiers can be directly compared. Interestingly, the u_{65} and u_{80} functions provides score which are numerically close to respectively the F_1 and F_2 metric, which have been used to score indeterminate classifications in [9], adopting an approach based on information retrieval.

In Figure 4 we compare the CMA utility (calculated using the u_{80} utility function) and the BMA accuracy. The utility produced by CMA is slightly higher on average than that of BMA; however the most striking feature of Fig.4 is that the CMA boxplots are much tighter than the BMA ones. This means that the utility yielded by CMA is not only higher on average, but also much more stable and predictable than that of BMA. The result do not change substantially if the u_{65} utility function is considered instead, apart from a slight shift downwards of the CMA boxplots.

5.4 The cost-sensitive setup

The classes of our problem are strongly skewed: about 4.5% and 95.5% of the instances are respectively presence and absence. It is unlikely that the two different kind of errors (false presence and false absence) have identical costs, as it is assumed by both the classification accuracy and the utility-discounted accuracy. To make the assessment more realistic, it is thus worth considering a cost-sensitive setup.

A simple measure of performance which accounts for costs is the AUC [20], namely the area under the *receiver operating characteristic* (ROC)

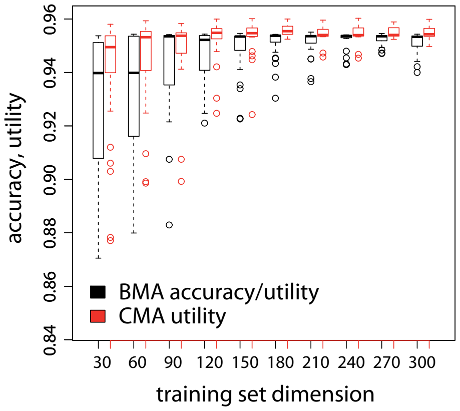


Figure 4: CMA utility compared to BMA accuracy, using the u_{80} utility function.

curve. Figure 5 shows that BMA achieves much higher AUC on the safe instances (determinately classified by CMA) than on the prior-dependent ones (indeterminately classified by CMA). This is a further favorable result for CMA.

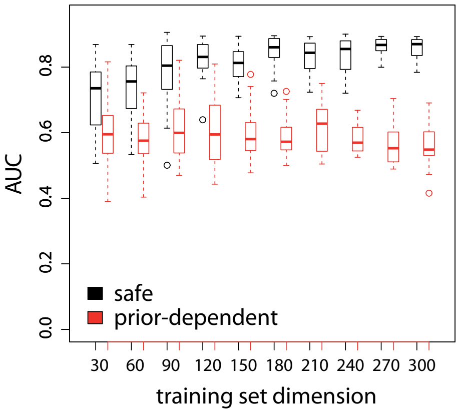


Figure 5: The AUC of BMA drops on the prior-dependent instances.

Yet, the AUC summarizes into a single scalar the whole area under the ROC curve, mixing the performance obtained under very different cost scenarios [20]. To provide a more detailed picture of the behavior of the classifier in the cost-sensitive setup, we then follow the approach of [12]. We introduce the *cost matrix*; in particular, we denote by $d(c_i, c_j)$ the cost of predicting class c_i when the actual class is c_j. The cost matrix is 2x2 since the problem has two classes (presence and absence), as shown in Table 2. Let us assume that the model is used to predict the presence/absence of burrows in a territory that has not yet been censused. If the model predicts the presence of a burrow in a given cell, an operator is sent to search for burrows, incurring the cost κ (this is a simplification, since the cost could vary for instance with the position of the cell to be surveyed). If a burrow is found, a gain ζ is obtained; overall, the negative cost (namely the reward) for having correctly predicted the presence is $\kappa - \zeta < 0$. If absence is predicted no survey is organized; thus, no costs are incurred regardless whether the considered cell contains or not a burrow.

	Actual	
Predicted	Absence	Presence
Absence	0	0
Presence	κ	$\kappa - \zeta$

Table 2: Cost matrix.

In the cost-sensitive setup, the classifier should return the class with the lowest expected cost rather than the most probable class. The expected cost of predicting class c_i is $\sum_{c_j \in \mathcal{C}} \pi_j d(c_i, c_j)$, where $\mathcal{C}$ denotes the set of classes and π_j is the posterior probability of class c_j, computed according to the logistic regression model. Given the above cost matrix, the expected cost of predicting absence is 0. Thus presence is predicted if the expected cost of doing so is negative:

$$\begin{aligned}\text{Expected cost (predicting presence)} < 0 \Leftrightarrow \\ \pi_1(\kappa - \zeta) + \pi_0(\kappa) < 0 \Leftrightarrow \\ \kappa - \pi_1\zeta < 0\end{aligned}$$

In other words, presence is predicted if its posterior probability is higher than the threshold $t = \kappa/\zeta$. Dealing with CMA, in some instances the posterior probability of presence might fluctuate below and above the threshold t depending on the prior probability assigned to the competing models. In this case, the decision should be suspended since the evidence coming from data is not strong enough to take a decision. However, we want CMA to take a decision. To this purpose, we consider the Γ-maximin approach [27], namely worst-case optimisation; this implies returning a prediction of *absence* on the prior-dependent instances. We also consider the opposite approach Γ-maximax, namely optimization of the best case; this implies returning a prediction of *presence* on the prior-dependent instances.

We perform experiments with different values of the threshold $t = \kappa/\zeta$. Moreover, to compare the results obtained with different t, we fixed $\zeta = 1$; it is indeed easy to prove that ζ is only a multiplicative factor in the computation of the total cost, so that its value does not influence the quality of the results. In Figure 7 we report the results for the case t=0.5 ($\zeta = 2\kappa$) and t=1/23 ($\zeta = 23\kappa$). The latter value, in which the threshold equals the marginal probability of presence, is referred to as Kolmogorov-Smirnov statistic in [15]. Given the rarity of presence, we do not consider values of ζ smaller than 2κ, namely $t > 0.5$. Figures 6 and 7 show the results for the prior-dependent instances only; on the instances which are not prior-dependent, BMA and CMA take the same decisions and thus incur the same costs. Given the cost matrix of Table 2, the Γ-maximin strategy incurs a cost of 0 on the prior-dependent instances. In the case $t = 1/2$ (Figure 6), Γ-maximin incur lower costs than if the decision is taken according to the single posterior probability computed by BMA. The highest costs are instead incurred adopting the Γ-maximax strategy. However, the situation is reversed in the case t=1/23 (Figure 7): Γ-maximax incurs the lowest costs, followed by BMA; Γ-maximin incurs instead the highest costs. Interestingly the differences among the costs incurred by the various policies generally decrease with the size of the training set. For the case $t = 1/5$ (not shown) the costs of all policies are almost equivalent, lying close to 0.

It cannot be predicted whether deciding according to either Γ-maximin or Γ-maximax will eventually incur lower or higher total costs, for the prior-dependent instances, than deciding according to BMA. Our viewpoint is that on the prior-dependent instances taking a decision should be preferably avoided, trying instead to acquire new information.

6 Conclusions

CMA has proven effective on the real-world case study of predicting the presence of the Alpine marmot. Some future extensions can be foreseen. The first is adopting maximality rather than interval-dominance for detecting the prior-dependent instances; this should decrease the number of instances indeterminately classified without compromising the robustness of the classifications. Secondly, one could allow the prior probability of inclusion of each covariate to vary within a different interval; this would however imply solving a more complex optimization problem to detect the upper and lower bounds of the logit interval. Eventually the current algorithms could be extended to deal with more than two classes; for this purpose, the base classifiers to be ensembled should be polytomous (rather than dychotomous) logistic regressors.

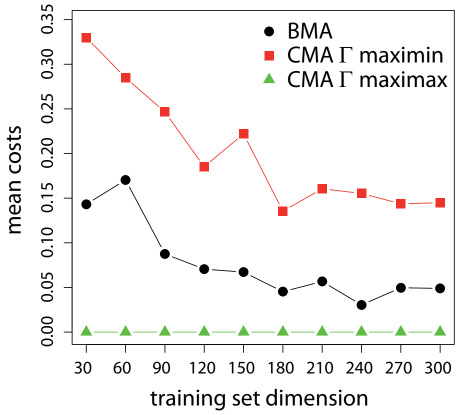


Figure 6: Mean costs incurred on the *prior-dependent* instances ($t = 1/2$).

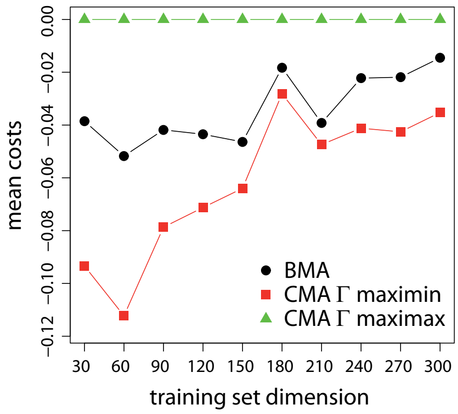


Figure 7: Mean costs incurred on the *prior-dependent* instances ($t = 1/23$).

Acknowledgments

The research in this paper has been partially supported by the Swiss NSF grants no. 200020-132252. The work has been performed during Andrea Mignatti's PhD, supported by Fondazione Lombardia per l'Ambiente (project SHARE- Stelvio). We thank the anonymous reviewers for their valuable insights.

References

[1] J. Abellán, R. Baker, F. Coolen, R. Crossman, and A. Masegosa. Classification with decision trees from a nonparametric predictive inference perspective. *Computational Statistics & Data Analysis*, in press, doi=10.1016/j.csda.2013.02.009, 2013.

[2] A Borgo. Habitat requirements of the Alpine marmot Marmota marmota in reintroduction areas of the Eastern Italian Alps. Formulation and validation of habitat suitability models. *Acta Theriologica*, 48(4):557–569, 2003.

[3] Kenneth P Burnham and David R Anderson. *Model selection and multi-model inference: a practical information-theoretic approach.* Springer, 2002.

[4] M. Cantini, C. Bianchi, N. Bovone, and D. Preatoni. Suitability study for the alpine marmot (marmota marmota marmota) reintroduction on the Grigne massif. *Hystrix, the Italian Journal of Mammalogy*, 9(1-2), 1997.

[5] M. Clyde and E. George. Model Uncertainty. *Statistical Science*, 19:81–94, 2004.

[6] G Corani and A Antonucci. Credal ensembles of classifiers. *Computational Statistics & Data Analysis*, in press, doi=10.1016/j.csda.2012.11.010, 2012.

[7] G Corani, A Antonucci, and M Zaffalon. Bayesian networks with imprecise probabilities: Theory and application to classification. In *Data Mining: Foundations and Intelligent Paradigms*, pages 49–93. Springer, 2012.

[8] G. Corani and M. Zaffalon. Credal model averaging: an extension of Bayesian model averaging to imprecise probabilities. *Proc. ECML-PKDD 2008 (Eur. Conf. on Machine Learning and Knowledge Discovery in Databases)*, pages 257–271, 2008.

[9] J. Del Coz and A. Bahamonde. Learning nondeterministic classifiers. *The Journal of Machine Learning Research*, 10:2273–2293, 2009.

[10] Alfred DeMaris. A tutorial in logistic regression. *Journal of Marriage and the Family*, pages 956–968, 1995.

[11] S. Destercke. A k-nearest neighbours method based on lower previsions. In *Proc. IPMU 2010 (Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods)*, pages 129–138. Springer, 2010.

[12] C. Elkan. The foundations of cost-sensitive learning. *Proc. Int. Joint Conference on Artificial Intelligence (IJCAI - 01)*, pages 973–978, 2001.

[13] C. Fernandez, E. Ley, and M. Steel. Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100(2):381–427, 2001.

[14] M. Guglielmin. Permaclim: a model for the distribution of mountain permafrost, based on climatic observations. *Geomorphology*, 51(4):245–257, April 2003.

[15] D. Hand. Evaluating diagnostic tests: The area under the ROC curve and the balance of errors. *Statistics in medicine*, 29(14):1502–10, 2010.

[16] Radu Herbei and Marten H Wegkamp. Classification with reject option. *Canadian Journal of Statistics*, 34(4):709–721, 2006.

[17] D. Lenti Boero. Long-term dynamics of space and summer resource use in the alpine marmot (Marmota marmota L.). *Ethology Ecology & Evolution*, 15(4):309–327, 2003.

[18] Haim Levy and Harry M Markowitz. Approximating expected utility by a function of mean and variance. *The American Economic Review*, 69(3):308–317, 1979.

[19] E. Ley and M.F.J. Steel. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24(4):651–674, 2009.

[20] Charles X Ling, Jin Huang, and Harry Zhang. Auc: a statistically consistent and more discriminating measure than accuracy. In *Proceedings of the 18th international joint conference on Artificial intelligence*, pages 519–524. Morgan Kaufmann Publishers Inc., 2003.

[21] W.A. Link and R.J. Barker. Model weights and the foundations of multimodel inference. *Ecology*, 87(10):2626–2635, 2006.

[22] B.C. López, I. Figueroa, J. Pino, A. López, and D. Potrony. Potential distribution of the alpine marmot in Southern Pyrenees. *Ethology Ecology & Evolution*, 21(3-4):225–235, 2009.

[23] C. Perrin and D. Berre. Socio-spatial Organization and Activity Distribution of the Alpine Marmot Marmota marmota: Preliminary Results. *Ethology*, 93:21–30, 1993.

[24] Adrian E Raftery. Bayesian model selection in social research. *Sociological methodology*, 25:111–164, 1995.

[25] A.E. Raftery and D. Madigan. Bayesian model averaging for linear regression models. *Journal of the American Statistical*, 92(437):179– 191, 1997.

[26] S. J Riley, S.D. DeGloria, and R. Elliot. A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of sciences*, 5(1-4):23–27, 1999.

[27] M. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007.

[28] B. Wintle, M. McCarthy, C. Volinsky, and R. Kavanagh. The use of Bayesian model averaging to better represent uncertainty in ecological models. *Conservation Biology*, 17(6):1579–1590, 2003.

[29] M. Zaffalon, G. Corani, and D. Maua. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282 – 1301, 2012.

[30] Marco Zaffalon. The naive credal classifier. *Journal of statistical planning and inference*, 105(1):5–21, 2002.

Coherent updating of 2-monotone previsions

Enrique Miranda
Dep. of Statistics and Operations Research
University of Oviedo
mirandaenrique@uniovi.es

Ignacio Montes
Dep. of Statistics and Operations Research
University of Oviedo
imontes@uniovi.es

Abstract

The conditions for a 2-monotone lower prevision to be uniquely updated to a conditional lower prevision are determined. Then a number of particular cases are investigated: completely monotone lower previsions, for which equivalent conditions in terms of the focal elements of the associated belief function are established; random sets, for which some conditions in terms of the measurable selections can be given; and minitive lower previsions, that are shown to correspond to the particular case of vacuous lower previsions.

Keywords. Coherent lower previsions, n-monotonicity, belief functions, minitive measures, natural extension, regular extension.

1 Introduction

The theory of imprecise probabilities contains a wide variety of mathematical models that are of interest in situations where it is unfeasible to determine the probability model associated to an experiment with certain guarantees. Under any of them, one important problem is that of updating the model under the light of new information. Unfortunately, this problem is far from settled, and quite a number of different rules have been proposed. Out of them, arguably some of the most popular are Dempster's rule of conditioning [11], regular extension [4] and natural extension [27].

In order to be able to choose one rule above the others, it is essential to have a clear interpretation of the mathematical model we are using. In this paper, we shall consider the behavioural approach championed by Peter Walley [27], that has its roots in the works on subjective probability by Bruno de Finetti [10]. This approach regards lower and upper probabilities as supremum and infimum betting rates, and focuses on a consistency notion between these betting rates called *coherence*.

When we move to the conditional case, there is also a notion of coherence that tells us if the conditional betting rates are compatible with the unconditional ones. However, this notion does not suffice to uniquely determine the conditional models from the unconditional ones. This was showed for instance in [20], where it was established that in general we may have an infinite number of conditional models compatible with the unconditional one, and that the smallest and greatest such models are determined by the procedures called *natural* and *regular* extension, respectively. In this paper, we investigate under which conditions there is only one conditional model that is coherent with the unconditional one.

Walley's theory is established in terms of lower and upper *previsions* (or expectations), because these are more informative than the lower and upper probabilities that can be considered as a particular case. We shall recall the basics from the theory of coherent lower previsions in Section 2. Then we shall focus on a particular case of coherent lower previsions: those satisfying the property of 2-monotonicity [2, 7]. Lower previsions with this property have the advantage of being uniquely determined by their restrictions to events (a 2-monotone lower probability) by means of the Choquet integral.

After establishing a necessary and sufficient condition for the uniqueness of the coherent extensions to the conditional case in Section 3, we focus on two particular cases of 2-monotone lower previsions. First, in Section 4 we consider completely monotone lower previsions, that correspond to the Choquet integral with respect to a belief function [7]; then we discuss minimum-preserving lower previsions in Section 5. Our results in this section illustrate one interesting fact: that the coherence between unconditional and conditional lower probabilities studied in [30] is not equivalent to the coherence of the respective lower previsions they determine by means of the Choquet integral.

Due to limitations of space, proofs have been omitted.

2 Preliminary concepts

2.1 Coherent lower previsions

Consider a possibility space Ω, that we shall assume in this paper to be *finite*. A *gamble* is a real-valued functional defined on Ω. We shall denote by $\mathcal{L}(\Omega)$ the set of all gambles on Ω. One instance of gambles are the indicators of events. Given a subset A of Ω, the indicator function of A is the gamble that takes the value 1 on the elements of A and 0 elsewhere. We shall denote this gamble by I_A, or by A when no confusion is possible.

A *lower prevision* is a functional $\underline{P}$ defined on a set of gambles $\mathcal{K} \subseteq \mathcal{L}(\Omega)$. Given a gamble f, $\underline{P}(f)$ is understood to represent a subject's supremum acceptable buying price for f, in the sense that for any $\epsilon > 0$ the transaction $f - \underline{P}(f) + \epsilon$ is acceptable for him.

Using this interpretation, we can derive a notion of coherence:

Definition 1. A lower prevision $\underline{P} : \mathcal{L}(\Omega) \to \mathbb{R}$ is called *coherent* if and only if it satisfies the following properties for every $f, g \in \mathcal{L}(\Omega)$ and every $\lambda > 0$:

(C1) $\underline{P}(f) \geq \min f$.

(C2) $\underline{P}(\lambda f) = \lambda \underline{P}(f)$.

(C3) $\underline{P}(f+g) \geq \underline{P}(f) + \underline{P}(g)$.

The interpretation of this notion is that the acceptable buying prices encompassed by $\{\underline{P}(f) : f \in \mathcal{L}(\Omega)\}$ are consistent with each other. In the particular case when $\underline{P}$ satisfies (C3) with equality for every $f, g \in \mathcal{L}(\Omega)$, it is called a *linear* prevision. Any coherent lower prevision is the *lower envelope* of the set of linear previsions that dominate it, i.e.,

$$\underline{P}(f) = \min\{P(f) : P \text{ linear prevision}, P \geq \underline{P}\}.$$

The conjugate functional $\overline{P}$ of a coherent lower prevision $\underline{P}$, given by $\overline{P}(f) = -\underline{P}(-f)$ for every $f \in \mathcal{L}(\Omega)$, is called a coherent *upper* prevision. It corresponds to the upper envelope of the set of linear previsions that dominate $\underline{P}$.

A coherent lower prevision defined only on indicators of events is called a *coherent lower probability*. In particular, the restriction of a linear prevision to indicators of events corresponds to a (finitely additive) probability measure. Hence, coherent lower previsions are simply lower envelopes of closed and convex sets of probability measures, and as such they can also be given a Bayesian sensitivity analysis interpretation.

One particular case of coherent lower previsions are the *vacuous* ones. They correspond to the case where we have the information that the outcome of the experiment belongs to some set A (and nothing else). In that case, our coherent lower prevision is given by

$$\underline{P}(f) = \min_{\omega \in A} f(\omega) \ \forall f \in \mathcal{L}(\Omega). \tag{1}$$

Although a linear prevision is uniquely determined by the probability measure that is its restriction to events, this is not the case for lower previsions: a coherent lower probability will have in general more than one coherent extension to the set of all gambles. This is the reason why the theory is established in terms of gambles instead of events. Interestingly, there are some cases where the restriction to events uniquely determines the coherent lower prevision. One particular case that shall be important in this paper is that where the restriction to events is 0–1-valued:

Lemma 1. *[27, Note 4, Section 3.2.6] Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}(\Omega)$ whose restriction to events is* 0–1*-valued. Then $\underline{P}$ is the unique coherent extension of its restriction to events, and it is given by*

$$\underline{P}(f) = \sup_{F : \underline{P}(F)=1} \inf_{\omega \in F} f(\omega);$$

moreover, the class $\{F \subseteq \Omega : \underline{P}(F) = 1\}$ is a filter.

This applies in particular for the vacuous lower previsions in Eq. (1).

2.2 Conditional lower previsions

Given a partition $\mathcal{B}$ of the possibility space Ω, a *conditional lower prevision* on $\mathcal{L}(\Omega)$ is a functional $\underline{P}(\cdot|\mathcal{B})$ on $\mathcal{L}(\Omega)$ that to any gamble f and any $B \in \mathcal{B}$ assigns the value $\underline{P}(f|B)$, that represents a subject's supremum acceptable buying price for f, if he comes to know later that the outcome of the experiment belongs to the subset B of Ω. Thus, $\underline{P}(\cdot|B)$ is a functional on $\mathcal{L}(\Omega)$ for every $B \in \mathcal{B}$. By putting all these values together, we end up with the gamble

$$\underline{P}(f|\mathcal{B}) := \sum_{B \in \mathcal{B}} I_B(f - \underline{P}(f|B)).$$

Similarly to conditions (C1)–(C3), we can establish a notion of coherence for conditional lower previsions.

Definition 2. A conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$ on $\mathcal{L}(\Omega)$ is *separately coherent* when

(SC1) $\underline{P}(f|B) \geq \min_{\omega \in B} f(\omega)$,

(SC2) $\underline{P}(\lambda f|B) = \lambda \underline{P}(f|B)$,

(SC3) $\underline{P}(f+g|B) \geq \underline{P}(f|B) + \underline{P}(g|B)$

for every $f, g \in \mathcal{L}(\Omega), \lambda > 0$ and $B \in \mathcal{B}$.

The behavioural interpretation of this notion is that the acceptable conditional buying prices encompassed by $\underline{P}(\cdot|B)$ are consistent with each other for every fixed set B in the partition $\mathcal{B}$. Together they imply $\underline{P}(B|B) = 1\ \forall B \in \mathcal{B}$.

If we start with a coherent lower prevision $\underline{P}$ and consider a partition $\mathcal{B}$ of the space Ω, there is in general not a unique way of updating $\underline{P}$ into a separately coherent conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$. This is related to the problem of conditioning on sets of probability zero, which has attracted a lot of attention in the literature [3, 13, 18]; see also [27, Chapter 6] for the approach considered in this paper. In the next section we detail how the conditional lower prevision may be derived and we formulate the problem we shall study in this paper.

2.3 Formulation of the problem

Consider now a coherent lower prevision $\underline{P}$ on $\mathcal{L}(\Omega)$, let $\mathcal{B}$ be a partition of Ω and assume we want to update $\underline{P}$ into a separately coherent conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$ on $\mathcal{L}(\Omega)$.

One strategy to derive $\underline{P}(\cdot|\mathcal{B})$ from $\underline{P}$ is to verify that the assessments present in these two lower previsions are compatible with each other. This gives rise to the concept of *joint coherence*, which is studied in much detail in [27, Chapters 6 and 7]. In this case, where we are dealing with finite spaces, we have the following characterisation:

Proposition 1. *[27, Theorem 6.5.4] Consider a coherent lower prevision $\underline{P}$ and a separately coherent conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$ on $\mathcal{L}(\Omega)$, where Ω is a finite space. They are jointly coherent when*

$$\underline{P}(B(f - \underline{P}(f|B))) = 0\ \forall f \in \mathcal{L}(\Omega), B \in \mathcal{B}. \qquad (2)$$

The above equation is called the *Generalised Bayes Rule*, because it reduces to the well-known Bayes' rule in the precise case. It holds trivially when $\overline{P}(B) = 0$, so any conditional lower prevision $\underline{P}(\cdot|B)$ is compatible with $\underline{P}$ in that case; on the other hand, if $\underline{P}(B) > 0$ then for every gamble f there is a unique real number μ such that $\underline{P}(B(f - \mu)) = 0$, so there is only one conditional lower prevision $\underline{P}(\cdot|B)$ that is compatible with $\underline{P}$.

The most interesting case is that where the conditioning event has zero lower probability and positive upper probability, i.e., that of $\underline{P}(B) = 0 < \overline{P}(B)$. In that case, there is usually an infinite number of conditional lower previsions that are compatible with $\underline{P}$; there were characterised in [20], where it was proven that they are bounded by the so-called natural and regular extensions.

Definition 3. Given $B \in \mathcal{B}$, the *natural extension* $\underline{E}(\cdot|B)$ induced by $\underline{P}$ is given by:

$$\underline{E}(f|B) := \begin{cases} \inf_{P \geq \underline{P}}\{P(f|B)\} & \text{if } \underline{P}(B) > 0 \\ \min_{\omega \in B} f(\omega) & \text{otherwise} \end{cases}$$

for any gamble $f \in \mathcal{L}(\Omega)$.

The natural extension is vacuous when the conditioning event has zero lower probability, and is uniquely determined by Eq. (2) otherwise. Although it produces a conditional lower prevision that is coherent with $\underline{P}$, it is arguably too uninformative. A more informative alternative is called the regular extension:

Definition 4. Given $B \in \mathcal{B}$, the *regular extension* $\underline{R}(\cdot|B)$ induced by $\underline{P}$ is given by:

$$\underline{R}(f|B) := \begin{cases} \inf_{P(B)>0, P \geq \underline{P}}\{P(f|B)\} & \text{if } \overline{P}(B) > 0 \\ \min_{\omega \in B} f(\omega) & \text{otherwise} \end{cases}$$

for any gamble $f \in \mathcal{L}(\Omega)$.

Hence, regular extension corresponds to applying Bayes' rule whenever possible on the set of precise models compatible with our conditional lower prevision, and to take then the lower prevision of the resulting set of conditional previsions. It has been proposed as an updating rule in a number of works in the literature [4, 8, 14, 15, 17, 28].

It turns out that the natural and the regular extensions characterise the set of conditional lower previsions that are jointly coherent with $\underline{P}$:

Proposition 2. *[20, Theorem 9] Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}(\Omega)$ and $\mathcal{B}$ a partition of Ω such that $\overline{P}(B) > 0$ for any $B \in \mathcal{B}$. Then a separately coherent conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$ is coherent with $\underline{P}$ if and only if $\underline{P}(f|B) \in [\underline{E}(f|B), \underline{R}(f|B)]$ for every $f \in \mathcal{L}(\Omega)$ and every $B \in \mathcal{B}$.*

In this paper we shall not deal with the case $\overline{P}(B) = 0$ because then any conditional model $\underline{P}(\cdot|B)$ satisfies the Generalised Bayes Rule with $\underline{P}$.

The conditional lower previsions determined by the natural and regular extension may not coincide when $\underline{P}(B) = 0 < \overline{P}(B)$ (see for instance Example 2 later on). In this paper, we are going to characterise their equality for one interesting particular case of coherent lower previsions: the 2-monotone ones. As particular cases, we shall consider completely monotone lower previsions, random sets and possibility measures.

3 Updating 2-monotone lower previsions

One important instance of coherent lower previsions are the n-monotone ones, that were first introduced by Choquet in [2]:

Definition 5. A coherent lower prevision $\underline{P}$ on $\mathcal{L}(\Omega)$ is called *n-monotone* if and only if

$$\underline{P}\left(\bigvee_{i=1}^{p} f_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,p\}} (-1)^{|I|+1} \underline{P}\left(\bigwedge_{i \in I} f_i\right) \quad (3)$$

for all $2 \leq p \leq n$, and all $f_1, \ldots, f_p$ in $\mathcal{L}(\Omega)$, where $\vee$ denotes the point-wise maximum and $\wedge$ the point-wise minimum.

In particular, a coherent lower probability $\underline{P} : \mathcal{P}(\Omega) \to [0,1]$ is n-monotone when

$$\underline{P}\left(\bigcup_{i=1}^{p} A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,p\}} (-1)^{|I|+1} \underline{P}\left(\bigcap_{i \in I} A_i\right) \quad (4)$$

for all $2 \leq p \leq n$, and all subsets $A_1, \ldots, A_p$ of Ω.

Although a coherent lower prevision is not determined uniquely by its restriction to events, it is when we require in addition the property of n-monotonicity, in the following sense: given a n-monotone lower probability, its natural extension is the only n-monotone extension to $\mathcal{L}(\Omega)$. It corresponds moreover to the Choquet integral [12] with respect to this fuzzy measure [7, 26], so we have that

$$\underline{P}(f) := (C)\int f d\underline{P} = \inf f + \int_{\inf f}^{\sup f} \underline{P}(f \geq t)dt$$

for every gamble f.

A coherent lower prevision on $\mathcal{L}(\Omega)$ that is n-monotone for all $n \in \mathbb{N}$ is called *completely monotone*, and its restriction to events is a *belief function*; its conjugate $\overline{P}$ is a plausibility function. One example of completely monotone coherent lower previsions are the vacuous ones in Eq. (1); another one is given by the linear previsions, that moreover satisfy Eq. (3) with equality for every n.

In particular, a coherent lower prevision $\underline{P}$ on $\mathcal{L}(\Omega)$ is 2-monotone if and only if it satisfies Eq. (3) for $n = 2$, that is, if and only if

$$\underline{P}(f \vee g) + \underline{P}(f \wedge g) \geq \underline{P}(f) + \underline{P}(g)$$

for every $f, g \in \mathcal{L}(\Omega)$. On the other hand, we deduce from Eq. (4) that a coherent lower probability on $\mathcal{P}(\Omega)$ is called 2-monotone whenever

$$\underline{P}(A \cup B) + \underline{P}(A \cap B) \geq \underline{P}(A) + \underline{P}(B) \; \forall A, B \subseteq \Omega.$$

In this section, we are going to determine under which conditions a 2-monotone lower prevision $\underline{P}$ on $\mathcal{L}(\Omega)$ can be uniquely updated to a conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$ that is coherent with $\underline{P}$, in the sense of Eq. (2). In order to do this, we shall use the formula for the conditional lower probability determined by regular extension:

Proposition 3. *[26, Theorem 7.2] Let $\underline{P}$ be a 2-monotone lower prevision on $\mathcal{L}(\Omega)$, and consider $B \subseteq \Omega$ such that $\overline{P}(B) > 0$. Then for any event A,*

$$\underline{R}(A|B) = \begin{cases} \frac{\underline{P}(A \cap B)}{\underline{P}(A \cap B) + \overline{P}(A^c \cap B)} & \text{if } \overline{P}(A^c \cap B) > 0, \\ 1 & \text{otherwise}, \end{cases} \quad (5)$$

and $\underline{R}(\cdot|B)$ is a 2-monotone lower probability.

Interestingly, we shall show in Example 3 later on that in general $\underline{R}(\cdot|B)$ need not be 2-monotone on gambles. As we shall see, we can only guarantee 2-monotonicity on gambles when the conditioning event has zero lower probability and positive upper probability.

To see that Eq. (5) does not hold without the assumption of 2-monotonicity, consider the following example:

Example 1. Consider $\Omega = \{a, b, c, d\}$ and let P_1, P_2 be the linear previsions determined by the mass functions p_1, p_2 given by

	a	b	c	d
p_1	0.5	0.5	0	0
p_2	0.25	0.25	0.25	0.25

It has been showed in [26, Section 6] that the lower envelope $\underline{P}$ of $\{P_1, P_2\}$ is a coherent lower prevision that is not 2-monotone. Consider $B = \{a, b\}$ and $A = \{a\}$. Then $\overline{P}(A^c \cap B) = \overline{P}(\{b\}) = 0.5 > 0$, and

$$\frac{\underline{P}(A \cap B)}{\underline{P}(A \cap B) + \overline{P}(A^c \cap B)} = \frac{0.25}{0.25 + 0.5} = \frac{1}{3};$$

on the other hand any $P \geq \underline{P}$ is given by $\alpha P_1 + (1 - \alpha)P_2$, where $\alpha \in [0, 1]$; since $P_1(\{a\}) = P_1(\{b\})$ and $P_2(\{a\}) = P_2(\{b\})$, it follows that any $P \geq \underline{P}$ must satisfy $P(\{a\}) = P(\{b\})$ too, whence $\underline{R}(A|B) = 0.5$. Hence, Eq. (5) does not hold. ⧫

From Proposition 3 we deduce the following:

Proposition 4. *Let $\underline{P}$ be a 2-monotone lower prevision on $\mathcal{L}(\Omega)$ and consider $B \subseteq \Omega$ such that $\underline{P}(B) = 0 < \overline{P}(B)$. Then for any gamble f*

$$\underline{R}(f|B) = \min_{\omega \in C} f(\omega),$$

where C is the smallest subset of B satisfying $\underline{R}(C|B) = 1$.

Interestingly, this shows that, if the lower prevision $\underline{P}$ satisfies 2-monotonicity, when the conditioning event B has zero lower probability and positive upper probability, the regular extension $\underline{R}(\cdot|B)$ is a completely monotone lower prevision, even if the lower prevision $\underline{P}$ we start from is not completely monotone.

Using these results, we can determine in which cases the natural and regular extensions coincide:

Proposition 5. *Let $\underline{P}$ be a 2-monotone lower prevision on $\mathcal{L}(\Omega)$, and consider $B \subseteq \Omega$ with $\overline{P}(B) > 0 = \underline{P}(B)$. The following are equivalent:*

1. *$\underline{E}(f|B) = \underline{R}(f|B)$ for every $f \in \mathcal{L}(\Omega)$.*
2. *$\underline{E}(A|B) = \underline{R}(A|B)$ for every $A \subseteq \Omega$.*
3. *$\overline{P}(\{\omega\}) > 0$ for every $\omega \in B$.*

We immediately deduce the following:

Theorem 1. *Let $\underline{P}$ be a 2-monotone lower prevision on $\mathcal{L}(\Omega)$, and let $\mathcal{B}$ be a partition of Ω. Then $\underline{E}(\cdot|\mathcal{B}) = \underline{R}(\cdot|\mathcal{B})$ if and only if $\overline{P}(\{\omega\}) > 0\ \forall\omega \in B \subseteq \Omega$ s.t. $\underline{P}(B) = 0 < \overline{P}(B)$.*

To see that this result cannot be extended to arbitrary coherent lower previsions, it suffices to consider the coherent lower prevision $\underline{P}$ in Example 1, $B = \{c, d\}$ and $A = \{c\}$: we get $\underline{E}(A|B) = 0 < 0.5 = \underline{R}(A|B)$.

4 Coherent updating of completely monotone lower previsions

We consider next the case where the lower prevision $\underline{P}$ on $\mathcal{L}(\Omega)$ is completely monotone.

One of the most important rules in that case is Dempster's rule of conditioning [11, 24], where, given a plausibility function $\overline{P}$ on $\mathcal{P}(\Omega)$ and a conditioning event B with $\overline{P}(B) > 0$, the conditional plausibility is defined by

$$\overline{P}(A|B) := \frac{\overline{P}(A \cap B)}{\overline{P}(B)}.$$

However, this conditional upper probability is not coherent with the unconditional upper probability $\overline{P}$ [31]; see also [27, Section 5.13] and [29]. Thus, Dempster's rule is not interesting from the behavioural point of view, and we shall focus in this section on the natural and the regular extensions instead.

Given a conditioning event B with $\overline{P}(B) > 0$, its regular extension is determined by Eq. (5). This formula has also been established in a few papers ([14, Theorem 3.4]; [15, Proposition 4]; see also [4, 11]). Moreover, it has been established in [14, 15, 25] that the restriction of $\underline{R}(\cdot|B)$ to events is a belief function for every $B \subseteq \Omega$ such that $\underline{P}(B) > 0$.

The equality between the natural and the regular extensions of $\underline{P}$ is characterised by Theorem 1. In this section, we give equivalent conditions in terms of the focal elements of $\underline{P}$.

Definition 6. [24] Given a belief function $\underline{P}$ on $\mathcal{P}(\Omega)$, its *Möbius inverse* $m : \mathcal{P}(\Omega) \to [0, 1]$ is given by

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \underline{P}(B)\ \forall A \subseteq \Omega.$$

It holds that $\underline{P}(A) = \sum_{B \subseteq A} m(B)$, and m is called a *basic probability assignment* within the evidential theory of Shafer. For the plausibility function $\overline{P}$ that is conjugate to $\underline{P}$, it holds that $\overline{P}(A) = \sum_{B \cap A \neq \emptyset} m(B)$ for every $A \subseteq \Omega$.

For the results in this section, it shall be interesting to work with the focal elements of the belief function:

Definition 7. [24] Given a belief function $\underline{P}$ with Möbius inverse m, a subset $B \subseteq \Omega$ is called a *focal element* when $m(B) > 0$. The union F of all the focal elements of $\underline{P}$ is called the *core* of $\underline{P}$.

We shall be particularly interested in those belief functions whose focal elements cover the possibility space Ω:

Definition 8. A belief function $\underline{P}$ with core F is called *full* when $F = \Omega$.

Since $\overline{P}(F^c) = \sum_{B \text{ focal}: B \cap F^c \neq \emptyset} m(B) = 0$, given a belief function that is not full, any set included in F^c will have zero upper probability. Equivalently, if $\underline{P}$ is a full belief function, any subset B of Ω has a positive upper probability.

Recall that for any conditioning event B, it holds that $\underline{E}(\cdot|B) = \underline{R}(\cdot|B)$ if $\underline{P}(B) > 0$ or $\overline{P}(B) = 0$. Hence, the natural and regular extensions will agree as soon as there is no conditioning event with zero lower probability and positive upper probability. This case is characterised by the following definition:

Definition 9. A belief function is called *non-atomic* if for every focal element B, it holds that $m(\{\omega\}) > 0$ for every $\omega \in B$.

The reason for this terminology is that given such a belief function there is no set B with $|B| \geq 2$ satisfying $\underline{P}(B) > 0$ and $\underline{P}(A) = 0$ for every $A \subsetneq B$. See [1, 19] for related concepts. Non-atomic and full belief functions can be characterised in the following way:

Proposition 6. *Let $\underline{P}$ be a belief function on $\mathcal{P}(\Omega)$.*

1. *$\underline{P}$ is non-atomic if and only if for any $B \subseteq \Omega$ either $\overline{P}(B) = 0$ or $\underline{P}(B) > 0$.*
2. *$\underline{P}$ is full if and only if for any $B \subseteq \Omega$, $\overline{P}(B) > 0$.*

3. *$\underline{P}$ is full and non-atomic if and only if $\underline{P}(B) > 0$ for every $B \subsetneq \Omega$.*

When the conditioning event B has zero lower probability and positive upper probability the equality between the natural and the regular extensions is characterised by Proposition 5: we need that $\overline{P}(\{\omega\}) > 0$ for every $\omega \in B$; in the case of belief functions, this is equivalent to $B \subseteq F$, the core of the belief function. From this we deduce the following result:

Proposition 7. *Let $\underline{P}$ be a completely monotone lower prevision on $\mathcal{L}(\Omega)$, and let μ denote the belief function that is the restriction of $\underline{P}$ to events. Then, $\underline{E}(\cdot|B) = \underline{R}(\cdot|B)$ for every $B \subseteq \Omega$ if and only if μ is either full or non-atomic.*

This result allows to provide an example where the natural and the regular extensions do not coincide:

Example 2. Consider $\Omega = \{a, b, c, d\}$, and let $\underline{P}$ be the completely monotone lower prevision given by

$$\underline{P}(f) = \min\{f(b), f(c)\} \; \forall f \in \mathcal{L}(\Omega).$$

The restriction to events of $\underline{P}$ is the belief function associated to the basic probability assignment m where

$$m(\{b, c\}) = 1 \text{ and } m(C) = 0 \text{ for every } C \neq \{b, c\}.$$

Obviously, this belief function is not full. If we take $B = \{a, b\}$ and $A = \{b\}$, then any probability $P \geq \underline{P}$ satisfying $P(B) > 0$ must satisfy $P(\{b\}) > 0$, because $P(\{a\}) \leq \overline{P}(\{a\}) = 0$. But then P will satisfy $P(A|B) = 1$, and from this we deduce that

$$\underline{R}(A|B) = 1 > 0 = \underline{E}(A|B),$$

where the last equality holds because $\underline{P}(B) = 0$. Hence, the natural and regular extensions do not coincide. ♦

Moreover, for completely monotone lower previsions we can give an alternative expression of the regular extension to that in Proposition 4.

Proposition 8. *Let $\underline{P}$ be a completely monotone lower prevision, and let F be the core of its associated belief function. Then for any $B \subseteq \Omega$ such that $\underline{P}(B) = 0 < \overline{P}(B)$,*

$$\underline{R}(f|B) = \min_{\omega \in B \cap F} f(\omega) \; \forall f \in \mathcal{L}(\Omega).$$

Let us recall again that the condition $\underline{P}(B) = 0 < \overline{P}(B)$ we consider in this theorem implies that the belief function is not non-atomic.

From Proposition 7 we immediately derive the following theorem.

Theorem 2. *Let $\underline{P}$ be a completely monotone lower prevision on $\mathcal{L}(\Omega)$ and let $\mathcal{B}$ be a partition of Ω. If the restriction to events μ of $\underline{P}$ is either full or non-atomic, then $\underline{E}(\cdot|\mathcal{B}) = \underline{R}(\cdot|\mathcal{B})$.*

Note that the sufficient condition in this theorem is not necessary: it may be that μ is neither full nor non-atomic and $\mu(B) > 0$ for every B in the partition $\mathcal{B}$, and then $\underline{E}(\cdot|\mathcal{B}) = \underline{R}(\cdot|\mathcal{B})$.

4.1 Random Sets

One context where completely monotone lower previsions arise naturally is that of measurable multi-valued mappings, or random sets [11, 23].

Definition 10. Let $(X, \mathcal{A}, P)$ be a probability space, $(\Omega, \mathcal{P}(\Omega))$ a measurable space, where Ω is finite, and $\Gamma : X \to \mathcal{P}(\Omega)$ a non-empty multi-valued mapping. It is called a *random set* when it satisfies the following measurability condition:

$$\Gamma_*(A) := \{x \in X : \Gamma(x) \subseteq A\} \in \mathcal{A} \quad \forall A \subseteq \Omega.$$

Its associated *lower probability* $P_{*\Gamma} : \mathcal{P}(\Omega) \to [0, 1]$ is a belief function and is given by

$$P_{*\Gamma}(A) = P(\Gamma_*(A)) \quad \forall A \subseteq \Omega. \tag{6}$$

The focal elements of $P_{*\Gamma}$ are given by

$$\{A \subseteq \Omega : P(\Gamma^{-1}(A)) > 0\},$$

and its Möbius inverse is given by $m = P \circ \Gamma^{-1}$. The conjugate plausibility measure is denoted by P^*_Γ and it is called the upper probability of the random set Γ. It satisfies

$$P^*_\Gamma(A) = 1 - P_{*\Gamma}(A^c) = P(\{x : \Gamma(x) \cap A \neq \emptyset\}),$$

where the set $\{x : \Gamma(x) \cap A \neq \emptyset\}$ is the *upper inverse* of A by Γ, and is usually denoted by Γ^*. The Choquet integral with respect to $P_{*\Gamma}$ is a completely monotone lower prevision on $\mathcal{L}(\Omega)$, and it corresponds to the natural extension of $P_{*\Gamma}$ from $\mathcal{P}(\Omega)$ to the set of all gambles. If we want to update this completely monotone lower prevision, we can use the natural or the regular extensions, that, by Proposition 7, coincide if and only if $P_{*\Gamma}$ is either full or non-atomic. These properties can be easily characterised in terms of the images of Γ:

Proposition 9. *Let $(X, \mathcal{A}, P)$ be a probability space, Ω a finite set and $\Gamma : X \to \mathcal{P}(\Omega)$ a random set with associated lower probability $P_{*\Gamma}$. Let F denote the core of $P_{*\Gamma}$.*

1. *$P_{*\Gamma}$ is full $\Leftrightarrow F = \Omega \Leftrightarrow P^*_\Gamma(B) > 0$ for all $B \subseteq X \Leftrightarrow P^*_\Gamma(\{\omega\}) > 0$ for all $\omega \in \Omega \Leftrightarrow P(\{x : \omega \in \Gamma(x)\}) > 0$ for all $\omega \in \Omega$.*

2. $P_{*\Gamma}$ is non-atomic $\Leftrightarrow \forall\omega \in F, P(\Gamma^{-1}(\omega)) > 0$.

Moreover, $\underline{E}(\cdot|B) = \underline{R}(\cdot|B)$ for all $B \subseteq \Omega$ if and only if $P_{*\Gamma}$ is either full or non-atomic.

One interesting interpretation of random sets is the *epistemic* one, where they are seen as models for the imprecise knowledge of a random variable [16]. In that case, our information about this random variable is provided by the *measurable selections* of Γ: those measurable mappings $U : X \to \Omega$ such that $U(x) \in \Gamma(x)\ \forall x \in X$. We shall denote by $S(\Gamma)$ the set of measurable selections of Γ and by $P(\Gamma)$ the set of the probability measures they induce on $\mathcal{P}(\Omega)$. This set is included in the class $\mathcal{M}(P_{*\Gamma})$ of probabilities that dominate $P_{*\Gamma}$. Although both sets do not coincide in general, when Ω is finite it can be checked that:

Proposition 10. *[21, Theorem 1] Let $\Gamma : X \to \mathcal{P}(\Omega)$ be a random set, where Ω is finite. Then $Ext(M(P_{*\Gamma})) \subseteq P(\Gamma)$ and $M(P_{*\Gamma}) = Conv(Ext(M(P_{*\Gamma})))$.*

Moreover, from [11], $M(P_{*\Gamma})$ has a finite number of extreme points, that are related to the permutations of the final space.

The epistemic interpretation can be carried on towards the regular extension, in the following sense:

Proposition 11. *Let $(X, \mathcal{A}, P)$ be a probability space, Ω a finite set and $\Gamma : X \to \mathcal{P}(\Omega)$ a random set with associated lower probability $P_{*\Gamma}$. Consider $B \subseteq \Omega$ with $P^*_\Gamma(B) > 0$. Then, for every $f \in \mathcal{L}(\Omega)$,*

$$\underline{R}(f \mid B) = \min\{P_U(f \mid B) : U \in S(\Gamma), P_U(B) > 0\}.$$

To conclude this section, we use random sets to establish that, even if the conditional lower probability derived from a completely monotone lower prevision by Generalised Bayes Rule is a belief function [14, 15], when we move from events to gambles we do not necessarily obtain a completely monotone lower prevision.

Example 3. Consider the probability space $(X, \mathcal{P}(X), P)$, where $X = \{a, b, c, d, e\}$, and P is the probability measure determined by the equalities $P(a) = P(b) = 1/8$, and $P(c) = P(d) = P(e) = 1/4$. Let Γ be the multi-valued mapping $\Gamma : X \to \mathcal{P}(\{1, 2, 3, 4\})$ given by $\Gamma(a) = \{1\}, \Gamma(b) = \{2\}, \Gamma(c) = \{1, 4\}, \Gamma(d) = \{2, 4\}, \Gamma(e) = \{3, 4\}$.

Let $P_{*\Gamma}$ denote the lower probability induced by this random set. This is a belief function, and the lower prevision $\underline{P}$ on $\mathcal{L}(\{1, 2, 3, 4\})$ given by $\underline{P}(f) = (C) \int f dP_{*\Gamma}$ is a completely monotone lower prevision.

It follows from Eq. (6) that

$$P_{*\Gamma}(\{1, 2, 3\}) = P(\{a, b\}) = \frac{1}{4} > 0.$$

As a consequence, the natural and regular extensions coincide, and we deduce from Proposition 11 that

$$\underline{R}(f|\{1, 2, 3\}) = \min\{P_U(f|\{1, 2, 3\}) : U \in S(\Gamma)\}. \tag{7}$$

Let us consider the gamble f on $\{1, 2, 3, 4\}$ given by $f(\omega) = 4 - \omega$ for all $\omega \in \{1, 2, 3, 4\}$. Then since $f = 1\, \mathbb{I}_{1,2,3} + 1\, \mathbb{I}_{1,2} + 1\, \mathbb{I}_1$, its Choquet integral with respect to $\underline{R}(\cdot|\{1, 2, 3\})$ would be

$$1 + \underline{R}(\{1, 2\}|\{1, 2, 3\}) + \underline{R}(\{1\}|\{1, 2, 3\}).$$

We deduce from Eq. (7) that

$$\underline{R}(\{1\}|\{1, 2, 3\}) = \frac{1}{6} \text{ and } \underline{R}(\{1, 2\}|\{1, 2, 3\}) = \frac{1}{2};$$

as a consequence, $(C) \int f\ d\underline{R}(\cdot|\{1, 2, 3\}) = 5/3$.

On the other hand, the smallest value of $\{P_U(f|\{1, 2, 3\}) : U \in S(\Gamma)\}$ is given by $7/4 > 5/3$. This means that $\underline{R}(f|\{1, 2, 3\}) > (C) \int f d\underline{R}(\cdot|\{1, 2, 3\})$.

But it has been established in [7, 26] that if we have a 2-monotone lower probability on all events (as is the case for $\underline{R}(\cdot|\{1, 2, 3\})$), the *only* 2-monotone extension to all gambles is the Choquet integral. This means that the conditional lower prevision $\underline{R}(\cdot|\{1, 2, 3\})$ is not 2-monotone on $\mathcal{L}(\{1, 2, 3\})$. ♦

5 Coherent updating of minimum-preserving previsions

We consider now the particular case of completely monotone lower previsions that are minimum-preserving, i.e., lower previsions $\underline{P}$ such that

$$\underline{P}(f \wedge g) = \min\{\underline{P}(f), \underline{P}(g)\}$$

for every pair of gambles f, g on Ω. They correspond to the Choquet integral with respect to their restriction to events, which is a necessity measure N. Their conjugate upper previsions $\overline{P}$ are the Choquet integral with respect to the possibility measure Π that is determined by N using duality, and are maximum-preserving.

From Proposition 7, we deduce the following:

Corollary 1. *Let $\underline{P}$ be a minimum-preserving coherent lower prevision. Then $\underline{E}(\cdot|B) = \underline{P}(\cdot|B)$ for all $B \subseteq \Omega$ if and only if either of the following conditions holds:*

(i) $\overline{P}(\{\omega\}) > 0$ for all $\omega \in \Omega$.

(ii) $\underline{P}(\{\omega\}) = 1$ *for some* $\omega \in \Omega$.

The result in Corollary 1 can be simplified further taking into account that de Cooman and Aeyels proved in [5] (see also [6]) that a coherent *upper* prevision $\overline{P}$ on $\mathcal{L}(\Omega)$ is maximum-preserving if and only if its restriction to events is a 0–1-valued possibility measure. Then, if we define $F := \{\omega : \overline{P}(\{\omega\}) = 1\}$, it turns out that F is the only focal element of the possibility measure $\overline{P}$, and $m(F) = 1$. Hence, $\overline{P}$ is the vacuous lower prevision on F, that is,

$$\underline{P}(f) = \min_{\omega \in F} f(\omega) \quad \forall f \in \mathcal{L}(\Omega).$$

Now, given a conditioning event $B \subseteq F$, there are a number of possibilities:

- $B \subseteq F^c$. Then $\overline{P}(B) = 0$ and both the natural and regular extensions are vacuous.
- $B \cap F \neq \emptyset \neq B \cap F^c$. Then $\underline{P}(B) = 0 < 1 = \overline{P}(B)$, whence $\underline{E}(\cdot|B)$ is vacuous on B and $\underline{R}(\cdot|B)$ is vacuous on $B \cap F$. Hence, in that case the natural and regular extensions do not coincide.
- $B \subseteq F$. Then both $\underline{E}(\cdot|B)$ and $\underline{R}(\cdot|B)$ are vacuous on B.

Note that in this case $\underline{P}$ is only non-atomic when F is a singleton (i.e., when $\underline{P}$ corresponds to the expectation operator with respect to a degenerate probability measure), and $\underline{P}$ is full if and only if $F = \Omega$, meaning that $\underline{P}$ corresponds to the vacuous model. Hence, we only have the equality between the natural and the regular extensions for all $B \subseteq \Omega$ in these two extreme cases.

We summarise the coherent updating of a minimum-preserving lower prevision in the following theorem.

Theorem 3. *Let $\underline{P}$ be a minimum-preserving lower prevision on $\mathcal{L}(\Omega)$, and consider a partition $\mathcal{B}$ of Ω. Consider $F \subseteq \Omega$ such that $\underline{P}(f) = \min_{\omega \in F} f(\omega) \forall f \in \mathcal{L}(\Omega)$. Given $B \in \mathcal{B}$ and $f \in \mathcal{L}(\Omega)$,*

1. $\underline{E}(f|B) = \begin{cases} \min_{\omega \in B} f(\omega) & \text{if } F \nsubseteq B \\ \min_{\omega \in F} f(\omega) & \text{if } F \subseteq B. \end{cases}$

2. $\underline{R}(f|B) = \begin{cases} \min_{\omega \in B \cap F} f(\omega) & \text{if } B \cap F \neq \emptyset \\ \min_{\omega \in B} f(\omega) & \text{if } B \cap F = \emptyset. \end{cases}$

3. *$\underline{E}(f|B) = \underline{R}(f|B)$ if and only if either $B \cap F = \emptyset$ or $B \cap F^c = \emptyset$.*

4. *A separately coherent conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$ is coherent with $\underline{P}$ if and only if*

$$\min_{\omega \in B} f(\omega) \leq \underline{P}(f|B) \leq \min_{\omega \in B \cap F} f(\omega)$$

for every $f \in \mathcal{L}(\Omega), B \in \mathcal{B}$ s.t $B \cap F \neq \emptyset$.

From Theorem 3, the bounds determined by natural and regular extension are both minimum-preserving, and as a consequence they correspond to the Choquet integral of their respective restrictions to events. However, it is easy to see that not every separately coherent conditional lower prevision between them is minimum-preserving.

5.1 Comparison with the updating of possibility measures

The results in this paper allow us to show one interesting phenomenon: that, even if a minimum-preserving lower prevision $\underline{P}$ is the natural extension of its restriction to events N, the coherence of N with a conditional lower probability $N(\cdot|\mathcal{B})$ is not equivalent to the coherence of the lower previsions $\underline{P}$, $\underline{P}(\cdot|\mathcal{B})$ that each of them determines by natural extension. This is the reason behind the apparent contradiction with the results in [30]: it is showed there that Dempster's rule is a coherent updating rule for updating a possibility measure, even if it can be more informative than the conditional possibility we obtain by regular extension.

To make this clearer, let us study the results in [30] in more detail. The authors consider two finite sets $\mathcal{X}$ and $\mathcal{Y}$, and let $\Omega = \mathcal{X} \times \mathcal{Y}$. They take a possibility measure $\Pi(\cdot,\cdot)$ on $\mathcal{P}(\Omega)$ and look for the smallest and greatest conditional possibility measures $\Pi(\cdot|Y)$ that satisfy coherence with Π. Note that, since we are dealing with upper previsions now, it follows from conjugacy and Proposition 2 that a conditional upper prevision $\overline{P}(\cdot|\mathcal{B})$ is coherent with $\overline{P}$ if and only if $\overline{P}(f|B) \in [\overline{R}(f|B), \overline{E}(f|B)]$ for every gamble f and every $B \subseteq \Omega$ s.t. $\overline{P}(B) > 0$, where $\overline{R}(\cdot|B)$ and $\overline{E}(\cdot|B)$ are the conjugate upper previsions of the regular and natural extensions, respectively.

In [30], the focus is on conditional upper *probabilities* instead of previsions, and in particular on those conditional possibility measures $\Pi(\cdot|Y)$ that satisfy coherence with the unconditional possibility measure Π. They prove in [30, Theorem 4] that the greatest such conditional possibility measure is given by natural extension, while the smallest such conditional possibility measure is determined by Dempster's rule, which produces the possibility measure associated to the following possibility distribution:

$$\pi_{DE}(x|y) = \begin{cases} \frac{\pi(x,y)}{\pi(y)} & \text{if } \pi(y) > 0 \\ 1 & \text{if } \pi(y) = 0. \end{cases}$$

Then in [30], it is advocated to use the harmonic mean between Dempster's rule and natural extension as an informative updating rule for updating a possibility measure Π. This harmonic mean determines the possibility measure defined by the possibility distribution $\pi_{HM}(x|y)$ given by

$$\begin{cases} \frac{2\pi(x,y)}{\pi(x,y)+\pi(y)+1-\max\{\pi(x,y),\Pi(\{y\}^c)\}} & \text{if } \pi(y) > 0 \\ 1 & \text{if } \pi(y) = 0. \end{cases}$$

However, this rule may be dominated by the regular extension, that produces the conditional possibility measure $\pi_{RE}(x|y)$ given by

$$\begin{cases} \frac{\pi(x,y)}{\pi(x,y)+1-\max\{\pi(x,y),\Pi(\{y\}^c)\}} & \text{if } \pi(y^c) < 1 \\ 0 & \text{if } \Pi(\{y\}^c) = 1, \pi(y) > \pi(x,y) = 0 \\ 1 & \text{otherwise}, \end{cases}$$

and as a consequence it is not a valid updating rule if we are working with upper previsions instead of upper probabilities. Consider the following example:

Example 4. Consider $\mathcal{X} = \{x_1, x_2\}, \mathcal{Y} = \{y_1, y_2\}$ and let Π be the possibility measure associated to the possibility distribution $\pi(x_1, y_1) = 0.3, \pi(x_1, y_2) = 1, \pi(x_2, y_1) = 0.5$ and $\pi(x_2, y_2) = 0.2$. Then it can be checked that the conditional possibility measure determined by the harmonic mean satisfies $\pi_{HM}(x_2|y_2) = 0.235$, whereas both the natural and the regular extensions produce $\pi_{NE}(x_2|y_2) = \pi_{RE}(x_2|y_2) = 0.285$. Thus, the conditional possibility measure determined by the harmonic mean is dominated by the one produced by regular extension, and as a consequence the conditional upper prevision determined by means of the Choquet integral with respect to $\Pi_{HM}(X|Y)$ is not coherent with the unconditional upper prevision associated to Π. ♦

6 Conclusions

In this work we have considered the problem of updating a coherent lower prevision into a conditional one, while preserving the property of coherence. This problem has a simple solution when the conditioning event has a positive lower probability, as showed by Walley in [27]: it suffices to apply Generalised Bayes Rule. However, when the conditioning event has zero lower probability and strictly positive upper probability, there may be an infinite number of coherent updated models. In that case, it becomes necessary to determine a rule to elicit the appropriate one for the problem at hand. Here, we have studied in which cases we can skip this situation, because the procedures of natural and regular extension give rise to the same updated model. We have considered the particular case when our unconditional model satisfies the property of 2-monotonicity, which guarantees that the lower prevision is the Choquet integral of the coherent lower probability that is its restriction to events, and we have obtained necessary and sufficient conditions for the equality between the natural and regular extensions. As particular cases, we have also considered the updating problem for completely monotone lower previsions, random sets and minimum-preserving previsions.

It is interesting to remark that the conditional lower probabilities determined by the natural and regular extension preserve the property of n-monotonicity from the unconditional model; in fact, when the conditioning event has zero lower probability and positive upper probability, they are moreover minimum-preserving. However, the conditional lower previsions they determine are not necessarily 2-monotone, even if we start from a completely monotone coherent lower prevision, as we have showed in Example 3. On the other hand, the properties of the natural and the regular extension are not shared in general by all the conditional models that are coherent with the unconditional one.

Finally, let us stress once again that, even if the property of 2-monotonicity means that the lower prevision is uniquely determined by its lower probability, the problem of coherently updating 2-monotone lower probabilities is not equivalent to that of updating 2-monotone lower previsions; this can be seen from the results in Section 5.1.

With respect to the open problems arising from this work, perhaps the most important one would be the extension of our results to infinite spaces. Although some work in this direction was already carried out in [20], we expect the problem to be much more difficult; one of the reasons is that the coherence condition between the unconditional and conditional lower previsions must take into account the property of conglomerability. See [27, Chapter 6] and [22] for more details. Another interesting line of research may be the extension of our work to the updating by means of several partitions. In that case, we should distinguish between the notions of weak and strong coherence studied by Walley in [27, Chapter 7].

Acknowledgements

The research in this paper was originated by some discussions with Didier Dubois during the 5th SIPTA School on Imprecise Probabilities. His insight and useful suggestions are much appreciated. We also acknowledge the financial support by project MTM2010-17844 and by the Science and Education Ministry FPU grant AP2009-1034.

References

[1] R. Aumann and L. S. Shapley. *Values of non-atomic games.* Princeton University Press, 1974.

[2] G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1953–1954.

[3] G. Coletti and R. Scozzafava. *Probabilistic logic in a coherent setting.* Kluwer, 2002.

[4] L. M. de Campos, M. T. Lamata, and S. Moral. The concept of conditional fuzzy measures. *International Journal of Intelligent Systems*, 5:237–246, 1990.

[5] G. de Cooman and D. Aeyels. Supremum preserving upper probabilities. *Information Sciences*, 118:173–212, 1999.

[6] G. de Cooman and E. Miranda. Lower previsions induced by filter maps. Submitted for publication, 2012.

[7] G. de Cooman, M. C. M. Troffaes, and E. Miranda. n-Monotone exact functionals. *Journal of Mathematical Analysis and Applications*, 347:143–156, 2008.

[8] G. de Cooman and M. Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*, 159(1-2):75–125, 2004.

[9] B. de Finetti. *Teoria delle Probabilità.* Einaudi, Turin, 1970.

[10] B. de Finetti. *Theory of Probability: A Critical Introductory Treatment*, volume 1. John Wiley & Sons, Chichester, 1974. English translation of [9].

[11] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[12] D. Denneberg. *Non-Additive Measure and Integral.* Kluwer Academic, Dordrecht, 1994.

[13] L. E. Dubins. Finitely additive conditional probabilities, conglomerability and disintegrations. *The Annals of Probability*, 3:88–99, 1975.

[14] R. Fagin and J. Y. Halpern. A new approach to updating beliefs. In P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, volume 6, pages 347–374. North-Holland, Amsterdam, 1991.

[15] J.-Y. Jaffray. Bayesian updating and belief functions. *IEEE Transactions on Systems, Man and Cybernetics*, 22:1144–1152, 1992.

[16] R. Kruse and K. D. Meyer. *Statistics with vague data.* D. Reidel Publishing Company, Dordrecht, 1987.

[17] V. P. Kuznetsov. *Interval Statistical Methods.* Radio i Svyaz Publ., 1991. (in Russian).

[18] I. Levi. *The enterprise of knowledge.* MIT Press, Cambridge, 1980.

[19] M. Marinacci and L. Montrucchio. Introduction to the mathematics of ambiguity. In I. Gilboa, editor, *Uncertainty in economic theory.* Routledge, New York, 2004.

[20] E. Miranda. Updating coherent lower previsions on finite spaces. *Fuzzy Sets and Systems*, 160(9):1286–1307, 2009.

[21] E. Miranda, I. Couso, and P. Gil. Upper probabilities and selectors of random sets. In P. Grzegorzewski, O. Hryniewicz, and M. A. Gil, editors, *Soft methods in probability, statistics and data analysis*, pages 126–133. Physica-Verlag, Heidelberg, 2002.

[22] E. Miranda, M. Zaffalon, and G. de Cooman. Conglomerable natural extension. *International Journal of Approximate Reasoning*, 53(8):1200–1227, 2012.

[23] H. T. Nguyen. *An introduction to random sets.* Chapman and Hall, 2006.

[24] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ, 1976.

[25] C. Sundberg and C. Wagner. Generalized finite differences and Bayesian conditioning of Choquet capacities. *Advances in Applied Mathematics*, 13(3):262–272, 1992.

[26] P. Walley. Coherent lower (and upper) probabilities. Technical Report Statistics Research Report 22, University of Warwick, Coventry, 1981.

[27] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.

[28] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996. With discussion.

[29] P. Walley. Measures of uncertainty in expert systems. *Artificial Intelligence*, 83(1):1–58, 1996.

[30] P. Walley and G. de Cooman. Coherence of rules for defining conditional possibility. *International Journal of Approximate Reasoning*, 21:63–107, 1999.

[31] P. Williams. On an new theory of epistemic probability. *British Journal for the Philosophy of Science*, 29:375–387, 1978.

Computing the Conglomerable Natural Extension

Enrique Miranda
University of Oviedo, Spain
mirandaenrique@uniovi.es

Marco Zaffalon
IDSIA, Lugano, Switzerland
zaffalon@idsia.ch

Abstract

Given a coherent lower prevision $\underline{P}$, we consider the problem of computing the smallest coherent lower prevision $\underline{F} \geq \underline{P}$ that is conglomerable, in case it exists. $\underline{F}$ is called the conglomerable natural extension. Past work has showed that $\underline{F}$ can be approximated by an increasing sequence $(\underline{E}_n)_{n\in\mathbb{N}}$ of coherent lower previsions. We close an open problem by showing that this sequence can be made of infinitely many distinct elements. Moreover, we give sufficient conditions, of quite broad applicability, to make sure that the point-wise limit of the sequence is $\underline{F}$ in case $\underline{P}$ is the lower envelope of finitely many linear previsions. In addition, we study the question of the existence of $\underline{F}$ and its relationship with the notion of marginal extension.

Keywords. Coherent lower previsions, conglomerability, conglomerable natural extension, natural extension, marginal extension.

1 Introduction

When the possibility space Ω is infinite and you express your beliefs through a coherent lower prevision $\underline{P}$, you may want to consider a partition $\mathcal{B}$ of Ω made of infinitely many conditioning events. In this case it may happen that $\underline{P}$ is not coherent, in Walley's sense, with any lower prevision conditional on $\mathcal{B}$; we say that $\underline{P}$ is not conglomerable.[1]

Conglomerability is a concern for Walley's theory, because its failure makes it impossible to update $\underline{P}$. More generally speaking, conglomerability should arguably be a rationality requirement for a probabilistic model under a dynamic interpretation of conditioning that relates present and future commitments, as detailed in [9].[2]

If we endorse conglomerability as a rationality requirement and consider a non-conglomerable coherent lower prevision $\underline{P}$, it becomes interesting to consider the *conglomerable natural extension* of $\underline{P}$, if it exists: that is, the weakest conglomerable coherent lower prevision $\underline{F}$ that extends $\underline{P}$. Thus, it plays the analogous role that the natural extension of a lower prevision (which avoids sure loss) plays with respect to coherence. Some recent work [5] has showed that $\underline{F}$ can be approximated though a sequence of coherent lower previsions $(\underline{E}_n)_{n\in\mathbb{N}}$ such that $\underline{P} \leq \underline{E}_1 \leq \underline{E}_2 \leq \cdots \leq \underline{E}_i \leq \cdots \leq \underline{F}$. It is known already that if the sequence becomes stable, that is, if $\underline{E}_{i-1} = \underline{E}_i$ for some i, then $\underline{E}_i = \underline{F}$; and, conversely, if the sequence breaks down, which means that $\underline{E}_i$ cannot be produced for some i, then $\underline{F}$ does not exist.

However, some fundamental questions have been left open with regard to the sequence $(\underline{E}_n)_n$. One of them is whether or not it may be infinite—without ever becoming stable. If that is the case, then the next question is whether or not the point-wise limit $\underline{Q}$ of the sequence equals $\underline{F}$. In fact, in principle it could be the case that $\underline{Q}$ is not conglomerable while $\underline{F}$ exists; this would mean that you should re-start a new sequence from $\underline{Q}$ in order to get to $\underline{F}$ (and possibly another, and another, and another, etc.).

After some introductory concepts we give in Section 2, we start a preliminary analysis in Section 3: we show that some basic procedures, like taking point-wise limits, or convex combinations, of conglomerable models do not preserve conglomerability in general. In Section 4 we discuss the question of the existence of $\underline{F}$ and its relationship with some pre-existing concepts about coherent lower previsions. In particular, Example 3 yields one more negative, and yet important, result: that $\underline{F}$ may not exist even when $\underline{P}$ avoids partial loss with its *conditional natural extension* $\underline{P}(\cdot|\mathcal{B})$, i.e., the model obtained by conditioning $\underline{P}$ in the least-committal way.

In Section 5 we close the first question mentioned above: we construct in Example 4 a model $\underline{P}$ whose related sequence $(\underline{E}_n)_n$ is infinite. In this case the limit $\underline{Q}$ of the sequence equals $\underline{F}$, which does not allow us to close the second question, which remains thus open.

In Section 6 we deepen the study, preliminarily started in [5], on the relationship between *marginal extension* and the conglomerable natural extension. We consider in particular

[1] We consider the case of a fixed partition $\mathcal{B}$ in this paper, so we do not deal with *full conglomerability*. See [6, Sections 6.8 and 6.9] for more details on the latter notion.

[2] On the other hand, there are also works on conditional previsions where conglomerability is not imposed [1, 3, 7]; we refer to [6, Section 6.8] for some discussion on this topic.

the relationship between $(\underline{E}_n)_n$ and the sequence $(\underline{M}_n)_n$, where $\underline{M}_n := \underline{E}_{n-1}(\underline{E}_{n-1}(\cdot|\mathcal{B}))$ is the marginal extension of $\underline{E}_{n-1}$ and its conditional natural extension $\underline{E}_{n-1}(\cdot|\mathcal{B})$. It turns out that $(\underline{M}_n)_n$ is also an increasing sequence of coherent lower previsions that is dominated by $\underline{F}$; however we show in Example 5 that the point-wise limit $\underline{Q}'$ of the sequence $(\underline{M}_n)_n$ may differ from $\underline{F}$. In addition, by detailing the relationships among $\underline{P}$, $\underline{Q}$, $\underline{Q}'$ and $\underline{F}$ we deduce in Proposition 8 that if $(\underline{E}_n(\cdot|\mathcal{B}))_n$ converges uniformly to the conditional natural extension $\underline{Q}(\cdot|\mathcal{B})$ of $\underline{Q}$, then $\underline{Q} = \underline{F}$.

In Section 7 we focus on the special case where $\underline{P}$ is dominated by a set of linear previsions with finitely many extreme points. This allows us to deduce two new simple conditions, which seem to be quite broadly applicable, that make sure that $(\underline{E}_n(\cdot|\mathcal{B}))_n$ converges uniformly to $\underline{Q}(\cdot|\mathcal{B})$, and hence, through Proposition 8, that $\underline{Q} = \underline{F}$. This analysis shows in particular that, when $\underline{P}$ is the lower envelope of two linear previsions, there is a procedure to determine whether $\underline{F}$ exists, and in this case we always have that $\underline{Q} = \underline{F}$.

We report our summary views in Section 8. Due to limitations of space, proofs have been omitted.

2 Introduction to Imprecise Probabilities

Let us introduce the basics of the theory of coherent lower previsions that we use in this paper. We refer to [6] for an in-depth study, and for a behavioural interpretation of the following notions in terms of buying and selling prices.

Consider a possibility space Ω. A *gamble* is a bounded map $f\colon \Omega \to \mathbb{R}$. The set of all gambles is denoted by $\mathcal{L}(\Omega)$, or simply by $\mathcal{L}$ when there is no ambiguity about the possibility space we are working with.

A *lower prevision* $\underline{P}$ is a real-valued functional defined on some set of gambles $\mathcal{K} \subseteq \mathcal{L}$. When the domain $\mathcal{K}$ of $\underline{P}$ is a linear space—closed under point-wise addition and multiplication by real numbers—$\underline{P}$ is called *coherent* when it satisfies the following conditions:

C1. $\underline{P}(f) \geq \inf f \; \forall f \in \mathcal{K}$;

C2. $\underline{P}(\lambda f) = \lambda \underline{P}(f) \; \forall f \in \mathcal{K}, \lambda \geq 0$;

C3. $\underline{P}(f+g) \geq \underline{P}(f) + \underline{P}(g) \; \forall f, g \in \mathcal{K}$.

Given a partition[3] $\mathcal{B}$ of Ω, a *conditional lower prevision* on $\mathcal{L}$ is a functional $\underline{P}(\cdot|\mathcal{B}) := \sum_{B\in\mathcal{B}} B\underline{P}(\cdot|B)$ such that for every set $B \in \mathcal{B}$, $\underline{P}(\cdot|B)$ is a lower prevision on $\mathcal{L}$. $\underline{P}(\cdot|\mathcal{B})$ is called *separately coherent* when $\underline{P}(\cdot|B)$ is coherent and $\underline{P}(B|B) = 1$ for every $B \in \mathcal{B}$. For every gamble f, $\underline{P}(f|\mathcal{B})$ is the gamble on Ω that takes the value $\underline{P}(f|B)$ on $\omega \in B$, and this for every $B \in \mathcal{B}$.

For every lower prevision $\underline{P}$ and every conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$, we use the notations: $G_{\underline{P}}(f) := f - \underline{P}(f)$, $G_{\underline{P}}(f|B) := B(f - \underline{P}(f|B))$ and $G_{\underline{P}}(f|\mathcal{B}) := f - \underline{P}(f|\mathcal{B}) = \sum_{B\in\mathcal{B}} G_{\underline{P}}(f|B)$. If we consider a coherent lower prevision $\underline{P}$ on $\mathcal{L}$ and a separately coherent conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$ on $\mathcal{L}$, they are called *coherent*[4] if and only if for every gamble f and every $B \in \mathcal{B}$,

$$\underline{P}(G_{\underline{P}}(f|\mathcal{B})) \geq 0, \qquad \text{(CNG)}$$

$$\underline{P}(G_{\underline{P}}(f|B)) = 0. \qquad \text{(GBR)}$$

This second condition is called the *generalised Bayes rule*, and if $\underline{P}(B) > 0$ it can be used to uniquely determine the value $\underline{P}(f|B)$: in that case there is only one value satisfying (GBR) with respect to $\underline{P}$. On the other hand, (CNG) is a *conglomerability* condition based on the behavioral idea that $G_{\underline{P}}(f|\mathcal{B})$ is a combination of (possibly infinitely many) acceptable transactions, and should be then an acceptable transaction, too.

A particular case of coherent $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ is that made of the *vacuous* unconditional and conditional lower previsions, given by $\underline{P}(f) = \inf_{\omega\in\Omega} f(\omega)$ and $\underline{P}(f|B) = \inf_{\omega\in B} f(\omega)$ for all $f \in \mathcal{L}$ and all $B \in \mathcal{B}$.

On the other hand, a coherent lower prevision $\underline{P}$ and a separately coherent conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$ on $\mathcal{L}$ are said to *avoid partial loss* (APL) when

$$\sup \left[G_{\underline{P}}(f) + G_{\underline{P}}(g|\mathcal{B}) \right] \geq 0 \qquad (1)$$

for every pair of gambles $f, g \in \mathcal{L}$. Eq. (1) holds whenever $\underline{P}(\cdot|\mathcal{B})$ is the vacuous conditional lower prevision irrespective of the coherent lower prevision $\underline{P}$, because in that case $G_{\underline{P}}(f|\mathcal{B}) \geq 0$ for any gamble f.

A particular case of coherent lower previsions is that of *linear* previsions. A linear prevision is a functional $P\colon \mathcal{L} \to \mathbb{R}$ satisfying conditions C1 and C2, and condition C3 with equality for all gambles $f, g \in \mathcal{L}$. Its restriction to $\mathcal{P}(\Omega)$, the powerset of Ω, is a finitely additive probability, and P is the corresponding expectation operator. The set of all linear previsions is denoted by $\mathbb{P}$. Given a lower prevision $\underline{P}$ on $\mathcal{K}$, its associated *credal set* is $\mathcal{M}(\underline{P}) := \{P \in \mathbb{P} \colon (\forall f \in \mathcal{K}) P(f) \geq \underline{P}(f)\}$, and each P in $\mathcal{M}(\underline{P})$ is said to *dominate* $\underline{P}$. A lower prevision for which $\mathcal{M}(\underline{P}) \neq \emptyset$ is said to *avoid sure loss*. It is coherent if and only if $\underline{P} = \min \mathcal{M}(\underline{P})$. Similarly, a *conditional linear prevision* is a functional $P(\cdot|\mathcal{B})$ on $\mathcal{L}$ such that $P(B|B) = 1$ and $P(\cdot|B)$ is a linear prevision for every $B \in \mathcal{B}$.

Given a coherent lower prevision $\underline{P}$, we define by

$$\underline{P}(f|B) := \begin{cases} \inf_{\omega\in B} f(\omega) & \text{if } \underline{P}(B) = 0 \\ \min\{P(f|B) : P \in \mathcal{M}(\underline{P})\} & \text{otherwise} \end{cases} \qquad (2)$$

[3] See also [7] for an alternative approach where the conditioning is made on a class of events that do not necessarily form a partition.

[4] See [6, Section 6.3.2] for a definition of coherence on more general domains, and also [6, Theorem 6.5.3].

its *conditional natural extension*. $\underline{P}(f|B)$ is a separately coherent lower prevision, defined for every $B \in \mathcal{B}$ and every $f \in \mathcal{L}$, which always satisfies (GBR) with $\underline{P}$. Thus, $\underline{P}, \underline{P}(\cdot|B)$ are coherent if and only if (CNG) holds for every gamble f. When that is the case, we say that $\underline{P}$ is a *conglomerably coherent lower prevision*. We refer to [6, Sections 6.8 and 6.9] for a thorough study of conglomerability. For the purposes of this paper, the most important property is that a conglomerably coherent lower prevision is one that can be updated to a conditional lower prevision while satisfying Walley's notion of coherence, so it is essential if we want to use Walley's approach in the conditional case.

Conglomerability holds trivially whenever $\underline{P}(B) = 0$ for all but a finite number of conditioning events $B \in \mathcal{B}$. Moreover, (CNG) always holds whenever the *support* of the gamble f, which is given by $S(f) := \{B \in \mathcal{B} : Bf \neq 0\}$ is finite. In particular, this means that conglomerability holds trivially for finite partitions.

3 Basic Properties of Conglomerability

Let us begin by making a preliminary study of conglomerably coherent lower previsions. Unlike the family of coherent lower previsions (see [6, Section 2.6]), the set of conglomerably coherent lower previsions is not closed under convex combinations or point-wise limits. We begin by focusing on this second property:

Example 1. Consider a partition $\mathcal{B}$ of Ω and two linear previsions P_1, P_2 on $\mathcal{L}$ such that P_1 is conglomerable and P_2 is not for a countable partition $\mathcal{B} := \{B_n : n \in \mathbb{N}\}$ such that $P_1(B_n), P_2(B_n) > 0$ for all n. (In this paper $\mathbb{N}$ denotes the set of positive natural numbers.)

Define Q_n on $\mathcal{L}$ by $Q_n(f) := P_2(f\mathbb{I}_{\cup_{i=1}^{n} B_i}) + P_1(f\mathbb{I}_{\cup_{i>n} B_i})$; it can easily be checked that Q_n is a linear prevision. Moreover, $Q_n(f|B_m)$ is equal to $P_2(f|B_m)$ if $m \leq n$ and to $P_1(f|B_m)$ if $m > n$, whence $Q_n(Q_n(f|\mathcal{B})) = Q_n(f)$.

This means that the linear prevision Q_n is conglomerable for every n. On the other hand, $\lim_n Q_n(f) = P_2(f)$ for every f, so the limit of the sequence $(Q_n)_n$ is not a conglomerable prevision.

The above comments also show that the coherence of an unconditional and a conditional lower prevision is not preserved by point-wise limits: since $Q_n(B_m) > 0$ for all $m, n \in \mathbb{N}$, we deduce that Q_n is coherent with its conditional natural extension $Q_n(\cdot|\mathcal{B})$, which is a linear prevision. However, the point-wise limit of the sequence $(Q_n)_n$, that is, the linear prevision P_2, is not coherent with its conditional natural extension $P_2(\cdot|\mathcal{B})$ because P_2 is not conglomerable. It also follows that $\lim_n Q_n(f|B_m) = P_2(f|B_m)$ for all $m \in \mathbb{N}, f \in \mathcal{L}$, whence $P_2(\cdot|\mathcal{B})$ is the limit of $Q_n(f|\mathcal{B})$. Thus $Q_n, Q_n(\cdot|\mathcal{B})$ are coherent for all n but their point-wise limits $P_2, P_2(\cdot|\mathcal{B})$ are not. ♦

Next, we investigate if the property of conglomerability is preserved by taking convex combinations. As discussed by Walley in [6, Theorem 6.9.1], a sufficient condition for a linear prevision P to be conglomerable is that it is countably additive on $\mathcal{B}$, in the sense that $\sum_{B \in \mathcal{B}} P(B) = 1$. This means in particular that a convex combination of two linear previsions P_1, P_2 that are countably additive on $\mathcal{B}$ will again be countably additive with respect to this partition, and as a consequence it will also be conglomerable.

However, there are also conglomerable linear previsions P that are not countably additive on $\mathcal{B}$ [6, Examples 6.6.4, 6.6.5], and they can be used to show that conglomerability is not necessarily preserved by convex combinations:

Example 2. Consider $\Omega := \mathbb{N} \cup -\mathbb{N}$, $B_n := \{-n, n\}$ and the partition $\mathcal{B} := \{B_n : n \in \mathbb{N}\}$. Let P_1, P_2 be two linear previsions whose restrictions to events satisfy

$$P_1(B_n) = \begin{cases} \frac{1}{2^n} & \text{if } n \text{ odd}, \\ 0 & \text{if } n \text{ even}, \end{cases} \qquad P_1(\{2n\}_{n\in\mathbb{N}}) = \tfrac{1}{3},$$

$$P_2(B_n) = \begin{cases} \frac{1}{2^{n-1}} & \text{if } n \text{ even}, \\ 0 & \text{if } n \text{ odd}, \end{cases} \qquad P_2(\{2n-1\}_{n\in\mathbb{N}}) = \tfrac{1}{3};$$

that is, P_1 (resp., P_2) is countably additive on $\cup_{n\in\mathbb{N}} B_{2n-1}$ (resp., $\cup_{n\in\mathbb{N}} B_{2n}$) and purely finitely additive on $\cup_{n\in\mathbb{N}} B_{2n}$ (resp., $\cup_{n\in\mathbb{N}} B_{2n-1}$). Assume moreover that $P_1(\{n\}) = P_1(\{-n\})$ and $P_2(\{n\}) = P_2(\{-n\})$ for every n.

For any gamble f on Ω, it holds that $P_1(G_1(f|\mathcal{B})) \geq P_1(G_1(f\mathbb{I}_{\cup_{n\in\mathbb{N}} B_{2n-1}}|\mathcal{B}))$, taking into account that $\underline{P}_1(\cdot|B_{2n})$ is vacuous for every n and as a consequence $G_1(f\mathbb{I}_{\cup n\in\mathbb{N}} B_{2n}) \geq 0$. Moreover, if we consider the set $D := \cup_{n\in\mathbb{N}} B_{2n}$ and the partition $\mathcal{B}' := \{D\} \cup \{B_{2n-1} : n \in \mathbb{N}\}$ of Ω, it follows that $\sum_{B' \in \mathcal{B}'} P_1(B') = 1$. Applying [6, Theorem 6.9.1], it follows that P_1 is conglomerable with respect to $\mathcal{B}'$, and from this we deduce that $P_1(G_1(f\mathbb{I}_{\cup_{n\in\mathbb{N}} B_{2n-1}}|\mathcal{B})) = P_1(G_1(f\mathbb{I}_{\cup_{n\in\mathbb{N}} B_{2n-1}}|\mathcal{B}')) \geq 0$. As a consequence, P_1 is conglomerable. Similarly, so is P_2. However, if we consider the linear prevision $P := 0.5P_1 + 0.5P_2$, it holds that $P(f|B_n) = \frac{f(n)+f(-n)}{2}$ $\forall n \in \mathbb{N}, f \in \mathcal{L}$. Given $f := 2\mathbb{I}_{-\mathbb{N}}$, it follows that $P(f|B_n) = 1$ for every n, whence $P(G(f|\mathcal{B})) = \frac{1}{3} - \frac{2}{3} < 0$, since $P_1(\mathbb{N}) = P_2(\mathbb{N}) = \frac{2}{3}$ by construction. This shows that P is not conglomerable. ♦

4 On the Existence of the Conglomerable Natural Extension

The above preliminary results illustrate the fact that conglomerably coherent lower previsions do not share many of the properties of coherent lower previsions. Another instance of this is that a lower prevision $\underline{P}$ that avoids sure loss has always a smallest dominating coherent lower prevision, but it may not have a dominating conglomerably coherent lower prevision. This is easy to see by means of a linear prevision P that is not conglomerable: any conglomerably coherent lower prevision $\underline{F}$ that dominates P should also coincide with P, because of linearity, and as a consequence such an $\underline{F}$ does not exist.

Although in Section 3 we have showed that the limit of a sequence of conglomerable lower previsions may not be conglomerable, it follows from [6, Theorem 6.9.3] that the lower envelope of a family of conglomerable lower previsions is again conglomerable. Hence, if $\underline{P}$ has a dominating conglomerable model, then there is also a smallest dominating conglomerable model. We shall refer to it as the conglomerable natural extension of $\underline{P}$.

Definition 1. Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}$ and let $\mathcal{B}$ be a partition of Ω. The *($\mathcal{B}$-)conglomerable natural extension* of $\underline{P}$ is the smallest coherent lower prevision $\underline{F} \geq \underline{P}$ that is conglomerable with respect to $\mathcal{B}$.

As we have showed before, the conglomerable natural extension of a lower prevision $\underline{P}$ may not exist. Taking this into account, it becomes interesting to provide sufficient conditions for its existence. We begin by investigating the relationships among a number of consistency notions from [6, Chapters 6 and 7]:

Proposition 1. *Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}$, $\mathcal{B}$ a partition of Ω, and $\underline{P}(\cdot|\mathcal{B})$ a separately coherent lower prevision. Consider the following possibilities:*

(a) *$\underline{P}, \underline{P}(\cdot|\mathcal{B})$ are coherent.*

(b) *$\underline{P}, \underline{P}(\cdot|\mathcal{B})$ are dominated by coherent $\underline{Q}, \underline{Q}(\cdot|\mathcal{B})$.*

(c) *The conglomerable natural extension of $\underline{P}$ exists.*

(d) *$\underline{P}, \underline{P}(\cdot|\mathcal{B})$ are dominated by $\underline{Q}, \underline{Q}(\cdot|\mathcal{B})$ that avoid partial loss.*

(e) *$\underline{P}, \underline{P}(\cdot|\mathcal{B})$ avoid partial loss.*

Then (a)⇒(b)⇒(d)⇔(e) *and* (b)⇒(c)*. If, in addition, $\underline{P}(\cdot|\mathcal{B})$ is the conditional natural extension of $\underline{P}$, then* (c) ⇒ (b) *holds as well, and if in particular $\underline{P}$ is linear then we have also that* (b) ⇒ (a) *and* (d) ⇒ (b)*, so all of them are equivalent conditions.*

Now, if we consider a coherent lower prevision $\underline{P}$, it follows that its conglomerable natural extension exists if and only if there is a coherent lower prevision $\underline{F} \geq \underline{P}$ that is conglomerable. Since conglomerability is equivalent to the coherence with the conditional natural extension, it follows that the conglomerable natural extension of $\underline{P}$ exists if and only if $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ are dominated by coherent $\underline{Q}, \underline{Q}(\cdot|\mathcal{B})$, where $\underline{P}(\cdot|\mathcal{B})$ denotes the conditional natural extension of $\underline{P}$. We deduce from Proposition 1 that the following implications hold:

$$\underline{P} \text{ conglomerable } \Rightarrow \underline{F} \text{ exists} \Rightarrow \underline{P}, \underline{P}(\cdot|\mathcal{B}) \text{ APL}, \quad (3)$$

where $\underline{F}$ is the conglomerable natural extension of $\underline{P}$, introduced in Definition 1. Moreover, $\underline{F}$ exists if and only if $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ *avoid conglomerable partial loss*, in the sense of [4, Definition 21]. The converses of the implications in (3) do not hold in general: on the one hand, there are previsions $\underline{P}$ that are not conglomerable but whose conglomerable natural extension exists (one instance is that in Example 4 later on). Next we show that the converse of the second implication does not hold either. In other words, the conditions of avoiding partial loss and avoiding conglomerable partial loss are not equivalent in general. In order to build this example, we need to define the notion of unconditional natural extension:

Definition 2. Let $\underline{P}$ be a coherent lower prevision and $\underline{P}(\cdot|\mathcal{B})$ be a separately coherent conditional lower prevision on $\mathcal{L}$. Their *unconditional natural extension* $\underline{E}_1$ is given on f by the supremum α such that

$$f - \alpha \geq G_{\underline{P}}(g) + G_{\underline{P}}(h|\mathcal{B}) \text{ for some } g, h \in \mathcal{L}.$$

Then $\underline{E}_1$ is a coherent lower prevision on $\mathcal{L}$ if and only if $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ avoid partial loss. Moreover, if $\underline{P}(\cdot|\mathcal{B})$ is the conditional natural extension of $\underline{P}$ and $\underline{E}_1(\cdot|\mathcal{B})$ is that of $\underline{E}_1$, then any coherent $\underline{Q}, \underline{Q}(\cdot|\mathcal{B})$ that dominate $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ must also dominate $\underline{E}_1, \underline{E}_1(\cdot|\mathcal{B})$. Thus, the conglomerable natural extensions of $\underline{P}$ and $\underline{E}_1$ coincide.

Example 3. Consider $\Omega := \mathbb{N} \cup -\mathbb{N}$, $B_n := \{n, -n\}$ and $\mathcal{B} := \{B_n : n \in \mathbb{N}\}$. Let P_1 be a σ-additive linear prevision on $\mathcal{L}$ determined by $P_1(n) := P_1(\{-n\}) := \frac{1}{2^{n+1}}$.

Let P be a finitely additive probability on $\mathcal{P}(\mathbb{N})$ satisfying $P(\{n\}) = 0$ for all n, $P(\{2n+1 : n \in \mathbb{N}\}) = 0$. We can use it to define a linear prevision P_2 on $\mathcal{L}$ whose restriction to events is the finitely additive probability given by $P_2(B) := \frac{3}{4}P(\Pi_1(B)) + \frac{1}{4}P(\Pi_2(B))$, where $\Pi_1(B) := B \cap \mathbb{N}$ and $\Pi_2(B) := -(B \cap -\mathbb{N})$. Define then the linear prevision $P_3 := \frac{1}{2}P_1 + \frac{1}{2}P_2$.

Let now P' be another finitely additive probability on $\mathcal{P}(\mathbb{N})$ such that $P'(\{n\}) = 0$ for all n, $P'(\{2n+1 : n \in \mathbb{N}\}) = 0.5$, so that $P'(\mathbb{I}_{\text{even}}) = 0.5$ too. Let P_4 be the linear prevision on $\mathcal{L}$ whose restriction to events is the finitely additive probability

$$P_4(B) := \frac{1}{4} \sum_{n \in B \cap \mathbb{N}} \frac{1}{2^n} + \frac{3}{4}P'(-(B \cap -\mathbb{N})).$$

Take $\underline{P} := \min\{P_3, P_4\}$. Then $\underline{P}(B_n) = \min\left\{\frac{1}{2^{n+1}}, \frac{1}{2^{n+2}}\right\} > 0 \ \forall n \in \mathbb{N}$, whence $\underline{P}(f|B_n) = \min\left\{\frac{f(n)+f(-n)}{2}, f(n)\right\} \forall f \in \mathcal{L}, n \in \mathbb{N}$.

Fix a gamble f and let $C := \cup_{n: f(n) < f(-n)} B_n$, so that $\underline{P}(f|B_n) = f(n)$ if $B_n \subseteq C$ and $\underline{P}(f|B_n) = \frac{f(n)+f(-n)}{2}$ otherwise. Then $G(f|\mathcal{B}) = G(f \cdot C|\mathcal{B}) + G(f \cdot C^c|\mathcal{B}) \geq G(f \cdot C^c|\mathcal{B})$ because $G(f|B_n) \geq 0$ if $B_n \subseteq C$.

Denote $P_\alpha := \alpha P_3 + (1-\alpha)P_4$. We are going to determine for which $\alpha \in [0,1]$ it holds that $P_\alpha(G(f|\mathcal{B})) \geq 0$ for all f. Taking into account the above observation, we can conclude that $P_\alpha(G(f|\mathcal{B})) \geq 0\ \forall f \in \mathcal{L}$ if and only if $P_\alpha(G(f|\mathcal{B})) \geq 0\ \forall f \in \mathcal{L}$ s.t. $f(n) \geq f(-n)\ \forall n$.

Take thus f s.t. $f(n) \geq f(-n)$ for all n (in this case C is empty). Then

$$\begin{cases} G(f|B_n)(n) = \frac{f(n)-f(-n)}{2} \geq 0, \\ G(f|B_n)(-n) = \frac{f(-n)-f(n)}{2} \leq 0. \end{cases} \tag{4}$$

If we denote $g := G(f|\mathcal{B})$, it holds that $g(n) = -g(-n)$, whence $P_1(g\mathbb{I}_{\mathbb{N}}) + P_1(g\mathbb{I}_{-\mathbb{N}}) = 0$. On the other hand, $P_2(g\mathbb{I}_{\mathbb{N}}) + P_2(g\mathbb{I}_{-\mathbb{N}}) = \frac{3}{4}P(g^+) + \frac{1}{4}P(g^-)$, where

$$\begin{array}{rcl} g^+ : \mathbb{N} \to \mathbb{R} & & g^- : \mathbb{N} \to \mathbb{R} \\ n \hookrightarrow g(n) & \text{and} & n \hookrightarrow g(-n) = -g(n). \end{array} \tag{5}$$

Thus $P_2(g\mathbb{I}_{\mathbb{N}}) + P_2(g\mathbb{I}_{-\mathbb{N}}) = \frac{1}{2}P(g^+) \geq 0$; as a consequence, $P_3(G(f|\mathcal{B})) \geq 0$ for every gamble f.

Now, if in particular we fix $n \in \mathbb{N}$ and let $f := 2\mathbb{I}_{\{2n+1,2n+3,\dots\}}$, then, using (4) again, $G(f|\mathcal{B}) = \mathbb{I}_{\{2n+1,2n+3,\dots\}} - \mathbb{I}_{\{-2n-1,-2n-3,\dots\}}$ and $P_1(G(f|\mathcal{B})) = 0 = P_2(G(f|\mathcal{B}))$, because we have chosen P such that $P(\{2n+1 : n \in \mathbb{N}\}) = 0$. Thus, $P_3(G(f|\mathcal{B})) = 0$.

On the other hand, for this gamble f we obtain that $P_4(G(f|\mathcal{B})) = \sum_{k \geq n} \frac{1}{2^{(2k+1)+2}} - \frac{3}{8} < 0$ for n big enough.

This implies that $P_\alpha(G(f|\mathcal{B})) < 0$ for all $\alpha \neq 1$. As a consequence $\{P_\alpha : P_\alpha(G(f|\mathcal{B})) \geq 0\ \forall f\} = P_3 = \underline{E}_1$, taking into account that $\mathcal{M}(\underline{P}) = \{P_\alpha : \alpha \in [0,1]\}$ and using [5, Proposition 13]. Since the natural extension $\underline{E}_1$ of $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ exists, it follows that $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ avoid partial loss. But P_3 is not conglomerable: given $g := 2\mathbb{I}_{-\mathbb{N}}$, we can use the expression of $P_3(\cdot|B_n)$ (available from that of $\underline{P}(\cdot|B_n)$) to see that $P_3(g|B_n) = \frac{[2\mathbb{I}_{-\mathbb{N}}](n)+[2\mathbb{I}_{-\mathbb{N}}](-n)}{2} = \frac{2}{2} = 1$, so that $G_{P_3}(g|\mathcal{B}) = -\mathbb{I}_{\mathbb{N}} + \mathbb{I}_{-\mathbb{N}}$ and $P_3(G(g|\mathcal{B})) = -\frac{1}{4} < 0$. Thus $P_3, P_3(\cdot|\mathcal{B})$ do not avoid partial loss, and applying (3) we deduce that the conglomerable natural extension of P_3 does not exist. But since P_3 is the natural extension of $\underline{P}, \underline{P}(\cdot|\mathcal{B})$, the conglomerable natural extension of $\underline{P}$ coincides with that of P_3. Hence, the conglomerable natural extension of $\underline{P}$ does not exist either. ⧫

We can get more, and different, results in the special case where the conditional natural extension of $\underline{P}$ is linear.

Proposition 2. *Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}$ and assume that its conditional natural extension is a linear prevision $P(\cdot|\mathcal{B})$. Then:*

(a) *$\underline{P}, P(\cdot|\mathcal{B})$ avoid partial loss if and only if $\underline{P}, P(\cdot|\mathcal{B})$ avoid conglomerable partial loss.*[5]

(b) *$\underline{P}$ is conglomerable if and only if it is a lower envelope of conglomerable linear models.*

From [6, Theorem 6.9.3], a lower envelope of a family of conglomerable lower previsions is again a conglomerable lower prevision; the converse is not true: [6, Example 6.6.9] shows that it may be that $\underline{P}$ is a conglomerably coherent lower prevision but no dominating model is. One interesting particular case where an assessment of conglomerability is compatible with an envelope theorem is when we are dealing with marginal extension models [6, Theorem 6.7.4]: any marginal extension is a conglomerable model that is a lower envelope of a family of conglomerable linear previsions. Proposition 2 provides an instance of this case.

5 Approximation by a Sequence

In [5], it was devised a procedure to approximate the conglomerable natural extension (if it exists) of a coherent lower prevision $\underline{P}$: we consider the sequence of coherent lower previsions $(\underline{E}_n)_n$, where $\underline{E}_0 := \underline{P}$ and for every $n \geq 1$, $\underline{E}_n$ is the (unconditional) natural extension of $\underline{E}_{n-1}, \underline{E}_{n-1}(\cdot|\mathcal{B})$, where $\underline{E}_{n-1}(\cdot|\mathcal{B})$ is the conditional natural extension of $\underline{E}_{n-1}$, given by Eq. (2).

Proposition 3. *[5] Assume that the conglomerable natural extension $\underline{F}$ of $\underline{P}$ exists. Then:*

1. *$(\underline{E}_n)_n$ is an increasing sequence of coherent lower previsions, and $(\underline{E}_n(\cdot|\mathcal{B}))_n$ is an increasing sequence of separately coherent conditional lower previsions.*
2. *Given their point-wise limits $\underline{Q}, \underline{Q}(\cdot|\mathcal{B})$, it holds that $\underline{Q}(\cdot|\mathcal{B})$ is the conditional natural extension of $\underline{Q}$.*
3. *$\underline{Q} \leq \underline{F}$, and $\underline{Q} = \underline{F} \Leftrightarrow \underline{Q}$ is conglomerable.*

Moreover, it was showed in [5, Example 5] that the sequence may not stabilise in the first step, or, in other words, that the natural extension of $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ does not always coincide with the conglomerable natural extension.

In terms of credal sets, we have the following:

Proposition 4. *[5, Propositions 13 and 14] Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}$, $\mathcal{B}$ a partition of Ω and $\underline{P}(\cdot|\mathcal{B})$ its conditional natural extension. Let $\underline{E}$ be the unconditional natural extension of $\underline{P}, \underline{P}(\cdot|\mathcal{B})$. Then*

$$\begin{aligned} \mathcal{M}(\underline{E}) &= \{P \in \mathcal{M}(\underline{P}) : P(G_{\underline{P}}(f|\mathcal{B})) \geq 0\ \forall f \in \mathcal{L}\} \\ &= \mathcal{M}(\underline{P}) \cap \mathcal{M}(\underline{M}),\ \textit{where}\ \underline{M} := \underline{P}(\underline{P}(\cdot|\mathcal{B})). \end{aligned}$$

In this section, we are going to study the above sequence in more detail. It follows that if the sequence stabilises in a finite number of steps, i.e., if $\underline{Q} = \underline{E}_n$ for some n, then $\underline{Q}$ is the conglomerable natural extension of $\underline{P}$. However, as we shall see later, it may happen that the sequence is infinite. In order to provide an example, we are going to give a tool first that will allow us to build sequences that can be made both conglomerable and non-conglomerable, depending on the choice of two parameters.

[5] This has essentially been showed already in [5, Proposition 15].

Proposition 5. *Let P_1 be a σ-additive probability on $\mathcal{L}(\mathbb{N})$ such that $P_1(\{n\}) > 0$ for all $n \in \mathbb{N}$; let P_2 be a finitely additive probability on $\mathcal{P}(\mathbb{N})$ such that $P_2(\{n\}) = 0$ for all $n \in \mathbb{N}$. We consider $\Omega := \mathbb{N} \cup -\mathbb{N}$ and $\mathcal{B} := \{B_n : n \in \mathbb{N}\}$, with $B_n := \{n, -n\}$. Given a gamble h in $\mathcal{L}(\Omega)$, we let h^+, h^- be derived from h as in Eq.* (5)*. Consider $\alpha, \beta \in [0,1]$ and let Q_1, Q_2 on $\mathcal{L}(\Omega)$ be given by*

$$Q_1(h) := \alpha P_1(h^+) + (1-\alpha)P_1(h^-) \text{ and}$$
$$Q_2(h) := \beta P_2(h^+) + (1-\beta)P_2(h^-).$$

Consider also $\gamma \in (0,1)$ and let $Q := \gamma Q_1 + (1-\gamma)Q_2$. Then Q is conglomerable $\Leftrightarrow \alpha = \beta$.

We exploit Proposition 5 to show that the sequence $(\underline{E}_n)_n$ may not stabilise in a finite number of steps.

Example 4. Consider $\Omega := \mathbb{N} \cup -\mathbb{N}$, $\mathcal{B} := \{B_n : n \in \mathbb{N}\}$, with $B_n := \{n, -n\}$, and the linear previsions on $\mathcal{L}(\Omega)$

$$\begin{aligned} P_1(\{n\}) &:= P_1(\{-n\}) := \frac{1}{2^{n+1}} \text{ for all } n \in \mathbb{N} \\ P_2(h) &:= \frac{1}{2}\sum_n h(n)\frac{1}{2^n} + \frac{1}{2}P(h^-) \\ P_3(h) &:= \frac{3}{4}P(h^+) + \frac{1}{4}P(h^-) \\ P_4(h) &:= \frac{1}{2}P_1(h) + \frac{1}{2}P_3(h), \end{aligned}$$

where P is a finitely additive probability on $\mathbb{N}$ s.t. $P(\{n\}) = 0$ for all $n \in \mathbb{N}$ and h^+, h^- are determined by Eq. (5). Given $\alpha \in [0,1]$, we set $Q_\alpha := \alpha P_2 + (1-\alpha)P_4$. It follows that

$$\begin{aligned} Q_\alpha(h) &= \frac{1}{2}\left[\frac{1+\alpha}{2}\tilde{P}_1(h^+) + \frac{1-\alpha}{2}\tilde{P}_1(h^-)\right] \\ &+ \frac{1}{2}\left[\frac{3-3\alpha}{4}P(h^+) + \frac{1+3\alpha}{4}P(h^-)\right], \end{aligned}$$

where we denote by $\tilde{P}_1$ the linear prevision given by $\tilde{P}_1(\{n\}) := \frac{1}{2^n}$ for all $n \in \mathbb{N}$. Proposition 5 yields:

$$Q_\alpha \text{ is conglomerable} \Leftrightarrow \frac{1+\alpha}{2} = \frac{3-3\alpha}{4} \Leftrightarrow \alpha = 0.2.$$

Let $\underline{P}$ be the lower envelope of the credal set $\{Q_\alpha : \alpha \in [a,b]\}$ for given a, b s.t. $0 < a < 0.2 < b < 1$. The conglomerable natural extension of $\underline{P}$ exists since $\underline{P} \leq Q_{0.2}$. We aim at analysing whether the sequence of coherent lower previsions $\underline{P}, \underline{E}_1, \underline{E}_2, \ldots$, originated by $\underline{P}$, yields the conglomerable natural extension in the limit and whether or not the sequence itself stabilises in a finite number of steps.

We start by detailing the form of the conditional natural extension of $\underline{P}$. Since $Q_\alpha(f|B_n) = \frac{1+\alpha}{2}f(n) + \frac{1-\alpha}{2}f(-n) \quad \forall f \in \mathcal{L}$ and $\underline{P}(B_n) > 0$, it follows from Eq. (2) and [6, Theorem 6.4.2] that for every gamble f,

$$\underline{P}(f|B_n) = \begin{cases} \frac{1+a}{2}f(n) + \frac{1-a}{2}f(-n) & \text{if } f(n) \geq f(-n) \\ \frac{1+b}{2}f(n) + \frac{1-b}{2}f(-n) & \text{if } f(n) \leq f(-n). \end{cases}$$

If we denote $A := \{n \in \mathbb{N} : f(n) \leq f(-n)\}$, then

$$\begin{cases} G_{\underline{P}}(f|B_n)(n) = \frac{1-b}{2}[f(n) - f(-n)] \leq 0 \\ G_{\underline{P}}(f|B_n)(-n) = \frac{1+b}{2}[f(-n) - f(n)] \geq 0 \end{cases}$$

whenever $n \in A$, and

$$\begin{cases} G_{\underline{P}}(f|B_n)(n) = \frac{1-a}{2}[f(n) - f(-n)] \geq 0 \\ G_{\underline{P}}(f|B_n)(-n) = \frac{1+a}{2}[f(-n) - f(n)] \leq 0 \end{cases}$$

when $n \notin A$. Now we would like to check for which values of α it is the case that $Q_\alpha(G_{\underline{P}}(f|\mathcal{B})) \geq 0$ for all $f \in \mathcal{L}$, because from Proposition 4 we have that $\mathcal{M}(\underline{E}_1) = \{Q_\alpha : Q_\alpha(G_{\underline{P}}(f|\mathcal{B})) \geq 0 \text{ for all } f \in \mathcal{L}\}$.

Given a gamble f, its associated set $A = \{n \in \mathbb{N} : f(n) \leq f(-n)\}$, and $C := \cup_{n \in A} B_n$, it holds that $f = f\mathbb{I}_C + f\mathbb{I}_{C^c}$, whence $G_{\underline{P}}(f|\mathcal{B}) = G_{\underline{P}}(\mathbb{I}_C f|\mathcal{B}) + G_{\underline{P}}(\mathbb{I}_{C^c} f|\mathcal{B})$. Denote $g' := G_{\underline{P}}(\mathbb{I}_C f|\mathcal{B}), g'' := G_{\underline{P}}(\mathbb{I}_{C^c} f|\mathcal{B})$. We proceed to determine when $Q_\alpha(g') \geq 0, Q_\alpha(g'') \geq 0$.

- Let us consider $Q_\alpha(g')$. If $n \notin A$, then $g'(-n) = g'(n) = 0$; if $n \in A$, then $g'(-n) = \frac{1+b}{2}[f(-n) - f(n)]$ and $g'(n) = \frac{1-b}{2}[f(n) - f(-n)]$. As a consequence, $g'(-n) = -\frac{1+b}{1-b}g'(n) \geq 0$. Then:

  $$P_2(g') = \sum_n g'(n)\frac{1}{2^{n+1}} + \frac{1}{2}P(g'^-) \text{ and}$$

  $$P_4(g') = \sum_n g'(n)\frac{1}{2^{n+1}} \cdot \frac{-b}{1-b} + P(g'^-) \cdot \frac{1}{4} \cdot \frac{2b-1}{1+b}.$$

  This implies that $Q_\alpha(g')$ is equal to

  $$\underbrace{\underbrace{\sum_n g'(n)\frac{1}{2^{n+1}}}_{\leq 0} \cdot \underbrace{\frac{\alpha - b}{1-b}}_{\leq 0}}_{\geq 0} + \underbrace{P(g'^-) \cdot \frac{1}{4}}_{\geq 0} \cdot \frac{3\alpha + 2b - 1}{1+b},$$

  so that $3\alpha + 2b - 1 \geq 0 \Rightarrow Q_\alpha(g') \geq 0$. On the other hand, if $3\alpha + 2b - 1 < 0$, we can always find g', by letting g'^- tend to 1 with $n \to \infty$, such that $P(g'^-) = 1$, using that P is a finitely additive probability that is not σ-additive. And this is compatible with making $\sum_n g'(n)\frac{1}{2^{n+1}}$ as small as we want by making the first m images equal to zero, where m is an arbitrary positive number: it holds that $\lim_m P(g'\mathbb{I}_{\cup_{n \geq m} B_n}) = P(g')$ while $\lim_m \sum_{n \geq m} g'(n)\frac{1}{2^{n-1}} = 0$. We conclude that we can always find some g' such that $Q_\alpha(g') < 0$ when $3\alpha + 2b - 1 < 0$.

- Let us focus on $Q_\alpha(g'')$. It holds that $g''(n) = -\frac{1-a}{1+a}g''(-n) \geq 0$. Then:

  $$P_2(g'') = \sum_n g''(n)\frac{1}{2^{n+1}} + \frac{1}{2}P(g''^-) \text{ and}$$

  $$P_4(g'') = \sum_n g''(n)\frac{1}{2^{n+1}} \cdot \frac{-a}{1-a} + P(g''^-) \cdot \frac{1}{4} \cdot \frac{2a-1}{1+a}.$$

This implies that $Q_\alpha(g'')$ is given by

$$\underbrace{\sum_n g''(n)\frac{1}{2^{n+1}}\cdot\frac{\alpha-a}{1-a}}_{\geq 0}+\underbrace{P(g''^{-})\cdot\frac{1}{4}}_{\leq 0}\cdot\frac{3\alpha+2a-1}{1+a},$$

so that $3\alpha+2a-1\leq 0\Rightarrow Q_\alpha(g'')\geq 0$. On the other hand, if $3\alpha+2a-1>0$, we can reason as in the case of $Q_\alpha(g')$ to conclude that we can always find some g'' such that $Q_\alpha(g'')<0$.

Let us consider the case where $3\alpha+2b-1\geq 0$ and $3\alpha+2a-1\leq 0$ (note that we can attain this case given that $3b+2b-1\geq 0$ and $3a+2a-1\leq 0$ if and only if $a\leq 0.2\leq b$). Then $Q_\alpha(g')\geq 0, Q_\alpha(g'')\geq 0$ and therefore $Q_\alpha(g)\geq 0$; using Proposition 4 we obtain that $Q_\alpha\in\mathcal{M}(\underline{E}_1)$. On the other hand, in the case where $3\alpha+2b-1<0$ or $3\alpha+2a-1\leq 0$, we know that there is g' s.t. $Q_\alpha(g')<0$, and g'' s.t. $Q_\alpha(g'')=0$ (it is enough to use an f, in the definition of g'', s.t. $f(n)=f(-n)$ for all $n\notin A$); applying again Proposition 4, we obtain that $Q_\alpha\notin\mathcal{M}(\underline{E}_1)$. Analogous considerations hold for the remaining cases.

Thus, recalling that $\mathcal{M}(\underline{P})=\{Q_\alpha:\alpha\in[a,b]\}$, with $0<a<0.2<b<1$, it follows that $\mathcal{M}(\underline{E}_1)$ is given by the linear previsions Q_α where $\alpha\in\left[\max\left\{a,\frac{1-2b}{3}\right\},\min\left\{\frac{1-2a}{3},b\right\}\right]$. Note that if $a<b$ then it must be the case that $[\max\{a,\frac{1-2b}{3}\},\min\{\frac{1-2a}{3},b\}]\subsetneq[a,b]$, because it is not possible that both $a\geq\frac{1-2b}{3}$ and $b\leq\frac{1-2a}{3}$ hold. This means that at least one of the two extreme points of $[a,b]$ must change. Moreover, note that the new interval will have still to contain the value 0.2 properly, in the sense that 0.2 will have to be an interior point of the new interval, because

$$a<0.2<b\Rightarrow\max\left\{a,\frac{1-2b}{3}\right\}<0.2\text{ and}$$
$$a<0.2<b\Rightarrow\min\left\{b,\frac{1-2a}{3}\right\}>0.2.$$

Thus, the infinite sequence $\underline{P},\underline{E}_1,\underline{E}_2,\ldots$ is in correspondence with an infinite sequence of intervals of strictly decreasing length, each one containing 0.2 properly.

Let us show now that 0.2 is actually the limit of this sequence. We must consider a number of cases:

- If in the passage from $\mathcal{M}(\underline{P})$ to $\mathcal{M}(\underline{E}_1)$ both extreme points of the interval change, then we go from $[a,b]$ to $[\frac{1-2b}{3},\frac{1-2a}{3}]$, and the length of the new interval is two thirds of the length of the previous one.

- Assume otherwise that that in the passage from $\mathcal{M}(\underline{P})$ to $\mathcal{M}(\underline{E}_1)$ only the left extreme of the interval $[a,b]$ changes (if it were the right extreme, we would eventually obtain analogous conclusions). We can then rewrite the interval as $[\max\{a,\frac{1-2b}{3}\},\min\{\frac{1-2a}{3},b\}]=[\frac{1-2b}{3},\min\{\frac{1-2a}{3},b\}]$. If we now do one more step, to get to $\mathcal{M}(\underline{E}_2)$, we see that the left extreme cannot change and hence the new interval will be $[\frac{1-2b}{3},\frac{1+4b}{9}]$. Hence, in two steps we go from $[a,b]$ to $[\frac{1-2b}{3},\frac{1+4b}{9}]$, and the length of the latter interval is $\frac{10b-2}{9}$. Now, since $a\leq\frac{1-2b}{3}$, we deduce that $3a+2b\leq 1$, and as a consequence $\frac{3}{2}\cdot\frac{10b-2}{9}=\frac{5b-1}{3}\leq b-a$. This means that the length of $[\frac{1-2b}{3},\frac{1+4b}{9}]$ is at most two thirds of the length of $[a,b]$.

By iterating the argument, we conclude that every two steps the length of the intervals decreases at least exponentially fast by $\frac{2}{3}$. As a consequence, given that 0.2 is always included in the intervals, the sequence $(\underline{E}_n)_n$ will converge towards $Q_{0.2}$, which, being conglomerable, is the conglomerable natural extension of $\underline{P}$. ♦

6 Conglomerability and Marginal Extension

The previous example shows that the sequence $(\underline{E}_n)_n$ may not stabilise in a finite number of steps. When $\underline{Q}$ does not coincide with $\underline{E}_n$ for any n, it is an open problem whether $\underline{Q}$ always coincides with the conglomerable natural extension or not. Here, we shall give a number of sufficient conditions for the equality $\underline{Q}=\underline{F}$. We shall show that one particular case of interest is that where $\underline{Q}$ is a marginal extension model and we are going to explore in more detail the connection between conglomerably coherent lower previsions and marginal extensions. We begin by proving an elementary and yet interesting result:

Proposition 6. *Let $\underline{P}$ be a coherent lower prevision on $\mathcal{L}$, $\mathcal{B}$ a partition of Ω and $\underline{P}(\cdot|\mathcal{B})$ the conditional natural extension of $\underline{P}$. Define $\underline{M}:=\underline{P}(\underline{P}(\cdot|\mathcal{B}))$. Then $\underline{M}\leq\underline{P}\Leftrightarrow\underline{P}$ conglomerable.*

It is possible to find examples that show that not every conglomerably coherent lower prevision is a marginal extension, or, in other words, that we do not necessarily have the equality $\underline{P}=\underline{M}$.

Next, we investigate the properties of the sequence of marginal extensions $(\underline{M}_n)_n$ associated to $(\underline{E}_n)_n$, where $\underline{M}_n:=\underline{E}_{n-1}(\underline{E}_{n-1}(\cdot|\mathcal{B}))$ for every $n>1$ and $\underline{M}_1:=\underline{P}(\underline{P}(\cdot|\mathcal{B}))$. It follows from Proposition 4 that $\mathcal{M}(\underline{E}_n)=\mathcal{M}(\underline{E}_{n-1})\cap\mathcal{M}(\underline{M}_n)$, so $\underline{M}_n\leq\underline{E}_n$ for all n. Since the sequence $(\underline{E}_n(\cdot|\mathcal{B}))_n$ is also increasing, we deduce that so is the sequence $(\underline{M}_n)_n$. Thus, $(\underline{M}_n)_n$ is an increasing sequence of conglomerable and coherent lower previsions that is dominated by $\underline{F}$, the conglomerable natural extension of $\underline{P}$. Moreover, if $\underline{E}_n$ is not conglomerable, then it cannot be $\underline{M}_n\geq\underline{P}$, because then it would be $\underline{M}_n=\underline{F}$, and therefore also $\underline{E}_n=\underline{F}$ would be conglomerable.

However, it may be that the conglomerable natural extension is not a marginal extension model, and therefore that the increasing sequence of marginal extensions stabilises on a model that is not the conglomerable natural extension, as the following example shows.

Example 5. Consider $\Omega := \mathbb{N} \cup -\mathbb{N}$, $B_n := \{n, -n\}$ and $\mathcal{B} := \{B_n : n \in \mathbb{N}\}$. Let P be a finitely additive probability on $\mathcal{P}(\mathbb{N})$ s.t. $P(\{n\}) = 0$ for all n, and P_1 a σ-additive probability on $\mathcal{P}(\Omega)$ s.t. $P_1(\{n\}) = P_1(\{-n\}) = \frac{1}{2^{n+1}}$ for all n. Consider also the linear previsions

$$\begin{aligned} P_2(h) &:= \frac{1}{2}\sum_n h(n)\frac{1}{2^n} + \frac{1}{2}P(h^-) \\ P_3(h) &:= \frac{3}{4}P(h^+) + \frac{1}{4}P(h^-) \\ P_4(h) &:= \frac{1}{2}P_1(h) + \frac{1}{2}P_3(h), \end{aligned}$$

where $h \in \mathcal{L}$ and h^+, h^- are derived by Eq. (5). Finally, let $\underline{P} := \min\{P_1, P_2, P_4\}$. Given $f := \mathbb{I}_{-\mathbb{N}}$, it holds that: $\underline{P}(f) = \min\left\{\frac{1}{2}, \frac{1}{2}, \frac{3}{8}\right\} = \frac{3}{8}$. In [5, Example 5] it is showed that the unconditional natural extension of $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ is given by

$$\underline{E}_1 = \min\left\{P_1, P_4, \frac{1}{3}P_2 + \frac{2}{3}P_4\right\},$$

that the conditional natural extension of $\underline{E}$ is given by

$$\underline{E}_1(h|B_n) = \min\left\{\frac{h(n) + h(-n)}{2}, \frac{2h(n) + h(-n)}{3}\right\},$$

and that $P_4(G_{\underline{E}}(h|\mathcal{B})) < 0$ for some h, so $\underline{E}_1$ is not conglomerable.

On the other hand, it can be showed that both $P_1(G_{\underline{E}_1}(\cdot|\mathcal{B})) \geq 0$ and $P_5(G_{\underline{E}_1}(\cdot|\mathcal{B})) \geq 0$. It follows from Proposition 4 that the unconditional natural extension $\underline{E}_2$ of $\underline{E}_1, \underline{E}_1(\cdot|\mathcal{B})$ is dominated by the lower envelope of $\{P_1, P_5\}$, from which we obtain that $\underline{E}_2(\cdot|B_n) \leq \min\{P_1(\cdot|B_n), P_5(\cdot|B_n)\}$ and in particular that $\underline{E}_2(h|B_n)$ is dominated by

$$\min\left\{\frac{h(n) + h(-n)}{2}, \frac{2h(n) + h(-n)}{3}\right\} = \underline{E}_1(h|B_n)$$

for every $h \in \mathcal{L}$ and every $n \in \mathbb{N}$, which implies that $\underline{E}_2(h|B_n) = \underline{E}_1(h|B_n)$ for every gamble h. Applying [5, Proposition 16], we deduce that $\underline{E}_2$ is conglomerable and therefore it is the conglomerable natural extension of $\underline{P}$.

Now, if we reconsider $f := \mathbb{I}_{-\mathbb{N}}$, then $\underline{E}_2(f|B_n) = \frac{1}{3}$ for all n, so if $\underline{E}_2$ was a marginal extension model, we would have $\underline{E}_2(f) = \underline{E}_2(\underline{E}_2(f|\mathcal{B})) = \underline{E}_2(\frac{1}{3}) = \frac{1}{3}$. But we know that $\underline{E}_2(f) \geq \underline{P}(f) = \frac{3}{8} > \frac{1}{3}$. This shows that the sequence of marginal extensions may not stabilise on the conglomerable natural extension. ♦

Let us study in more detail the sequence $(\underline{M}_n)_n$ of marginal extensions. We begin by characterising their relationship with $\underline{Q}$ in terms of credal sets.

Proposition 7. *Let* $\underline{Q} := \lim_n \underline{E}_n$ *and let* $\underline{Q}' := \lim_n \underline{Q}(\underline{E}_n(\cdot|\mathcal{B}))$. *Then* $\mathcal{M}(\underline{Q}') = \cap_n \mathcal{M}(\underline{M}_n)$, *whence:*

1. $\mathcal{M}(\underline{Q}) = \mathcal{M}(\underline{P}) \cap (\cap_n \mathcal{M}(\underline{M}_n)) = \mathcal{M}(\underline{P}) \cap \mathcal{M}(\underline{Q}')$.
2. $\underline{Q}'$ *conglomerable* $\Leftrightarrow \underline{Q}' = \underline{Q}(\underline{Q}(\cdot|\mathcal{B}))$.

Thus, the limit of the increasing sequence $(\underline{M}_n)_n$ is the coherent lower prevision $\underline{Q}' = \lim_n \underline{Q}(\underline{E}_n(\cdot|\mathcal{B}))$. Taking this into account, we can establish a sufficient condition for the conglomerable natural extension to be the limit of the sequence of marginal extensions:

Proposition 8. *Let* $\underline{Q}, \underline{Q}'$ *be given as in Proposition 7, and consider the following possibilities:*

(a) $\underline{Q}(\cdot|\mathcal{B})$ *is the uniform limit of* $(\underline{E}_n(\cdot|\mathcal{B}))_n$.

(b) $\underline{Q} = \underline{Q}' = \underline{F}$.

(c) $\underline{Q}'$ *is conglomerable.*

(d) $\underline{Q}$ *is conglomerable.*

(e) $\underline{Q} = \underline{F}$.

Then (a) $\Rightarrow$ (c) $\Rightarrow$ (d) $\Leftrightarrow$ (e) *and* (b) $\Rightarrow$ (c). *If in particular* $\underline{Q}' \geq \underline{P}$, *then:*

1. (b) $\Leftrightarrow$ (c) $\Leftrightarrow \underline{Q}' = \underline{Q}(\underline{Q}(\cdot|\mathcal{B}))$.
2. (d) $\Leftrightarrow$ (e) $\Leftrightarrow \underline{Q} = \underline{Q}(\underline{Q}(\cdot|\mathcal{B}))$.
3. (a) $\Rightarrow$ (b) $\Leftrightarrow$ (c) $\Rightarrow$ (d) $\Leftrightarrow$ (e).

7 The Finitary Case: Sufficient Conditions

As we have showed in Example 4, the sequence $(\underline{E}_n)_n$ of coherent lower previsions that provides a lower bound on the conglomerable natural extension may not stabilise in a finite number of steps. On the other hand, in Proposition 8 we have showed that a sufficient condition for $(\underline{E}_n)_n$ to converge towards the conglomerable natural extension is the uniform convergence of the sequence of conditional lower previsions. In this section, we give two sufficient conditions for this uniform convergence.

We focus on the case of an initial lower prevision $\underline{P}$ characterised by an associated credal set $\mathcal{M}(\underline{P})$ that contains *finitely many* extreme points. We call this a *finitary* model, or a finitary lower prevision.

In other words, we consider finitely many linear previsions $P_1, \ldots, P_k$ on $\mathcal{L}$ and let $\underline{P} := \min\{P_1, \ldots, P_k\}$. Then $\mathcal{M}(\underline{P}) = \{P_{\bar{\alpha}} : \bar{\alpha} \in \Delta\}$, where $\Delta := \{(\alpha_1, \ldots, \alpha_k) : \alpha_i \geq 0\ \forall i, \sum_{i=1}^k \alpha_i = 1\}$ is the $(k-1)$-dimensional simplex, and simplifying the notation by letting $P_{\bar{\alpha}} := \alpha_1 P_1 + \cdots + \alpha_k P_k$, with $\bar{\alpha} := (\alpha_1, \ldots, \alpha_k)$. We consider as usual a partition $\mathcal{B}$ of Ω and the sequence $(\underline{E}_n)_n$ of coherent lower previsions that we use to approximate the conglomerable natural extension $\underline{F}$ of $\underline{P}$ (provided that it exists), and $\underline{Q} = \lim_n \underline{E}_n$. We aim at giving sufficient conditions for $\underline{Q}$ to coincide with $\underline{F}$.

If there is $m \in \mathbb{N}$ such that $\underline{E}_m = \underline{E}_{m-1}$, then $\lim_n \underline{E}_n = \underline{E}_m = \underline{F}$ and in particular $\underline{Q} = \underline{F}$. Otherwise, if the sequence never stabilises, then $\underline{E}_n \lneq \underline{E}_{n+1}$ for all n, whence $\mathcal{M}(\underline{E}_n) \supsetneq \mathcal{M}(\underline{E}_{n+1})$. For each natural number n, we have that $\mathcal{M}(\underline{E}_n) = \{P_{\bar{\alpha}} : \bar{\alpha} \in \Delta_n\}$, where Δ_n is a closed and convex subset of Δ.

Hence $(\Delta_n)_n$ is a strictly decreasing sequence of closed and convex subsets of Δ; since Δ is a compact subset of $\mathbb{R}^k$, we deduce that $\lim_n \Delta_n =: \Delta'$ is a compact subset of Δ, that determines moreover $\underline{Q} = \lim_n \underline{E}_n$.

Next, we are going to use these sets to give a sufficient condition for the uniform convergence of the sequence of conditional natural extensions. One important issue here is that of the positivity of the lower probabilities of the conditioning events: as we have showed in (2), $\underline{Q}(f|B)$ can only be non-vacuous when $\overline{Q}(B) > 0$, and similarly for $\underline{E}_n$. Then it may be that $\overline{Q}(B) > 0$ for all B in $\mathcal{B}$ while for every n there is an infinity of B for which $\underline{E}_n(B) = 0$, thus preventing the uniform convergence. Our next result shows that for finitary models this is not an issue:

Lemma 9. *If $\underline{P} = \min\{P_1, \dots, P_k\}$, then there is some natural number n such that, for every $B \in \mathcal{B}$, $\underline{Q}(B) > 0 \Rightarrow \underline{E}_n(B) > 0$.*

Since the conglomerable natural extension of $\underline{P}$ coincides with that of $\underline{E}_n$ for every $n \in \mathbb{N}$, we are going to assume that $\underline{P}(B) > 0$ whenever $\underline{Q}(B) > 0$; otherwise, it suffices to start the sequence at the n for which the condition in Lemma 9 holds.

Let us give now two sufficient conditions for the uniform convergence of the sequence $(\underline{E}_n(\cdot|\mathcal{B}))_n$.

Theorem 10. *Under any of the following conditions:*

1. $\exists N > 0$ *s.t.* $\frac{\overline{P}(B)}{\underline{P}(B)} < N \; \forall B \in \mathcal{B}$,

2. $\exists \nu > 0$ *s.t.* $\min_{i=1}^k \alpha_i \geq \nu > 0 \; \forall \bar{\alpha} \in \Delta'$,

$\underline{Q}(f|\mathcal{B})$ is the uniform limit of $(\underline{E}_n(f|\mathcal{B}))_n \; \forall f \in \mathcal{L}$ and therefore $\underline{Q}$ is the conglomerable natural extension of $\underline{P}$.

It can be checked that neither of these sufficient conditions is necessary for the limit to be conglomerable.

Remark 1. The second of these sufficient conditions is particularly revealing in the binary case, where we consider the lower envelope of two linear previsions, $\underline{P} := \min\{P_1, P_2\}$. If we denote $P_\alpha := \alpha P_1 + (1-\alpha)P_2$, then we can identify each Δ_n with a subset of $[0,1]$:

$$\begin{aligned} \mathcal{M}(\underline{P}) &: = \{P_\alpha : \alpha \in [0,1]\}, \\ \mathcal{M}(\underline{E}_n) &: = \{P_\alpha : \alpha \in [a_n, b_n]\} \text{ and} \\ \mathcal{M}(\underline{Q}) &: = \{P_\alpha : \alpha \in [a,b]\}, \end{aligned}$$

where $0 \leq a_n \leq b_n \leq 1$ for all n, and $(a_n)_n \uparrow a$, $(b_n)_n \downarrow b$. There are a number of possibilities:

- If $a = b = 1$, then $\underline{Q} = P_1$, so the conglomerable natural extension exists if and only if it coincides with $\underline{Q} = P_1$.

- If $a = b = 0$, then $\underline{Q} = P_2$, so the conglomerable natural extension exists if and only if it coincides with $\underline{Q} = P_2$.

- If $a, b \in (0,1)$, then Theorem 10 implies that $(\underline{E}_n(f|\mathcal{B}))_n$ converges uniformly to $\underline{Q}(f|\mathcal{B})$, and as a consequence $\underline{Q}$ is the conglomerable natural extension of $\underline{P}$.

- If $a = 0$ and $b \in (0,1)$, then we can deduce from Theorem 10 that $(P_{b_n}(f|\mathcal{B}))_n$ converges uniformly to $P_b(f|\mathcal{B})$ for every gamble f, and from this we deduce that $\underline{Q}$ is the conglomerable natural extension of $\underline{P}$. A similar result applies when $a \in (0,1)$ and $b = 1$.

This means that if we consider a binary model $\underline{P} = \min\{P_1, P_2\}$ and that the conglomerable natural extension of $\underline{P}$ exists, then it necessarily coincides with $\underline{Q}$. ♦

8 Conclusions

The importance of the conglomerable natural extension can be appreciated when one realises that it is the analog, for a theory of probability based on conglomerability, of the deductive closure in logic. Unfortunately, this paper shows that such a closure is not finitary, in the sense that to compute the conglomerable natural extension $\underline{F}$ of a coherent lower prevision $\underline{P}$, one might have to create an infinite sequence $(\underline{E}_n)_n$ of distinct approximating coherent lower previsions.

Moreover, at the moment it is still an open problem whether the point-wise limit $\underline{Q}$ of such a sequence actually attains $\underline{F}$ in general. However, in the special case where $\underline{P}$ is the envelope of finitely many linear previsions, this paper gives sufficient conditions for $\underline{Q} = \underline{F}$ that seem to have quite broad applicability. This gives reasons to believe that $\underline{Q}$ will equal $\underline{F}$ in many cases of practical interest.

Yet, solving the mentioned problem in general seems to us the most important question, and a very difficult one too, that should be addressed by future research.

Acknowledgements

This work was partly supported by project MTM2010-17844, by the Swiss NSF grants nos. 200020_134759 / 1, 200020_137680 / 1, and by the Hasler foundation grant n. 10030.

References

[1] B. de Finetti. Sulla proprietà conglomerativa delle probabilità subordinate. *Rendiconti del Reale Instituto Lombardo*, 63:414–418, 1930.

[2] B. de Finetti. *Teoria delle Probabilità*. Einaudi, Turin, 1970.

[3] B. de Finetti. *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons, Chichester, 1974–1975. English translation of [2], two volumes.

[4] E. Miranda and M. Zaffalon. Conglomerable coherence. *International Journal of Approximate Reasoning*, 2013. Accepted for publication.

[5] E. Miranda, M. Zaffalon, and G. de Cooman. Conglomerable natural extension. *International Journal of Approximate Reasoning*, 53(8):1200–1227, 2012.

[6] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[7] P. M. Williams. Notes on conditional previsions. Technical report, School of Mathematical and Physical Science, University of Sussex, UK, 1975. Reprinted in [8].

[8] P. M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44:366–383, 2007. Revised journal version of [7].

[9] M. Zaffalon and E. Miranda. Probability and time. *Artificial Intelligence*, 198:1–51, 2013.

Modeling Uncertainty in First-Order Logic: A Dempster-Shafer Theoretic Approach

Rafael C. Núñez
Department of Electrical and Computer Engineering
University of Miami
nunez@umiami.edu

Matthias Scheutz
Department of Computer Science
Tufts University
mscheutz@cs.tufts.edu

Kamal Premaratne
Department of Electrical and Computer Engineering
University of Miami
kamal@miami.edu

Manohar N. Murthi
Department of Electrical and Computer Engineering
University of Miami
mmurthi@miami.edu

Abstract

First order logic lies at the core of many methods in mathematics, philosophy, linguistics, and computer science. Although important efforts have been made to extend first order logic to the task of handling uncertainty, there is still a lack of a consistent and unified approach, especially within the Dempster-Shafer (DS) theory framework. In this work we introduce a systematic approach for building belief assignments based on first order logic formulas. Furthermore, we outline the foundations of *Uncertain Logic*, a robust framework for inference and modeling when information is available in the form of first order logic formulas subject to uncertainty. Applications include data fusion, rule mining, credibility estimation, and crowd sourcing, among many others.

Keywords. Uncertain Logic, Uncertain Reasoning, Probabilistic Logic, Dempster-Shafer Theory, Belief Theory.

1 Introduction

Natural language processing, artificial intelligence, and graph analysis are among a number of applications that heavily rely on first order logic formulations. Due to its capability for representing knowledge for inference systems, first order logic has been gradually enriched to handle imperfections in real-life data. Some approaches include fuzzy logic and probabilistic logic [1]. These solutions, however, are not well suited for handling scenarios characterized by ranges of uncertainty, or that require modeling evidence in a very strict manner to minimize the risk of inference results leading to wrong conclusions.

Dempster-Shafer theory [2] provides an ideal modeling tool to address this problem. However, although significant effort has been dedicated to modeling uncertainty in logic under DS theory, there is still a need for a unified approach that is consistent with basic logic operations and that provides the support for handling variables and quantifiers. To address this problem we introduce *Uncertain Logic*, which is the extension of first order logic into DS theory. Consider, for example, an expression of the form: $\exists x :\ \varphi(x),$ with uncertainty $[\alpha, \beta]$, where $\varphi(x)$ is a logic predicate that depends on the variable x. The uncertain logic framework allows us to model this sentence, and to combine it with similar ones in order to solve various inference problems. When $\alpha = \beta$, uncertain logic renders probabilistic results. When $\alpha = \beta \in \{0, 1\}$, uncertain logic converges to first order logic. Unlike existing DS models for logic that, in general, cannot guarantee logic consistency for a plurality of logic constructs, uncertain logic preserves this consistency, and can grow to incorporate logic rules and properties without loss of uncertainty measures. By preserving this consistency, it is possible to seamlessly move between the logic and DS domains, and to incorporate both the strength of first order logic for information representation, inference, and resolution, and the strength of DS for representing and manipulating uncertainty in the data.

1.1 Existing Methods for Handling Uncertainty in Logic

The need for reasoning under the presence of uncertainty has lead to important work aimed at providing logic reasoning with uncertainty management capabilities. Research in this area encompasses a number of aims, such as the investigation of the source and meaning of uncertainty, the enrichment of logic systems with appropriated formalisms for uncertainty management (e.g., semantics, axioms), and the creation of appropriate models and operators to quantify the propagation of uncertainty in reasoning and inference problems.

Relevant foundational work, with emphasis on analyzing the source and representation of uncertainty in logic systems, can be found in [3]. In this work, the author introduces two different approaches to giving semantics to first-order logics of probability, the first one incorporating probability in the domain (for problems involving statistical information), and the second one assigning probabilities to possible worlds. This work is extended in [4], where the author further discusses the use of a “possible-worlds” framework to represent and reason about uncertainty. Then, quantification of the uncertainty is accomplished by assigning a probability distribution to the possible worlds. In addition, the author discusses the importance of considering time in the inference process, i.e., possible words should describe states at each time point of interest. The work in [5] provides insight on how to process and combine data-driven (e.g., information obtained from observed events) and knowledge-driven (e.g., information provided by domain experts) using different logic systems.

In addition to first-order logic, uncertain representations of logic systems have been extended to other types of logic. For example, the work in [6] introduces a multi-agent epistemic logic able to represent and merge partial beliefs of multiple agents. This logic system is based on possibility theory [7], and enhances epistemic logic with parametric models to obtain lower bounds on the degree of belief of agents. Similarly, an axiomatization of a modal logic using fuzzy sets and DS belief functions for measuring probabilities of modal necessity is presented in [8].

When addressing quantification and propagation of uncertainty in logic reasoning systems, one of the most important approaches is probabilistic logic [9]. Probabilistic logic provides a generalization of logic in which the truth values of sentences are probability values (between 0 and 1). A related approach, possibilistic logic [10], defines mechanisms (based on possibility theory) to associate classical logic formulas with weights. These weights represent lower bounds of necessity degrees. Other approaches that extend logic reasoning to address uncertain scenarios are many-valued and fuzzy logics. Many-valued logics do not restrict the number of truth values of propositions to two. The interpretation of the truth values depends on the actual application. Fuzzy logic can be seen as a type of many-valued logic. Fuzzy logic is based on the theory of fuzzy sets [11]. In fuzzy logic, the imprecision in probabilities is modeled through membership functions defined on the sets of possible probabilities and utilities.

Although useful in some applications, these approaches are sometimes limited by the way they model uncertainty, or simply by the complexity of the problem formulation. Extensions of these approaches could be strengthened by adding more flexibility in assigning probabilities (e.g., through intervals) and a more rigorous method of assigning probability measures (e.g., one that does not require defining priors or membership functions).

Regarding the use of intervals as means of representing uncertainty, it appears in several methods, such as possibility theory [12] and DS theory. The latter, in addition, incorporates a rigorous methodology for assigning probabilistic measures based on available evidence [13]. Given the direct relation that exists between DS theory and probability (DS belief and plausibility measures correspond precisely to probabilistic inner and outer measures [13]), it is possible to simplify DS models to probabilistic models. Considering these advantages, a number of researchers have studied the relation of DS theory and logic. In [14], DS theory is formulated in terms of propositional logic, enabling certain logic reasoning operations in the DS framework. Insight into the relationship between DS theory and probabilistic logic is presented in [14]. A belief-function logic that uses DS models and operations to quantify and estimate uncertainty of logic formulas is introduced in [15]. This logic system allows non-zero belief assignments to the empty set, relies on Dempster's combination rule as the method for quantifying the propagation of uncertainty, and is used in deduction systems where the logic formulas are in Skolemized normal conjunctive form. An application of this system for inference is described in [16]. Further analysis on DS-based logic is presented in [17]. A detailed study on uncertain implication rules is in [18]. This latter work, however, is not focused on ensuring consistency with classical logic, but on modeling causal probabilistic relations.

In spite of existing research to provide logic with uncertainty modeled by DS, efforts to date can be improved by ensuring consistency with classical logic and reducing the number of assumptions needed for the logic systems to work. For example, most of the existing methods are based on Dempster's Combination Rule, which, as it is shown in this manuscript, is not necessarily well suited for logical reasoning. In addition, inference processes could benefit from eliminating the condition that logic formulas need to be expressed in normal conjunctive form or as implication rules, as well as eliminating the need for allowing non-zero belief assignments to the empty set in a DS model.

1.2 Our Contribution: Uncertain Logic

To address these issues, and with emphasis on methods to quantify uncertainty propagation, we introduce uncertain logic. Uncertain logic deals with logic propositions whose truth is uncertain. The level of uncertainty is modeled with DS theory. Uncertain logic allows reasoning and inference using (conventional) first order logic inference rules, but also allows for appending uncertainty to the inference process.

To describe the uncertain logic framework, we start in Section 2 with an overview of DS theory. Basic definitions and notation of uncertain logic are then introduced in Section 3. A set of uncertain logic operators and quantifiers are described in Sections 4 and 5, respectively. Finally, inference in uncertain logic is introduced in Section 6.

2 DS Theory: Basic Definitions

DS Theory is defined for a discrete set of elementary events related to a given problem. This set is called the *Frame of Discernment* (FoD). In general, a FoD is defined as $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$, and has a finite cardinality

$N = |\Theta|$. Elements (or singletons) $\theta_i \in \Theta$ represent the lowest level of discernible information. The power set of Θ is defined as a set containing all the possible subsets of Θ, i.e., $2^\Theta = \{A : A \subseteq \Theta\}$. The cardinality of the power set of Θ is 2^N. Next we introduce some basic definitions of DS Theory, as required for building uncertain logic models. For additional details on DS Theory, we refer the reader to [1, 2].

2.1 Basic Belief Assignment

A Basic Belief Assignment (BBA) or *mass assignment* is a mapping $m_\Theta(\cdot) : 2^\Theta \to [0,1]$ such that: $\sum_{A \subseteq \Theta} m_\Theta(A) = 1$ and $m_\Theta(\emptyset) = 0$. The BBA measures the support assigned to proposition $A \subseteq \Theta$. Masses in DS theory can be assigned to any singleton or non-singleton (e.g., $\{\theta_1, \theta_2\}$, $\{\theta_1, \theta_3\}$, $\{\theta_1, \theta_2, \theta_3\}$) proposition. A belief function is called Bayesian if each focal element in Θ is a singleton. The subsets A such that $m(A) > 0$ are referred to as focal elements of the BBA. The set of focal elements is the core $\mathcal{F}_\Theta$. The triple $\{\Theta, \mathcal{F}_\Theta, m_\Theta(\cdot)\}$ is referred to as *Body of Evidence* (BoE).

2.2 Belief and Plausibility

Given a BoE $\{\Theta, \mathcal{F}, m\}$, the *belief* function $\mathrm{Bel} : 2^\Theta \to [0,1]$ is defined as: $\mathrm{Bel}_\Theta(A) = \sum_{B \subseteq A} m_\Theta(B)$. $\mathrm{Bel}(A)$ represents the total belief that is committed to A without also being committed to its complement A^C. The *plausibility* function $\mathrm{Pl} : 2^\Theta \to [0,1]$ is defined as: $\mathrm{Pl}_\Theta(A) = 1 - \mathrm{Bel}_\Theta(A^C)$. It corresponds to the total belief that does not contradict A. The *uncertainty* of A is: $[\mathrm{Bel}_\Theta(A), \mathrm{Pl}_\Theta(A)]$.

2.3 Combination Rules

Dempster Combination Rule (DCR). For two focal sets $C \subseteq \Theta$ and $D \subseteq \Theta$ such that $B = C \cap D$, and two BBAs $m_j(\cdot)$ and $m_k(\cdot)$, the combined $m_{jk}(B)$ is given by: $m_{jk}(B) = \frac{1}{1-K_{jk}} \sum_{C \cap D = B; B \neq \emptyset} m_j(C)\, m_k(D)$, where $K_{jk} = \sum_{C \cap D = \emptyset} m_j(C)\, m_k(D) \neq 1$ is referred to as the *conflict* between the two BBAs; $K_{jk} = 1$ identifies two totally conflicting BBAs for which DCR-based fusion cannot be carried out.

Conditional Fusion Equation (CFE). A combination rule that is robust when confronted with conflicting evidence is the *Conditional Fusion Equation (CFE)* [19], which is based on the DS theoretic conditional approach [20]. The CFE combines M BBAs as [19]: $m(B) = \sum_{i=1}^{M} \sum_{A_i \in \mathcal{A}_i} \gamma_i(A_i)\, m_i(B|A_i)$, where $\sum_{i=1}^{M} \sum_{A_i \in \mathcal{A}_i} \gamma_i(A_i) = 1$. Here $\mathcal{A}_i = \{A \in \mathcal{F}_i : \mathrm{Bel}_i(A) > 0\}$, $i = 1, \ldots, M$. The conditionals are computed using Fagin-Halperns' Rule of Conditioning [21].

3 From Propositional Logic to Uncertain First-Order Logic

Propositional Logic. Recall that a *proposition* is simply a statement such as "this is an introduction to uncertain logic". We will represent formulas in propositional logic by lower case greek letters (e.g., φ, ψ). In propositional logic, a proposition can be obtained from other propositions using connectives like $\wedge$ (and), $\vee$ (or), $\neg$ (not), and $\Longrightarrow$ (implies). Through (classical) *inference*, propositions can be derived from a given a set of propositions (called *premises*) using (classical) "rules of inference" such as "modus ponens".

Predicate Logic. Predicate logic allows us to look into the structure of propositions. For example, the fact that some entity a is *above* another entity b would be expressed as `Above(a, b)`, where "Above" is a two-place predicate symbol and "a" and "b" are individual constants. For the remainder of this paper, we will assume finite domains for the interpretation of predicate logic formulas (i.e., individual variables ranges over a finite number of entities).[1]

First-Order Logic. First Order Logic extends predicate logic by the *universal quantifier* ($\forall$) and the *existential quantifier* ($\exists$). Quantified formulas provide a more flexible way of talking about all objects in the domain (i.e., elements in our universe of discourse) or of asserting a property of an individual object.

Uncertain Logic. Uncertain logic deals with propositions $(\varphi_1, \varphi_2, \ldots)$ whose truth is uncertain. The level of uncertainty is modeled with DS theory and is bounded in the range $[0,1]$. In general, we will consider formulas with k free variables that range over individuals from some finite domain $\Theta_X = \{x_1, \ldots, x_n\}$, with $n \geq 1$, for example

$$\varphi(x), \text{ with uncertainty } [\alpha, \beta], \tag{1}$$

where $\varphi(x)$ is a formula with the only free variable x ranging over elements in Θ_X and $[\alpha, \beta]$ it the corresponding uncertainty interval with $0 \leq \alpha \leq \beta \leq 1$. [2]

To emphasize the fact that uncertain logic models uncertainty of the true value of a proposition, we define the *logical FoD* as follows.

Definition 1 (Logical FoD) *Given a logic proposition $\varphi(x)$ with x ranging over entities in Θ_X, and a true-false FoD $\Theta_{t\text{-}f} = \{\mathbf{1}, \mathbf{0}\}$, the logical FoD $\Theta_{\varphi(x) \times \{\mathbf{1},\mathbf{0}\}}$ is given by:*

$$\Theta_{\varphi(x) \times \{\mathbf{1},\mathbf{0}\}} = \{\varphi(x) \times \mathbf{1}, \varphi(x) \times \mathbf{0}\}. \tag{2}$$

[1] When referring to propositional and predicate logic, we follow the conventions and definitions provided in [22] and [23].

[2] We can define this first-order logic expression more formally as follows: Consider a quantifier-free first-order formula $\varphi(x)$ from a (not necessarily finite) set of formulas Φ in some first-order language L with x being the only free variable in φ. Moreover, let $\Theta_X = \{x_1, \ldots, x_n\}$ be a non-empty set of individuals under observation with respect to formulas in Φ. Throughout this paper, we may represent the logic formula $\varphi(x/x_i)$, the property expressed by φ for the individual $x_i, i = 1, \ldots, n$, with the abbreviated notation $\varphi(x_i)$, i.e., $\varphi(x_i) \equiv \varphi(x/x_i)$. In addition, the DS models that we define for a quantifier-free first-order formula $\varphi(x)$ extend to the sets of formulas $\varphi(x_i), i = 1, \ldots, n$, defined on the corresponding logical FoDs. This extension is used in Section 5, where we define models for existential and universal quantifiers.

When no confusion can arise, we will employ the following notation:

$$\varphi(x) \equiv \varphi(x)\times\mathbf{1}; \varphi(\overline{x}) \equiv \varphi(x)\times\mathbf{0}; \Theta_{\varphi(x)} \equiv \Theta_{\varphi(x)\times\{\mathbf{1},\mathbf{0}\}}.$$

A DS theoretic model that would capture the information in (1) is:

$$\begin{aligned} \varphi(x): \;\; m(\varphi(x)) &= \alpha; \\ m(\varphi(\overline{x})) &= 1-\beta; \\ m(\Theta_{\varphi(x)}) &= \beta-\alpha, \end{aligned} \tag{3}$$

defined over the logical FoD $\{\varphi(x), \varphi(\overline{x})\}$. In order to simplify the arguments in the mass assignments, we may use the following alternate notation:

$$\varphi(x): \; m_\varphi(x) = \alpha; m_\varphi(\overline{x}) = 1-\beta; m_\varphi(\Theta_{\varphi(x)}) = \beta-\alpha. \tag{4}$$

Semantics. In classical logic there are two truth values, “true” and “false”. An expression that is true for all interpretations is called a tautology (“$\top$”). An expression that is not true for any interpretation is a contradiction (“$\perp$”). Two expressions are semantically equivalent if they take on the same truth value for all interpretations.

In uncertain logic we extend these definitions. The truth value of an expression corresponds to the support that is projected into the true-false FoD, $\Theta_{t\text{-}f} = \{\mathbf{1}, \mathbf{0}\}$. A BBA (3) defined by $[\alpha, \beta] = [1, 1]$ corresponds to the classical logical truth. A BBA (3) defined by $[\alpha, \beta] = [0, 0]$ corresponds to the classical logical falsehood.

The notions of tautology and contradiction in uncertain logic are extended following an approach similar to that in [24]. In particular, given a generic proposition ψ characterized by the uncertainty interval $\sigma = [\alpha, \beta]$, we define a σ-tautology as $\top_\sigma \equiv \psi \vee \neg\psi$, and a σ-contradiction as $\perp_\sigma \equiv \psi \wedge \neg\psi$. It follows that $\top \equiv \top_{\sigma=[1,1]}$, and $\perp \equiv \perp_{\sigma=[0,0]}$.

4 Uncertain Logic Operators

The AND and OR operators are, together with the logical negation, the basic operators in classical logic. This is also the case in uncertain logic, as any other operator can be defined using combinations of these three basic operators. In order to ensure consistency with classical logic, uncertain logic operators should satisfy at least the following: (a) $(\varphi_1(x) \vee \varphi_2(x))$ and $\neg(\neg\varphi_1(x) \wedge \neg\varphi_2(x))$ must have identical DS theoretic models; (b) $(\varphi_1(x) \wedge \varphi_2(x))$ and $\neg(\neg\varphi_1(x) \vee \neg\varphi_2(x))$ have identical DS theoretic models; (c) in the general case, the DS model for AND and OR operations are distinct; (d) in the absence of uncertainty, uncertain logic models converge to those of conventional logic; (e) in a probabilistic scenario (i.e., $\alpha = \beta$), uncertain logic models are also probabilistic; (f) Uncertain logic AND and OR operators must be idempotent, commutative, associative, and distributive.

4.1 Uncertain Logic Negation

Consider a logical FoD $\Theta_{\varphi(x)} = \{\varphi(x), \varphi(\overline{x})\}$ and a BBA $m_\varphi(\cdot)$ defined as:

$$m_\varphi(x) = \alpha; \;\; m_\varphi(\overline{x}) = 1-\beta; \;\; m_\varphi(\Theta_{\varphi(x)}) = \beta-\alpha. \tag{5}$$

A complementary BBA for (5) is given by [25]:

$$m^c_\varphi(x) = 1-\beta; \; m^c_\varphi(\overline{x}) = \alpha; \; m^c_\varphi(\Theta_{\varphi(x)}) = \beta-\alpha. \tag{6}$$

Based on the complementary BBA, we can define an uncertain logic negation as follows.

Definition 2 (Logical Not in Uncertain Logic) *Given an uncertain proposition $\varphi(x)$ as defined in (1), and its corresponding DS model defined by (4), the logical negation of $\varphi(x)$ is given by:*

$$\neg\varphi(x), \; \textit{with uncertainty} \; [1-\beta, 1-\alpha]. \tag{7}$$

We utilize the complementary BBA corresponding to (4) as the DS theoretic model for $\neg\varphi(x)$, i.e.,

$$\begin{aligned} \neg\varphi(x): \;\; m^c_\varphi(x) &= 1-\beta; \\ m^c_\varphi(\overline{x}) &= \alpha; \\ m^c_\varphi(\Theta_{\varphi(x)}) &= \beta-\alpha. \end{aligned} \tag{8}$$

Definition 2 satisfies an important property: Given a proposition $\varphi(x)$, the BBA corresponding to its double-negation is the same model as the one associated with $\varphi(x)$. In other words, Definition 2 satisfies $\neg\neg\varphi(x) = \varphi(x)$, which is a basic property in (classical) logic.

4.2 Uncertain Logic AND/OR

Definition 3 (Logical And & Or in Uncertain Logic) *Suppose that we have M logic propositions, each providing a statement of the following type regarding the truth of x with respect to the proposition $\varphi_i(\cdot)$:*

$$\varphi_i(x), \; \textit{with uncertainty} \; [\alpha_i, \beta_i], \; i = 1, \ldots, M. \tag{9}$$

The corresponding DS theoretic models are $\varphi_i(x): m_{\varphi_i}(x) = \alpha_i$; $m_{\varphi_i}(\overline{x}) = 1-\beta_i$; $m_{\varphi_i}(\Theta_{\varphi_i(x)}) = \beta_i - \alpha_i$, for $i = 1, 2, \ldots, M$. We propose to utilize the following DS theoretic models for the logical AND and OR of the statements in (9)*:*

$$\begin{aligned} &\bigwedge_{i=1}^{M} \varphi_i(x): \;\; m(\cdot) = \bigcap_{i=1}^{M} m_{\varphi_i}(\cdot); \\ \textit{and} \;\; &\bigvee_{i=1}^{M} \varphi_i(x): \;\; m(\cdot) = \left(\bigcap_{i=1}^{M} m^c_{\varphi_i}(\cdot)\right)^c, \end{aligned} \tag{10}$$

where $\bigcap$ denotes an appropriate fusion operator.[3]

[3] A similar model can be obtained for the case of AND/OR operations of a set of expressions $\{\varphi(x_i)\}$ with uncertainty $[\alpha_i, \beta_i]$, $x_i \in \{x_1, x_2, \ldots, x_n\}$. In this case, $\bigwedge_{i=1}^n \varphi(x_i): \; m(\cdot) = \bigcap_{i=1}^n m_\varphi(\cdot)$, and $\bigvee_{i=1}^n \varphi(x_i): \; m(\cdot) = \left(\bigcap_{i=1}^n m^c_\varphi(\cdot)\right)^c$. This case represents AND/OR models applied to the truthfulness of elements $\{x_i\}$ satisfying a property φ, whereas (10) analyzes the case of x satisfying multiple properties $\{\varphi_i\}$.

Table 1: DCR-Based Logical AND and OR. Note that the DS models for AND and OR are identical, which suggests that DCR is not an appropriate fusion operator for consistent logic operations. Note that, in both cases, the masses should be normalized by $1-K$, with $K = 1 - \sum_{A\in\mathcal{F}} m(A) = \alpha_1(1-\beta_2) + (1-\beta_1)\alpha_2$.

Focal Set	$\varphi_1(x) \wedge \varphi_2(x)$	$\varphi_1(x) \vee \varphi_2(x)$
x	$\alpha_1\beta_2 + (\beta_1-\alpha_1)\alpha_2$	$\alpha_1\beta_2 + (\beta_1-\alpha_1)\alpha_2$
$\overline{x}$	$(1-\beta_1)(1-\alpha_2) + (\beta_1-\alpha_1)(1-\beta_2)$	$(1-\beta_1)(1-\alpha_2) + (\beta_1-\alpha_1)(1-\beta_2)$
$\Theta_{(\varphi_1\cdot\varphi_2)(x)}$	$(\beta_1-\alpha_1)(\beta_2-\alpha_2)$	$(\beta_1-\alpha_1)(\beta_2-\alpha_2)$

4.3 DCR-Based Uncertain Logic

When the fusion operator $\bigcap$ in (10) is DCR, the AND operation in this model is equivalent to the conjunctive rule of combination in [17]. In this subsection we go further and explore the viability of using DCR as the fusion operator in uncertain logic.

Consider the two-source/two-propositions (*i.e.*, $M = 2$) case. Table 1 contains the DCR-based logical AND and OR operations for this case. Notice that the mass assignments for the AND operation (*i.e.*, $\varphi_1(x) \wedge \varphi_2(x)$) are exactly the same as the ones obtained for the OR operation (*i.e.*, $\varphi_1(x) \vee \varphi_1(x)$). Having identical models for both AND and OR operators suggests that, although DCR may work as a fusion operator for certain operations, it does not render models that satisfy important properties for all the logical operations defined in this paper. More particularly, DCR-based uncertain logic does not satisfy the "uniqueness of the model" property. As an alternative, we propose using a more appropriate fusion strategy, such as the CFE, which is analyzed next.

4.4 CFE-Based Uncertain Logic

Recall (from Section 2) that CFE-based fusion requires the definition of coefficients $\gamma_i(\cdot)$. For uncertain logic, we introduce the Logic Consistent (LC) strategy, which ensures consistency with logical operations.

Definition 4 (Logic Consistent (LC) Strategy) *For the case $M = 2$ in (10), let us define $\underline{\alpha} = \min(\alpha_1,\alpha_2)$; $\underline{\beta} = \min(\beta_1,\beta_2)$; $\overline{\alpha} = \max(\alpha_1,\alpha_2)$; $\overline{\beta} = \max(\beta_1,\beta_2)$; $\delta_1 = \beta_1 - \alpha_1$; $\delta_2 = \beta_2 - \alpha_2$; $\underline{\delta} = \underline{\beta} - \underline{\alpha}$; and $\overline{\delta} = \overline{\beta} - \overline{\alpha}$. Then select the CFE parameters as follows:*

$$\gamma_1(x) = \gamma_2(x) \equiv \gamma(x); \quad \gamma_1(\overline{x}) = \gamma_2(\overline{x}) \equiv \gamma(\overline{x});$$
$$\gamma_1(\Theta) = \gamma_2(\Theta) \equiv \gamma(\Theta),$$

where the CFE parameters $\gamma(x)$, $\gamma(\overline{x})$, and $\gamma(\Theta)$ are selected in the following manner.

a. Logical AND:

- *If $\delta_1 + \delta_2 \neq 0$:*

$$\gamma(x) = \frac{\underline{\alpha}(\beta_1+\beta_2) - \underline{\beta}(\alpha_1+\alpha_2)}{2(\delta_1+\delta_2)};$$
$$\gamma(\overline{x}) = \frac{1}{2} - \frac{\underline{\beta}(2-\alpha_1-\alpha_2) - \underline{\alpha}(2-\beta_1-\beta_2)}{2(\delta_1+\delta_2)};$$
$$\gamma(\Theta) = \frac{\underline{\delta}}{\delta_1+\delta_2}.$$

- *If $\delta_1 + \delta_2 = 0$, i.e., $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$:*

$$\gamma(x) = \frac{\underline{\alpha} - \gamma(\Theta)\,(\alpha_1+\alpha_2)}{2};$$
$$\gamma(\overline{x}) = \frac{(1-\underline{\alpha}) - \delta(\Theta)\,(2-\alpha_1-\alpha_2)}{2};$$
$$\gamma(\Theta) = \textit{arbitrary}.$$

b. Logical OR:

- *If $\delta_1 + \delta_2 \neq 0$:*

$$\gamma(x) = \frac{1}{2} - \frac{\overline{\beta}(2-\alpha_1-\alpha_2) - \overline{\alpha}(2-\beta_1-\beta_2)}{2(\delta_1+\delta_2)};$$
$$\gamma(\overline{x}) = \frac{\overline{\alpha}(\beta_1+\beta_2) - \overline{\beta}(\alpha_1+\alpha_2)}{2(\delta_1+\delta_2)};$$
$$\gamma(\Theta) = \frac{\overline{\delta}}{\delta_1+\delta_2}.$$

- *If $\delta_1 + \delta_2 = 0$, i.e., $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$:*

$$\gamma(x) = \frac{\overline{\alpha} - \gamma(\Theta)\,(\alpha_1+\alpha_2)}{2};$$
$$\gamma(\overline{x}) = \frac{(1-\overline{\alpha}) - \delta(\Theta)\,(2-\alpha_1-\alpha_2)}{2};$$
$$\gamma(\Theta) = \textit{arbitrary}.$$

When used for the AND operation, the LC strategy renders the following BBA (see Appendix A for the derivation of this BBA):

$$\begin{aligned} \varphi_1(x) \wedge \varphi_2(x):\quad & m(x) && = \underline{\alpha}; \\ & m(\overline{x}) && = 1 - \underline{\beta}; \text{ and} \\ & m(\Theta_{(\varphi_1\wedge\varphi_2)(x)}) && = \underline{\beta} - \underline{\alpha}. \end{aligned} \tag{11}$$

When used for the OR operation, the LC strategy renders the following BBA:

$$\begin{aligned} \varphi_1(x) \vee \varphi_2(x):\quad & m(x) && = \overline{\alpha}; \\ & m(\overline{x}) && = 1 - \overline{\beta}; \text{ and} \\ & m(\Theta_{(\varphi_1\vee\varphi_2)(x)}) && = \overline{\beta} - \overline{\alpha}. \end{aligned} \tag{12}$$

In general, the CFE-based models for the logical AND and OR are not identical (the exception would be a particular combination of uncertainty parameters $[\alpha, \beta]$ rendering identical models), as is the case when DCR is used. Therefore, CFE-based fusion is better suited for uncertain logic than DCR. Indeed, referring to the conditions at the beginning of Section 4, the CFE-based operations are *consistent*

Table 2: CFE-Based AND/OR Operations: Uncertainty parameters are defined so that they represent complete certainty on the truth (or falseness) of each proposition.

Parameters		$m_{\varphi_1 \wedge \varphi_2}(\cdot)$			$m_{\varphi_1 \vee \varphi_2}(\cdot)$		
$[\alpha_1, \beta_1]$	$[\alpha_2, \beta_2]$	x	$\overline{x}$	Θ	x	$\overline{x}$	Θ
$[0,0]$	$[0,0]$	0	1	0	0	1	0
$[0,0]$	$[1,1]$	0	1	0	1	0	0
$[1,1]$	$[1,1]$	1	0	0	1	0	0

Table 3: CFE-Based Logical AND/OR Operations: Probabilistic Scenario ($[\alpha_i, \beta_i] = [\alpha_i, \alpha_i],\ i \in \{1,2\}$).

Logical AND	Logical OR
$m(x) = \underline{\alpha}$ $m(\overline{x}) = 1 - \underline{\alpha}$ $m(\Theta_{(\varphi_1 \wedge \varphi_2)(x)}) = 0$	$m(x) = \overline{\alpha}$ $m(\overline{x}) = 1 - \overline{\alpha}$ $m(\Theta_{(\varphi_1 \vee \varphi_2)(x)}) = 0$

with classical logic. Referring to the same conditions, (a) and (b) can be verified by checking Definition 3; (c) is verified by (11) and (12) above; (d) is proved in Table 2; (e) is shown in Table 3; (f) is proved in Appendix B.

4.5 Other Uncertain Logic Operators

Based on the uncertain logic definitions and operators described above, it is possible to extend them and create new operators. As an example, consider implication rules.

Definition 5 (Logical Implication in Uncertain Logic) *Given two logic statements $\varphi_1(\cdot)$ and $\varphi_2(\cdot)$, an implication rule in propositional logic has the property:*

$$\begin{aligned}\varphi_1(x_i) \implies \varphi_2(y_j) &= \neg\varphi_1(x_i) \vee \varphi_2(y_j) \\ &= \neg\left(\varphi_1(x_i) \wedge \neg\varphi_2(y_j)\right),\end{aligned}$$

where $x_i \in \Theta_X$ and $y_j \in \Theta_Y$. Consider the case where the antecedent $\varphi_1(x_i)$ and/or the consequent $\varphi_2(y_j)$ are/is uncertain. Furthermore, suppose that said uncertainty is represented via the DS theoretic models $m_X(\cdot)$ and $m_Y(\cdot)$ over the logical FoDs $\{\varphi(x_i), \varphi(\overline{x_i})\}$ and $\{\varphi(y_j), \varphi(\overline{y_j})\}$, respectively. Then, the implication rule $\varphi_1(x_i) \implies \varphi_2(y_j)$ is taken to have the following DS theoretic model:

$$\begin{aligned}m_{\varphi_X \to \varphi_Y}(\cdot) &= (m_X^c \vee m_Y)(\cdot) \\ &= (m_X \wedge m_Y^c)^c(\cdot),\end{aligned} \tag{13}$$

over the FoD $\{\varphi(x_i), \varphi(\overline{x_i})\} \times \{\varphi(y_j), \varphi(\overline{y_j})\}$.

5 Uncertain Logic Quantifiers

We define existential and universal quantifiers in uncertain logic as follows.

Definition 6 (Existential Quantifier in Uncertain Logic) *Consider the statement:*

$$\exists x \ \ \varphi(x), \ \textit{with uncertainty } [\alpha, \beta], \tag{14}$$

where $x \in \Theta_X = \{x_1, x_2, \dots, x_N\}$. Let us define an extended logical FoD $\Theta_{X'} = \{\varphi(x_1), \varphi(x_2), \dots, \varphi(x_N)\} \times \{\mathbf{1}, \mathbf{0}\}$. Then, we define the DS theoretic model for (14) as:

$$\bigvee_{i=1}^{N} \varphi(x_i), \tag{15}$$

over the FoD $\Theta_{X'}$, subject to the constraint:

$$\begin{aligned}m(\mathbf{1}) &= \sum\nolimits_{i=1}^{N} m_\varphi(x_i) = \alpha; \\ m(\mathbf{0}) &= \sum\nolimits_{i=1}^{N} m_\varphi(\overline{x_i}) = 1 - \beta; \\ m(\Theta_{X'}) &= \beta - \alpha.\end{aligned} \tag{16}$$

This model is an alternative to Skolemization [23]. This model, however, does not rule out the use of Skolemization, as there might be scenarios where the latter technique is a better alternative. Note that if the uncertainty of at least one of the propositions $\varphi(x_i)$ in (15) is $[\alpha, \beta]$, and the uncertainty of every other proposition is $[0,0]$ (or, in general, $[\alpha_j, \beta_j]$, with $\alpha_j \leq \alpha$, $\beta_j \leq \beta$, and $i \neq j$), then the DS model corresponding to (15) is equivalent to the DS model corresponding to (14) when the OR operations are computed as indicated by Definitions 3 and 4. Also, although an infinite number of solutions satisfy (16), a useful solution (e.g., for existential instantiation on inference problems) is given by $m_\varphi(x_i) = \alpha$; $m_\varphi(\overline{x_i}) = 1 - \beta$; and $m_\varphi(\{x_i, \overline{x_i}\}) = \beta - \alpha, i = 1, 2, \dots, N$. This solution can be proven by successively applying the idempotency property to the OR operator.

Definition 7 (Universal Quantifier in Uncertain Logic) *Consider the statement:*

$$\forall x \ \ \varphi(x), \ \textit{with uncertainty } [\alpha, \beta], \tag{17}$$

where $x \in \Theta_X = \{x_1, x_2, \dots, x_N\}$. Then, we define the DS theoretic model for (17) as:

$$\bigwedge_{i=1}^{N} \varphi(x_i), \tag{18}$$

over the FoD $\Theta_{X'} = \{\varphi(x_1), \varphi(x_2), \dots, \varphi(x_N)\} \times \{\mathbf{1}, \mathbf{0}\}$, subject to the constraint:

$$\begin{aligned}m(\mathbf{1}) &= \sum\nolimits_{i=1}^{N} m_\varphi(x_i) = \alpha; \\ m(\mathbf{0}) &= \sum\nolimits_{i=1}^{N} m_\varphi(\overline{x_i}) = 1 - \beta; \\ m(\Theta_{X'}) &= \beta - \alpha.\end{aligned} \tag{19}$$

Note that if the uncertainty of every proposition $\varphi(x_i)$ in (18) is $[\alpha, \beta]$, then the DS model corresponding to (18) is equivalent to the DS model corresponding to (17) when the AND operations are computed as indicated by Definitions 3 and 4. Also, although an infinite number of solutions satisfy (19), a useful solution (e.g., for universal instantiation on inference) is given by $m_\varphi(x_i) = \alpha$; $m_\varphi(\overline{x_i}) = 1 - \beta$; and $m_\varphi(\{x_i, \overline{x_i}\}) = \beta - \alpha, i = 1, 2, \dots, N$. This solution can be proven by applying idempotency to the AND operator.

6 Inference in Uncertain Logic

Inference in uncertain logic shares the fundamental principles of classical logic, and adds the possibility of attaching, tracking, and propagating uncertainties that may arise on premises and/or rules. Due to the extensive number of methods for logic inference, the scope of this section is limited to the introduction of some of the most fundamental inference rules, along with some basic examples that illustrate uncertain logic inference. For an extended definition of these rules and their application for inference in the context of classical logic, we refer the reader to [22].

Modus Ponens (MP). This rule states that, whenever the logic sentences $\varphi \implies \psi$ and φ have been established, then it is acceptable to infer the sentence ψ as well. MP extends to uncertain logic as follows. Consider:

$$\begin{aligned} &\varphi_1(x), \text{ with uncertainty } [\alpha_1, \beta_1]; \\ &\varphi_2(y), \text{ with uncertainty } [\alpha_2, \beta_2]; \text{ and} \\ \varphi_1(x) \implies &\varphi_2(y), \text{ with uncertainty } [\alpha_R, \beta_R]. \end{aligned} \tag{20}$$

Then, given the uncertain premises $\varphi_1(x) \implies \varphi_2(y)$ and φ_1, MP allows us to infer the uncertain expression $\varphi_2(y)$. Note that, if the uncertainty parameters $[\alpha_2, \beta_2]$ are unknown, their value should be obtained by applying the methodology introduced in Section 4 above. It can be shown that uncertain MP (as well as the inference rules introduced this section) lead to $\top_\sigma$, with $\sigma = [\max(\alpha_R, 1-\beta_R), \max(\alpha_R, 1-\beta_R)]$.

To better understand MP in uncertain logic, consider an example where $\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 1$. By using the model in Definition 5, we can obtain $\alpha_R = \beta_R = 1$. Furthermore, given the $\varphi_1(x) \implies \varphi_2(y)$ and $\varphi_1(x)$, then we can infer $\varphi_2(y)$ with uncertainty $[\alpha_2 = \beta_2] = [1, 1]$. This case represents a scenario with no uncertainty.

Now consider a scenario where there is uncertainty in the rule, in such a way that $[\alpha_R, \beta_R] = [0.5, 1.0]$, and assume that we have a model for the uncertainty of $\varphi_1(x)$ such that $\alpha_1 = \beta_1 = 1$. Then, MP allows us to infer $\varphi_2(y)$, with the uncertainty $[\alpha_2, \beta_2]$ obtained from the equations $\alpha_R = \max(1-\beta_1, \alpha_2)$ and $\beta_R = \max(1-\alpha_1, \beta_2)$. Solving these equations we obtain $\alpha_2 = 0.5$ and $\beta_2 = 1$.

Modus Tolens (MT). This rule states that, if we know that $\varphi \implies \psi$, then we can infer $\neg\varphi$ if we believe that ψ is false. MT extends to uncertain logic as follows. Assume that the uncertainty on each of the expressions involved in MP are defined by (20). Then, given the uncertain premises $\varphi_1(x) \implies \varphi_2(y)$ and $\neg\varphi_2$, MT allows us to infer the uncertain expression $\neg\varphi_1(y)$. As with MP above, if the uncertainty parameters $[\alpha_2, \beta_2]$ are unknown, their value should be obtained by applying the methodology introduced in Section 4.

Other rules of inference. Uncertain logic can be extended by incorporating new rules of inference that already exist in conventional logic inference. Some examples of new rules of inference are: AND elimination (AE), AND introduction (AI), universal instantiation (UI), and existential instantiation (EI). The definition of these rules of inference is straightforward based on their definition for conventional logic, and is not included in this manuscript.

Example. Consider the following problem, originally introduced in [22]. We know that horses are faster than dogs and that there is a greyhound that is faster than every rabbit. We know that Harry is a horse and that Ralph is a rabbit. We also know that greyhounds are dogs and that our speed relationship is transitive. Then:

$$\forall x\, \forall y\ \text{Horse}(x) \wedge \text{Dog}(y) \Rightarrow \text{Faster}(x, y) \tag{21a}$$

$$\exists y\ \text{Greyhound}(y) \wedge (\forall z\ \text{Rabbit}(z) \Rightarrow \text{Faster}(y, z)) \tag{21b}$$

$$\forall y\ \text{Greyhound}(y) \Rightarrow \text{Dog}(y) \tag{21c}$$

$$\forall x\, \forall y\, \forall z\ \text{Faster}(x, y) \wedge \text{Faster}(y, z) \Rightarrow \text{Faster}(x, z) \tag{21d}$$

$$\text{Horse}(\text{Harry}) \tag{21e}$$

$$\text{Rabbit}(\text{Ralph}). \tag{21f}$$

Using these logic statements, it can be inferred that Harry is faster than Ralph (i.e., Faster(Harry, Ralph)) [22].

Now, let us introduce uncertain logic operations by assuming that the logic premise (21a) is uncertain, with uncertainty $[\alpha_1, \beta_1]$, and that there is no uncertainty in premises (21b)-(21f). This represents some uncertainty in the sentence "horses are faster than dogs", which may occur if we consider cases such as sick or old horses compared to healthy dogs. The steps that are used for inferring Faster(Harry, Ralph), as well as the uncertainty in each of the steps of this process are in Table 4. It is easy to verify that, if $\alpha_1 = \beta_1 = 1$. The initial steps in the inference process are simply the reproduction of (21a)-(21f) as premises 1 to 6. Steps 7 to 13 can be obtained from applying EI, AI, UI, and MP rules to premises 2 to 6. In our initial example (only the first premise is uncertain), the uncertainty in premises 2 to 6 is $[\alpha_i, \beta_i] = [1, 1], i = 2, 3, \ldots, 6$. Uncertain logic operations become relevant in steps 14 to 19. For example, the uncertainty in premise 16 is obtained from solving the system of equations shown in the corresponding row in Table 4. This system of equations is derived from Definition 5. As a consequence, any change in the uncertainty $[\alpha_1, \beta_1]$ directly affects $[\alpha_{16}, \beta_{16}]$. Figure 1 illustrates the result in a probabilistic scenario. Note that, for us to be able to conclude "Faster(Harry, Ralph)" given the initial uncertainty, α_4 must be larger than α_1. Similar results can be further verified by modifying uncertainties on the premises, whose values can be computed as indicated in Table 4.

7 Conclusions

We have introduced *Uncertain Logic*, a DS theoretic approach for first order logic operations. Uncertain logic provides support for handling variables and quantifiers, in addition to fundamental logic operations (i.e., $\neg, \wedge, \vee$). The framework introduced in this paper allows systematic generation of mass assignments based on uncertain

Table 4: Steps followed for the inference of the sentence Faster(Harry, Ralph) based on the premises defined in (21). The uncertainty is obtained from applying uncertain logic definitions and rules to the example described in Section 6.

	Logic Formula	Premises & Rule	Uncertainty
1	$\forall x\,\forall y$ Horse$(x) \wedge$ Dog$(y) \Rightarrow$ Faster(x,y)	Δ	$[\alpha_1,\beta_1]$
2	$\exists y$ Greyhound$(y) \wedge (\forall z$ Rabbit$(z) \Rightarrow$ Faster$(y,z))$	Δ	$[\alpha_2,\beta_2]=[1,1]$
3	$\forall y$ Greyhound$(y) \Rightarrow$ Dog(y)	Δ	$[\alpha_3,\beta_3]=[1,1]$
4	$\forall x\,\forall y\,\forall z$ Faster$(x,y) \wedge$ Faster$(y,z) \Rightarrow$ Faster(x,z)	Δ	$[\alpha_4,\beta_4]$
5	Horse(Harry)	Δ	$[\alpha_5,\beta_5]$
6	Rabbit(Ralph)	Δ	$[\alpha_6,\beta_6]=[1,1]$
7	Greyhound(Greg) $\wedge$ $(\forall z$ Rabbit$(z) \implies$ Faster(Greg, z))	2, EI	$[\alpha_7,\beta_7]=[1,1]$
8	Greyhound(Greg)	7, AE	$[\alpha_8,\beta_8]=[1,1]$
9	$\forall z$ Rabbit$(z) \implies$ Faster(Greg, z)	7, AE	$[\alpha_9,\beta_9]=[1,1]$
10	Rabbit(Ralph) $\implies$ Faster(Greg, Ralph)	9, UI	$[\alpha_{10},\beta_{10}]=[1,1]$
11	Faster(Greg, Ralph)	10, 6, MP	$[\alpha_{11},\beta_{11}]=[1,1]$
12	Greyhound(Greg) $\implies$ Dog(Greg)	3, UI	$[\alpha_{12},\beta_{12}]=[1,1]$
13	Dog(Greg)	12, 8, MP	$[\alpha_{13},\beta_{13}]=[1,1]$
14	Horse(Harry) $\wedge$ Dog(Greg) $\implies$ Faster(Harry, Greg)	1, UI	$[\alpha_{14},\beta_{14}]=[\alpha_1,\beta_1]$
15	Horse(Harry) $\wedge$ Dog(Greg)	5, 13, AI	$[\alpha_{15},\beta_{15}]=[\alpha_5,\beta_5]$
16	Faster(Harry, Greg)	14, 15, MP	$[\alpha_{16},\beta_{16}]$ obtained from solving $\begin{cases}\alpha_{14}=\max(1-\beta_{15},\alpha_{16})\\ \beta_{14}=\max(1-\alpha_{15},\beta_{16})\end{cases}$
17	Faster(Harry, Greg) $\wedge$ Faster(Greg, Ralph) $\implies$ Faster(Harry, Ralph)	4, UI	$[\alpha_{17},\beta_{17}]=[\alpha_4,\beta_4]$
18	Faster(Harry, Greg) $\wedge$ Faster(Greg, Ralph)	16, 11, AI	$[\alpha_{18},\beta_{18}]=[\alpha_{16},\beta_{16}]$
19	Faster(Harry, Ralph)	17, 18, MP	$[\alpha_{19},\beta_{19}]$ obtained from solving $\begin{cases}\alpha_{17}=\max(1-\beta_{18},\alpha_{19})\\ \beta_{17}=\max(1-\alpha_{18},\beta_{19})\end{cases}$

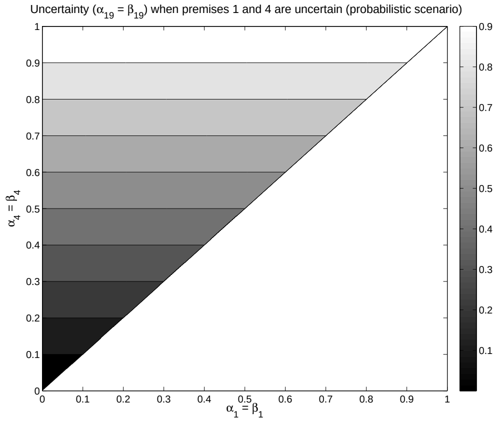


Figure 1: Uncertainty in Premise 19 of Table 4.

first order logic formulas. Furthermore, by using appropriate fusion operators, higher-level applications are possible within this framework, such as inference and resolution based on uncertain data models.

Acknowledgements

This work is based on research supported by the US Office of Naval Research (ONR) via grants #N00014-10-1-0140. and #N00014-11-1-0493, and the US National Science Foundation (NSF) via grant #1038257.

Appendix A. BBA for LC CFE-based AND

Based on the definition of the CFE fusion operator:

$$m(x) = \gamma_1(x) + \gamma_1(\Theta)m_1(x) + \gamma_2(x) + \gamma_2(\Theta)m_2(x). \quad (22)$$

Substituting the CFE coefficients for the AND operation, as indicated by Definition 4, in (22):

$$m(x) = 2\gamma(x) + 2\gamma(\Theta)(\alpha_1 + \alpha_2).$$

- When $\delta_1 + \delta_2 \neq 0$:

$$\begin{aligned} m(x) &= \frac{\underline{\alpha}(\beta_1+\beta_2) - \underline{\beta}(\alpha_1+\alpha_2)}{\delta_1+\delta_2} + \frac{\underline{\delta}(\alpha_1+\alpha_2)}{\delta_1+\delta_2} \\ &= \tfrac{1}{\delta_1+\delta_2}(\underline{\alpha}\beta_1 + \underline{\alpha}\beta_2 - \alpha_1\underline{\beta} - \alpha_2\underline{\beta} + \underline{\delta}(\alpha_1+\alpha_2)). \end{aligned}$$

  Since $\underline{\delta} = \underline{\beta} - \underline{\alpha}$:

$$\begin{aligned} m(x) &= \tfrac{1}{\delta_1+\delta_2}(\underline{\alpha}\beta_1 + \underline{\alpha}\beta_2 - \alpha_1\underline{\beta} - \alpha_2\underline{\beta} \\ &\qquad + \alpha_1\underline{\beta} + \alpha_2\underline{\beta} - \alpha_1\underline{\alpha} - \alpha_2\underline{\alpha}) \\ &= \tfrac{1}{\delta_1+\delta_2}(\underline{\alpha}\beta_1 + \underline{\alpha}\beta_2 - \alpha_1\underline{\alpha} - \alpha_2\underline{\alpha}) \\ &= \tfrac{1}{\delta_1+\delta_2}(\underline{\alpha}(\beta_2 - \alpha_2 + \beta_1 - \alpha_1)). \qquad (23) \end{aligned}$$

  Substituting $\delta_1 = \beta_1 - \alpha_1$ and $\delta_2 = \beta_2 - \alpha_2$ in (23): $m(x) = \underline{\alpha}$.

- When $\delta_1 + \delta_2 = 0$, and making $\gamma(\Theta) = 0$:

$$m(x) = 2\gamma(x) = \underline{\alpha}.$$

The mass $m(\overline{x})$ is given by:

$$m(\overline{x}) = \gamma_1(\overline{x}) + \gamma_1(\Theta)m_1(\overline{x}) + \gamma_2(\overline{x}) + \gamma_2(\Theta)m_2(\overline{x}). \quad (24)$$

Substituting the CFE coefficients as indicated by Definition 4 in (24):

$$m(\overline{x}) = 2\gamma(\overline{x}) + 2\gamma(\Theta)(2 - \beta_1 - \beta_2).$$

- When $\delta_1 + \delta_2 \neq 0$:

$$\begin{aligned} m(\overline{x}) &= \frac{\delta_1 + \delta_2 - \underline{\beta}(2-\alpha_1-\alpha_2) + \underline{\alpha}(2-\beta_1-\beta_2)}{\delta_1+\delta_2} \\ &\quad + \frac{\underline{\delta}(2-\beta_1-\beta_2)}{\delta_1+\delta_2} \\ &= \tfrac{1}{\delta_1+\delta_2}(\delta_1 + \delta_2 - \underline{\beta}(2-\alpha_1-\alpha_2) \\ &\qquad + \underline{\alpha}(2-\beta_1-\beta_2) + \underline{\delta}(2-\beta_1-\beta_2)). \end{aligned}$$

Since $\underline{\delta} = \underline{\beta} - \underline{\alpha}$:

$$\begin{aligned} m(\overline{x}) &= \tfrac{1}{\delta_1+\delta_2}(\delta_1 + \delta_2 - \underline{\beta}(2 - \alpha_1 - \alpha_2) \\ &\qquad + \underline{\alpha}(2 - \beta_1 - \beta_2) + (\underline{\beta} - \underline{\alpha})(2 - \beta_1 - \beta_2)) \\ &= \tfrac{1}{\delta_1+\delta_2}(\delta_1 + \delta_2 - \underline{\beta}(2 - \alpha_1 - \alpha_2 - 2 + \beta_1 + \beta_2)) \\ &= \tfrac{1}{\delta_1+\delta_2}(\delta_1 + \delta_2 - \underline{\beta}(\beta_1 - \alpha_1 + \beta_2 - \alpha_2)). \end{aligned}$$

Substituting $\delta_1 = \beta_1 - \alpha_1$ and $\delta_2 = \beta_2 - \alpha_2$ in (24):

$$m(x) = 1 - \underline{\beta}.$$

- When $\delta_1 + \delta_2 = 0$, and making $\gamma(\Theta) = 0$:

$$m(\overline{x}) = 2\gamma(\overline{x}) = 1 - \underline{\alpha} = 1 - \underline{\beta}.$$

Finally, $m(\Theta) = 1 - m(x) - m(\overline{x}) = \underline{\beta} - \underline{\alpha}$.

Appendix B. Properties of the LC CFE-based Uncertain Logic operations

Consider logic expressions of the form $\varphi(x_i)$, with $1 \leq i \leq N$. Then, the following properties are satisfied:

1. *Idempotency*: This property is defined by: $\varphi_i(x) \wedge \varphi_i(x) = \varphi_i(x) \vee \varphi_i(x) = \varphi_i(x)$. In this case:

$$\begin{aligned} m_\wedge(x) &= \underline{\alpha} = \min(\alpha_i, \alpha_i) = \alpha_i \\ &= \max(\alpha_i, \alpha_i) = \overline{\alpha} = m_\vee(x); \\ m_\wedge(\overline{x}) &= 1 - \underline{\beta} = 1 - \min(\beta_i, \beta_i) = 1 - \beta_i \\ &= 1 - \max(\beta_i, \beta_i) = 1 - \overline{\beta} = m_\vee(\overline{x}); \\ m_\wedge(\Theta) &= \underline{\beta} - \underline{\alpha} = \beta_i - \alpha_i \\ &= \overline{\beta} - \overline{\alpha} = m_\vee(\Theta). \end{aligned}$$

2. *Commutativity*: This property refers to satisfying: $\varphi_1(x) \wedge \varphi_2(x) = \varphi_2(x) \wedge \varphi_1(x)$, and $\varphi_1(x) \vee \varphi_2(x) = \varphi_2(x) \vee \varphi_1(x)$. Let us call $m_{\varphi_i \wedge \varphi_j}(\cdot)$ the BBA resulting from $\varphi_i(x) \wedge \varphi_j(x)$, $i = \{1, 2\}$. Then, for the AND operation:

$$\begin{aligned} m_{\varphi_1\wedge\varphi_2}(x) &= \min(\alpha_1, \alpha_2) \\ &= \min(\alpha_2, \alpha_1) = m_{\varphi_2\wedge\varphi_1}(x) \\ m_{\varphi_1\wedge\varphi_2}(\overline{x}) &= 1 - \min(\beta_1, \beta_2) \\ &= 1 - \min(\beta_2, \beta_1) = m_{\varphi_2\wedge\varphi_1}(\overline{x}) \\ m_{\varphi_1\wedge\varphi_2}(\Theta) &= \min(\beta_1, \beta_2) - \min(\alpha_1, \alpha_2) \\ &= \min(\beta_2, \beta_1) - \min(\alpha_2, \alpha_1) \\ &= m_{\varphi_2\wedge\varphi_1}(\Theta). \end{aligned}$$

   A proof for commutativity for the logical OR operation is obtained by following a similar procedure.

3. *Associativity*: The associative property is defined by: $\varphi_1(x) \wedge [\,\varphi_2(x) \wedge \varphi_3(x)\,] = [\varphi_1(x) \wedge \varphi_2(x)\,] \wedge \varphi_3(x)$, and $\varphi_1(x) \vee [\,\varphi_2(x) \vee \varphi_3(x)\,] = [\,\varphi_1(x) \vee \varphi_2(x)\,] \vee \varphi_3(x)$. Let us call $\varphi_4(\cdot)$ the model generated by $\varphi_2(x) \wedge \varphi_3(x)$, and $\varphi_5(\cdot)$ the model generated by $\varphi_1(x) \wedge \varphi_2(x)$. Also, let us call $m_{\varphi_i\wedge\varphi_j}(\cdot)$ the BBA resulting from $\varphi_i(x) \wedge \varphi_j(x)$, $i = \{1, \ldots, 5\}$. Our goal (for the AND operation) is to show that the model for $\varphi_1(\cdot) \wedge \varphi_4(\cdot)$ is equivalent to the model for $\varphi_5(\cdot) \wedge \varphi_3(\cdot)$:

$$\begin{aligned} m_{\varphi_1\wedge\varphi_4}(x) &= \min(\alpha_1, \min(\alpha_2, \alpha_3)) \\ &= \min(\min(\alpha_1, \alpha_2), \alpha_3) = m_{\varphi_5\wedge\varphi_3}(x) \\ m_{\varphi_1\wedge\varphi_4}(\overline{x}) &= 1 - \min(\beta_1, \min(\beta_2, \beta_3)) \\ &= 1 - \min(\min(\beta_1, \beta_2), \beta_3) \\ &= m_{\varphi_5\wedge\varphi_2}(\overline{x}) \\ m_{\varphi_1\wedge\varphi_4}(\Theta) &= \min(\beta_1, \min(\beta_2, \beta_3)) \\ &\quad - \min(\alpha_1, \min(\alpha_2, \alpha_3)) \\ &= \min(\min(\beta_1, \beta_2), \beta_3) \\ &\quad - \min(\min(\alpha_1, \alpha_2), \alpha_3) = m_{\varphi_5\wedge\varphi_3}(\Theta). \end{aligned}$$

   A proof for associativity for the logical OR operation is obtained by following a similar procedure.

4. *Distributivity*: Distributive operations satisfy: $\varphi_1(x_i) \wedge [\,\varphi_2(x_j) \vee \varphi_3(x_k)\,] = [\,\varphi_1(x_i) \wedge \varphi_2(x_j)\,] \vee [\,\varphi_1(x_i) \wedge \varphi_3(x_j)\,]$, and $\varphi_1(x_i) \vee [\,\varphi_2(x_j) \wedge \varphi_3(x_k)\,] = [\,\varphi_1(x_i) \vee \varphi_2(x_j)\,] \wedge [\,\varphi_1(x_i) \vee \varphi_3(x_j)\,]$. Let us call $\varphi_4(\cdot)$ the model generated by $\varphi_1(x) \wedge [\varphi_2(x) \vee \varphi_3(x)]$, and $\varphi_5(\cdot)$ the model generated by $[\varphi_1(x) \wedge \varphi_2(x)] \vee [\varphi_1(x_)) \wedge \varphi_3(x_))]$. Our goal is to show that the model for $\varphi_4(\cdot)$ is equivalent to the model for $\varphi_5(\cdot)$. In general, these two models are:

$$\begin{aligned} m_{\varphi_4}(x) &= \min(\alpha_1, \max(\alpha_2, \alpha_3)); \\ m_{\varphi_4}(\overline{x}) &= 1 - \min(\beta_1, \max(\beta_2, \beta_3)); \\ m_{\varphi_4}(\Theta) &= \min(\beta_1, \max(\beta_2, \beta_3)); \\ &\qquad - \min(\alpha_1, \max(\alpha_2, \alpha_3)); \text{ and} \end{aligned}$$

$$\begin{aligned} m_{\varphi_5}(x) &= \max(\min(\alpha_1, \alpha_2), \min(\alpha_1, \alpha_3)); \\ m_{\varphi_5}(\overline{x}) &= 1 - \max(\min(\beta_1, \beta_2), \min(\beta_1, \beta_3)); \\ m_{\varphi_5}(\Theta) &= \max(\min(\beta_1, \beta_2), \min(\beta_1, \beta_3)) \\ &\qquad - \max(\min(\alpha_1, \alpha_2), \min(\alpha_1, \alpha_3)). \end{aligned}$$

   Now, consider the focal set x. We have three cases (other possible cases are equivalent to these three after applying the commutativity rule): (a) $\alpha_1 \leq \alpha_2 \leq \alpha_3$; (b) $\alpha_2 \leq \alpha_1 \leq \alpha_3$; and (c) $\alpha_2 \leq \alpha_3 \leq \alpha_1$. The mass associated to the focal set x is:

   (a) $m_{\varphi_4}(x) = \alpha_1 = m_{\varphi_5}(x)$;

   (b) $m_{\varphi_4}(x) = \alpha_1 = m_{\varphi_5}(x)$; and

   (c) $m_{\varphi_4}(x) = \alpha_3 = m_{\varphi_5}(x)$;

   i.e., $m_{\varphi_4}(x) = m_{\varphi_5}(x)$ in all the cases. For the focal set $\overline{x}$ we also have three basic cases: (a) $\beta_1 \leq \beta_2 \leq \beta_3$; (b) $\beta_2 \leq \beta_1 \leq \beta_3$; and (c) $\beta_2 \leq \beta_3 \leq \beta_1$; which render:

   (a) $m_{\varphi_4}(\overline{x}) = 1 - \beta_1 = m_{\varphi_5}(\overline{x})$;

   (b) $m_{\varphi_4}(\overline{x}) = 1 - \beta_1 = m_{\varphi_5}(\overline{x})$; and

   (c) $m_{\varphi_4}(\overline{x}) = 1 - \beta_3 = m_{\varphi_5}(\overline{x})$;

   Based on the cases above, it can be shown that also $m_{\varphi_4}(\Theta) = m_{\varphi_5}(\Theta)$, proving distributivity for the logical AND operation. A proof for distributivity for the logical OR operation is obtained by following a similar procedure.

References

[1] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[2] G. Shafer, *A mathematical theory of evidence*. Princeton University Press, 1976.

[3] J. Y. Halpern, "An analysis of first-order logics of probability," *Artificial Intelligence*, vol. 46, pp. 311–350, December 1990.

[4] J. Y. Halpern, "A logical approach to reasoning about uncertainty: a tutorial," in *Discourse, Interaction and Communication* (X. Arrazola, K. Korta, and F. Pelletier, eds.), Philosophical Studies Series, Springer, 1998.

[5] D. Dubois, P. Hájek, and H. Prade, "Knowledge-driven versus data-driven logics," *Journal of Logic, Language and Information*, vol. 9, pp. 65–89, January 2000.

[6] L. Boldrin and A. Saffiotti, "A modal logic for fusing partial belief of multiple reasoners," *Journal of Logic and Computation*, vol. 9, no. 1, pp. 81–103, 1999.

[7] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy sets and systems*, vol. 1, no. 1, pp. 3–28, 1978.

[8] L. Godo, P. Hájek, and F. Esteva, "A fuzzy modal logic for belief functions," *Fundam. Inform.*, vol. 57, no. 2-4, pp. 127–146, 2003.

[9] N. J. Nilsson, "Probabilistic logic," *Artificial Intelligence*, vol. 28, pp. 71–88, February 1986.

[10] D. Dubois and H. Prade, "Possibilistic logic: a retrospective and prospective view," *Fuzzy Sets and Systems*, vol. 144, no. 1, pp. 3–23, 2004.

[11] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

[12] D. Dubois and H. Prade, *Possibility Theory: An Approach to Computerized Processing of Uncertainty (traduction revue et augmente de "Thorie des Possibilits")*. New York: Plenum Press, 1988.

[13] R. Fagin and J. Y. Halpern, "Uncertainty, belief, and probability," in *Proceedings of the 11th international joint conference on Artificial intelligence - Volume 2*, IJCAI'89, (San Francisco, CA, USA), pp. 1161–1167, Morgan Kaufmann Publishers Inc., 1989.

[14] G. M. Provan, "A logic-based analysis of Dempster-Shafer theory," *International Journal of Approximate Reasoning*, vol. 4, no. 5-6, pp. 451–495, 1990.

[15] A. Saffiotti, "A belief-function logic," in *Proceedings of the tenth national conference on Artificial intelligence*, AAAI'92, pp. 642–647, AAAI Press, 1992.

[16] A. Saffiotti and E. Umkehrer, "Inference-driven construction of valuation systems from first-order clauses," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 24, no. 11, pp. 1611–1624, 1994.

[17] B. Ristic and P. Smets, "Target identification using belief functions and implication rules," *IEEE transactions on Aerospace and Electronic Systems*, vol. 41, July 2005.

[18] A. Benavoli, L. Chisci, A. Farina, and B. Ristic, "Modelling uncertain implication rules in evidence theory," in *11th International Conference on Information Fusion*, 2008.

[19] T. Wickramarathne, K. Premaratne, and M. N. Murthi, "Consensus-Based Credibility Estimation of Soft Evidence for Robust Data Fusion," in *Proceedings of the 2nd International Conference on Belief Functions, Compiègne, France, 9-11 May 2012*, vol. 164 of *Advances in Soft Computing*, pp. 301–309, Springer, 2012.

[20] K. Premaratne, M. N. Murthi, J. Zhang, M. Scheutz, and P. H. Bauer, "A Dempster-Shafer theoretic conditional approach to evidence updating for fusion of hard and soft data," in *12th International Conference on Information Fusion, 2009. FUSION '09*, pp. 2122 –2129, July 2009.

[21] R. Fagin and J. Y. Halpern, "A new approach to updating beliefs," in *Uncertainty in Artificial Intelligence*, pp. 347–374, Elsevier Science Publishers, 1991.

[22] M. R. Genesereth and N. J. Nilsson, *Logical foundations of artificial intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1987.

[23] E. A. Bender, *Mathematical methods in artificial intelligence*. Los Alamitos, CA, USA: IEEE Computer Society Press, 1996.

[24] G.-J. Wang and Y. Leung, "Integrated semantics and logic metric spaces," *Fuzzy Sets Syst.*, vol. 136, pp. 71–91, May 2003.

[25] D. Dubois and H. Prade, "A set-theoretic view of belief functions," in *Classic Works of the Dempster-Shafer Theory of Belief Functions* (R. Yager and L. Liu, eds.), vol. 219 of *Studies in Fuzziness and Soft Computing*, pp. 375–410, Springer Berlin / Heidelberg, 2008.

Characterizing Coherence, Correcting Incoherence

Erik Quaeghebeur
SYSTeMS Research Group, Ghent University & Decision Support Systems Group, Utrecht University
Erik.Quaeghebeur@UGent.be

Abstract

Lower previsions defined on a finite set of gambles can be looked at as points in a finite-dimensional real vector space. Within that vector space, the sets of sure loss avoiding and coherent lower previsions form convex polyhedra. We present procedures for obtaining characterizations of these polyhedra in terms of a minimal, finite number of linear constraints. As compared to the previously known procedure, these procedures are more efficient and much more straightforward. Next, we take a look at a procedure for correcting incoherent lower previsions based on pointwise dominance. This procedure can be formulated as a multi-objective linear program, and the availability of the finite characterizations provide an avenue for making these programs computationally feasible.

Keywords. Coherence, avoiding sure loss, linear constraint, polytope, enumeration, projection, multi-objective linear programming, incoherence, dominance.

1 Introduction

In the theory of coherent lower previsions (for an overview, see Walley 1991 or Miranda 2008), its coherence condition takes a central role: it defines which models—lower previsions—are fully rational, meaning that they do not implicitly encode commitments—in terms of buying prices for gambles—that are more demanding than the ones explicitly made. The consequences of this criterion have been extensively studied both in the unconditional and the conditional case, in finite and infinite spaces.

In this paper, we study the coherence criterion for unconditional lower previsions defined on a finite set of gambles, which in turn are essentially defined on a finite possibility space. What can we still add in this restricted setting? Results that make new numerical applications feasible, namely, procedures for obtaining a characterization of coherence in terms of a minimal, finite number of linear constraints that are more efficient than the existing one. These results are presented in Section 4. Note that our procedures give an answer to the question “Which lower previsions are coherent?”, and should not be confused with verification procedures, which deal with the question “Is this specific lower prevision coherent?”. Of course, the characterization our procedures generate can be used for verification purposes, but this may be reasonable only if many verifications need to be performed

One may wonder what new kinds of applications are possible once we have a minimal linear constraints characterization? In Section 5, we provide one example in a proposal for a method to correct an incoherent lower prevision downward to make it coherent. Similarly to natural extension, this method is formulated in terms of pointwise dominance of lower previsions.

Because of the finitary context of this paper and its aim to be an enabler for numerical applications, it is advantageous to reformulate a number of variants of the coherence criterion and the related criterion of avoiding sure loss in matrix terms; we do this in Section 3.

We make use of polytope theory concepts throughout this paper. We also make use of multi-objective linear programming both for our downward correction method as well as for some of our procedures to obtain the minimal linear constraints characterization. Therefore, we start out with short primers on these topics in Section 2.

2 Primers

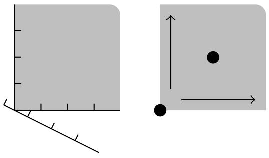
2.1 Polytope Theory Essentials

Let us review some concepts and techniques from polytope theory (for more information, see, e.g., Grünbaum 1967, Ziegler 1995, or Fukuda 2004). Any convex polyhedron in a n-dimensional space can be described in two ways:

As an *H-representation* $\{x \in \mathbb{R}^n : Ax \leq b\}$: A set of k linear constraints (inequalities/half-spaces) defined by a matrix A in $\mathbb{R}^{k\times n}$ and a column vector b in $\mathbb{R}^k$; denoted compactly as $[A,b]$, where the comma denotes horizontal concatenation of matrices.

As a *V-representation* $\{x \in \mathbb{R}^n : x = V\mu \wedge \mu \geq 0 \wedge w^\mathsf{T}\mu = 1\}$: A set of ℓ points and rays, defined by a matrix V in $\mathbb{R}^{n\times\ell}$ and a row vector w in $\mathbb{R}^\ell$, with the zero components indicating rays; denoted compactly as $[V;w]$, where the semicolon denotes vertical concatenation of matrices.

The two representations are dual in the sense that $[A^\mathsf{T};b^\mathsf{T}]$ is the V-representation of some polyhedron and $[V^\mathsf{T}, w^\mathsf{T}]$ is the H-representation of some—possibly different—polyhedron. This duality is also present in the algorithms of polytope theory.



On the right, we give a simple 2-dimensional polyhedron, in gray, in both a visual H- and V-representation.

H- and V-representations may contain redundant constraints and points or rays, i.e., those that are implied by the other constraints or the other points or rays. Non-redundant extreme points or rays are called vertices and extreme rays. In our illustration, there is one redundant constraint in the H-representation and one redundant point in the V-representation. Let i be the total number of constraints or points and j the non-redundant number; redundancy removal algorithms essentially require solving i linear programming problems of size $n \times j$ (Clarkson 1994).

Moving between the H- and V-representations is done using vertex enumeration algorithms and the dual facet enumeration algorithms. There are enumeration algorithms with a complexity linear in n, k, and ℓ (Avis & Fukuda 1992). Nevertheless, enumeration is inherently highly complex, as ℓ can be exponential in k and vice versa.

Projecting a polyhedron is straightforward in V-representation: project the vertices and then remove the redundant ones. However, in H-representation the best technique depends on the polyhedron's properties: the classical approach, Fourier–Motzkin elimination, is inefficient and on top of that generates a lot of redundant constraints; another approach, block elimination, is inefficient when the number of vertices is high, which is common. The equality set projection approach is claimed to be useful in such cases (Jones et al. 2004), but our input data caused errors in the available code (Kvasnica et al. 2006).

Below, we assume that the output of enumeration and projection algorithms is minimal, i.e., non-redundant.

2.2 Multi-Objective Linear Programming

We here give a brief introduction to multi-objective linear programming (for more information, see, e.g., Ehrgott 2005). We assume familiarity with standard, single objective linear programming (if not, have a quick look at a standard reference such as Bertsimas & Tsitsiklis 1997).

Any *multi-objective linear program* can be put in the following form:

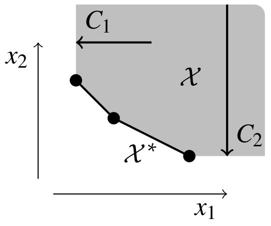
$$\begin{aligned} &\text{maximize } \ y = Cx, \\ &\text{subject to } \ Ax \leq b \ \text{ and } \ x \geq 0. \end{aligned} \tag{1}$$

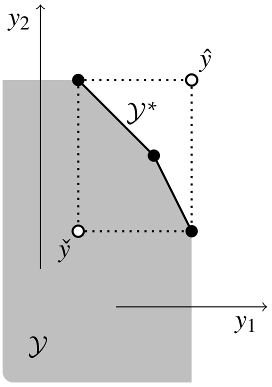
In this program, x denotes the n-dimensional real *optimization vector*, y is the m-dimensional *objective vector*, and $Ax \leq b$ is a set of k *linear constraints*; so we assume $C \in \mathbb{R}^{m\times n}$, $A \in \mathbb{R}^{k\times n}$, and $b \in \mathbb{R}^k$ as given. Vector inequalities should be read as follows: $x \geq z \Leftrightarrow \min(x-z) \geq 0$ and $x > z \Leftrightarrow x \geq z \wedge x \neq z$. Here, min (max) selects its argument vector's minimum (maximum) component value.

Whereas in single objective linear programming, with $m = 1$, all optimization vectors x are completely ordered by the single objective, whenever $m > 1$, they are only partially ordered through the standard ordering of the objective vectors. Consequently, whereas in single objective linear programming all optimal solutions are equivalent from the objective value point of view, in multi-objective linear programming there are in general multiple sets of incomparable 'Pareto' optimal (or 'efficient')—i.e. *C-undominated*—solutions.

The sets of feasible optimization and objective vectors are $\mathcal{X} := \{x \in \mathbb{R}^n : Ax \leq b \wedge x \geq 0\}$ and $\mathcal{Y} := \{Cx : x \in \mathcal{X}\}$, respectively. Furthermore, $\mathcal{X}^* := \{x \in \mathcal{X} : (\forall z \in \mathcal{X} : Cx \not< Cz)\}$ is the set of C-undominated solutions, and so $\mathcal{Y}^* := \{Cx : x \in \mathcal{X}^*\}$ is the set of undominated objective vectors. The sets of extreme points of the sets of undominated solutions and objectives are $\operatorname{ext}\mathcal{X}^*$ and $\operatorname{ext}\mathcal{Y}^*$, respectively.

Let us give a simple graphical illustration (with $n = m = 2$) below right to clarify the concepts just introduced. The sets $\mathcal{X}$ and $\mathcal{Y}$ are shaded gray. The sets $\mathcal{X}^*$ and $\mathcal{Y}^*$ are shown as black lines. The members of $\operatorname{ext}\mathcal{X}^*$ and $\operatorname{ext}\mathcal{Y}^*$ are shown as black dots. The vectors C_1 and C_2—rows of C—that point towards higher objective vector component values are drawn free: only their direction and magnitude matter.





In the picture of the objective vector space, we have included the so-called *ideal point* $\hat{y}$ and *nadir point* $\check{y}$, the upper and lower envelopes of $\mathcal{Y}^*$, respectively. They provide bounds on the values attained by the undominated objective vector components.

The main computational tasks are, in non-decreasing order of complexity:

M1. Finding the ideal point $\hat{y}$, which can be done by solving a linear program maximizing each of the components of y separately.

M2. Finding the nadir point $\check{y}$ (for algorithms, see Ehrgott & Tenfelde-Podehl 2003 and Alves & Costa 2009).

M3. Finding the extreme points $\operatorname{ext}\mathcal{Y}^*$ and the whole set $\mathcal{Y}^*$ of undominated objective vectors (for algorithms, see Benson 1998 and Ehrgott et al. 2012; these are relatively efficient only if m is small compared to n).

M4. Finding the extreme points $\operatorname{ext}\mathcal{X}^*$ of the set of optimal optimization vectors (for algorithms, MOLP simplex solvers, see, e.g., Evans & Steuer 1973, Strijbosch et al. 1991, or Ehrgott 2005, Sec. 7).

M5. Finding the whole set $\mathcal{X}^*$ of optimal optimization vectors (for algorithms, based on post-processing the MOLP simplex solver output, see, e.g., Yu & Zeleny 1975 or Isermann 1977).

3 Matrix Formulations of Avoiding Sure Loss and Coherence

Consider a finite possibility space Ω and a finite set of gambles $\mathcal{K} \subset \mathbb{R}^{\Omega}$ on this possibility space. The elements of $\mathcal{K}$ can be looked at as vectors; we group them as columns in a gamble matrix $K \in \mathbb{R}^{\Omega\times\mathcal{K}}$. We use the same notation for scalars and constant vectors; the identity matrix is denoted $\mathbb{I}$; there will be no ambiguity in this paper because we leave their size implicit. The columns of $K^{\top}$ are the degenerate previsions, so $\{K^{\top}\mu : \mu \geq 0 \wedge 1^{\top}\mu = 1\}$ is the set of linear previsions. Any lower prevision $\underline{P}$ defined on $\mathcal{K}$ can be looked at as a column vector in $\mathbb{R}^{\mathcal{K}}$. Similarly, min and max can also thought of at as column vectors in $\mathbb{R}^{\mathcal{K}}$.

A lower prevision $\underline{P}$ on $\mathcal{K}$ is said to *avoid sure loss* (cf., e.g., Walley 1991, §2.4) if and only if

$$\forall\lambda \geq 0 : \ \underline{P}^{\top}\lambda \leq \max(K\lambda), \qquad (2)$$

or, based on dominance by a linear prevision (cf. Walley 1991, §3.3.3(a)), if

$$\exists\mu \geq 0 : \ \underline{P} \leq K^{\top}\mu \ \wedge \ 1^{\top}\mu = 1, \qquad (3)$$

or, by introducing slack variables, if

$$\exists\mu,\nu \geq 0 : \ \underline{P} = K^{\top}\mu - \mathbb{I}\nu \ \wedge \ 1^{\top}\mu = 1. \qquad (4)$$

This last form shows that the set of all sure loss avoiding lower previsions is a convex polyhedron by providing a V-representation

$$\begin{bmatrix} V \\ w \end{bmatrix} := \begin{bmatrix} K^{\top} & -\mathbb{I} \\ 1^{\top} & 0^{\top} \end{bmatrix}. \qquad (5)$$

Now, let $\mathcal{S}$ denote the set of matrices obtained from the identity matrix by changing at most one 1 to -1. Then a lower prevision $\underline{P}$ on $\mathcal{K}$ is called *coherent* (cf., e.g., Walley 1991, §2.5) if and only if

$$\forall S \in \mathcal{S} : \forall\lambda \geq 0 : \ \underline{P}^{\top}S\lambda \leq \max(KS\lambda), \qquad (6)$$

or, by formal analogy to Equation (3) and because $S^{\top} = S$, if

$$\forall S \in \mathcal{S} : \exists\mu_S \geq 0 : \ S\underline{P} \leq SK^{\top}\mu_S \ \wedge \ 1^{\top}\mu_S = 1, \qquad (7)$$

or, by introducing slack variables and because $S^{-1} = S$, if

$$\forall S \in \mathcal{S} : \exists\mu_S,\nu_S \geq 0 : \ \underline{P} = K^{\top}\mu_S - S\nu_S \ \wedge \ 1^{\top}\mu_S = 1. \qquad (8)$$

This last form shows that the set of all coherent lower previsions is an intersection of $|\mathcal{K}|+1$ convex polyhedra with V-representations

$$\begin{bmatrix} V_S \\ w_S \end{bmatrix} := \begin{bmatrix} K^{\top} & -S \\ 1^{\top} & 0^{\top} \end{bmatrix}, \qquad (9)$$

and therefore is a convex polyhedron. Furthermore, coherence implies that $\min \leq \underline{P} \leq \max$ (cf. Walley 1991, §2.6.1(a)), so the set of coherent lower previsions is a bounded polyhedron, i.e., a *polytope*.

We will later on in this paper use the Lower Envelope Theorem (see, e.g., Walley 1991, §2.6.3):

Theorem. The lower envelope $\underline{P}$ of a subset $\mathcal{Q}$ of the coherent lower previsions on a set of gambles $\mathcal{K}$ is coherent. (So $\underline{P}f := \inf_{\underline{Q}\in\mathcal{Q}} \underline{Q}f$ for each gamble f in $\mathcal{K}$.) ◁

We give a proof based on Equation (7)—a version of the coherence criterion a shallow search of ours left unencountered in the literature:

Proof. By coherence of the $\underline{Q}$ in $\mathcal{Q}$, we have a vector $\mu_{\underline{Q},S}$ such that $S\underline{Q} \leq SK^{\top}\mu_{\underline{Q},S}$ for each S in $\mathcal{S}$. By the lower envelope definition, for $S := \mathbb{I}$, we have $\underline{P} \leq \underline{Q} \leq K^{\top}\mu_{\underline{Q},\mathbb{I}}$ for any $\underline{Q}$ in $\mathcal{Q}$. For other S, let g_S denote the gamble corresponding to the -1 diagonal component in S. Let $\underline{Q}_S$ be a coherent lower prevision from $\mathcal{Q}$ such that $\underline{P}g_S = \underline{Q}_S g_S$. Then $S\underline{P} \leq S\underline{Q}_S \leq SK^{\top}\mu_{\underline{Q}_S,S}$. □

In the literature on verification procedures—which are typically formulated in the more general conditional context—, there is a clear separation between algorithms based on criteria formulations of the type of Equations (2) and (6) (cf. Walley et al. 2004), and those of the type of Equations (3)–(4) and (7)–(8) (see, e.g., Vicig 1996 and Biazzo & Gilio 2000). This separation is also present in the characterization procedures we present; the latter type leads to the procedures in Section 4.1, the former to those in Section 4.2.

4 Computing Constraints Efficiently

Building on earlier work with lower probabilities (Walley 1991, App. A; Quaeghebeur & De Cooman 2008; Quaeghebeur 2009), we presented a procedure for obtaining characterizations of the polytope of coherent lower previsions in terms of a minimal, finite number of linear constraints (Quaeghebeur 2010). However, the procedure is such that a relatively large number of redundant constraints are generated, which at a later step need to be removed—a computationally demanding task. Moreover, the procedure and its derivation is somewhat involved.

It is possible to derive procedures in a more direct way. Some of these more direct procedures turn out to be computationally more efficient as well, resulting in running times that are up to an order of magnitude shorter.

What are our concrete goals? We wish to find minimal H-representations for the set of all lower previsions $\underline{P}$

A. that avoid sure loss ($[\Lambda_{\mathrm{A}}, \alpha_{\mathrm{A}}]$),
B. that avoid sure loss and for which $\underline{P} \geq \min$ ($[\Lambda_{\mathrm{B}}, \alpha_{\mathrm{B}}]$),
C. that are coherent ($[\Lambda_{\mathrm{C}}, \alpha_{\mathrm{C}}]$).

So for each goal, we want to obtain a block matrix $[\Lambda, \alpha]$ that stands for the linear constraints $\Lambda\underline{P} \leq \alpha$.

These goals are formulated based on experimental results from earlier work (Quaeghebeur & De Cooman 2008; Quaeghebeur 2009, 2010): For coherence, we observed that the V-representations have a much larger size than the H-representations, and to such a degree that it currently seems impractical to generate and use them. We observed that avoiding sure loss with lower bound constraints leads to a smaller H-representation than plain avoiding sure loss. As the lower bound constraints are uncontroversial in most contexts, it may be useful to use this combination as a 'lighter' proxy for plain avoiding sure loss.

Below, we first discuss the direct procedures and follow this up with a look at improved versions of our earlier, involved approach. We close the section with a short discussion of our numerical experiments.

4.1 Straightforward Procedures

The straightforward procedures for Goal A go as follows:

A1. Apply a facet enumeration algorithm to the V-representation of the polyhedron of lower previsions that avoid sure loss in Equation (5) to obtain $[\Lambda_{\mathrm{A}}, \alpha_{\mathrm{A}}]$.

A2. As can be seen from Equation (3), we know an H-representation for pairs $[\underline{P}; \mu_{\mathbb{I}}]$ of which the $\underline{P}$-components are lower previsions that avoid sure loss:

$$\begin{bmatrix} A_{\mathbb{I},\underline{P}} & A_{\mathbb{I},\mu_{\mathbb{I}}} & b_0 \end{bmatrix} := \begin{bmatrix} \mathbb{I} & -K^{\mathsf{T}} & 0 \\ & -\mathbb{I} & 1 \\ & 1^{\mathsf{T}} & 1 \\ & -1^{\mathsf{T}} & -1 \end{bmatrix}. \tag{10}$$

Project this H-representation onto the $\underline{P}$-part to obtain $[\Lambda_{\mathrm{A}}, \alpha_{\mathrm{A}}]$.

The straightforward procedures for Goal B build on those for Goal A:

B1. Start from the resulting H-representation of Procedure A1 and add the lower bound constraints to it, i.e., the block row $[-\mathbb{I}, -\min(K)^{\mathsf{T}}]$, where the minimum is taken column-wise. Because some constraints may have become redundant because of this, perform redundancy removal to obtain $[\Lambda_{\mathrm{B}}, \alpha_{\mathrm{B}}]$.

B2. Idem as Procedure B1, but now starting from the H-representation resulting from Procedure A2.

The straightforward procedures for Goal C are based on the similarities of the underlying problem with that of Goal A:

C1. Recall that the polytope of coherent lower previsions is the intersection of $|\mathcal{S}| = |\mathcal{K}| + 1$ polyhedra, one for each value of S. So apply a facet enumeration algorithm to the V-representation as given in Equation (9) for each S to obtain the corresponding H-representations $[A_S, b_S]$. An H-representation of the intersection polyhedron of polyhedra given as H-representations is the vertical concatenation of these matrices. (Intersection of polyhedra in V-representation, or mixed representations is not straightforward.) Perform redundancy removal on this concatenation H-representation to obtain $[\Lambda_{\mathrm{C}}, \alpha_{\mathrm{C}}]$.

C2. As can be seen from Equation (7), for each S we also know an H-representation for pairs $[\underline{P}; \mu_S]$ of which the $\underline{P}$-component belongs to the polyhedron corresponding to S already mentioned in Procedure C1:

$$\begin{bmatrix} A_{S,\underline{P}} & A_{S,\mu_S} & b_0 \end{bmatrix} := \begin{bmatrix} S & -SK^{\mathsf{T}} & 0 \\ & -\mathbb{I} & 1 \\ & 1^{\mathsf{T}} & 1 \\ & -1^{\mathsf{T}} & -1 \end{bmatrix}. \tag{11}$$

Project this H-representation onto the $\underline{P}$-part to obtain the H-representation $[A_S, b_S]$ already encountered in Procedure C1, the remainder of which is to be followed here as well.

C3. Equation (7) also shows that we can actually create a single H-representation for pairs $[\underline{P}; \mu]$ of which the $\underline{P}$-components are coherent lower previsions:

$$\begin{bmatrix} A_{\underline{P}} & A_{\mu} & b \end{bmatrix} := \begin{bmatrix} A_{\mathbb{I},\underline{P}} & A_{\mathbb{I},\mu_{\mathbb{I}}} & & & b_0 \\ \vdots & & \ddots & & \vdots \\ A_{S_g,\underline{P}} & & & A_{S_g,\mu_{S_g}} & b_0 \\ \vdots & & & \ddots & \vdots \end{bmatrix}, \tag{12}$$

where $S_g \in \mathcal{S}$, with g in $\mathcal{K}$, has negative diagonal g-component. Projecting this H-representation onto the $\underline{P}$-part again gives us $[\Lambda_{\mathrm{C}}, \alpha_{\mathrm{C}}]$. Because of the block diagonal structure of the set of columns to be removed by projection, this procedure is essentially identical to Procedure C2 from the computational point of view.

Comparing the two main procedure types, enumeration-based (A1, B1, C1) and projection-based (A2, B2, C2, C3), our numerical experiments showed that the enumeration-based ones were faster by at least an order of magnitude. It is not yet clear whether this is inherent or whether this is due to the fact that the enumeration implementation used (the double description method of Fukuda & Prodon 1996) is efficient, and the facet projection implementations used (Fourier–Motzkin and block elimination) are not.

4.2 A More Involved Type of Procedure

All of the procedures in the previous section were based on Equations (3)–(4) and (7)–(8). In these expressions, $\underline{P}$ appears free, i.e., without being multiplied by a variable vector such as λ, this in contrast to the other expressions characterizing avoiding sure loss and coherence, Equations (2) and (6). This allowed us to consider $\underline{P}$ as variable as well, directly leading to the straightforward procedures.

In our earlier work (Quaeghebeur 2010), we created a procedure starting from the expressions with bound $\underline{P}$. It is, by the standard set by the best performing of the straightforward procedures, inefficient. However, it is possible to create bound-$\underline{P}$-based procedures that are relatively efficient; we present the ones we found here, as the techniques used might be useful in other contexts as well.

We first make an assumption, namely that all gambles are non-constant and non-negative with zero minimum, or NNZM. In Appendix 4.3 immediately following this section we show that for coherent lower previsions this assumption is non-limiting and how to move between general gamble sets and NNZM gamble sets. The assumption is, however, limiting for lower previsions that only avoid sure loss. Note that $\underline{P} \geq \min$ becomes $\underline{P} \geq 0$ for an NNZM set of gambles $\mathcal{K}$; i.e., positivity constraints.

We do not develop procedures for Goal A here and move straight to Goal B, which because of the limiting nature of the NNZM assumption must be seen as preparation for the procedures for Goal C:

B3. We can rewrite Equation (2) as

$$\forall \gamma \in \mathbb{R} : \forall \lambda \geq 0 : \ \max(K\lambda) = \gamma \ \Rightarrow \ \underline{P}^{\mathsf{T}}\lambda \leq \gamma, \quad (13)$$

which, because $\mathcal{K}$ is NNZM, can be normalized to

$$\forall \lambda \geq 0 : \ \max(K\lambda) = 1 \ \Rightarrow \ \underline{P}^{\mathsf{T}}\lambda \leq 1. \quad (14)$$

Now, again because $\mathcal{K}$ is NNZM, $K\lambda$ is pointwise strictly increasing in λ. So we know that the feasible set $\{\lambda \geq 0 : K\lambda \leq 1\}$ is bounded and that apart from 0, all its vertices satisfy $\max(K\lambda) = 1$. So in our procedure, we first vertex enumerate

$$\begin{bmatrix} A & b \end{bmatrix} := \begin{bmatrix} K & 1 \\ -\mathbb{I} & 0 \end{bmatrix}, \quad (15)$$

and then use this V-representation $[V; w]$ for the λ's to construct an H-representation $[V^{\mathsf{T}}, w^{\mathsf{T}}]$ for lower previsions. Add positivity constraints $[-\mathbb{I}, 0]$; then after redundancy removal we obtain $[\Lambda_{\mathrm{B}}, \alpha_{\mathrm{B}}]$.

B4. Because we assume $\mathcal{K}$ is NNZM, $\underline{P} \geq 0$, so we know that all pointwise dominated vertices of the feasible set $\{\lambda \geq 0 : K\lambda \leq 1\}$ encountered in Procedure B3 result in redundant constraints (cf. the implicand in Equation (14)). So we can use the MOLP

$$\begin{aligned} &\text{maximize } \lambda, \\ &\text{subject to } K\lambda \leq 1 \text{ and } \lambda \geq 0, \end{aligned} \quad (16)$$

to select only the undominated vertices. Gather them as columns in a matrix $\hat{V}$ and construct the H-representation $[\hat{V}^{\mathsf{T}}, 1]$ to replace $[V^{\mathsf{T}}, w^{\mathsf{T}}]$ of Procedure B3.

B5. Because $K\lambda$ is pointwise strictly increasing in λ, we can replace the MOLP (16) by

$$\begin{aligned} &\text{maximize } K\lambda, \\ &\text{subject to } K\lambda \leq 1 \text{ and } \lambda \geq 0. \end{aligned} \quad (17)$$

We are now ready to present the procedures for Goal C, which strongly parallel those for Goal B:

C4. We can rewrite Equation (6) as

$$\begin{aligned} &\forall S \in \mathcal{S} : \forall \lambda \geq 0 : \forall \gamma \in \mathbb{R} : \\ &\qquad \max(KS\lambda) = \gamma \ \Rightarrow \ \underline{P}^{\mathsf{T}}S\lambda \leq \gamma, \end{aligned} \quad (18)$$

which, because $\mathcal{K}$ is NNZM and only a single column of KS is non-positive, but with zero maximum, can be normalized and rewritten as

$$\forall S \in \mathcal{S} : \forall \kappa \in \mathbb{R}^{\mathcal{K}} : \\ S\kappa \geq 0 \Rightarrow \begin{cases} \max(K\kappa) = 1 & \Rightarrow \ \underline{P}^{\mathsf{T}}\kappa \leq 1, \\ \max(K\kappa) = 0 & \Rightarrow \ \underline{P}^{\mathsf{T}}\kappa \leq 0. \end{cases} \quad (19)$$

Now, again because $\mathcal{K}$ is NNZM, $K\kappa$ is pointwise monotone strictly increasing in κ. So we know that the set $\{S\kappa \geq 0 : K\kappa \leq 1\}$ is bounded and that apart from 0, all its vertices satisfy $\max(K\kappa) = 1$. We also know that the set $\{0 \leq S\kappa \leq 1 : K\kappa \leq 0\}$ is bounded and that all its vertices satisfy $\max(K\kappa) = 0$. So the procedure consists in, for every S in $\mathcal{S}$, vertex enumerating

$$\begin{bmatrix} A_{S,0} & b_{S,0} \end{bmatrix} := \begin{bmatrix} K & 0 \\ -S & 0 \\ S & 1 \end{bmatrix}, \quad \begin{bmatrix} A_{S,1} & b_{S,1} \end{bmatrix} := \begin{bmatrix} K & 1 \\ -S & 0 \end{bmatrix}; \quad (20)$$

then use the resulting V-representations $[V_{S,1}; w_{S,1}]$ and $[V_{S,0}; w_{S,0}]$ to construct the H-representations $[V^{\mathsf{T}}{}_{S,1}, w^{\mathsf{T}}{}_{S,1}]$ and $[V^{\mathsf{T}}{}_{S,0}, 0]$. Vertically concatenate these H-representations for every S to obtain an H-representation for the set of coherent lower previsions on $\mathcal{K}$ and apply redundancy removal to obtain $[\Lambda_{\mathrm{C}}, \alpha_{\mathrm{C}}]$.

Entirely analogously to what was done in Procedures B4 and B5, we can use MOLPs to generate undominated vertex versions of $[V_{S,\gamma}; w_{S,\gamma}]$ for all S in $\mathcal{S}$ and γ in $\{0, 1\}$:

C5. The κ-variant:

$$\begin{aligned} &\text{maximize } \kappa, \\ &\text{subject to } K\kappa \leq \gamma, \ S\kappa \geq 0 \text{ and, if } \gamma = 0, \ S\kappa \leq 1. \end{aligned} \quad (21)$$

C6. The $K\kappa$-variant:

$$\begin{aligned} &\text{maximize } K\kappa, \\ &\text{subject to } K\kappa \leq \gamma, \ S\kappa \geq 0 \text{ and, if } \gamma = 0, \ S\kappa \leq 1. \end{aligned} \quad (22)$$

In principle, the MOLP-based procedures (B4, B5, C5, and C6) should be more efficient than the vertex enumeration ones (B3, C4), as for both the same polytope needs to be mapped, but for the MOLPs only in part, which also results in less redundant constraints to be removed later on. In our numerical experiments, the vertex enumeration variant turned out to be quite efficient: the number of redundant constraints it produces is about the same as the number of non-redundant ones; for our earlier procedure, this quickly grew beyond a difference of an order of magnitude. However, the results for Procedures B4 and C5 were not as good: the M3-solver at our disposal (Löhne 2012) could not deal in reasonable time with sets of gambles that the enumeration-based procedures digested almost instantly (its author explained that it was not designed for large objective vectors). Procedures B5 and C6 could not be tested due to an apparent lack of publicly available M4-solvers.

4.3 Appendix: the NNZM Assumption & Coherence

Given a general set of gambles $\mathcal{K}$, let $\bar{\mathcal{K}}$ be the subset of constant gambles and $\check{\mathcal{K}}$ the subset of non-constant gambles. Let $\bar{b}$ be the vector with the values of the constant gambles and $\hat{\mathcal{K}}$ an NNZM set of gambles associated with $\check{\mathcal{K}}$. The restrictions of a lower prevision $\underline{P}$ on $\bar{\mathcal{K}} \cup \check{\mathcal{K}} \cup \hat{\mathcal{K}}$ to these sets are $\bar{\underline{P}}$, $\check{\underline{P}}$, and $\hat{\underline{P}}$. (Properties of coherent lower previsions used here can be found in Walley 1991, §2.6.1(b),(c).)

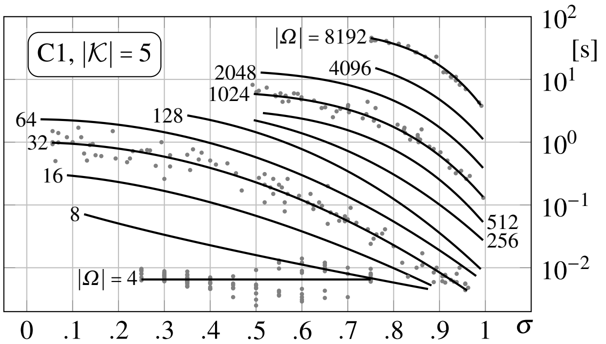
If $\underline{P}$ is coherent, we know that $\underline{P}\beta = \beta$ for any constant gamble β and so the constraints are $\bar{\underline{P}} = \bar{b}$. For any other gamble f in $\mathcal{K}$ we have the linking constraint $\check{\underline{P}}f - \hat{\underline{P}}(f - \min f) = \min f$. Fix $\hat{\mathcal{K}} := \{f - \min f : f \in \check{\mathcal{K}}\}$; this set is NNZM. Let $\hat{A}\hat{\underline{P}} \leq \hat{b}$ be the constraints for the polytope of coherent lower previsions $\hat{\underline{P}}$ on $\hat{\mathcal{K}}$, then, using the linking constraints, the corresponding constraints for $\check{\underline{P}}$ on $\check{\mathcal{K}}$ are $\hat{A}\check{\underline{P}} \leq \hat{b} + \hat{A}\min$. So the full H-representation of the set of coherent lower previsions $[\bar{\underline{P}};\check{\underline{P}}]$ on $\mathcal{K}$ is

$$\begin{bmatrix} A_{\mathcal{K}} & b_{\mathcal{K}} \end{bmatrix} := \begin{bmatrix} \mathbb{I} & & \bar{b} \\ -\mathbb{I} & & -\bar{b} \\ & \hat{A} & \hat{b}+\hat{A}\min \end{bmatrix}. \tag{23}$$

4.4 Quantitative Results of Numerical Experiments

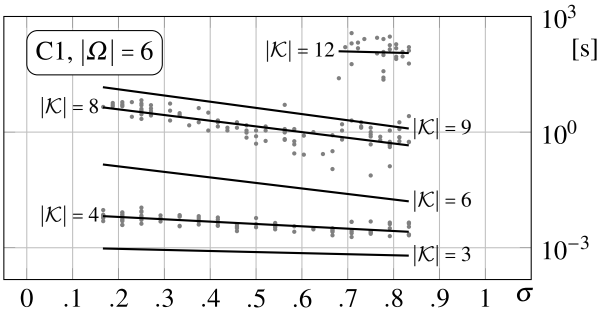
Above, we have already mentioned some qualitative evaluations and comparisons of the different procedures. Here we present more quantitative results. Our CPU-bound numerical (floating point) experiments were run on an Intel i7-2620M processor. (The Python scripts we developed are publicly available: Quaeghebeur, *pycohconstraints*.)

Our experiments showed that the *sparsity* σ, i.e., the fraction of zero components in the gamble matrix K, has an important influence on the running times of our procedures. The graph below indicates that the running time of Procedure C1 decreases exponentially as a function of the sparsity. The approximate equidistance of the curves of doubling possibility space cardinality $|\Omega|$ indicates that the running time increases approximately linearly as a function of $|\Omega|$. The curves are least-squares fits to the data points obtained from randomly generated NNZM gamble sets with values taken from $\{0,\ldots,9\}$. To give an idea of the variance, we have also plotted the data points for $|\Omega|$ in $\{4,32,1024,8192\}$ as gray dots.



The same gamble sets were also processed using Procedure C4; the running times were typically 1.5 times, but sometimes 4 times longer. The other procedures were orders of magnitude too slow for reliable testing.

In the graph below, the approximate equidistance of the lines for $|\mathcal{K}|$ in $\{3,6,9\}$ and for $|\mathcal{K}|$ in $\{4,8,12\}$, respectively, indicates that the running time of Procedure C1 increases (at least) exponentially as a function of $|\mathcal{K}|$. Again to give an idea of the variance, we have plotted the data points for $|\mathcal{K}|$ in $\{4,8,12\}$ as gray dots.



5 Correcting Incoherent Lower Previsions

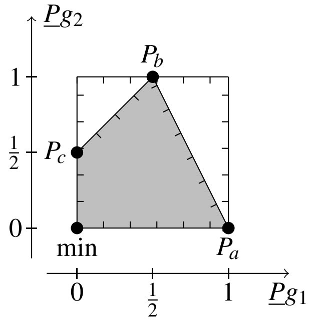
Now that we have procedures for obtaining minimal linear constraint characterizations for lower previsions that avoid sure loss or are coherent, we are ready to look at what lies beyond: sure loss and other forms of incoherence.

Automatic methods for learning lower previsions from data ideally produce coherent lower prevision, but some may not—possibly for good reasons. Also, when eliciting lower previsions from experts—but not in imprecise probability theory—, it is not reasonable to expect the result to be coherent or perhaps even avoid sure loss. For incoherent, but sure loss avoiding lower previsions, we can apply natural extension to perform a pointwise upward correction that makes explicit all implicit commitments. This is appropriate when the user of the automatic method or the elicitee provide informed consent. Otherwise a conservative, downward correction may be more acceptable.

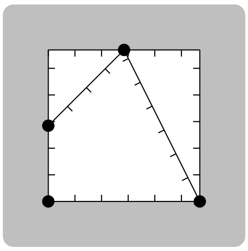
Downward changes of a lower prevision imply a reduction in both explicit and implicit commitments. When it is not possible to decide on the changes with input from the user or the elicitee, automatic downward correction methods are an option, after informed consent. We here propose one such automatic downward correction method.

5.1 Forms of Incoherence

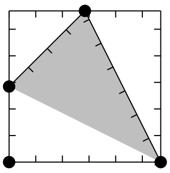
Let us briefly give a categorization of the possible forms of incoherence. To this end, consider a two-gamble example on a possibility space $\{a,b,c\}$: consider the set $\mathcal{K} := \{g_1, g_2\}$, with $g_1 := [1; 1/2; 0]$ and $g_2 := [0; 1; 1/2]$. Using a procedures from Section 4, we have obtained the constraints, drawn using bestubbled lines, delimiting the shaded convex polytope of coherent lower previsions. Its vertices have been named: the vacuous lower prevision min and for every atom ω in $\{a,b,c\}$ the degenerate prevision $P_\omega := [g_1\omega; g_2\omega]$, the columns of $K^\top$.



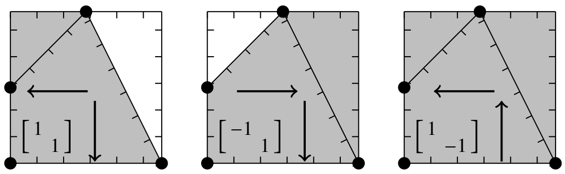
We recalled at the end of Section 3 that coherent lower previsions $\underline{P}$ are bounded, i.e., that $\min \le \underline{P} \le \max$. Our first category of incoherent previsions are those that are out of bounds. On the right, we shaded the magnitude-wise smallest part of this unbounded region in gray.



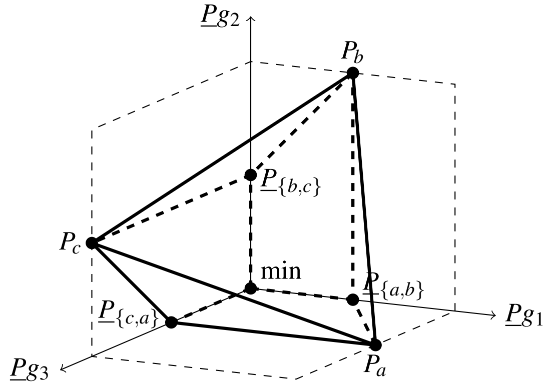
Equations (3)–(4) and (7)–(8) showed us that the convex set of linear previsions can take a central role in both the definitions of avoiding sure loss and coherence. For our example, it is in gray on the right.



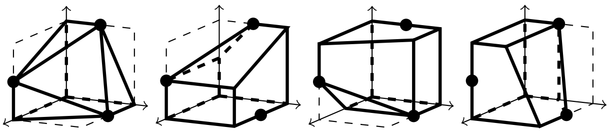
More concretely, Equation (8) made it clear that the polytope of coherent lower previsions is an intersection of polyhedra corresponding to avoiding sure S-loss—i.e., S-dominance by a linear prevision—, one for each S in $\mathcal{S}$. Below, we show, in gray, the part of these polyhedra within bounds, accompanied by their respective S-matrix and the extreme rays of the dominance cone it implies. With each S there corresponds a set whose members incur sure S-loss. The set of incoherent lower previsions is their union.


$\begin{bmatrix} 1 & \\ & 1 \end{bmatrix}$
$\begin{bmatrix} -1 & \\ & 1 \end{bmatrix}$
$\begin{bmatrix} 1 & \\ & -1 \end{bmatrix}$

To get a feel for what constellations can occur when faced with larger sets of gambles, we extend our two-gamble example with a gamble $g_3 := [1/2; 0; 1]$. Below, we give the polytope of coherent lower previsions. It is bounded by the cuboid defined by the min and max points. Its edges in the coordinate planes are shown using thin dashed lines. The new vertices can be characterized for g in $\{g_1, g_2, g_3\}$ by $\underline{P}_A g := \min_{\omega \in A} g\omega$. The range of values attained by the vertex lower previsions is $\{0, 1/2, 1\}$.


$\underline{P}g_2$
P_b
$\underline{P}_{\{b,c\}}$
P_c
min
$\underline{P}_{\{a,b\}}$
$\underline{P}g_1$
$\underline{P}_{\{c,a\}}$
$\underline{P}g_3$
P_a

Below, we furthermore give the $|\mathcal{S}| = |\mathcal{K}| + 1 = 4$ sets of lower previsions that avoid sure S-loss.



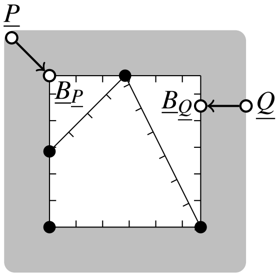
This illustration shows that some lower previsions within bounds may incur sure S-loss for all S; max, for example.

5.2 Bringing Lower Previsions Within Bounds

Correcting a lower prevision $\underline{P}$ that is out of bounds to one that is within bounds is trivial: We replace it by the pointwise closest such lower prevision $\underline{B_P}$, so for every gamble f in $\mathcal{K}$ we have

$$\underline{B_P} f := \begin{cases} \min f & \underline{P}f \le \min f, \\ \max f & \underline{P}f \ge \max f, \\ \underline{P}f & \text{otherwise.} \end{cases} \qquad (24)$$

This correction method may produce both downward and upward pointwise changes.



From now on we assume that all lower previsions are within bounds.

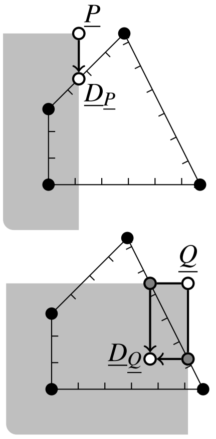
5.3 Maximal Dominated Coherent Lower Previsions

Our proposal for the downward correction of an incoherent lower prevision $\underline{P}$ is the lower envelope of the maximal coherent lower previsions dominated by $\underline{P}$. In other words, it is the nadir point $\underline{D}_P$ of the MOLP (cf. Section 2.2):

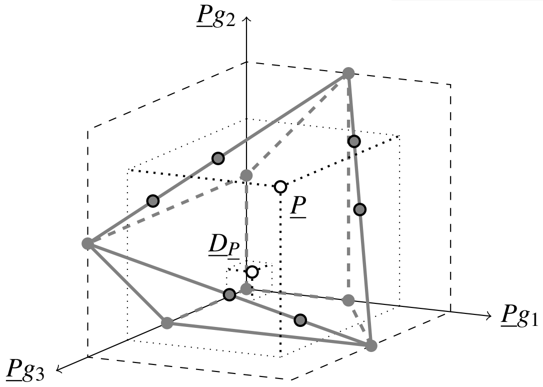
$$\begin{aligned} &\text{maximize } \underline{Q}, \\ &\text{subject to } \Lambda_{\mathrm{C}}\underline{Q} \leq \alpha_{\mathrm{C}} \text{ and } \underline{Q} \leq \underline{P}. \end{aligned} \tag{25}$$

This proposal is essentially the same as the specific so-called *prudential correction* $\overline{P}_H$ mentioned by Pelessoni & Vicig (2003, §3.4). They generalize the interval-probability concept *F-Hülle* (see Weichselberger 2001, 342ff. and 375ff.; translated as *F-cover* in Weichselberger 2000). However, they only aim to apply this correction when sure loss is avoided; we make no such restriction.

On the right, the method is illustrated for two incoherent lower previsions that are within bounds; extreme maximal dominated coherent lower previsions are shown as gray-filled dots.



We should not conclude from these illustrations that the extreme maximal coherent lower previsions dominated by the given incoherent lower prevision can always be reached by reducing single components; a graphical counterexample is given below.



The lower prevision $\underline{D}_P$ satisfies the necessary requirements:

i. It is a downward correction as a lower envelope of lower previsions dominated by $\underline{P}$.
ii. It is coherent by the Lower Envelope Theorem.

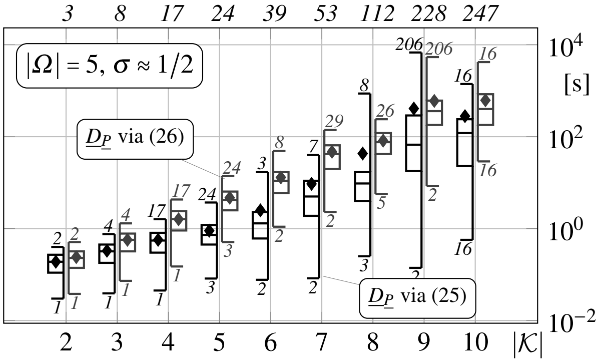
Furthermore, as a nadir point it has a number of further desirable properties:

iii. The correction it embodies is neutral in the sense that no tradeoff between corrections for the different components of $\underline{P}$ is made; this makes it especially suited for unguided corrections.
iv. It is the maximal such neutral correction—the vacuous lower prevision min is another—and therefore preserves as much of the commitments expressed by $\underline{P}$ as possible.
v. The set of coherent lower previsions dominated by an incoherent lower prevision $\underline{P}$ is non-decreasing with pointwise increasing $\underline{P}$. So the more incoherent a lower prevision, the more imprecise its correction.

It is actually not necessary to calculate $[\Lambda_{\mathrm{C}}, \alpha_{\mathrm{C}}]$ in order to find $\underline{D}_P$, because we have a full constraint based characterization of coherence with the H-representation (12). So an alternative to the MOLP (25) is the following MOLP:

$$\begin{aligned} &\text{maximize } \underline{Q}, \\ &\text{subject to } A_{\underline{Q}}\underline{Q} + A_\mu \mu \leq b \text{ and } \underline{Q} \leq \underline{P}, \end{aligned} \tag{26}$$

where we use the notation of Equation (12). (Weichselberger 2001, 468ff, also proposes an as of yet untested algorithm that is essentially based on a representation such as the one given by Equation (12).) This problem has $(|\mathcal{K}|+1)\cdot|\Omega|$ more variables than the MOLP (25), which has $|\mathcal{K}|$ variables. It has $(|\mathcal{K}|+1)\cdot(|\mathcal{K}|+|\Omega|+2)$ constraints, whereas the MOLP (25) typically has of the order of $3\cdot|\mathcal{K}|$ constraints. This results in a greater average running time for the nadir computation using the alternative MOLP, even if we take the setup time—calculating $[\Lambda_{\mathrm{C}}, \alpha_{\mathrm{C}}]$ (cf. Section 4.4) versus generating $[A_{\underline{Q}}, A_\mu, b]$ (about 10^{-3}s)—into account. This can be seen in the graphical summary of the results of our numerical experiments, which we are going to describe next. (The Octave/Matlab scripts we developed are publicly available: Quaeghebeur, *mcohconstraints*.)



In this experiment, for each value of $|\mathcal{K}|$ in $\{2,\ldots,10\}$, we generated about 10 NNZM gamble sets $\mathcal{K}$—as in Section 4.4—with sparsity σ fixed at approximately $1/2$, on a possibility space Ω with $|\Omega| = 5$. Next, we calculated the corresponding $[\Lambda_{\mathrm{C}}, \alpha_{\mathrm{C}}]$—using Procedure C1—and generated the corresponding $[A_{\underline{Q}}, A_\mu, b]$. Finally, for each $\mathcal{K}$, we generated about 10 incoherent lower previsions within bounds to correct. This we did using both the MOLP (25) and the MOLP (26), resulting in about 100 computation time samples per $|\mathcal{K}|$ for each of both approaches. Each of these sample sets is summarized using a box plot indicating minimum, lower quartile, median, upper quartile, and maximum; its arithmetic mean is indicated with a lozenge. Black left-leaning box plots are used for the results obtained with the MOLP (25); darkgray right-leaning ones for those obtained with the MOLP (26).

With the M3-solver we used (Löhne 2012), average computation time seems to increase exponentially as a function of $|\mathcal{K}|$. Surprisingly, the number of extreme maximal dominated coherent lower previsions is not a major factor. This is illustrated by the number of these extreme points found for the minimum and maximum computation times—put in italics near the respective box plot whiskers—and the maximum number of extreme points in the sample—listed in italics at the top edge of the plot axis.

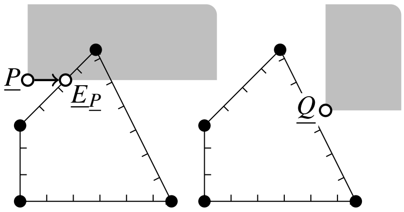
The M3-solver does compute all these extreme points, so we suspect that it is highly inefficient for the task at hand. Therefore we believe substantial efficiency gains can be achieved by switching to an M4-solver, which we expect to be *output sensitive*, i.e., to depend on the number of extreme points. Nadir point calculation algorithms that do not need to calculate all these extreme points (e.g., Alves & Costa 2009) should provide a further increase in efficiency. Because elicited lower previsions can be expected to generally be closer to coherent than our randomly generated ones, we also expect them to generally dominate less extreme points and thus, because of output sensitivity, be faster to correct. We already observed this phenomenon for randomly generated sure loss avoiding lower previsions.

5.4 Least Dominating Coherent Lower Prevision

For completeness's sake, let us also have a look at upward correction using the MOLP approach. Given an incoherent lower prevision $\underline{P}$, we consider the set of minimal pointwise dominating coherent lower previsions; this is the solution to the following MOLP:

$$\begin{aligned} &\text{minimize } \underline{E}_{\underline{P}}, \\ &\text{subject to } \Lambda_{\mathrm{C}}\underline{E}_{\underline{P}} \leq \alpha_{\mathrm{C}} \text{ and } \underline{E}_{\underline{P}} \geq \underline{P}. \end{aligned} \tag{27}$$

Because of the Lower Envelope Theorem, there is only one such $\underline{E}_{\underline{P}}$, so we may replace this vector objective by the scalar objective $\sum_{g\in\mathcal{K}} \underline{E}_{\underline{P}}g$, reducing the problem to a plain LP. This coherent lower prevision $\underline{E}_{\underline{P}}$ is the one least dominating $\underline{P}$, to wit, its natural extension (cf. Walley 1991, §3.1). This plain LP method for obtaining it is illustrated on the right.



Again, we can use the H-representation (12) to formulate an alternative to the MOLP (27):

$$\begin{aligned} &\text{minimize } \underline{E}_{\underline{P}}, \\ &\text{subject to } A_{\underline{E}_{\underline{P}}}\underline{E}_{\underline{P}} + A_\mu \mu \leq b \text{ and } \underline{E}_{\underline{P}} \geq \underline{P}. \end{aligned} \tag{28}$$

Thanks to the block structure of the constraint matrix, it is straightforward to deduce some well-known facts:

i. It is necessary that $\underline{P}$ avoids sure loss for a solution $\underline{E}_{\underline{P}}$ to exist (cf. right-hand side illustration above).
ii. For each gamble g in $\mathcal{K}$, we can calculate the corresponding natural extension component $\underline{E}_{\underline{P}}g$ separately as $\max\{g^\top\mu : \underline{P} \leq K^\top\mu \wedge \mu \geq 0 \wedge 1^\top\mu = 1\}$.

These facts raise the currently still open question of whether there exist specific classes of incoherent lower previsions $\underline{P}$ for which the calculation of $\underline{D}_{\underline{P}}$ can be simplified, e.g., to separate calculations for each component.

6 Conclusions

We hope that you are now convinced of the fact that the availability of a finite, minimal linear constraints characterization of coherence opens doors for many new numerical applications dealing with the set of coherent lower previsions. In our application, downward correction of incoherent lower previsions, we saw that it proved useful to keep the running time of the inherently computationally complex implementation of our proposed method a bit in check. We determined that currently, sets of up to 5 gambles can be dealt with sufficiently fast even for interactive applications. In a domain where complex systems are often decomposed into smaller ones linked in some network structure, this is not overly restrictive.

We also hope that this paper has kindled your interest in the application of multi-objective linear programming to imprecise probability problems. We believe that beyond the two applications of them presented in this paper, there are bound to be more in our research field because of the common underlying assumption that incomparability should be modeled, not avoided.

There are some unfinished strands in this paper:

i. Testing an efficient projection implementation (cf. Kvasnica et al. 2006).
ii. Finding and testing a MOLP simplex solver (cf. M4) and a nadir computation algorithm (cf. M2).
iii. Theoretically investigate whether $\underline{D}_{\underline{P}}$ can be calculated more efficiently if $\underline{P}$ satisfies some additional conditions beyond being within bounds.

We hope these are picked up by us, or others, in the future.

Acknowledgments

Thanks to the reviewers for pointing out very relevant literature. Thanks to J. De Bock for providing me with extracts from Weichselberger's 2001 tome and Th. Augustin for helping me interpret them correctly. Thanks to G. de Cooman and N. Huntley for sharing their thoughts about the downward correction idea. Thanks to A. P. Pedersen for pointers to potentially relevant belief revision texts, which, I am sorry to say, I could not yet follow up on.

References

Alves M.J. & Costa J.P. (2009). An exact method for computing the nadir values in multiple objective linear programming. *European Journal of Operational Research* 198.2, 637–646. DOI: 10.1016/j.ejor.2008.10.003.

Avis D. & Fukuda K. (1992). A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. *Discrete & Computational Geometry* 8.1, 295–313. DOI: 10.1007/BF02293050. URL: http://www.digizeitschriften.de/dms/img/?PPN=GDZPPN000365548.

Benson H.P. (1998). An outer approximation algorithm for generating all efficient extreme points in the outcome set of a multiple objective linear programming problem. *Journal of Global Optimization* 13.1, 1–24. DOI: 10.1023/A:1008215702611.

Bertsimas D. & Tsitsiklis J.N. (1997). *Introduction to linear optimization*. Athena Scientific.

Biazzo V. & Gilio A. (2000). A generalization of the fundamental theorem of de Finetti for imprecise conditional probability assessments. *International Journal of Approximate Reasoning* 24.2–3, 251–272. DOI: 10.1016/S0888-613X(00)00038-4.

Clarkson K.L. (Nov. 1994). More output-sensitive geometric algorithms. *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, 695–702. DOI: 10.1109/SFCS.1994.365723.

Ehrgott M. (2005). *Multicriteria optimization*. 2nd ed. Springer.

Ehrgott M. & Tenfelde-Podehl D. (2003). Computation of ideal and nadir values and implications for their use in MCDM methods. *European Journal of Operational Research* 151.1, 119–139. DOI: 10.1016/S0377-2217(02)00595-7.

Ehrgott M., Löhne A. & Shao L. (2012). A dual variant of Benson's "outer approximation algorithm" for multiple objective linear programming. *Journal of Global Optimization* 52.4, 757–778. DOI: 10.1007/s10898-011-9709-y.

Evans J.P. & Steuer R.E. (1973). A revised simplex method for linear multiple objective programs. *Mathematical Programming* 5.1, 54–72. DOI: 10.1007/BF01580111.

Fukuda K. (2004). *Frequently asked questions in polyhedral computation*. URL: http://www.ifor.math.ethz.ch/~fukuda/polyfaq.

Fukuda K. & Prodon A. (1996). Double description method revisited. *Combinatorics and Computer Science*. Ed. by Deza, Euler & Manoussakis. Vol. 1120. Lecture Notes in Computer Science. Springer-Verlag, 91–111. DOI: 10.1007/3-540-61576-8_77. URL: http://www.ifor.math.ethz.ch/~fukuda/cdd_home.

Grünbaum B. (1967). *Convex polytopes*. London: Interscience Publishers.

Isermann H. (1977). The enumeration of the set of all efficient solutions for a linear multiple objective program. *Operational Research Quarterly* 28.3, 711–725. JSTOR: 3008921.

Jones C.N., Kerrigan E.C. & Maciejowski J.M. (June 2004). *Equality Set Projection: A new algorithm for the projection of polytopes in halfspace representation*. Tech. rep. CUED/F-INFENG/TR.463. Department of Engineering, University of Cambridge. URL: http://www-control.eng.cam.ac.uk/~cnj22/docs/resp_mar_04_15.pdf.

Kvasnica M., Grieder P. & Baotić M. (2006). *Multi-Parametric Toolbox (MPT), version 2.6.3*. URL: http://control.ee.ethz.ch/~mpt.

Löhne A. (Nov. 2012). *bensolve, version 1.2*. URL: http://ito.mathematik.uni-halle.de/~loehne/index_en_dl.php.

Miranda E. (2008). A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning* 48.2, 628–658. DOI: 10.1016/j.ijar.2007.12.001.

Pelessoni R. & Vicig P. (2003). Imprecise previsions for risk measurement. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 11.4, 393–412. DOI: 10.1142/S0218488503002156.

Quaeghebeur E. *mcohconstraints: Matlab/Octave functions for generating coherence and avoiding sure loss constraints for lower previsions*. URL: http://github.com/equaeghe/mcohconstraints.

– *pycohconstraints: Python code for generating coherence constraints for lower previsions*. URL: http://github.com/equaeghe/pycohconstraints.

– (2009). Learning from samples using coherent lower previsions. PhD thesis. Ghent University. HDL: 1854/LU-495650.

– (2010). Characterizing the set of coherent lower previsions with a finite number of constraints or vertices. *UAI-10: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. Ed. by Spirtes & Grünwald. AUAI Press, 466–473. HDL: 1854/LU-984156.

Quaeghebeur E. & De Cooman G. (Sept. 2008). Extreme lower probabilities. *Fuzzy Sets and Systems* 159.16, 2163–2175. DOI: 10.1016/j.fss.2007.11.020. HDL: 1854/LU-429244.

Strijbosch L.W., van Doorne A.G. & Selen W.J. (1991). A simplified MOLP algorithm: The MOLP-S procedure. *Computers & Operations Research* 18.8, 709–716. DOI: 10.1016/0305-0548(91)90008-F.

Vicig P. (1996). An algorithm for imprecise conditional probability assessments in expert systems. *IPMU '96: Proceedings of the Sixth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. (Granada, Spain), 61–66.

Walley P. (1991). *Statistical reasoning with imprecise probabilities*. Vol. 42. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Walley P., Pelessoni R. & Vicig P. (2004). Direct algorithms for checking consistency and making inferences from conditional probability assessments. *Journal of Statistical Planning and Inference* 126.1, 119–151. DOI: 10.1016/j.jspi.2003.09.005.

Weichselberger K. (2000). The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning* 24.2–3, 149–170. DOI: 10.1016/S0888-613X(00)00032-3.

– (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als Umfassendes Konzept*. Heidelberg: Physica-Verlag.

Yu P.L. & Zeleny M. (1975). The set of all nondominated solutions in linear cases and a multicriteria simplex method. *Journal of Mathematical Analysis and Applications* 49.2, 430–468. DOI: 10.1016/0022-247X(75)90189-4.

Ziegler G.M. (1995). *Lectures on polytopes*. Springer.

On Sharp Identification Regions for Regression Under Interval Data

Georg Schollmeyer
Department of Statistics, LMU Munich
georg.schollmeyer@stat.uni-muenchen.de

Thomas Augustin
Department of Statistics, LMU Munich
thomas.augustin@stat.uni-muenchen.de

Abstract

The reliable analysis of interval data (coarsened data) is one of the most promising applications of imprecise probabilities in statistics. If one refrains from making untestable, and often materially unjustified, strong assumptions on the coarsening process, then the empirical distribution of the data is imprecise, and statistical models are, in Manski's terms, partially identified. We first elaborate some subtle differences between two natural ways of handling interval data in the dependent variable of regression models, distinguishing between two different types of identification regions, called *Sharp Marrow Region (SMR)* and *Sharp Collection Region (SCR)* here. Focusing on the case of linear regression analysis, we then derive some fundamental geometrical properties of SMR and SCR, allowing a comparison of the regions and providing some guidelines for their canonical construction. Relying on the algebraic framework of adjunctions of two mappings between partially ordered sets, we characterize SMR as a right adjoint and as the monotone kernel of a criterion function based mapping, while SCR is indeed interpretable as the corresponding monotone hull. Finally we sketch some ideas on a compromise between SMR and SCR based on a set-domained loss function.

Keywords. partial identification, imprecise probabilities, interval data, sharp identification regions, coarse data, adjunctions, partially ordered sets, linear regression model, best linear predictor, set-domained loss function.

1 Introduction

The methodology of imprecise probabilities offers powerful methods for reliable handling of *coarse(ned) data*, see, e.g., the ISIPTA contributions by [18, 45, 41, 38, 42, 7, 20]. The term coarsened data, or epistemic data imprecision, is an umbrella term, comprising all situations where data are not observed in the resolution intended in the subject matter context. This means, there is a certain true precise value $y \in \mathcal{Y}$ of a generic variable Y of material interest, but we only observe a set $A \supseteq \{y\}$. An extreme special case of coarse data are *missing data*, where the missingness of value y_i of unit i can be interpreted as having observed the whole sample space $\mathcal{Y}$. In the case where A is an interval $[\underline{\mathrm{y}}, \overline{\mathrm{y}}]$ for $\underline{\mathrm{y}}, \overline{\mathrm{y}} \in \mathbb{R}$ coarse data are commonly called *interval data.*

Before turning to the formal framework, two issues with fundamental importance for practical applications shall be recalled.
First of all, it must be stressed that the term 'coarse' is a relative term. Whether data are coarse or not depends on the specified sample space, and therefore on the subject matter context to be investigated. If, for instance, the sample space is taken to consist of some a priori specified ranges for income data, and that is all what is needed, then data are not coarse, while if precise income values are of interest, the data are coarse.[1]
Secondly, it is important to emphasize that coarse data typically are not just the result of sloppy research, like an insufficient study design or improper data handling. On the contrary, coarse data are an integral part of data collection, in particular in social surveys. Interval data arise naturally from the use of categories in order to avoid refusals in the case of sensitive questions, and are a means to model roughly rounded responses (see, e.g., [24]). Coarsened categorical data are, for instance, produced by matching data sets with not fully overlapping categories, are the direct outcome of data protection by some anonymization techniques (see, e.g., [13]), or may be produced

[1] Indeed, even unions of intervals may constitute precise observations, for instance as the response to the question 'When did you live in Munich?', measured in years. Then $\{[1986; 1991] \cup [1997; 2000]\}$ is a precise observation in the sample space of all finite unions of closed intervals $[a, b]$ with $a, b \in \mathbb{N}_+$. (See in particular the distinction between *conjunctive* and *disjunctive* random sets in [14, Section 1.4], from which also this example is adopted.)

by the combination of spaces with given marginals by Frechèt bounds (see, e.g., [19]). Another prototypic setting is the case of systematically missing data, arising from treatment evaluations in non-randomized designs like observational studies.[2]

By confining themselves to precise probabilities, traditional statistical approaches to cope with coarse data are inevitably forced to try to escape the imprecision in the data eventually. An immediate way in the case of interval data $[\underline{y}_i, \overline{y}_i]$ for each unit $i = 1, \ldots, n$ in the sample is to replace each interval by the corresponding central value $\mathring{y}_i = (\underline{y}_i + \overline{y}_i)/2$, and then to proceed with a standard analysis based on that fictitious sample. More sophisticated approaches add complex, typically untestable assumptions, either to explicitly model the coarsening process by a precise model, or to characterize idealized situations where the coarsening can be included in standard likelihood and Bayesian inference without biasing the analysis systematically.[3]

In recent years, awareness in statistics and econometrics has grown that such strong assumptions quite often cannot be justified by substantive arguments, and thus the – too high – price for the seemingly precise result of the statistical analysis is the loss of credibility of the conclusions, and in the end consequentially the practical relevance of the statistical analysis.[4] In the light of this, it is of particular importance to develop approaches that reflect the underlying imprecision in the data properly, resulting in potentially imprecise, but reliable results. The fascinating insight, corroborated by a variety of applications mainly in econometrics (see the exemplary references below), is that in many studies these results are still enough to answer important substantive science questions, and if not, the scientist is alerted that strong conclusions drawn from the data may be mere artefacts.

Related approaches, considering all possible data compatible with the observed set of values, have been developed almost independently in different settings, ranging from reliable computing and interval analysis in engineering (e.g., [29]) and extensions of generalized Bayesian inference [10, 46] to reliable descriptive statistics in social sciences ([32, Chapter 17f], [30]). This cautious way to proceed is closely related to set-based (profile-)likelihood approaches ([48, 7]) and to the methodology of partial identification, in particular propagated by Manski (e.g., [23]) in econometrics, and to systematic sensitivity analysis (e.g. [43]) in biometrics, where a general framework for imprecise data models, i.e. sets of observationally equivalent statistical models, has been developed. In these models instead of single valued parameters one obtains so-called **identification regions**, i.e. sets of all parameters compatible with the data. On the inferential side, there has been important progress in the development of appropriate confidence procedures (see, e.g., [5, 27, 6]), and computational techniques have maturated to the extent that routine use of basic procedures has become feasible (e.g., [8, 1, 39, 34]). As a result, applied contributions are now rather common and are particularly influential in econometrics and allied fields see, e.g., [28] for an analysis of income poverty measures based on coarsened survey data, [21] for a study of the German reform of unemployment compensation based on register data and [26] for an analysis of treatment effects in observational studies with an illustration based on the National Longitudinal Survey of Youth.

The paper is organized as follows. After some basic definitions (Section 2), we emphasize in Section 3 the distinction between different understandings and goals of regression models, leading to two different types of identification regions, called SMR and SCR here. Section 4 formulates some basic geometrical properties, while sections 5 applies an algebraic framework for investigating mappings between partially ordered sets. We recall the basic concepts needed here, and explain them exemplary in the context of Dempster-Shafer-Theory and by describing coherent lower previsions as hulls. Then SMR and SCR are characterized as the monotone kernel and monotone hull of a criterion function based mapping, respectively. Finally, Section 6 suggests another type of identification regions that is based on a strict set-valued perspective, relying on a loss function depending on sets of parameters, while Section 7 concludes. Proofs for the propositions and additional illustrations of the different identification regions can be found in a homonymous technical report ([33]), that is going to appear soon.

2 Basic Definitions

Let Θ be a parameter space and $P := \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ a corresponding statistical model on a measurable space $(\Omega, \mathcal{F})$ with the associated observable random variables $X, \underline{Y}, \overline{Y}$ and the unobserved random variable Y.

[2] To evaluate effects of treatment or intervention A over treatment B, in principle, it would be necessary to have information from a parallel universe, so-to-say, i.e. to know in addition how the units treated with A *would* have reacted *if* they had been given the treatment B and vice versa.
This question has in particular attracted intensive attention in the partial identification literature in econometrics (see, for instance, the survey [40] or the instructive case study [37].

[3] Most prominent is here Little and Rubin's [22] classification, distinguishing situations of *missingness completely at random (MCAR)* or *missing at random (MAR)* from *missing not at random (MNAR)* settings, where a systematic bias has to be expected. This classification has been extended to coarsening by [16].

[4] See Manski's Law of Decreasing Credibility [23, p. 1].

We are interested in the relationship between X and Y, but we have no full information about Y, we only know that the unobserved variable Y is related to the observed $\underline{Y}$ and $\overline{Y}$ in the sense that Y fulfills a certain relation, for example $\mathbb{P}(\underline{Y} \leq Y \leq \overline{Y}) = 1$ or $\mathbb{E}(\underline{Y} \mid X) \leq \mathbb{E}(Y \mid X) \leq \mathbb{E}(\overline{Y} \mid X)$[5]. In the sequel, we assume the second condition with the additional assumption that $\mathbb{E}(\underline{Y} \mid x)$ and $\mathbb{E}(\overline{Y} \mid x)$ are continuous in x. With $\mathbb{P}$ we denote the unknown true model and with $\mathbb{E}$ the corresponding expectations. The expectations for a model $\mathbb{P}_\theta$ are denoted with $\mathbb{E}_\theta$. The joint distribution of the random variables $X, Y, \underline{Y}, \overline{Y}$ under a model P_θ is denoted with $F_\theta^{X,Y,\underline{Y},\overline{Y}}$ (or short F_θ) and the joint distribution under the true model $\mathbb{P}$ is denoted with $F^{X,Y,\underline{Y},\overline{Y}}$ (or short F). Analogously, the distribution of a subset of random variables, eg. $\{X, Y\}$ is denoted with $F_\theta^{X,Y}$ and $F^{X,Y}$ respectively. For arbitrary random variables like e.g. $X, Z, \underline{Y}, \overline{Y}$ we denote their joint distribution with $F^{X,Z,\underline{Y},\overline{Y}}$. Because Y is not observable, we do not have the full information about Y, which generally leads to partially identified models, which we define in the sequel: Two parameters $\theta_1, \theta_2 \in \Theta$ are undistinguishable (i.e. $\theta_1 \sim \theta_2$) if the corresponding models $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ are empirically undistinguishable, which means that the distributions of the observable variables are the same. A statistical model P is called **point-identified**, if any two different parameters θ_1 and θ_2 are empirically distinguishable. Otherwise it is called **partially identified**.

Example 1 *The simple linear model with interval outcomes:* $\Theta = B \times R$ *with* $B = \mathbb{R}^2$ *the actually interesting parameter space and* $R = \mathbb{R}^\Omega \times \mathbb{R}^\Omega_{\geq 0} \times \mathbb{R}^\Omega_{\geq 0}$ *describing the error-terms and the coarsening-process: For* $\theta = (\beta, (\varepsilon, \delta_l, \delta_u)) \in \Theta$ *the associated variables are defined as* $Y = X\beta + \varepsilon$, $\quad \underline{Y} = X\beta + \varepsilon - \delta_l$ *and* $\overline{Y} = X\beta + \varepsilon + \delta_u$ *with* $\varepsilon, \delta_l, \delta_u$ *measurable and* ε *with existing conditional expectations* $\mathbb{E}(\varepsilon \mid x) = 0$. *The coarsening process is modeled by the random variables* δ_l *and* δ_u *that are nonnegative, which ensures* $\underline{Y} \leq Y \leq \overline{Y}$. *By abuse of notation we identify the random variable* X *with the matrix* $(1, X)$ *to use matrix notations like above, if useful. Furthermore, in the sequel we assume* X *as a fixed random variable with support* $\mathbb{R}$ *and therefore omit it in the parameter space* Θ. *It is clear that this model is only partially identified. For example* $((\beta_0, \beta_1), (\varepsilon, 0, 1)) \sim ((\beta_0 + 1, \beta_1), (\varepsilon, 1, 0))$. *Moreover, the quotient space* $\Theta_{/\sim}$ *is not of the form* $\Theta_{/\sim} = B_{/\sim_B} \times R_{/\sim_R}$ *for some relations* $\sim_B$ *and* $\sim_R$, *so we must factorize the whole space* Θ *and not only the interesting part* B *to make the model point-identified.*

3 Two Types of Identification Regions

There are two ideal type senses of what a statistical model is and what it should render. One can assume a statistical model as the exact true underlying probabilistic structure, from which one only has to know all details and then one knows the exact distributions of all involved random variables and can make inferences with this knowledge. In contrast one can see a statistical model not as a truth, but as a rough approximation of truth and use it as a parsimonious tool to predict for example future observations of some variables or to get a rough insight into the real underlying structure that is actually more complex. As examples for this differentiation one could see firstly the estimation of the intercept and the slope of a linear model and secondly the problem of finding the best linear predictor in the sense of [2], which makes predictions that are linear in the covariates, but the underlying model needs not to be linear. The main difference is here that in the first case we really assume a linear model and rely on it, whereas in the second case we use the linearity of the predictions only to have a parsimonious model for predictions or explanations, but we assume nothing about the true statistical model. These views lead to different problem formulations, which we want to state now as we need it in our context. In order to efficiently tackle our goal, we leave the statistical perspective and join Manski ([23, p. 7]), who recommends that problems of identification become much clearer when one firstly separates non-identifiability from sample variation, and assumes all distributions to be known for the analytic treatment[6] (later on then sample counterparts may be constructed in the usual way). In particular, we also assume that the distribution of Y is known (and we have no variables $\underline{Y}$ and $\overline{Y}$) and later we generalize this to the case of an unobserved Y, which leads to different sharp identification regions that are then our objects of interest. The first problem statement is:

Given the distribution $F^{X,Y}$ of (X, Y), which is an element of the class $\{F_\theta^{X,Y} \mid \theta \in \Theta\}$, find all θ, such that $(X, Y) \sim F_\theta^{X,Y}$, which is equivalent to find all θ with $L(F_\theta^{X,Y}, F^{X,Y}) = 0$ for an arbitrary distance-function $L(\cdot, \cdot)$ or a similar function, which is zero if and only if both arguments are equal. Here we think of a kind of loss function and introduce this equivalent formulation to indicate the analogy to the second problem formulation:

Given the distribution $F^{X,Y}$, which is an element of the class $\{F_\theta^{X,Y} \mid \theta \in \Theta\}$, find all θ, such that $L(F_\theta^{X,Y}, F^{X,Y})$ is minimal. In contrast to the first

[5] This means $\forall x : \mathbb{E}(\underline{Y} \mid x) \leq \mathbb{E}(Y \mid x) \leq \mathbb{E}(\overline{Y} \mid x)$.

[6] The identification regions arising in this limit case are called *sharp* identification regions.

problem, this problem definition is also meaningful if $F^{X,Y} \notin \{F_\theta^{X,Y} \mid \theta \in \Theta\}$, i.e. if the model is not correctly specified. If the model is correctly specified, then both problems are often essentially the same in the sense that for example for a linear model, the BLUE-estimator and the best linear predictor (with a quadratic loss-function) are solving different tasks, but the parameter estimates are identical. The actual problem is now that $F^{X,Y}$ is unknown. One part of the problem is that also if we could observe Y, we could not know the exact distribution of Y and so we have to estimate it. In particular, we cannot decide with certainty, if $F^{X,Y}$ is an element of the class $\{F_\theta^{X,Y} \mid \theta \in \Theta\}$, and so the two problem formulations are moving together a little bit. The other part of the problem is that the variable Y, we are actually interested in, is not observable. As argued above for the moment we only address this second part of the problem and assume that we know the exact distribution of all observable variables. Later in section 5.2 we also address the other part. If now Y is unobserved, we can generalize the two problems by applying them to all possible Y that are consistent with $(\underline{Y}, \overline{Y})$.This leads to different regions of parameters that were proposed in different papers: The region related to the first problem was introduced slightly differently in [9] and the other region was proposed as the sharp identification region for the best linear predictor in [2].

Definition 1 *Let $P = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ be a statistical model with the corresponding joint distributions $\{F_\theta^{X,Y} \mid \theta \in \Theta\}$ and $X, \underline{Y}, \overline{Y}$ given random variables. The **sharp marrow region (SMR)** is defined as:*

$$SMR = \{\theta \in \Theta \mid \mathbb{E}(\underline{Y} \mid X) \le \mathbb{E}_\theta(Y \mid X) \le \mathbb{E}(\overline{Y} \mid X)\}$$

Note that the Y in the definition is the Y coming from the model $\mathbb{P}_\theta$, not the Y from the true model. If the model is correctly specified (or if at least $SMR \neq \emptyset$), this region can also be written as:[7]

$$SMR = \operatorname*{argmin}_{\theta \in \Theta} \left[\min_{Z \in \mathbb{E}([\underline{Y}, \overline{Y}] \mid X)} L\left(F_\theta^{X,Y}, F^{X,Z}\right) \right]$$

with an arbitrary loss function L. Here with $\mathbb{E}([\underline{Y}, \overline{Y}] \mid X)$ we denote the set of all random variables Z fulfilling $\mathbb{E}(\underline{Y} \mid X) \le \mathbb{E}(Z \mid X) \le \mathbb{E}(\overline{Y} \mid X)$. This equivalent characterization of SMR is valid because a parameter $\theta \in \Theta$ is in SMR if and only if there exists a $Z \in \mathbb{E}([\underline{Y}, \overline{Y}] \mid X)$ with $F_\theta^{X,Y} = F^{X,Z}$ or equivalently $L(F_\theta^{X,Y}, F^{X,Z}) = 0$. From the above representation of SMR we can see that SMR can be written as the solution of a decision problem with a minimin decision rule.

*The **sharp collection region (SCR)** is defined as:*

$$SCR := \bigcup_{Z \in [\underline{Y}, \overline{Y}]} \operatorname*{argmin}_{\theta \in \Theta} L\left(F_\theta^{X,Y}, F^{X,Z}\right).$$

With $[\underline{Y}, \overline{Y}]$ we denote the set of all random variables Y that lie between $\underline{Y}$ and $\overline{Y}$ for all $\omega \in \Omega$.

A first comparison of this two regions that emphasizes the case of misspecification and interpretational problems for the sharp marrow region in this case can be found in [31]: While the interpretation of the sharp collection region as the collection of all best linear predictors is clear, the interpretation of SMR seems to be not so useful under misspecification, especially if SMR is empty.[8] From an empty SMR we can conclude, that the model is misspecified, but not more. Furthermore in the above-mentioned paper the authors make clear that a tight SMR "cannot be viewed as an indicator that the underlying model contains a lot of information about the true but partially identified parameter."[9]

4 Geometrical Properties of Identification Regions

From now on, we concentrate on the case of a linear model like in example 1 and the classical quadratic loss function. Since we are only interested in the components (β_0, β_1) of an element $\theta = ((\beta_0, \beta_1), (\varepsilon, \delta_l, \delta_u)) \in SMR$, by abuse of notation, we also denote the set $\{(\beta_0, \beta_1) \mid ((\beta_0, \beta_1), (\varepsilon, \delta_l, \delta_u)) \in SMR\}$ as the sharp marrow region (analogously for the sharp collection region). Then we have $SMR = \left\{\beta \in B | \mathbb{E}(\underline{Y} \mid X) \le X\beta \le \mathbb{E}(\overline{Y} \mid X)\right\}$ and $SCR = \{\operatorname*{argmin}_{\beta \in B} \ \mathbb{E}((X\beta - Y)^2) \mid Y \in [\underline{Y}, \overline{Y}]\}$.

Remark 4.1 *The sharp marrow region is always a subset of the sharp collection region, and this is the reason for calling it sharp marrow region, it is the marrow of all truly linear models that fit to some $Z \in [\underline{Y}, \overline{Y}]$. In contrast, the sharp collection region collects the best fitting parameters for every possible $Z \in [\underline{Y}, \overline{Y}]$.*

It is easy to see that the sharp marrow region is convex and closed. Furthermore, all convex, compact sets can be represented as a sharp marrow region:

Proposition 4.1 [10] *Let $I \subset \mathbb{R}^2$ be a compact convex set. Then there exist random variables $\underline{Y}, \overline{Y}$ such that $SMR(\underline{Y}, \overline{Y}) = I$, namely: $\underline{Y} = \min\{X\beta \mid \beta \in I\}$, $\overline{Y} = \max\{X\beta \mid \beta \in I\}$.*

[7] Note that we use the set-valued definition of argmin.

[8] But compare the remarks in the next to last paragraphs of chapters 5.1 and 5.2.

[9] [31, p. 202].

[10] Proofs for all propositions are given in [33].

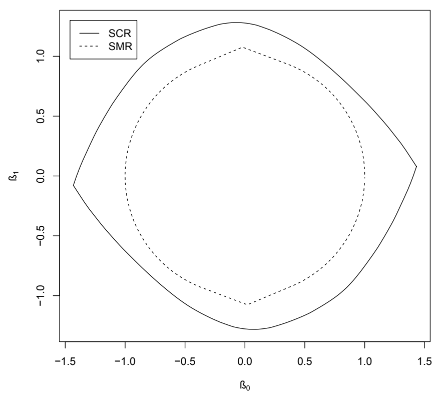


Figure 1: SCR and SMR with unsmooth boundary

For the sharp collection region, the situation is more complicated. To analyze this, we need some definitions from geometry (cf. [47]):

Definition 2 *The Minkowski sum*

$$M = \bigoplus_{i=1}^{n} l_i = \left\{ \sum_{i=1}^{n} p_i \,\middle|\, p_i \in l_i \right\}$$

of n line segments $l_i \subseteq \mathbb{R}^d$ is called a ***zonotope****. A zonotope is a convex, compact and centrally symmetric polytope with finite many extreme points and centrally symmetric facets. A closed, centrally symmetric convex set $Z \subseteq \mathbb{R}^d$ is called a* ***zonoid*** *if it can be approximated arbitrarily closely by zonotopes (w.r.t. a metric, e.g. the Hausdorff distance). For $d = 2$ the zonoids are exactly the closed, centrally symmetric convex sets (see, e.g., [3]).*

Proposition 4.2 *Let* $\mathbb{E}(\underline{\mathrm{Y}}), \mathbb{E}(\overline{\mathrm{Y}}), \mathbb{E}(\underline{\mathrm{Y}} \cdot X), \mathbb{E}(\overline{\mathrm{Y}} \cdot X)$ *be finite and $\mathbb{V}ar(X) \neq 0$. Then the sharp collection region is a zonoid.*

Now, the question arises, if every zonoid can be represented as a sharp collection region. At first glance this seems to be not the case. By looking at examples of (estimates of) sharp collection regions, like that in figure 1 one observes that this regions often have two points on its boundary at which the boundary is not smooth. Note that the situation for SMR is similar, if X has finite or compact support, see figure 1. Fortunately one can prove that every zonotope Z in general position (and, by looking on suitable limit-processes, also every zonoid) can be represented as a sharp collection region if we define the distribution of $X, \underline{\mathrm{Y}}$ and $\overline{\mathrm{Y}}$ in a certain way[11]. The last question is now, how independently from each other the regions SMR and SCR can be generated.

Proposition 4.3 *Let $I = SCR(\underline{\mathrm{Y}}^*, \overline{\mathrm{Y}}^*) \subseteq \mathbb{R}^2$ be a zonoid and $E \subseteq SMR(\underline{\mathrm{Y}}^*, \overline{\mathrm{Y}}^*)$ an arbitrary compact convex set. Then for every $\varepsilon > 0$ there exist random variables $\underline{\mathrm{Y}}, \overline{\mathrm{Y}}$ such that:*

$$\begin{aligned} d(SCR(\underline{\mathrm{Y}}, \overline{\mathrm{Y}}), I) &\leq \varepsilon \\ d(SMR(\underline{\mathrm{Y}}, \overline{\mathrm{Y}}), E) &\leq \varepsilon, \end{aligned}$$

where d is a metric on subsets of $\mathbb{R}^2$, e.g. the Hausdorff distance.

Proof: For $\varepsilon > 0$ define $S \subseteq \mathbb{R}$ such that the distance from any point x to S and the distance from x to S^C and $\mathbb{P}(X \in S)$ tends to zero as ε goes to zero. Then set $(\underline{\mathrm{Y}}, \overline{\mathrm{Y}})$ to $(\underline{\mathrm{Y}}^*, \overline{\mathrm{Y}}^*)$ if $x \in S^C$ and if $x \in S$ set $(\underline{\mathrm{Y}}, \overline{\mathrm{Y}})$ to the random variables $(\underline{\mathrm{Z}}, \overline{\mathrm{Z}})$ that generate E. Then $SMR((\underline{\mathrm{Y}}, \overline{\mathrm{Y}})) \longrightarrow SMR((\underline{\mathrm{Z}}, \overline{\mathrm{Z}})) = E$ and $SCR((\underline{\mathrm{Y}}, \overline{\mathrm{Y}})) \longrightarrow SCR((\underline{\mathrm{Y}}^*, \overline{\mathrm{Y}}^*)) = I$. ■

5 An Algebraic View on Identification Regions

In the next section, we want to look at SMR and SCR as mappings. To analyze the algebraic structure of these mappings, we need some facts about adjunctions. Adjunctions arise in many contexts and often make life a bit easier, see the next examples. For an introduction to partially ordered sets and adjunctions see, e.g., [11, 15].

Definition 3 *Let $(P, \leq)$ and $(Q, \sqsubseteq)$ be partially ordered sets. A pair (f, g) of mappings $f : P \longrightarrow Q$ and $g : Q \longrightarrow P$ is called* ***adjunction****, if:*

$$\forall p \in P \forall q \in Q : \quad p \leq g(q) \iff f(p) \sqsubseteq q.$$

In this case, f is called ***left adjoint*** *and g is called* ***right adjoint****.*

Lemma 5.1 *Let (f, g) be an adjunction. Then the following holds:*

A1 *$g \circ f$ is extensive and $f \circ g$ is intensive, i.e.:* $\forall p \in P, q \in Q : g(f(p)) \geq p \quad \& \quad f(g(q)) \sqsubseteq q.$

A2 *f and g are order-preserving (monotone).*

A3 *$f \circ g \circ f = f$ and $g \circ f \circ g = g$ and thus $f \circ g$ and $g \circ f$ are idempotent.*

A4 *From A1 - A3 it follows that $g \circ f$ is a closure operator and $f \circ g$ is a kernel operator.*[12]

A5 *The adjoints f and g are determining each other unambiguously.*

A6 *f preserves existing joins and g preserves existing meets.*

[11] The main difference to SMR is that there we could construct regions for every arbitrary X with support $\mathbb{R}$.

[12] A closure operator is a monotone, extensive and idempotent mapping and a kernel operator is a monotone, intensive and idempotent one.

To illustrate the concept of adjunctions, we apply it to two areas of the theory of imprecise probability.

Example 2 *Dempster-Shafer-Theory*[13]*:*
In [12] we have the multivalued mapping $\Gamma : X \longrightarrow 2^S$ with which we can associate a set-domained version

$$\tilde{\Gamma} : (2^X, \subseteq) \longrightarrow (2^S, \subseteq) : A \mapsto \bigcup_{a \in A} \Gamma(a).$$

Furthermore we have the operator

$$\begin{array}{rcl} {}_* : (2^S, \subseteq) & \longrightarrow & (2^X, \subseteq) : \\ T & \mapsto & T_* := \{x \in X \mid \Gamma(x) \subseteq T\}. \end{array}$$

Then it is obvious that the pair $(\tilde{\Gamma},\ {}_*)$ is an adjunction because both $A \subseteq T_*$ and $\tilde{\Gamma}(A) \subseteq T$ are meaning exactly that all $a \in A$ are mapped to subsets of T. From this, the ∞-monotonicity of a belief function $B = P \ \circ \ {}_*$ with P a probability-measure follows immediately, since P is ∞-monotone and ${}_*$ is meet-preserving: $B(\bigcup_{i=1}^{k} T_i) = P((\bigcup_{i=1}^{k} T_i)_*) \geq P(\bigcup_{i=1}^{k} (T_i)_*)$ $\geq \sum_{J \neq \emptyset} (-1)^{|J|+1} P(\bigcap_{i \in J} (T_i)_*) = \sum_{J \neq \emptyset} (-1)^{|J|+1} P((\bigcap_{i \in J} T_i)_*)$ $= \sum_{J \neq \emptyset} (-1)^{|J|+1} B(\bigcap_{i \in J} T_i)$.

Furthermore, it is clear that also the composition of a belief function and ${}_*$ or another meet-preserving mapping is ∞-monotone.

Example 3 *Lower Coherent Previsions*[14]*:*
With $(\mathbb{R}^{\mathcal{L}(\Omega)}, \leq)$ the set of all previsions that are defined on all gambles and avoid sure loss, equipped with the dominance relation $\underline{P}_1 \leq \underline{P}_2 :\iff \forall X \in \mathcal{L}(\Omega) : \underline{P}_1(X) \leq \underline{P}_2(X)$ and $(2^{\mathscr{P}(\Omega)}, \supseteq)$ the set of all nonempty sets of finitely additive probability-measures on Ω with the ordinary superset relation, we can construct the following adjunction:

$$\begin{array}{rcl} f : (\mathbb{R}^{\mathcal{L}(\Omega)}, \leq) & \longrightarrow & (2^{\mathscr{P}(\Omega)}, \supseteq) : \underline{P} \mapsto \mathcal{M}(\underline{P}) \\ g : (2^{\mathscr{P}(\Omega)}, \supseteq) & \longrightarrow & (\mathbb{R}^{\mathcal{L}(\Omega)}, \leq) : M \mapsto \underline{P}_M \end{array}$$

with $\mathcal{M}(\underline{P}) = \{p \in \mathscr{P}(\Omega) \mid \forall X \in \mathcal{L}(\Omega) : p(X) \geq \underline{P}(X)\}$, where $\mathscr{P}(\Omega)$ is the set of all finitely additive probability-measures and $\underline{P}_M : \mathcal{L}(\Omega) \longrightarrow \mathbb{R} : X \mapsto \inf_{p \in M} p(X)$. In this language, because of the lower envelope theorem[15]*, coherent lower previsions are exactly the hulls*[16] *of the closure operator $g \circ f$, which maps a lower prevision that avoids sure loss to its natural extension. It is now easy to see that the natural extension of a prevision $\underline{P}$ is the lowest coherent prevision that dominates $\underline{P}$: If $\underline{P}_2 \geq \underline{P}$ is another coherent prevision that dominates $\underline{P}$, then it is a hull $(g \circ f)(Q)$ for some Q and with the idempotence and the monotonicity of $g \circ f$ we have $\underline{P}_2 = (g \circ f)(Q) = (g \circ f \circ g \circ f)(Q) \geq (g \circ f)(\underline{P})$, where the right hand side is the natural extension of $\underline{P}$.*

5.1 SMR as a Right Adjoint

Proposition 5.2 *Let $(\mathscr{Y}, \leq)$ be the set of pairs of numeric random variables $\mathcal{Y} = (\underline{\mathrm{Y}}, \overline{\mathrm{Y}})$, equipped with the relation $\leq$ defined by*

$$\begin{array}{rcl} \mathcal{Y}_1 \leq \mathcal{Y}_2 : & \iff & \mathbb{E}(\overline{\mathrm{Y}}_1 \mid X) \leq \mathbb{E}(\overline{\mathrm{Y}}_2 \mid X) \quad \& \\ & & \mathbb{E}(\underline{\mathrm{Y}}_1 \mid X) \geq \mathbb{E}(\underline{\mathrm{Y}}_2 \mid X). \end{array}$$

This means that if $\mathcal{Y}_1 \leq \mathcal{Y}_2$, the observable variables $(\underline{\mathrm{Y}}_1, \overline{\mathrm{Y}}_1)$ are more informative than $(\underline{\mathrm{Y}}_2, \overline{\mathrm{Y}}_2)$ or equally informative, because from $(\underline{\mathrm{Y}}_1, \overline{\mathrm{Y}}_1)$ we can learn more or the same about the conditional expectations of the unobserved variable Y, we are actually interested in. The mapping

$$\begin{array}{rl} SMR : (\mathscr{Y}, \leq) & \longrightarrow (2^B, \subseteq) : \\ (\underline{\mathrm{Y}}, \overline{\mathrm{Y}}) & \mapsto \{\beta \mid \mathbb{E}(\underline{\mathrm{Y}} | X) \leq X\beta \leq \mathbb{E}(\overline{\mathrm{Y}} | X)\} \end{array}$$

is a right adjoint. The corresponding left adjoint is the prediction-operator[17]*:*

$$\begin{array}{rcl} PR : (2^B, \subseteq) & \longrightarrow & (\mathscr{Y}, \leq) : \\ \Gamma & \mapsto & \left(\inf_{\beta \in \Gamma} X\beta, \sup_{\beta \in \Gamma} X\beta \right). \end{array}$$

Because SMR is a right adjoint, it has the properties $A1 - A6$. The monotonicity $A2$ means that $SMR(\mathcal{Y})$ is more informative if $\mathcal{Y}$ is more informative. The idempotence $A3$ means that if we estimate, predict and then estimate again, we get the same information as if we had only estimated one time. Analogously if we predict, estimate and then predict once more, we get the same prediction as we would get, if we predicted only once. This property is often satisfied by classical estimators, for example the classical least squares estimator has an idempotent prediction matrix. Because $PR \circ SMR$ is a kernel operator, we can now give a clear interpretation of SMR, which is also valid in the misspecified case: The sharp marrow region is the largest region for which the corresponding predictions constitutes the largest inner approximation of the conditional expectations[18]. This interpretation may be not so useful in the misspecified situation, but it is clearly stated.

[13] For an introduction, see [12] and [35].

[14] For an introduction, see [44].

[15] See [44, p. 134].

[16] Hulls are the images of a closure operator and similarly kernels are the images of a kernel operator.

[17] Here, the empty infimum is defined as ∞ and the empty supremum is defined as $-\infty$.

[18] An empty SMR means, that there is no inner approximation induced by the prediction of a set of parameters.

The monotonicity is also shared by SCR, but SCR is no right adjoint, since it is not meet-preserving, because the intersection of two zonoids is generally not a zonoid. Furthermore, generally only $(SCR \circ PR \circ SCR)(\mathcal{Y}) \supset SCR(\mathcal{Y})$ holds, which means that we generally loose information if we predict and estimate once more.

5.2 SMR and SCR as a Kernel and a Hull

In [9] a criterion function based identification region is proposed. The criterion function (see Prop. 5.3) is based on a generalization of the expected squared errors to the expected squared minimal errors. The proposed sharp identification region is the argmin of this criterion function and it is very similar to SMR, but it is not monotone. It shows up that SMR is the highest lower and SCR is the lowest upper monotone approximation of this region.

Definition 4 *Let $E : (P, \leq) \longrightarrow (Q, \sqsubseteq)$ be a mapping. The monotone hull of E is defined as:*

$$H(E) \quad : \quad (P, \leq) \longrightarrow (Q, \sqsubseteq) : X \mapsto \bigvee_{Y \leq X} E(Y).$$

The monotone kernel of E is defined as:

$$K(E) \quad : \quad (P, \leq) \longrightarrow (Q, \sqsubseteq) : X \mapsto \bigwedge_{Y \geq X} E(Y).$$

These set-valued mappings are both order-preserving. Furthermore, the mapping $E \mapsto H(E)$ is a closure operator and the mapping $E \mapsto K(E)$ is a kernel operator, thus indeed $H(E)$ is a hull and $K(E)$ is a kernel. In particular, $H(E)$ is the lowest order-preserving mapping that is higher than E. Analogously, $K(E)$ is the highest order-preserving mapping that is lower than E.

Proposition 5.3 *Let the criterion function $Q : B \to \mathbb{R}$ be defined as*

$$Q(\beta) = \int \left(\mathbb{E}(\underline{\mathrm{Y}}\,|x) - x\beta\right)_+^2 + \left(\mathbb{E}(\overline{\mathrm{Y}}\,|x) - x\beta\right)_-^2 d\mathbb{P}(x)$$

$$= \int \min_{Y \in [\underline{\mathrm{Y}}, \overline{\mathrm{Y}}]} (\mathbb{E}(Y \mid x) - x\beta)^2 d\mathbb{P}(x).$$

Then the criterion function based mapping

$$\begin{array}{rcl} E_Q : (\mathscr{Y}, \leq) & \longrightarrow & (2^B, \subseteq) : \\ (\underline{\mathrm{Y}}, \overline{\mathrm{Y}}) & \mapsto & \operatorname*{argmin}_{\beta \in B} Q(\beta) \end{array}$$

is a source of SMR and SCR:

$$SMR = K(E_Q) \qquad \text{and} \qquad SCR = H(E_Q).$$

From all above, the region SMR seems to be (at least in algebraic terms) a more satisfying region, but note that this region assumes that the model is in fact linear, which is generally untestable in this context. But the linearity assumption could be understood differently, firstly as an assumption on the true model and secondly as something like a regularization or simplification method to avoid overfitting or to have a parsimonious model. The first case points to the sharp marrow region and the second seemingly to the sharp collection region, but the parsimoniousness is decreasing if we allow for sets of parameters β instead of a single parameter and it is not a matter of course, if the SCR, constructed as the union of all reasonable best linear predictors, is still a useful model of the data.[19]

The region SCR can be estimated from samples in a consistent, monotone and nonpartial way. With nonpartial we mean that no pair $\underline{\mathrm{y}} \leq \overline{\mathrm{y}}$ of data would lead to the empty set as the estimate for SCR. One possibility is the estimator proposed in [2]. In contrast, also a nonempty SMR cannot be estimated in such a way[20]. To see this, take a sample $(\underline{\mathrm{y}}, \overline{\mathrm{y}}) = (e^{-x^2}, e^{-x^2})$, $(\underline{\mathrm{z}}_1, \overline{\mathrm{z}}_1) = (0, \overline{\mathrm{y}}) \geq (\underline{\mathrm{y}}, \overline{\mathrm{y}})$ and $(\underline{\mathrm{z}}_2, \overline{\mathrm{z}}_2) = (\underline{\mathrm{y}}, 1) \geq (\underline{\mathrm{y}}, \overline{\mathrm{y}})$. If an estimator $\hat{SMR}$ is consistent and monotone then for n large enough it should satisfy $\hat{SMR}((\underline{\mathrm{y}}, \overline{\mathrm{y}})) \subseteq \hat{SMR}((\underline{\mathrm{z}}_1, \overline{\mathrm{z}}_1)) \cap \hat{SMR}((\underline{\mathrm{z}}_2, \overline{\mathrm{z}}_2)) \approx \{(0,0)\} \cap \{(1,0)\} = \emptyset$. Furthermore SMR could not be estimated robustly in the sense that if one has a mixture in the sense of the proof of Proposition 4.3 then for ε small enough it is not clear what should be the estimated SMR, because that part of the data from the smaller region could be outliers or not, which would lead to different regions.

6 An Identification Region Based on a Set-Domained Loss Function

Now we try to establish a region, which could be understood as a compromise between SMR and SCR. The idea here is that we look on loss functions that are dependent on sets of parameters instead of single parameters. So in a sense we take the fact seriously that the region is a whole set that constitutes an imprecise probability structure. We do not look explicitly at every point of the set and then temporarily forget that the envisaged point is only one point of the set and maltreat it with a classical method. Instead, we see the set as a whole and do not look

[19] In terms of parsimoniousness SMR is comparable to SCR and in fact SMR sometimes describes the data better, e.g., if $\mathcal{Y} = PR(\Gamma)$ for some Γ because then we have $PR(SMR(\mathcal{Y})) = \mathcal{Y}$ but generally only $PR(SCR(\mathcal{Y})) > \mathcal{Y}$.

[20] Note that the estimator proposed in [9] assumes a finite support of X and is not monotone.

into it too deeply. We will construct a distance function between the set of conditional expectations of Y that cannot be refuted and the set of conditional expectations that are predicted by a set Γ of parameters. Here we do not assume that the true model is a linear one (if we would make this assumption, then we would get the region SMR again). Since we have to measure the distance between the two sets $\mathcal{A}(\mathcal{Y}) := \{(x, \mathbb{E}(Y \mid x)) \mid Y \in [\underline{\mathrm{Y}}, \overline{\mathrm{Y}}], x \in \mathbb{R}\}$ and $\mathcal{B}(\Gamma) := \{(x, x\beta) \mid \beta \in \Gamma, x \in \mathbb{R}\}$, we could use for example the Hausdorff distance

$$d_H(\mathcal{A}, \mathcal{B}) = \max\left\{\sup_{a \in \mathcal{A}} \inf_{b \in \mathcal{B}} d(a,b), \quad \sup_{b \in \mathcal{B}} \inf_{a \in \mathcal{A}} d(a,b)\right\}$$

with some metric d of $\mathbb{R}^2$, which possibly takes the distribution of X into account and weights the distance according to the density $f(x)$. For a fixed x we have the possible conditional expectations of Y and the conditional expectations that are predicted by the parameter set Γ. Thus, both point sets are matched in a sense. Because the Hausdorff distance does not match the points of the two sets but compares all points of the two sets to each other, this distance seems to be a little bit counterintuitive. Thus, we propose a slightly different matched distance:

$$\begin{aligned} d_M(\mathcal{A}, \mathcal{B}) : &= \int \left(\sup_{(x, a_2) \in \mathcal{A}} a_2 - \sup_{(x, b_2) \in \mathcal{B}} b_2 \right)^2 + \\ & \left(\inf_{(x, a_2) \in \mathcal{A}} a_2 - \inf_{(x, b_2) \in \mathcal{B}} b_2 \right)^2 d\mathbb{P}(x). \end{aligned}$$

Now we can define a set-domained loss function as

$$L_S(\mathcal{Y}, \Gamma) \quad = \quad d_M\left(\mathcal{A}(\mathcal{Y}), \mathcal{B}(\Gamma)\right)$$

and construct the sharp identification region for the minimizers of the set-domained loss (short: sharp setloss region) $SSR := \bigcup\limits_{\Gamma \subseteq B} \operatorname{argmin} L_S(\mathcal{Y}, \Gamma)$. Note that the argmin is not always unique, so that we have to take the union of all sets that minimizes L_S. To compute SSR one can look at the space $\mathscr{K} = \{PR(\Gamma) \mid \Gamma \subseteq B\}$ of all pairs of random variables $(\underline{\mathrm{Z}}, \overline{\mathrm{Z}})$ that are predicted by some set Γ. Since the predicted variables are only dependent on x, we treat them as functions from $\mathbb{R}$ to $\mathbb{R}$. The set $\mathscr{K}$ is then exactly the set of all $(\underline{\mathrm{Z}}, \overline{\mathrm{Z}})$ satisfying $\forall x_3 \notin [x_1, x_2]$:

$$\overline{\mathrm{Z}}(x_1) + (x_3 - x_1) \cdot \frac{\overline{\mathrm{Z}}(x_2) - \overline{\mathrm{Z}}(x_1)}{x_2 - x_1} \in [\underline{\mathrm{Z}}(x_3), \overline{\mathrm{Z}}(x_3)] \quad \&$$

$$\underline{\mathrm{Z}}(x_1) + (x_3 - x_1) \cdot \frac{\underline{\mathrm{Z}}(x_2) - \underline{\mathrm{Z}}(x_1)}{x_2 - x_1} \in [\underline{\mathrm{Z}}(x_3), \overline{\mathrm{Z}}(x_3)].$$

That implies particularly that $\overline{\mathrm{Z}}$ is convex and $\underline{\mathrm{Z}}$ is concave. The task is now to find a pair $(\underline{\mathrm{Z}}^*, \overline{\mathrm{Z}}^*) \in \mathscr{K}$ that minimizes

$$\int (\underline{\mathrm{Z}}(x) - \underline{\mathrm{Y}}(x))^2 + (\overline{\mathrm{Z}}(x) - \overline{\mathrm{Y}}(x))^2 d\mathbb{P}(x).$$

This problem is nothing else than the problem of finding the projection of $(\underline{\mathrm{Y}}, \overline{\mathrm{Y}})$ on $\mathscr{K}$ and since $\mathscr{Y}$ is a Hilbert space and $\mathscr{K}$ is a closed convex set, this projection is unique. The candidate for the sharp setloss region is then $SMR((\underline{\mathrm{Z}}^*, \overline{\mathrm{Z}}^*))$. Because of $(\underline{\mathrm{Z}}^*, \overline{\mathrm{Z}}^*) = PR(\Gamma)$ for some Γ, we have $PR(SMR((\underline{\mathrm{Z}}^*, \overline{\mathrm{Z}}^*)) = PR(SMR(PR(\Gamma))) = PR(\Gamma) = (\underline{\mathrm{Z}}^*, \overline{\mathrm{Z}}^*)$, which means that our region predicts exactly $(\underline{\mathrm{Z}}^*, \overline{\mathrm{Z}}^*)$. Furthermore, every other set that also predicts $(\underline{\mathrm{Z}}^*, \overline{\mathrm{Z}}^*)$ has to be a subset of our region and thus we have $SSR = SMR((\underline{\mathrm{Z}}^*, \overline{\mathrm{Z}}^*))$. From the construction of SSR it is also clear that the compositions $PR \circ SSR$ and $SSR \circ PR$ are also idempotent. To estimate the region SSR from a sample, we can analogously project the pair of vectors $(\underline{\mathrm{y}}, \overline{\mathrm{y}})$ on the set of pairs of vectors $(\underline{\mathrm{z}}, \overline{\mathrm{z}})$ satisfying $\forall x_k \notin [x_i, x_j] : \overline{\mathrm{z}}_i + (x_k - x_i)\frac{\overline{\mathrm{z}}_j - \overline{\mathrm{z}}_i}{x_j - x_i} \in [\underline{\mathrm{z}}_k, \overline{\mathrm{z}}_k] \quad \& \quad \underline{\mathrm{z}}_i + (x_k - x_i)\frac{\underline{\mathrm{z}}_j - \underline{\mathrm{z}}_i}{x_j - x_i} \in [\underline{\mathrm{z}}_k, \overline{\mathrm{z}}_k]$. With $\theta = (\underline{\mathrm{z}}_1, \ldots, \underline{\mathrm{z}}_n, \overline{\mathrm{z}}_1, \ldots \overline{\mathrm{z}}_n)$ this problem can be written as the minimization of $\theta' Q \theta + c'\theta$ subject to $A\theta \geq 0$ for a (positive definite) matrix Q, a matrix A and a vector c. To compute the solution, one can use for example the algorithm proposed in [25]. To compute the final set $SMR((\underline{\mathrm{z}}^*, \overline{\mathrm{z}}^*))$, one can use standard linear programming techniques. The method can be robustified by modifying the loss function, but then, the solution may be not unique anymore. The minimization problem would get nonlinear, but the dimension of the problem would be n, which is maybe still acceptable[21]. Another idea is to allow only special sets of parameters. Here especially sets of sets of parameters that are closed under Minkowski convex-combinations are interesting, because this would ensure the uniqueness of the solution, because then the set of predictions made by such sets is convex. Such sets of sets are e.g. the set of all zonoids or the set of all zonotopes that are generated by line-segments that have a special angle. The minimization of L_S is then still tractable if the set of sets is parametrizable with a not too high number of parameters. An advantage of using special sets is that these sets are possibly better interpretable, especially if one has a higher number of covariates. For example an arbitrary high dimensional convex point set represented by all its extreme points is harder to figure out than a high dimensional ellipsoid represented by its location and the direction and spread of all main axes.

[21] Note that the naive robustification of SCR seems to be not so easy, because one has to look at the robust estimates for all $y \in [\underline{\mathrm{y}}, \overline{\mathrm{y}}]$ and this is not as easy as the computation of the image of $[\underline{\mathrm{y}}, \overline{\mathrm{y}}]$ under a linear mapping.

7 Concluding Remarks

We have worked out some differences between two types of identification regions in regression analysis under interval data, and discussed some of their properties. Indeed, SMR, relying so-to-say on the marrow of the regression model, and SCR, taking in a collection procedure all potential combinations of data points equally seriously, can be characterized as the monotone kernel and the monotone hull of a criterion function based mapping.

Furthermore, we sketched an appealing, rigorously set-based compromise, whose properties have still to be investigated in more detail. Other topics of further research include the additional inclusion of coarse covariates and an extension to generalized linear models. For generalized linear predictors in [36] a characterization of the sharp collection region is already given. If also covariates are interval-valued, the description of SCR becomes more complicated and a reformulation relying on roots of likelihood-based score-functions seems promising.[22] For the sharp marrow region the crucial role the conditional expectation $\mathbb{E}(Y|X)$ plays in the definition of SMR provides an immediate, promising link. Another direction of future research might be the analysis of models with instrumental variables. For this case a sharp characterization of SCR in terms of the support function of the identified set as well as some asymptotics of corresponding estimates can be found in [4].

Acknowledgements

The authors wish to thank the anonymous reviewers and Marco Cattaneo, Ulrich Pötter and Andrea Wiencierz for their very helpful comments and suggestions.

References

[1] Beresteanu, A., Molchanov, I., & Molinari, F. (2011): Sharp identification regions in models with convex moment predictions, *Econometrica*, 79, 1785–1821.

[2] Beresteanu, A., & Molinari, F. (2008): Asymptotic properties for a class of partially identified models, Econometrica, 76, 763–814.

[3] Bolker, E.D. (1971): The zonoid problem, *The American Mathematical Monthly*, 78, 529–531.

[4] Bontemps, C., Magnac, T., & Maurin, E. (2012): Set identified linear models, *Econometrica*, 80, 1129–1155.

[5] Bugni, F. A. (2010): Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set, *Econometrica*, 78, 735–753.

[6] Canay, I. A. (2010): El inference for partially identified models: Large deviations optimality and bootstrap validity, *Journal of Econometrics*, 156, 408–425.

[7] Cattaneo, M. E. G. V. & Wiencierz, A. (2012): Likelihood-based imprecise regression, *International Journal of Approximate Reasoning*, 53, 1137–1154, extended version of the 2011 ISIPTA paper. see F. P. A. Coolen, G. de Cooman, T. Fetz, & M. Oberguggenberger (eds.): *ISIPTA '11: Proc. Seventh Int. Symp. on Imprecise Probability: Theories and Applications*, 119–128, Innsbruck.

[8] Černý, M. & Rada M. (2011): On the possibilistic approach to linear regression with rounded or interval-censored data. *Measurement Science Review*, 11, 34–40.

[9] Chernozhukov, V., Hong, H., & Tamer, E. (2007): Estimation and confidence regions for parameter sets in econometric models, *Econometrica*, 75, 1243–1284.

[10] de Cooman, G. & Zaffalon, M. (2004): Updating beliefs with incomplete observations. *Artificial Intelligence*, 159, 75–125.

[11] Davey, B.A. & Priestley, H.A. (2002): *Introduction to Lattices and Order*, 2nd edition. Cambridge UP.

[12] Dempster, A.P. (1967): Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38, 325–339.

[13] Dobra, A.& Fienberg, S. (2000): Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences of the United States of America*, 97, 11885–11892.

[14] Dubois, D. (1986): Belief structure, possibility theory and decomposable confidence measures on finite sets. *Computers and Artificial Intelligence*, 5, 403–416.

[15] Heijmans, H.J.A.M., (1994): *Morphological image operators*, Academic Press, Boston.

[16] Heitjan, D. & Rubin, D. (1991): Ignorability and coarse data. *The Annals of Statistics*, 19, 2244–2253.

[17] Holmes, R. N. (1975): *Geometric Functional Analysis and its Applications.* Springer, New York.

[18] Horowitz, J. L. & Manski, C. F. (2001): Imprecise identification from incomplete data. In: G. de Cooman, T. Fine, & T. Seidenfeld (eds.): *ISIPTA'01: Proc. Second Int. Symp. on Imprecise Probabilities and Their Applications*, 213–218, Maastricht, Shaker.

[19] Kwerel, S. (1983): Fréchet Bounds, In: S. Kotz & N. Johnson (eds.): *Encyclopedia of Statistical Sciences: Volume 3*, 203–209. Wiley, New York.

[22] See [34], who also developed algorithms for calculating SCR's like regions in a GLM based on an exponential model.

[20] Küchenhoff, H., Augustin, T. & Kunz, A. (2012): Partially identified prevalence estimation under misclassification using the kappa coefficient. *International Journal of Approximate Reasoning*, 53, 1168–1182, extended version of the 2011 ISIPTA paper. see F. P. A. Coolen, G. de Cooman, T. Fetz, & M. Oberguggenberger (eds.): *ISIPTA '11: Proc. Seventh Int. Symp. on Imprecise Probability: Theories and Applications*, 237–246, Innsbruck.

[21] Lee, S. & Wilke, R. A. (2009): Reform of unemployment compensation in Germany: A nonparametric bounds analysis using register data, *Journal of Business & Economic Statistics*, 27, 193–205.

[22] Little, R. & Rubin, D. (1987): *Statistical Analysis with Missing Data.* Wiley, New York.

[23] Manski, C. F. (2003): *Partial Identification of Probability Distributions.* Springer, New York.

[24] Manski, C. F. & Molinari, F. (2010): Rounding probabilistic expectations in surveys. *Journal of Business & Economic Statistics*, 28, 219–231.

[25] Meyer, M. C. (2013): A simple new algorithm for quadratic programming with applications in statistics. *Communications in Statistics*, 42, 1126–1139.

[26] Molinari, F. (2010): Missing treatments, *Journal of Business & Economic Statistics*, 28, 82–95.

[27] Moon, H. R. & Schorfheide, F. (2012): Bayesian and frequentist inference in partially identified models, *Econometrica*, 80, 755–782.

[28] Nicoletti, C., Peracchi, F., & Foliano, F. (2011): Estimating income poverty in the presence of missing data and measurement error, *Journal of Business & Economic Statistics*, 29, 61–72.

[29] Nguyen, H., Kreinovich, V., Wu, B., & Xiang, G. (2011): *Computing Statistics under Interval and Fuzzy Uncertainty: Applications to Computer Science and Engineering.* Springer, Berlin/Heidelberg.

[30] Pötter, U. (2008): *Statistical Models of Incomplete Data and Their Use in Social Sciences*, Ruhr-Universität Bochum (Habilitation Thesis).

[31] Ponomareva, M., & Tamer, E. (2011): Misspecification in moment inequality models: back to moment equalities?. *Econometrics Journal*, 14, 186-203.

[32] Rohwer, G. & Pötter, U. (2001): *Grundzüge der sozialwissenschaftlichen Statistik.* Juventa, Weinheim.

[33] Schollmeyer, G., & Augustin, T. (2013): On Sharp Identification Regions for Regression Under Interval Data. Technical Report 143, *Department of Statistics*, LMU Munich.

[34] Seitz, M. (2012): Estimating partially identified parameters in generalized linear models under interval data (in German), Master Thesis, Department of Statistics, LMU Munich.

[35] Shafer, G. (1976): *A Mathematical Theory of Evidence.* Princeton U. P.

[36] Stoye, J. (2007): Bounds on generalized linear predictors with incomplete outcome data. *Reliable Computing*, 13, 293–302.

[37] Stoye, J. (2009a): Partial identification and robust treatment choice: An application to young offenders. *Journal of Statistical Theory and Practice*, 3, 239–254.

[38] Stoye, J. (2009b): Statistical inference for interval identified parameters. In: T. Augustin, F. Coolen, S. Moral, & M. Troffaes (eds.): *ISIPTA '09: Proc. Sixth Int. Symp. on Imprecise Probability: Theories and Applications*, 395–404, Durham, UK, SIPTA.

[39] Stoye, J. (2010): Partial identification of spread parameters. *Quantitative Economics*, 1, 323–357.

[40] Tamer, E. (2010): Partial identification in econometrics. *Annual Review of Economics*, 2, 167–195.

[41] Utkin, L. V. & Augustin, T. (2007): Decision making under imperfect measurement using the imprecise Dirichlet model. *International Journal of Approximate Reasoning*, 44, 322–338; extended version of the 2005 ISIPTA paper. see see F. Cozman, R. Nau, & T. Seidenfeld (eds.): *ISIPTA '05: Proc. Fourth Int. Symp. on Imprecise Probabilities and Their Applications*, Manno.

[42] Utkin, L. V. & Coolen, F.P.A. (2011): Interval-valued regression and classification models in the framework of machine learning. In: F. P. A. Coolen, G. de Cooman, T. Fetz, & M. Oberguggenberger (eds.): *ISIPTA '11: Proc. Seventh Int. Symp. on Imprecise Probability: Theories and Applications*, 371–380, Innsbruck.

[43] Vansteelandt, S., Goetghebeur, E., Kenward, M., & Molenberghs, G. (2006): Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16, 953–979.

[44] Walley, P. (1991): *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London.

[45] Zaffalon, M. (2005): Conservative rules for predictive inference with incomplete data. In: F. Cozman, R. Nau, & T. Seidenfeld (eds.): *ISIPTA '05: Proc. Fourth Int. Symp. on Imprecise Probabilities and Their Applications*, 406–415, Manno.

[46] Zaffalon, M. & Miranda, E. (2009): Conservative inference rule for uncertain reasoning under incompleteness. *Journal of Artificial Intelligence Research*, 34, 757–821.

[47] Ziegler, G.M. (2008): *Lectures on Polytopes.* Graduate Texts in Mathematics, vol. 152, Springer, New York.

[48] Zhang, Z. (2010): Profile likelihood and incomplete data. *International Statistical Review*, 78, 102-116.

Two theories of conditional probability and non-conglomerability

Teddy Seidenfeld **Mark J. Schervish** **Joseph B. Kadane**
Carnegie Mellon University
teddy@stat.cmu.edu mark@stat.cmu.edu kadane@stat.cmu.edu

Abstract

Conglomerability of conditional probabilities is suggested by some (e.g., Walley, 1991) as necessary for rational degrees of belief. Here we give sufficient conditions for non-conglomerability of conditional probabilities in the de Finetti/Dubins sense. These sufficient conditions cover familiar cases where P(·) is a continuous, countably additive probability. In this regard, we contrast the de Finetti/Dubins sense of conditional probability with the more familiar account of regular conditional distributions, in the fashion of Kolmogorov.

Keywords. Non-conglomerability, conditional probability, κ-additive probability, regular conditional distribution.

1 Introduction

Consider a finitely, but not necessarily countably additive probability P(·) defined on a sigma-field of sets $\mathcal{B}$, each set a subset of the sure-event Ω. In other terms, <Ω, $\mathcal{B}$, P> is a (finitely additive) measure space.

We begin by reviewing the theory of conditional probability that we associate with de Finetti (1974) and Dubins (1975).

Let B, C, D, E, F, G ∈ $\mathcal{B}$, with B ≠ ∅ and F ∩ G ≠ ∅.
Definition 1. A *conditional probability* P(· | B) satisfies the following three conditions:

(i) P(C ∪ D | B) = P(C | B) + P(D | B), whenever B ∩ C ∩ D = ∅;
(ii) P(B | B) = 1.

In order to regulate conditional probability given a non-empty *null* event, i.e., one that itself may be of unconditional or conditional probability 0, we require the following.

(iii) P(E ∩ F | G) = P(E | F ∩ G)P(F | G).

Throughout, we follow the usual identification of unconditional probability with conditional probability given the sure-event, P(·) = P(· | Ω).

This account of conditional probability is not the usual theory from contemporary Mathematical Probability, which we associate with Kolmogorov (1956). That theory, instead, defines conditional probability through regular conditional distributions, as follows.

Let $\mathcal{A}$ be a sub-σ-field of $\mathcal{B}$.
Definition 2. P(· | $\mathcal{A}$) is a *regular conditional distribution* [rcd] on $\mathcal{B}$, given $\mathcal{A}$ provided that:

1. For each ω ∈ Ω, P(·|$\mathcal{A}$)(ω) is a countably additive probability on $\mathcal{B}$.
2. For each B ∈ $\mathcal{B}$, P(B| $\mathcal{A}$)(·) is an $\mathcal{A}$-measurable function.
3. For each A ∈ $\mathcal{A}$,

$$P(A \cap B) = \int_A P(B \mid \mathcal{A})(\omega)\, dP(\omega).$$

That is, P(B| $\mathcal{A}$) is a version of the Radon-Nikodym derivative of P(· ∩ B) with respect to P(·).

Definition 3: An *$\mathcal{A}$-atom* is the intersection of all elements of $\mathcal{A}$ that contain a given point ω of Ω.

When P(A) > 0 and ω ∈ A ∈ $\mathcal{A}$ and A is an $\mathcal{A}$-atom, then

$$P(B \mid \mathcal{A})(\omega) = P(A \cap B) / P(A).$$

The theory of conditional probability that we use here differs from the received theory of Kolmogorovian regular conditional distributions in at least five ways:

(1) The theory of regular conditional distributions requires that probabilities and conditional probabilities are countably additive. The theory of conditional probability from Definition 1 requires only that probability is finitely additive. In this note we bypass this difference by exploring countably additive conditional probabilities.

(2) When P(A) = 0 and A is not empty, a regular conditional probability given A is relative also to a sub-sigma field $\mathcal{A}$ ⊆ $\mathcal{B}$, where A ∈ $\mathcal{A}$. By contrast, in the theory of conditional probability, P(· | A), depends solely on the event A and not on any sub-field that embeds it. Example 2, below, illustrates this difference.

(3) Dubins (1975) establishes that for each set Ω there is a *full* conditional probability function, P(B|A), defined whenever A ≠ ∅ and B are elements of $\mathcal{B}$, the powerset of Ω. However, some countably additive probabilities do not admit regular conditional distributions relative to a particular sub-sigma field $\mathcal{A} \subseteq \mathcal{B}$, even when each sigma-field, $\mathcal{A}$ and $\mathcal{B}$, is countably generated. The canonical example of a measure space that admits no rcd's is obtained by extending the σ-field of Borel sets on [0,1] under Lebesgue measure, μ, with the addition of one non-measurable set.

Denote the initial measure space by <[0,1], $\mathcal{B}$, μ>. A familiar maneuver allows an extension of $\mathcal{B}$ to a larger σ-field of sets, $\mathcal{B}'$, generated by adding one Lebesgue non-measurable set to $\mathcal{B}$, and an extension of μ to a countably additive probability μ′ over $\mathcal{B}'$. However, there is no rcd $\mu'(\cdot|\mathcal{A})(\omega)$ on $\mathcal{B}'$ given $\mathcal{A}$ when, e.g., $\mathcal{A} = \mathcal{B}$. (See Halmos, 1950, p. 211; Billingsley, 1986, Exercise 33.13; Breiman, 1968, p.81; Doob, 1953, p. 624; or Loeve, 1955, p. 370 for variations on this common theme.) Though, for each B ∈ $\mathcal{B}$, the extended measure space has Radon-Nikodym derivatives P(B |·) satisfying condition 3, above, these resist assembly of pointwise probabilities into a countably additive probability distribution over $\mathcal{B}'$ as required by condition 1.

In our (2001, Corollary 1) we show that, quite generally, a measure space admitting rcd's can be extended to another measure space admitting rcd's if and only if the latter lies within the measure completion of the former. In rejoinder to the existence problem, however, a sufficient condition for rcd's to exist on $\mathcal{B}$ (given any sub σ-field $\mathcal{A}$) is that $\mathcal{B}$ be isomorphic under a 1-1 measurable mapping to the σ-field of a random variable. (See, Billingsley 1986, T.33.3; or Breiman, 1968, T. 4.30.)

(4) Blackwell (1955), Blackwell and Ryll-Nardzewski (1963) and Blackwell and Dubins (1975) introduce an additional constraint, *propriety* of an rcd, matching condition (ii) of de Finetti/Dubins' theory of conditional probabilities.

Definitions 4:

- An rcd $P(\cdot|\mathcal{A})(\omega)$ on $\mathcal{B}$ given $\mathcal{A}$, is *proper at* ω if $P(A|\mathcal{A})(\omega) = 1$ whenever $\omega \in A \in \mathcal{A}$.
- $P(\cdot|\mathcal{A})(\omega)$ is *improper at* ω, otherwise.
- $P(\cdot|\mathcal{A})$ is *proper* if $P(\cdot|\mathcal{A})(\omega)$ is proper for each ω∈Ω.

Definition 5: Say that a probability distribution is *extreme* if its range is the two point set {0,1}.

Theorem 1 (Blackwell and Dubins, 1975) When $\mathcal{B}$ is a countably generated σ-field, *no* rcd on $\mathcal{B}$ given $\mathcal{A}$ is proper if there exists *some* extreme probability on $\mathcal{A}$ supported by no $\mathcal{A}$-atom belonging to $\mathcal{A}$.

In other words, provided there exists even one extreme probability on $\mathcal{A}$ which is supported by none of its $\mathcal{A}$-atoms, then the sub-σ-field $\mathcal{A}$ is anomalous for all rcd's on $\mathcal{B}$ given $\mathcal{A}$ in that they are improper, each and every one! However, this result does not identify at how many points, ω, or how badly, the rcd is improper. The following result addresses that question.

Assume that $\mathcal{A}$ is an atomic sub-σ-field of $\mathcal{B}$, with $\mathcal{A}$-atoms *a*. Denote by *a*(ω) that $\mathcal{A}$-atom containing the point ω.

Theorem 2 (our 2001): Let P be an extreme probability on $\mathcal{A}$ that is not supported by any of its $\mathcal{A}$-atoms. If an rcd $P(\cdot|\mathcal{A})(\omega)$ on $\mathcal{B}$ given $\mathcal{A}$ exists, there is one where $P\{\omega: P(a(\omega)|\mathcal{A})(\omega) = 0\} = 1$. And, if $\mathcal{B}$ is countably generated, then this rcd is unique.

Theorem 2 asserts that when $\mathcal{B}$ is countably generated and the antecedent of Theorem 1 is satisfied, then almost surely with respect to P, the rcd's on $\mathcal{B}$ given $\mathcal{A}$ are maximally improper, in two senses simultaneously:

- The set of points where propriety fails has measure 1 under P.
- For P-almost all points ω, $P(a(\omega)|\mathcal{A})(\omega) = 0$ when propriety requires that $P(a(\omega)|\mathcal{A})(\omega) = 1$.

The following Corollary applies Theorem 2 when conditioning on the sub-sigma field associated with de Finetti's theorem on exchangeability.

Let $\Omega = \{0,1\}^{\aleph_0}$; let $\mathcal{B}$ = the Borel subsets of Ω; and let P be a symmetric probability, in the sense of Hewitt and Savage (1955) defined as follows. Let *T* be an arbitrary finite permutation of the positive integers, i.e., a permutation of the coordinates of Ω that leaves all but finitely many places fixed. For B ∈ $\mathcal{B}$, given *T*, define the set $T^1B = \{\omega: T(\omega) \in B\}$.

Definitions 6:

- P is called a *symmetric probability* if $P(T^1B) = P(B)$, for each B ∈ $\mathcal{B}$ and *T*.
- If $B = T^1B$ for all (finite) permutations *T*, B is called *a symmetric event*.

Let $\mathcal{A}$ be the sub-σ-field of $\mathcal{B}$ generated by the class $\boldsymbol{T}$ of all finite permutations of the coordinates of Ω, i.e., $\mathcal{A}$ is the σ-field of the symmetric events. $\mathcal{A}$ is atomic, with $\mathcal{A}$-atoms comprised by a countable set of sequences, each pair of sequences in the same atom differing by some finite permutation of its coordinates. In all but two cases the $\mathcal{A}$-atoms are countably infinite sets. The two exceptions are the two constant sequences, which are singleton sets.

Corollary (see our 2001). Each rcd $P(\cdot|\mathcal{A})(\omega)$ on $\mathcal{B}$ given $\mathcal{A}$, for a symmetric probability P, satisfies

$$P\{\omega: P(a(\omega) \mid A)(\omega) = 0)\} = 1,$$

provided that $P(<0,0,0,\ldots.>) = P(<1,1,1\ldots.>) = 0$.

For additional related results see (Berti and Rigo, 2007)

(5) Our focus in this paper is on a fifth feature that distinguishes the de Finetti/Dubins theory of conditional probability and the Kolmogorovian theory of regular conditional probability. This aspect of the difference involves *conglomerability* of conditional probability functions.

Let $E \in \mathcal{B}$, let $\boldsymbol{N}$ be an index set and let $\pi = \{h_\nu: \nu \in \boldsymbol{N}\}$ be a partition of the sure event where the conditional probabilities, $P(E \mid h_\nu)$, are well defined for each $\nu \in \boldsymbol{N}$.

Definition 7: The conditional probabilities $P(E \mid h_\nu)$ are *conglomerable* in π provided that, for each event $E \in \mathcal{B}$ and arbitrary real constants k_1 and k_2,

if $k_1 \leq P(E \mid h_\nu) \leq k_2$ for each $\nu \in \mathbf{N}$, then $k_1 \leq P(E) \leq k_2$.

In our (1984) we show that if P is merely finitely additive (i.e., if P is finitely but not countably additive) with conditional probabilities that satisfy Definition 1, then P fails conglomerability in some countable partition. That is, for each merely finitely additive probability P there is an event E, an $\varepsilon > 0$, and a countable partition $\pi = \{h_n: n = 1, \ldots\}$, where $P(E) > P(E \mid h_n) + \varepsilon$ for each $h_n \in \pi$.

The following illustrates a failure of conglomerability for a merely finitely additive probability P in a countable partition $\pi = \{h_n: n \in \{1, 2, \ldots\}\}$, where each element of the partition is not null, i.e., $P(h_n) > 0$, n = 1, 2, … .

Then, apart from the requirement of countable additivity, both theories agree on the relevant conditional probabilities: $P(E \mid h_n) = P(E \cap h_n)/P(h_n)$ is well defined. Thus, the failure of conglomerability in this example is due to the failure of countable additivity, rather than to a difference in how conditional probability is defined.

***Example* 1** (Dubins, 1975): Let $\mathbf{\Omega} = \{(i, n): i \in \{1, 2\}$ and $n \in \{1, 2, \ldots\}\}$ and let $\mathcal{B}$ be the powerset of $\mathbf{\Omega}$.
Let event $E = \{\{1, n\}: n \in \{1, 2, \ldots\}\}$ and events $h_n = \{\{1,n\}, \{2,n\}\}$, n = 1, ... Observe that the h_n form a partition: $\pi = \{h_n: n \in \{1, 2, \ldots\}\}$.
Partially define the (unconditional) probability P by

(a) $P(\{(i, n)\}) = 1/2^{n+1}$ if i = 1, n = 1, 2, ...
(b) $P(\{(i, n)\}) = 0$ if i = 2, n = 1, 2, ...
(c) $P(E) = 0.5$.

So P is countably additive given E, and strongly finitely additive given E^c. (A finitely additive probability is strongly finitely additive if there is a countable partition of the sure event each of whose elements is null.)
Clearly, $P(h_n) = P(\{(1,n)\}) + P(\{2,n)\}) = 1/2^{n+1} > 0$ for each $n \in \{1, 2, \ldots\}$.
But P is not conglomerable in π, as:
$P(E \mid h_n) = P(E \cap h_n)/P(h_n) = 1$, for each $n \in \{1, 2, \ldots\}$, whereas $P(E) = 0.5$. Example 1
In our (1996), we discuss this example in connection with the value of information.

The non-conglomerability of Example 1 extends to a non-trivial IP class, $\mathcal{P}$. Let $\mathcal{P}$ be the set of all finitely additive conditional probabilities whose unconditional probabilities $P(\cdot) = P(\cdot \mid \Omega)$ satisfy conditions (*a*), (*b*) and (*c*) of Example 1. The class $\mathcal{P}$ is convex in the usual sense, applied to unconditional probabilities. That is, assume $\mathcal{P}$ contains two finitely additive conditional probabilities $P_1(\cdot \mid \cdot)$ and $P_2(\cdot \mid \cdot)$ with unconditional probabilities, respectively, $P_1(\cdot)$ and $P_2(\cdot)$. Let $0 \leq x \leq 1$. Then there is a finitely additive conditional probability $P_3(\cdot \mid \cdot)$ in $\mathcal{P}$ whose unconditional probability $P_3(\cdot)$ satisfies $P_3(\cdot) = xP_1(\cdot) + (1-x)P_2(\cdot)$. The cardinality of $\mathcal{P}$ is $2^{|\Re|}$, where $|\Re|$ is the cardinality of the continuum. This follows as $\mathcal{P}$ includes all finitely additive probabilities P where $P(\cdot \mid E^c)$ is a non-principal ultrafilter probability on the positive integers and there are $2^{|\Re|}$-many such non-principal ultrafilters.) Each $P \in \mathcal{P}$ fails conglomerability in π exactly as in Example 1. Hence, with respect to lower and upper unconditional and conditional probabilities, the IP-set $\mathcal{P}$ fails to be conglomerable in the partition π. In Section 5 we give sufficient conditions for an IP set of countably additive probabilities to experience non-conglomerability in an uncountable partition.

2 Non-conglomerable σ-additive probability

The focus of this note is non-conglomerability for countably additive probabilities. In the appendix to our (1986) we show that for a continuous, countably additive probability defined on the continuum, and assuming conditional probabilities that satisfy Definition 1 rather than regular conditional distributions, then non-conglomerability results by considering continuum-many different partitions of the continuum. These alternative partitions are generated by sets of equivalent (non-linearly transformed) random variables.
Conglomerability cannot be satisfied in all the partitions.

Here we generalize that result to a large class of countably additive probabilities, P, that are not κ-additive for some uncountable cardinal κ, by identifying for each such P specific partitions where P fails to be conglomerable.

In the following presentation, let α, β, and γ be ordinals and λ and κ cardinals.

Definitions 8:

- A probability P is *κ-additive* if, for each increasing γ-sequence of measurable events, $\{E_\alpha: \alpha < \gamma \leq \kappa\}$, where $E_\alpha \subseteq E_\beta$ whenever $\alpha < \beta < \gamma$, then $P(\cup_{\alpha<\gamma} E_\alpha) = \sup_{\alpha<\gamma} P(E_\alpha)$.

That is, with $\gamma \leq \kappa$, P is κ-additive provided that probability is continuous from below over γ-long sequences that approximate events from below. This definition agrees with the usual definition of countable additivity; let $\kappa = \aleph_0$.

- Say that P is *not κ-additive* when, for some event E and increasing γ-sequence that approximates E from below, $P(\cup_{\alpha<\gamma} E_\alpha) > \sup_{\alpha<\gamma} P(E_\alpha)$.
- If P is κ-additive for each cardinal κ, then call P *perfectly additive*.

Consider a countably additive probability P that is not κ-additive for some cardinal κ. Since the cardinals below a given cardinal form a well-ordered set, we consider the least cardinal κ for which P is not κ-additive. And since we assume that P is countably additive, then κ is some uncountable cardinal – unless P is perfectly additive. Thus, assume that for an uncountable cardinal κ, P is not κ-additive but is λ-additive for each cardinal $\lambda < \kappa$.

We make the following two structural assumptions on the measurable sets $\mathcal{B}$.

- We take the measure completion of P. Each subset of a P-null event is measurable.

That is, if $E \in \mathcal{B}$ with $P(E) = 0$ and $F \subseteq E$ then $F \in \mathcal{B}$.

We require also that $\mathcal{B}$ includes sufficiently many events.

- If E is not P-null and $|E| = \kappa$, then E can be partitioned into two measurable sets of the same cardinality

That is, if $P(E) > 0$ then there exits $E_1, E_2 \in \mathcal{B}$, $E_1 \cap E_2 = \emptyset$, $E_1 \cup E_2 = E$, with $|E_1| = |E_2|$.
Note that, given the first assumption, the second structural assumption can be satisfied in a variety of ways. For example, assume that when E is a κ-sized non-null event, $P(E) > 0$, then there is a κ-sized, null sub-event: There exists $E_1 \subset E$, $|E_1| = \kappa$, and $P(E_1) = 0$.

These two assumptions provide for a rich space of measurable events while stopping short of requiring P to be defined on a power set, which otherwise would require κ to be greater than a weakly inaccessible cardinal, by Ulam's [1930] result for real-valued measurable cardinals.

Here we identify a simple condition involving *tiers* of points that ensures P fails to be conglomerable in a partition of cardinality κ.

Definition 9: A *tier* τ is a (measurable) set of points such that for each pair of points $\{\omega_i, \omega_j\} \subset \tau$ $(i \neq j)$

$$0 < P(\{\omega_i\} \mid \{\omega_i, \omega_j\}) < 1.$$

Proposition: Let P be σ-additive but not κ-additive ($\kappa \geq \aleph_1$), having conditional probabilities defined relative to non-empty sets in $\mathcal{B}$, $P(\cdot \mid \mathcal{B})$, and which satisfies the two structural assumptions on $\mathcal{B}$ identified above. If there is an uncountable tier τ of points, $|\tau| \geq \kappa$ with $P(\tau) > 0$, then P fails to be conglomerable in a partition π with $|\pi| = \kappa$.

Thus, rather than thinking that non-conglomerability is an anomalous feature of finite but not countably additive probabilities, and arises solely with finitely but not countably additive probabilities in countable partitions, here we argue for a different conclusion: Let $P(\cdot \mid \cdot)$ be a conditional probability according to Definition 1. Non-conglomerability of P's conditional probabilities occurs in a partition whose cardinality $|\pi| = \kappa$ matches the κ-non-additivity of P.

We summarize: Let P be defined on a measurable space $<\boldsymbol{\Omega}, \mathcal{B}>$, where $\mathcal{B}$ includes each of the points of the space, $\boldsymbol{\Omega} = \{\omega_\alpha: \alpha < \kappa\}$, with α ranging over all ordinals less than κ. That is, without loss of generality, assume $\boldsymbol{\Omega}$ has cardinality κ and where, if a measurable event E is null, i.e., whenever $P(E) = 0$, then $\mathcal{B}$ includes each subset of E, and where κ-sized non-null events can be split into two measurable κ-sized events. Then if some tier of points is not null, P fails to be conglomerable in a partition of cardinality κ.

Since P is not perfectly additive, it follows that κ is a regular cardinal: it has cofinality κ. Otherwise, κ is singular with cofinality$(\kappa) = \lambda < \kappa$. Then, using this λ-sequence which is cofinal in κ, as P is λ-additive for each $\lambda < \kappa$, P would be κ-additive as well.

3 Proof of the Proposition

Suppose there exists a tier of points τ, $|\tau| = \kappa$, with $P(\tau) > 0$. Then $P(\{\omega\}) = 0$ for each $\omega \in \tau$, because $P(\tau) > 0$ and P is λ-additive for each cardinal $\lambda < \kappa$. Partition τ into two disjoint sets, $T_0 \cap T_1 = \emptyset$ with $T_0 \cup T_1 = \tau$; each with cardinality κ, $|T_0| = |T_1| = \kappa$; and label them so that $P(T_0) \leq P(T_1) = d > 0$.

We identify a partition of cardinality κ where P fails to be conglomerable, which we write as $\pi = \{h_\alpha: \alpha < \kappa\} \cup \{h'_\beta: \beta < \gamma \leq \kappa\}$, where $\{h_\alpha: \alpha < \kappa\} \cap \{h'_\beta: \beta < \gamma \leq \kappa\} =$

$\varnothing$, and where $P(T_1 \mid h) < d/2$ for each $h \in \pi$. Possibly the second set, $\{h'_\beta: \beta < \gamma \leq \kappa\}$, is empty, as we explain below. Each element $h \in \pi$ is a finite set. Each element h_α contains exactly one point from T_1, and some positive finite number of points from T_0, selected to insure that $P(T_1 \mid h) < d/2$. If the second set, $\{h'_\beta: \beta < \gamma \leq \kappa\}$, is not empty, each $h'_\beta = \{\omega_\beta\}$ is a singleton with $\omega_\beta \in \Omega - T_1$. So, if $\{h'_\beta: \beta < \gamma \leq \kappa\}$ is not empty, then $P(T_1 \mid h'_\beta) = 0$ for each h'_β. Next we establish the existence of such a measurable partition π.

By the Axiom of Choice, consider a κ-long well ordering of T_1, $\{\omega_1, \omega_2, \ldots, \omega_\beta, \ldots\}$ with ordinal indices $0 < \beta < \kappa$. Define π by induction. As each of T_0, T_1 is a subset of the tier τ, consider the countable partition of T_0 into sets
$\rho_{1,n} = \{\omega \in T_0: (n\text{-}1)/n \leq P(\{\omega_1\} \mid \{\omega_1, \omega\}) < n/(n+1)\}$ for n = 1, 2 … .

Observe that $\cup_n \rho_{1,n} = T_0$. Since $|T_0| = \kappa \geq \aleph_1$, by the pigeon-hole principle, consider the least n* such that ρ_{1,n^*} is infinite. Let $U_1 = \{\omega_{1,1}, \ldots, \omega_{1,m}\}$ be m-many points chosen from ρ_{1,n^*}. Note that $P(\{\omega_1\} \mid U_1 \cup \{\omega_1\}) \leq n^*/(m+n^*)$. Choose m sufficiently large so that $n^*/(m+n^*) < d/2$. Let $h_1 = U_1 \cup \{\omega_1\}$.

For ordinals $1 < \beta < \kappa$, define h_β, by induction, as follows. Denoting $T_{0,1} = T_0$, and let $T_{0,\beta} = T_0 - (\cup_{0<\alpha<\beta}\, h_\alpha)$. Since, for each α, $0 < \alpha < \beta$, by hypothesis of induction h_α is a finite set, then $|\cup_{0<\alpha<\beta}\, h_\alpha| < \kappa$. So, $|T_{0,\beta}| = \kappa$. Since $T_{0,\beta}$ is a subset of τ, just as above, consider the countable partition of $T_{0,\beta}$ into sets
$\rho_{\beta,n} = \{w \in T_{0,\beta}: (n\text{-}1)/n \leq P(\{\omega_\beta\} \mid \{\omega_\beta, \omega\}) < n/(n+1)\}$ for n = 1, 2, … . Again, by the pigeon-hole principle, consider the least integer n* such that ρ_{β,n^*} is infinite. Let $U_\beta = \{\omega_{\beta,1}, \ldots, \omega_{\beta,m}\}$ be m-many points chosen from ρ_{β,n^*}. Note that

$$P(\{\omega_\beta\} \mid U_\beta \cup \{\omega_\beta\}) \leq n^*/(m+n^*).$$

Choose m sufficiently large that $n^*/(m+n^*) < d/2$. Let $h_\beta = U_\beta \cup \{\omega_\beta\}$. Observe that $T_1 \subset \cup_{0<\beta<\kappa}\, h_\beta$ and that for each $0 < \beta < \kappa$, $P(T_1 \mid h_\beta) < d/2$. In order to complete the partition π, consider a catch-all set with all the remaining points $\omega_\beta \in \Omega - \cup_{0<\beta<\kappa}\, h_\beta$. Note that each such ω_β is not a member of T_1, if any such points exist. Add each such point $\{\omega_\beta\} = h'_\beta$ as a separate partition element of π. Thus, if there are any such points, $P(T_1 \mid h'_\beta) = 0 < d/2$.

Hence, P is not conglomerable in π as $P(T_1) = d > 0$, yet for each $h \in \pi$, $P(T_1 \mid h) < d/2$.◊ Proposition

4 An Example of the Proposition

Next, we illustrate the Proposition and with it also the difference (2) between the theory of conditional probability according to Definition 1 and the theory of regular conditional distributions.

Example 2: Let $<\Omega, \mathcal{B}>$ be the measurable space of Lebesgue measurable subsets of the half-open unit interval of real numbers: $\Omega = [0,1)$ and $\mathcal{B}$ is its algebra of Lebesgue measurable subsets. Let P be the uniform, countably additive probability with constant density function $f(\omega) = 1$ for each real number $0 \leq \omega < 1$, and $f(\omega) = 0$ otherwise. So $P(\{\omega\}) = 0$ for each $\omega \in \Omega$. Evidently P is not κ-additive, because $\kappa = |\Omega| = |\Re|$.

Consider the uniform density function f to identify conditional probability given finite sets as uniform over those finite sets, as well. That is, when $F = \{\omega_1, \ldots, \omega_k\}$ is a finite subset of Ω with k-many points, let $P(\cdot \mid F)$ be the perfectly additive probability that is uniform on these k-many points. These conditional probabilities create a single tier, $\tau = \Omega$, because $P(\{\omega_1\} \mid \{\omega_1, \omega_2\}) = 0.5$ for each pair of points in Ω.

Consider the two events $E = \{\omega: 0 \leq \omega < 0.9\}$ and its complement with respect to Ω, $E^c = \{\omega: 0.9 \leq \omega < 1\}$, where $P(E) = 0.9$. Let g be the 1-1 (continuous) map between E and E^c defined by $g(\omega) = 0.9 + \omega/9$, for $\omega \in E$. Consider the κ-size partition of Ω by pair-sets, $\pi = \{\{\omega, g(\omega)\}: \omega \in E\}$. By assumption, $P(\{\omega\} \mid \{\omega, g(\omega)\}) = 1/2$ for each pair $\{\omega, g(\omega)\} \in \pi$. But then P is not conglomerable in π.◊Example 2

The theory of regular conditional distributions treats the example differently. We continue Example 2 from that point of view.

Example 2 (continued) Consider the measure space $<\Omega, \mathcal{B}, P>$ as above. Let the random variable $X(\omega) = \omega$, so that $X \sim U[0,1)$, X has the uniform distribution on Ω. In order to consider conditional probability given the pair of points $\{\omega, g(\omega)\}$, let

$$g(X) = (X/9) + 0.9 \qquad \text{if } 0 \leq X < 0.9$$

$g(X) = 9(X - 0.9)$ if $0.9 \le X < 1$.
Define the random variable

$$Y(\omega) = X(\omega) + g(X(\omega)) - 0.9.$$

Observe that $Y \sim U[0, 1.0)$. Also, note that Y is 2-to-1 between Ω and [0.0, 1.0). That is $Y = y$ entails that either $\omega = 0.9y$ or $\omega = 0.1(y + 9)$.

Let the sub-sigma field $\mathcal{A}$ be generated by the random variable Y. The regular conditional distribution relative to this sub-sigma field, $P(\mathcal{B} \mid \mathcal{A})(\omega)$, is a real-valued function defined on Ω that is $\mathcal{A}$-measurable and satisfies the integral equation

$$\int_A P(B \mid \mathcal{A})(w)\, dP(\omega) = P(A \cap B)$$

whenever $A \in \mathcal{A}$ and $B \in \mathcal{B}$.

In our case, then $P[B \mid \mathcal{A}](\omega)$ almost surely satisfies:

$$P(X = 0.9Y \mid Y)(\omega) = 0.9$$

and $$P(X = 0.1(Y + 9.0) \mid Y)(\omega) = 0.1.$$

Thus, relative to the random variable Y, this regular conditional distribution assigns conditional probabilities as if $P(\{\omega\} \mid \{\omega, g(\omega)\}) = 0.9$ for almost all pairs $\{\omega, g(\omega)\}$ with $0 \le \omega < 0.9$. However, just as in the Borel "paradox" (Kolmogorov, 1956), for a particular pair $\{\omega, g(\omega)\}$, the evaluation of $P(\{\omega\} \mid \{\omega, g(\omega)\})$ is not determinate and is defined only relative to which sub-sigma field $\mathcal{A}$ embeds it.

For an illustration of this last feature of the received theory of regular conditional distributions, consider a different pair of complementary events with respect to Ω. Let $F = \{\omega: 0 \le \omega < 0.5\}$ and $F^c = \{\omega: 0.5 \le \omega < 1\}$. So, $P(F) = 0.5$.

Let $f(X) = 1.0 - X$ if $0 < X < 1$.
$= 0$ if $X = 0$.

Analogous to the construction above, let

$$Z(\omega) = |X(\omega) - f(X(\omega))|.$$

So Z is uniformly distributed, $Z \sim U[0, 1)$, and is 2-to-1 from Ω onto [0, 1). Consider the sub-sigma field $\mathcal{A}'$ generated by the random variable Z. Then the regular conditional distribution $P(\mathcal{B} \mid \mathcal{A}')(\omega)$, almost surely satisfies:

$$P(X = 0.5 - Z/2 \mid Z \ne 0)(\omega) = 0.5$$

and $$P(X = 0.5 + Z/2 \mid Z \ne 0)(\omega) = 0.5$$

and for convenience,

$$P(X = 0 \mid Z = 0) = P(X = 0.5 \mid Z = 0) = 0.5.$$

However, $g(.09) = .91 = f(.09)$ and $g(.91) = .09 = f(.91)$. That is, $Y = 0.1$ if and only if $Z = 0.82$. So in the received theory, it is permissible to have

$$P(\omega = .09 \mid Y = 0.1) = 0.9$$

as evaluated with respect to the sub-sigma field generated by Y, and also to have

$$P(\omega = .09 \mid Z = 0.82\}) = 0.5$$

as evaluated with respect to the sub-sigma field generated by Z, even though the conditioning events are the same event.◊ Example 2 (continued)

5 Non-conglomerability for an IP Bounded Density Ratio model

Our focus in this note is on non-conglomerability for a single, σ-additive but non-κ-additive probability P that has conditional probabilities according to Definition 1, and where some non-null tier τ (i.e., $P(\tau) > 0$) is composed of null points from Ω. We highlight this case as we think it typifies how conditional probabilities given finite set of points are associated with familiar continuous statistical models. Thus, we have demonstrated non-conglomerability in a particular partition for what we judge is the usual interpretation of conditional probabilities from a single continuous, countably additive probability distribution.

The *Proposition* applies to each element of an IP model, when that model uses conditional probabilities from a countably additive, continuous probability that satisfy Definition 1. This puts pressure, we think on those who (e.g., Walley, 1991) appear to require conglomerability in arbitrary partitions as a condition for coherent IP degrees of belief. Here is a Corollary to the Proposition illustrating the point.

Let $\mathcal{P}$ be a set of countably additive, but not κ-additive probabilities. Assume each $P \in \mathcal{P}$ is defined on a common measurable space $\{\Omega, \mathcal{B})$, where the points of Ω are the atoms of $\mathcal{B}$, and where each P has conditional probabilities $P(\cdot \mid \cdot)$ satisfying Defintion 1. Assume that $\mathcal{P}$ satisfies the following *Bounded Density Ratio* [BDR] condition, which is a weakened variant of DeRobertis and Hartigan's (1981) *Density Ratio* model:

- **BDR** There exist a set $T \subseteq \Omega$ where,

(1) T can be partitioned into two sets T_0, T_1 with

$$|T_0| = |T_1| = \kappa \text{ and } \mathit{Inf}_{P \in \mathcal{P}}[P(T_1)] = d > 0.$$

(2) For each pair, $\omega_\alpha \ne \omega_\beta \in T$,

$$\mathit{Sup}_{P \in \mathcal{P}}\,[\, P(\{\omega_\alpha\} \mid \{\omega_\alpha, \omega_\beta\})\,] < 1$$

Note that the BDR condition requires only that the probability distributions that belong to $\mathcal{P}$ have bounded relative densities with respect to pairs of atoms from $\mathcal{B}$. As a consequence of the BDR condition, with respect to each $P \in \mathcal{P}$, the distinguished P-non-null set T belongs to one P-non-null tier.

Corollary: When $\mathcal{P}$ is an IP BDR model, then $\mathcal{P}$ fails to be conglomerable. Specifically, there exits a κ-sized partition by finite sets, $\pi = \{h_\alpha: |h_\alpha| < \aleph_0, \alpha < \kappa\}$ where

$Sup_{\,h\in\pi,\, P\in\mathcal{P}}[\, P(T_1 \mid h)\,] \;<\; d = Inf_{\,P\in\mathcal{P}}[\, P(T_1)\,]$.

Proof: The proof of the Corollary parallels the proof of the Proposition, with one change. That difference is in the sets $\rho_{\beta,n}$. For the Corollary, denoting these by $\rho'_{\beta,n}$, we define them inductively as follows.
Let $\rho'_{1,n} = \{\omega \in T_0$:

$(n\text{-}1)/n \;\leq\; Sup_{\,P\in\mathcal{P}}[P(\{\omega_1\} \mid\{\omega_1, \omega\}) \;<\; n/(n+1)\}$

for n = 1, 2, … . By BDR(2), the sets $\{\rho'_{1,n}: n = 1, 2, \ldots\}$ partition T_0.

Consider the least n* such that ρ'_{1,n^*} is infinite. Let $U_1 = \{\omega_{1,1}, \ldots, \omega_{1,m}\}$ be m-many points chosen from ρ'_{1,n^*}. Note that for each $P \in \wp$ $P(\{\omega_1\} \mid U_1 \cup \{\omega_1\}) \leq n^*/(m+n^*)$. Choose m sufficiently large so that $n^*/(m+n^*) < d/2$. Let $h_1 = U_1 \cup \{\omega_1\}$. So,

$Sup_{\,P\in\mathcal{P}}[\, P(T_1 \mid h_1)\,] \;\leq\; d/2$

Define h_β, by induction, just as in the proof of the Proposition. For $\beta < \kappa$, define $T_{0,\beta} = T_0 - (\cup_{0<\alpha<\beta}\, h_\alpha)$. Consider the countable partition of the set $T_{0,\beta}$ into sets $\rho'_{\beta,n} = \{\omega \in T_{0,\beta}$:

$P(n\text{-}1)/n \;\leq\; Sup_{\,P\in\mathcal{P}}[\, P(\{\omega_\beta\} \mid\{\omega_\beta, \omega\})\,] \;<\; n/(n+1)\}$

for n = 1, 2, … . The proof of the Corollary then follows the proof of the Proposition, resulting in the required partition π. ◊ Corollary

6 Concluding Remarks

In a different paper (2012), we investigate the question of non-conglomerability for a single countably additive but κ-non-additive probability where no set of P-null points forms a P-non-null tier. Though the mathematics for analyzing this case is rather different from the reasoning used in the Proposition presented here, we point the reader to some interesting features about tiers that we use to address this other case.

Definition 10: Consider the relation, ~, of relative-non-nullity on pairs of points in Ω. That is, for two different points, $\omega_1 \neq \omega_2$ they bear the relation $\omega_1 \sim \omega_2$ provided that
$0 < P(\{\omega_1\}\mid \{\omega_1, \omega_2\}) < 1$.
We make ~ into an equivalence relation by stipulating that, for each point ω, ω ~ ω.

Next we state and prove an elementary fact.

Fact: ~ is an equivalence relation.
Proof: Only transitivity requires verification. Assume $\omega_1 \sim \omega_2 \sim \omega_3$. That is, assume
$0 < P(\{\omega_1\} \mid \{\omega_1, \omega_2\}), P(\{\omega_2\}\mid \{\omega_2, \omega_3\}) < 1$.
Then by (iii) of Definition 1 for conditional probability:
$P(\{\omega_1\}\mid \{\omega_1, \omega_2, \omega_3\}) =$
$\quad P(\{\omega_1\}\mid \{\omega_1, \omega_2\})\, P(\{\omega_1, \omega_2\} \mid \{\omega_1, \omega_2, \omega_3\})$.
Also, $P(\{\omega_3\}\mid \{\omega_1, \omega_2, \omega_3\}) =$
$\quad P(\{\omega_3\} \mid \{\omega_2, \omega_3\})\, P(\{\omega_2, \omega_3\} \mid \{\omega_1, \omega_2, \omega_3\})$.
Now argue indirectly by cases.
If $P(\{\omega_1\} \mid \{\omega_1, \omega_3\}) = 0$,
then $P(\{\omega_1\} \mid \{\omega_1, \omega_2, \omega_3\}) = 0$
and $P(\{\omega_1, \omega_2\} \mid \{\omega_1, \omega_2, \omega_3\}) = 0$,
since, by assumption. $P(\{\omega_1\} \mid \{\omega_1, \omega_2\}) > 0$.
Then $P(\{\omega_2\}\mid \{\omega_1, \omega_2, \omega_3\}) = 0 = P(\{\omega_2\} \mid \{\omega_2, \omega_3\})$,
which contradicts $\omega_2 \sim \omega_3$.
If $P(\{\omega_1\} \mid \{\omega_1, \omega_3\}) = 1$,
then $0 = P(\{\omega_3\} \mid \{\omega_1, \omega_3\}) = P(\{\omega_3\} \mid \{\omega_1, \omega_2, \omega_3\})$.
Then $0 = P(\{\omega_2, \omega_3\} \mid \{\omega_1, \omega_2, \omega_3\})$,
since $0 < P(\{\omega_3\}\mid \{\omega_2, \omega_3\})$.
So, $0 = P(\{\omega_2\}\mid \{\omega_1, \omega_2, \omega_3\}) = P(\{\omega_2\} \mid \{\omega_1, \omega_2\})$,
which contradicts $\omega_1 \sim \omega_2$.
Hence $0 < P(\{\omega_1\} \mid \{\omega_1, \omega_3\}) < 1$, as required. ◊ Fact

Thus, the equivalence relation ~ partitions Ω into disjoint *tiers* τ of relative non-null pairs of points. For each pair of points $\{\omega_1, \omega_2\}$ that belong to different tiers, $\omega_i \in \tau_i$ (i = 1, 2), when $\tau_1 \neq \tau_2$, then $P(\{\omega_1\} \mid \{\omega_1, \omega_2\}) \in \{0,1\}$. If $P(\{\omega_2\} \mid \{\omega_1, \omega_2\}) = P(\{\omega_3\} \mid \{\omega_2, \omega_3\}) = 1$, then $P(\{\omega_3\} \mid \{\omega_1, \omega_3\}) = 1$. Thus, the tiers are linearly ordered by the relations ↑, ↓ defined as:
Definitions 11:

- $\tau_1 \uparrow \tau_2$ if for each pair $\{\omega_1, \omega_2\}$, $\omega_i \in \tau_i$ (i = 1, 2), $P(\{\omega_2\} \mid \{\omega_1, \omega_2\}) = 1$.

The reverse ordering also is linear. We express this as

- $\tau_2 \downarrow \tau_1$ if for each pair $\{\omega_1, \omega_2\}$, $\omega_i \in \tau_i$ (i = 1, 2), $P(\{\omega_2\} \mid \{\omega_1, \omega_2\}) = 1$.

That is, $\tau_2 \downarrow \tau_1$ if and only if $\tau_1 \uparrow \tau_2$.

Next, consider the possibly empty set of P-non-null points. Let $\tau^* = \{\omega: P(\omega) > 0)$. Evidently, when $\varnothing \neq \tau^* \neq \tau$, then $\tau^* \downarrow \tau$, and τ* is the top element in the linear order of tiers.

We note that this linear order of tiers plays an important role in Dubins (1975) proof of the existence of fully defined finitely additive conditional probabilities, i.e., where $\mathcal{B}$ is the powerset of Ω and P(B | A) is well-defined whenever $\varnothing \neq$A, B are elements of $\mathcal{B}$. Also, it appears in both Levi's (1980, §5.5) and Regazzini's (1985) strengthened version of de Finetti's criterion of coherence for conditional previsions. Levi and Regazzini strengthen de Finetti's coherence criterion for a called off gamble given a null event in order to have coherent conditional previsions that satisfy Definition 1.

Under additional structural assumptions about $\mathcal{B}$, including measurability of the intervals of tiers formed under $\downarrow$, in our (2012) we extend the Proposition to include non-conglomerability for such cases as well. This permits us to conclude that the anomalous phenomenon of non-conglomerability is a result of adopting the de Finetti/Dubins theory of conditional probability instead of the rival Kolmogorovian theory of regular conditional distributions. Non-conglomerability is not a result primarily of the associated debate over whether probability is allowed to be merely finitely additive rather than satisfying countable additivity.

Restated, our conclusion is the observation that (subject to structural assumptions on the algebra $\mathcal{B}$) even when P is λ-additive for each $\lambda < \kappa$, if P is not κ-additive and has conditional probabilities that satisfy Definition 1, then P will experience non-conglomerability in a κ-sized partition. And then such conditional probabilities will not satisfy condition (3) of the theory of regular conditional distributions.

On the other hand, regular conditional distributions avoid non-conglomerability by allowing conditional probability to depend upon a sub-sigma field, rather than being defined given an event. And, occasionally, they avoid non-conglomerability by abandoning the requirement of *Propriety*, which is clause (ii) of Definition 1 of the de Finetti/Dubins theory of conditional probabilities.

Evidently, some countably additive continuous IP models that use the theory of conditional probabilities associated with Definition 1 require non-conglomerability in specific, uncountable partitions. We think this is a better alternative than using IP models with conditional probabilities based on the theory of regular conditional distributions. In future work on IP models with conditional probabilities, we hope to address the following question:

- With respect to a given IP model that use conditional probabilities, in the sense of Definition 1, in which partitions is non-conglomerability mandated?

Acknowledgments

In writing this paper we have benefitted by helpful comments from Jeremy Avigad, Jessi Cisewski, Paul Pedersen, Rafael Stern, Wilfried Sieg and three anonymous referees.

References

Berti, P. and Rigo, P. (2007) "0-1 Laws for Regular Conditional Distributions," *Ann. Prob.* **35**: 649-662.

Billingsley, P. (1986) *Probability and Measure*, 2nd ed. N.Y.: John Wiley.

Blackwell, D. (1955) "On a class of probability spaces," *Proc. Third Berkeley Symp. Math. Statist. Prob.*, pp 1-6. University of California Press.

Blackwell, D. and Ryll-Nardzewski, C. (1963) "Non-existence of everywhere proper conditional distributions," *Annals of Mathematical Statistics* **34**, 223-225.

Blackwell, D. and Dubins, L.E. (1975), "On existence and non-existence of proper, regular, conditional distributions," *Annals of Probability* **3**, 741-752.

Breiman, L. (1968) *Probability*. Mass.: Addison-Wesley.

deFinetti, B. (1974) *Theory of Probability*. N.Y: Wiley.

DeRobertis, L, and Hartigan, J. (1981) "Bayesian inference using intervals of measures," Ann. Stat. **9**: 235-244.

Dubins, L.E. (1975) "Finitely Additive conditional probabilities, conglomerability, and disintegrations," *Ann. Prob*. **3**, 89-99.

Doob, J.L.(1953) *Stochastic Processes*. N.Y.: Wiley.

Halmos, P. (1950) *Measure Theory*. NY: van Nostrand.

Hewitt, E. and Savage, L.J. (1955) "Symmetric measures on Cartesian products," *A.M.S. Trans.* **80**, 470-501.

Kadane, J.B., Schervish, M.J., and Seidenfeld, T. (1986) "Statistical implications of finitely additive probability," in *Bayesian Inference and Decision Techniques with Applications*. Eds. P.K Goel and A.Zellner, New York: Elsevier, pp. 59-76.

Kadane, J.B., Schervish, M.J., and Seidenfeld, T. (1996) Reasoning to a Foregone Conclusion. *J. American Statistical Association* **91**: 1228-1235.

Kolmogorov, A.N. (1956) *Foundations of the Theory of Probability*. New York: Chelsea.

Levi, I. (1980) *The Enterprise of Knowledge*. Cambridge, MA: MIT Press.

Loeve, M. (1955) *Probability Theory*. NY: van Nostrand.

Regazzini, E. "Finitely additive conditional probabilities," *Milan J. Mathematics*, **55** (1):69–89, 1985.

Schervish, M.J., Seidenfeld, T, and Kadane, J.B. (1984), "The extent of non-conglomerability of finitely additive probability," *Z.War*. **66**, 205-226.

Seidenfeld, T., Schervish, M.J. and Kadane, J.B. (2001) "Improper Regular Conditional Distributions," *Annals of Probability* **29**:1612-1624.

Seidenfeld, T., Schervish, M.J., and Kadane, J.B. (2012) "Non-conglomerability for countably additive measures that are not κ-additive," Technical Report available at http://www.hss.cmu.edu/philosophy/faculty-seidenfeld.php

Ulam, S., (1930) Zur Masstheorie in der allgemeinen Mengenlehre, *Fund. Math*. **16**: 140-150.

Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Conflict and Ambiguity: Preliminary Models and Empirical Tests

Michael Smithson
The Australian National University, Canberra, Australia
Michael.Smithson@anu.edu.au

Abstract

The proposition that conflict and ambiguity are distinct kinds of uncertainty remains debatable, although there is substantial behavioral and some neurological evidence favoring this claim. Recently formal decisional models that combine ambiguity and conflict have been proposed. This paper presents empirical tests of four hypotheses and five models of uncertainty judgments under ambiguity and conflict, via comparisons between pairs of conflicting and ambiguous interval estimates by a sample of 395 adults. The main findings are as follows.

1. Human judges see conflict even in nested intervals with identical midpoints and symmetrically differing endpoints.
2. Identical envelopes of intervals may not be perceived as equally conflictive. Moreover, sets of intervals whose average widths are identical may not be perceived as equally ambiguous.
3. Perceived degree of conflict does not necessarily covary with the magnitudes of the differences between corresponding pairs of interval endpoints. Indeed, a nested pair of intervals may be regarded as more conflictive than a non-nested overlapping pair whose pairs of endpoints differ identically to the nested pair.
4. Judgments of degrees of conflict and ambiguity both contribute independently to judgments of overall uncertainty. However, judgments of ambiguity and conflict appear to be positively correlated.

None of the models pass all empirical tests, but specific suggestions for improving the models are derived from the findings.

Keywords. Uncertainty, ambiguity, conflict, judgment, decision.

1 Introduction

Whether conflict and ambiguity are distinct kinds of uncertainty remains an open question, as does their joint impact on judgments of overall uncertainty. There is behavioral evidence (Smithson 1999, Cabantous 2007, Cabantous et al. 2011, Baillon et al. 2012) and some neurological evidence (Pushskarskaya et al. 2013) in favor of the notion that conflict and ambiguity are separate. However, there are generalized probability frameworks that deal in sets of probabilities, where this distinction appears unnecessary or irrelevant.

Recently formal models of decision making under conflict and ambiguity have been proposed (Gajdos & Vergnaud 2012) that include separate parameters to represent orientations towards conflictive and ambiguous uncertainties. Such models can differ in important ways that are amenable to empirical tests by human judges. In so doing, we must simultaneously investigate judges' understandings of the terms "conflict" and "ambiguity" and how those understandings translate into judgments of uncertainty. Thus, this study is in the genre of the literature on people's numerical interpretations of verbal probability expressions. Here, we shall examine simple comparisons between interval estimates, where the intervals may or may not overlap, and we will focus on four questions:

1. Do nested intervals (special case: identical midpoints) imply no conflict?
2. Do identical envelopes of intervals imply equal conflict and/or equal ambiguity? What about identical interval averages?
3. Does conflict covary with the magnitudes of the differences between corresponding pairs of interval endpoints?
4. Do judgments of degrees of conflict and ambiguity both contribute independently to judgments

of overall uncertainty?

The rationale for questions 1-3 is that conventional pooling rules for sets of quantitative estimates may yield “yes” and “no” answers to these questions. For example, two equally credible interval estimates $[1, 7]$ and $[3, 5]$ may be averaged to yield a pooled interval estimate $[2, 6]$, the same result if both interval estimates were identical intervals $[2, 6]$. So this example could be interpreted as answering “yes” to questions 1 and the average interval version of 2.

A second example, two interval estimates $[1, 5]$ and $[3, 7]$, also may be averaged to yield $[2, 6]$. This example would seem to answer “yes” to question 3 when we compare it to the first example, because in both examples the magnitude of the difference between the lower endpoints is 2 and so is the difference between the upper endpoints. The same comparison also answers “yes” to the identical envelopes version of question 2. But now consider the pair of intervals $[0, 4]$ and $[4, 8]$. Averaging them yields $[2, 6]$ again, despite the fact that their lower and upper endpoints differ by 4 instead of 2. Given that both intervals have the same widths as those in the second example so they are equally ambiguous, it would seem that the degree of conflict does not covary with these differences and this example says “no” to question 3.

A more risk-averse pooling rule that stipulates taking the minimum of the lower endpoints and the maximum of the upper endpoints of equally credible interval estimates says “no” to questions 1 and 2. Pooling intervals $[2, 6]$ and $[3, 5]$ with this rule yields $[2, 6]$, the same result if both interval estimates were identical intervals $[2, 6]$. Clearly the first pair of intervals is, on average, less ambiguous than the second, so perhaps the first pair has some degree of conflict whereas the second identical pair, of course, does not. The lesser ambiguity is then compensated by the greater conflict to yield the same overall uncertainty in the pooled interval. So we have “no” to questions 1 and the identical envelopes version of question 2.

The rationale for question 4 stems from behavioral evidence (Smithson 1999, Cabantous 2007 and Baillon et al. 2012) and recent neurological evidence (Pushkarskaya et al. 2013) that people treat uncertainty arising from conflicting information as distinct from uncertainty arising from ambiguity. Even granting this claim, it is not clear how people combine the two kinds of uncertainty if asked to evaluate the overall uncertainty of a prospect.

Simple empirical tests of all four questions can be constructed by two-alternative forced-choice experiments in conjunction with simple models incorporating each hypothesis. In the next section we shall see that reasonable models of ambiguity and conflict can be constructed to yield “yes” and “no” answers to questions 1-3.

2 Models

Suppose that K judges provide estimates of a quantity of the form $[p_{k1}, p_{k2}, \ldots, p_{kJ}]$, where the p_{kj} are order statistics: $p_{k1} < p_{k2} < \ldots < p_{kJ}$. The simplest setup of this kind, which we shall consider, has two judges, each of whom provides a lower and upper estimate, so that $K = 2$ and $J = 2$.

The k^{th} judge’s assessment is ambiguous or vague insofar as the p_{kj} diverge in some sense from one another, and we will consider functions $A(p_{kj})$ to measure ambiguity. Likewise, judges’ assessments may conflict with one another insofar as their assessments differ in some sense from each other, and we will also consider functions $C(p_{kj})$ to measure conflict. Finally, a decision maker (DM) who is given these judges’ assessments may have a subjective appraisal of the combined uncertainty resulting from both ambiguity and conflict that weighs these two uncertainty components according to their relative aversiveness to the DM. We will therefore investigate uncertainty functions $S(\alpha, \theta, C(p_{kj}), A(p_{kj}))$ that are monotonically increasing in $C(p_{kj})$ and $A(p_{kj})$, where α is the conflict weight and θ is the ambiguity weight.

2.1 Model Types

2.1.1 Variance Component Models

A natural uncertainty metric for both ambiguity and conflict could be variance. Ambiguity effects on judgments and decisions have been explained in terms of variance (Rode et al. 1999), and conflict also has implications for variability in outcomes. The ambiguity of each judge’s estimates can be measured by

$$A_k = \sum_{j=1}^{J} (p_{kj} - \overline{p_{k.}})^2 \Big/ J, \tag{1}$$

so that the total ambiguity is just the within-judge component of the variance of the p_{kj}:

$$A = \sum_{k=1}^{K} A_k / K.$$

An intuitively plausible candidate for measuring conflict, then, is the between-judge variance component:

$$C_1 = \sum_{k=1}^{K} (\overline{p_{k.}} - \overline{p_{..}})^2 \Big/ K \tag{2}$$

However, an alternative conflict measure is the variance among the order-statistics of the same rank:

$$C_2 = \sum_{k=1}^{K}\sum_{j=1}^{J}(p_{kj} - \overline{p_{.j}})^2 \Big/ JK. \tag{3}$$

I shall refer to the first model as variance component model 1 (VC1) and the second as VC2. The conflict function in equation 3 differs from that in equation 2 in an important way, because when $\overline{p_{k.}}$ are identical for all K judges, $C_1 = 0$ whereas this is not true for C_2. Thus, VC1 predicts that a pair of interval estimates with identical midpoints will not be perceived as conflictive, whereas VC2 predicts that they will be.

A DM's degree of concern or disutility about ambiguity is represented by a weight, θ, that takes values in the closed unit interval. Likewise, the DM's degree of concern about disagreement or conflict is represented by a weight, α, whose domain also is the unit interval. There are several ways these weights may be employed to combine the ambiguity and conflict measures to construct a measure of overall uncertainty. The simplest is a weighted sum:

$$S(\alpha, \theta, A, C_j) = \theta A + \alpha C_j, \tag{4}$$

where $j = 1, 2$.

2.1.2 Distance Models

Distance models are related to variance models and provide another potential metric for both ambiguity and conflict. A distance model evaluates ambiguity and conflict in terms of distances between order statistics. The ambiguity of the k^{th} judge can be expressed as

$$A_k = \sum_{j_1=1}^{J}\sum_{j_2=1}^{J}|p_{kj_1} - p_{kj_2}|^n / J^2 \tag{5}$$

where $n > 0$ ($n = 2$ is the Euclidean special case). As before, the total ambiguity then is simply

$$A = \sum_{k=1}^{K} A_k / K.$$

Conflict between the judges may be evaluated in two ways. First, we may sum those differences over the ranks and take the absolute value of that sum:

$$C_1 = 2\sum_{k_1=1}^{K}\sum_{k_2=k_1+1}^{K}\left|\sum_{j=1}^{J}(p_{k_1 j} - p_{k_2 j})^n\right| \Big/ K\,(K-1). \tag{6}$$

Second, we may sum the absolute differences between pairs of order-statistics of the same rank:

$$C_2 = 2\sum_{k_1=1}^{K}\sum_{k_2=k_1+1}^{K}\sum_{j=1}^{J}|p_{k_1 j} - p_{k_2 j}|^n \Big/ K\,(K-1). \tag{7}$$

I shall refer to the first model as distance model 1 (D1) and the second as D2. As with the previous pair of models, D1 predicts that a pair of interval estimates with identical midpoints will not be perceived as conflictive, whereas D2 predicts that they will be.

As with the variance models, it is possible to combine A and C_j in a weighted sum to produce an overall evaluation of total uncertainty. The result is equation (4) with the weights expressing degrees of disutility regarding ambguity and conflict, and the distance model versions of A and C_j substituted for the variance models versions.

2.1.3 The Gajdos-Vergnaud Model

Gajdos and Vergnaud (2012) develop a model of decision making under ambiguity and conflict based on the Schmeidler-Gilboa (1989) maxmin framework. For the sake of simplicity, I present only the two-state, two-judge special case of their model, and modify their notation to be compatible with the notation used for the other models in this paper. They intended their model to apply to probability judgments; here I extend it to judgments of magnitudes.

In the Gajdos-Vergnaud (GV) model, the α and θ weights are used to modify the order statistics of each judge. The θ parameter contracts the $[p_{k1}, p_{k2}]$ interval around its midpoint at a rate $1-\theta$, yielding lower and upper bounds

$$\begin{aligned}\pi_{k1} &= p_{k1}(1+\theta)/2 + p_{k2}(1-\theta)/2, \\ \pi_{k2} &= p_{k1}(1-\theta)/2 + p_{k2}(1+\theta)/2.\end{aligned} \tag{8}$$

Gajdos and Vergnaud do not define an ambiguity measure along the lines of those in this paper, but as with the variance and distance models we may construct one by summing the differences $\pi_{k2} - \pi_{k1}$. It can be shown that this ambiguity measure is identical to the distance model's ambiguity measure divided by 2 when $n = 1$.

The GV model treats α as contracting the pairs of interval endpoints p_{kj} and p_{mj} around their mean at the rate $1-\alpha$. Thus, the order statistics are modified in the following way:

$$\begin{aligned}\gamma_{kj} &= p_{kj}(1+\alpha)/2 + p_{mj}(1-\alpha)/2, \\ \gamma_{mj} &= p_{mj}(1+\alpha)/2 + p_{kj}(1-\alpha)/2.\end{aligned} \tag{9}$$

Again, Gajdos and Vergnaud do not define a conflict measure but one may be defined by summing the absolute values of the differences $\gamma_{kj} - \pi_{mj}$. It can be shown that this conflict measure is identical to the distance model's C_2 measure divided by 2 when $n = 1$.

If we evaluate overall uncertainty by summing the ambiguity and conflict measures, clearly we obtain an uncertainty measure identical to that in D2 when $n = 1$. An alternative evaluation of overall uncertainty is suggested by the maxmin decisional model incorporated into the GV framework. In the development of the GV decisional model, the order statistics are transformed by one parameter and then those results are transformed in turn by the second parameter, according to equations (8) and (9). It is not difficult to show that this procedure is commutative, so that if the α transformation occurs before or after the θ transformation the result is the same. Thus, we may define our alternative uncertainty measure by

$$S(\alpha, \theta, A, C) = \max_{k,j}(\gamma_{kj}) - \min_{k,j}(\gamma_{kj}). \qquad (10)$$

As will be demonstrated, this measure does not behave identically to the measure for D2.

3 Method

Hypotheses and the models were tested via an online experiment. The online study was reviewed and approved by the Australian National University Human Research Ethics Committee. The participant sample consisted of 508 North American adults (205 women, 189 men, 1 unspecified; with mean age = 39.95, sd = 15.04), recruited through Qualtrics, of which 395 cases were found to be trustworthy data. Four comparisons between two pairs of estimates, $\{P_1, Q_1\}$ and $\{P_2, Q_2\}$, were used to test questions 1-3, their results also lending insight into question 4. Comparisons 2 and 3 test question 1, Comparisons 3 and 4 test question 2, and Comparisons 2-4 partially test question 3. These comparisons are graphed in Figure 1. Participants were presented with both the graphs and verbal statements of the estimate pairs. They were asked to choose which pair of estimates exhibited more agreement, which exhibited more ambiguity, and which made them feel more uncertain about the quantity being estimated.

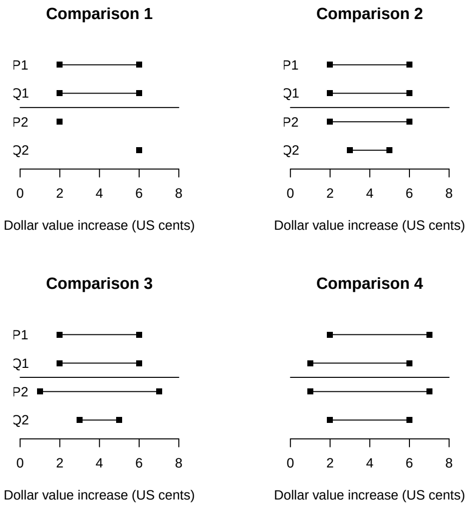


Figure 1: Four Pairs of Judgments

There were two conditions, differing solely on the nature of the estimate. In one condition they were told that the estimates were experts' predictions of the change in global average temperature by the year 2040 (in degrees Celsius). In the other, they were told the estimates were experts' predictions of the change in the value of the Australian dollar against the American dollar in the next 5 years (in US cents). Both scenarios are fictitious, and participants were advised of this in a debriefing at the end of the online survey. Neither of these scenarios is based on expert predictions. The global average temperature scenarios actually are over-estimates of warming, according to the IPCC (2007) report, and the estimates reported therein do not disagree as much as the estimates in some of these scenarios do. Genuine forecasts of currency fluctuations seldom range farther into the future than 6 months to one year, and near-term predictions for the Australian dollar's exchange-rate against the US dollar are mixed with some predicting a decline and others predicting an increase. The goal here was to provide identical numbers under the guise of very different topics, to ascertain whether topics might influence perceptions of conflict or ambiguity. All of this having been said, the topic of the estimates turned out to make no significant difference to people's choices, so from here on these two conditions are ignored.

An example of the text of the first condition is presented here.

> In this section, we want you to make some judgments about estimates of the increase in average temperature by the year 2040. You will be presented with two pairs of estimates from refereed climate science forecasts. We are interested in which pair you think has the greatest uncertainty.
> Expert P1: By 2040 global average temperature will have increased by 2-6 degrees Celsius
> Expert Q1: By 2040 global average temperature will have increased by 2-6 degrees Celsius
> Expert P2: By 2040 global average tem-

> perature will have increased by 2 degrees Celsius
> Expert Q2: By 2040 global average temperature will have increased by 6 degrees Celsius
> Taken together, which pair of experts do you think is in more agreement?
> Taken together, which pair of experts do you think is more vague?
> Taken together, which pair of experts makes you more uncertain about the temperature increase?

The models presented in the previous section all agree that in Comparison 1, $\{P_1, Q_1\}$ is more ambiguous and less conflictive than $\{P_2, Q_2\}$, and whether one is rated as more uncertain overall depends on the magnitudes of the α and θ parameters. For Comparison 2, all models agree that $\{P_1, Q_1\}$ is more ambiguous than $\{P_2, Q_2\}$, but GV, D2, and VC2 rate $\{P_1, Q_1\}$ as less conflictive than $\{P_2, Q_2\}$ whereas D1 and VC1 rate them as equally conflictive. In Comparison 3 the models make the same predictions about conflict as in Comparison 2, but while VC1 and VC2 rate $\{P_1, Q_1\}$ as less ambiguous than $\{P_2, Q_2\}$, GV, D1 and D2 rate them as equally ambiguous. Finally, for Comparison 4, the models' predictions regarding ambiguity are the same as in Comparison 3, but D1 and VC1 rate $\{P_1, Q_1\}$ as more conflictive than $\{P_2, Q_2\}$ whereas GV, D2, and VC2 rate them as equally conflictive.

Overall uncertainty predictions from the models are not determined for all four comparisons because they may vary with the α and θ parameters. Nevertheless, for every model at least two comparisons yield fixed outcomes. In Comparison 2, GV, D1 and VC1 rate $\{P_1, Q_1\}$ as more uncertain than $\{P_2, Q_2\}$. In Comparison 3, all models except D1 rate $\{P_1, Q_1\}$ as less uncertain than $\{P_2, Q_2\}$; D1 rates them as equally uncertain. In Comparison 4, GV amd D1 rate $\{P_1, Q_1\}$ as more uncertain than $\{P_2, Q_2\}$, VC1 amd DVC2 rate $\{P_1, Q_1\}$ as less uncertain than $\{P_2, Q_2\}$, and D2 rates them as equally uncertain.

4 Results

4.1 Questions 1-3

Regarding question 1, in Comparisons 2 and 3 large majorities of respondents chose the nested interval pair as being more conflictive than the identical interval pair. For Comparison 2, 83.8% made this choice (95% confidence interval (CI) = $[79.8\%, 87.1\%]$); and for Comparison 3, 87.6% made this choice (95% CI = $[84.0\%, 90.5\%]$). These figures are similar to the percentage choosing the two pointwise estimates in Comparison 1 as more conflictive than the identical intervals (84.3%). An unexpected finding was that in Comparison 4, 61.5% chose the nested interval pair as more conflictive than the non-nested, overlapping pair (95% CI = $[56.6\%, 66.2\%]$). These results all strongly suggest that nested interval estimates are perceived as conflictive even when they have identical midpoints.

The finding regarding conflict in Comparison 4 also addresses questions 2 and 3, indicating that neither identical envelopes nor equal differences between pairs of endpoints will ensure that pairs of estimates will be regarded as equally conflicting. The finding for Comparison 3 demonstrates that identical average interval widths for pairs of estimates also will not ensure that they are perceived as equally conflicting.

Question 2 applied to ambiguity, on the other hand, yielded mixed results. In Comparison 4, where the pairs have identical envelopes, we cannot rule out the possibility that respondents were evenly split on which pair is the more ambiguous (95% CI = $[47.0\%, 56.8\%]$). However, in Comparison 3 where the average interval widths are the same for both pairs, 78.0% chose the nonidentical pair of intervals as more ambiguous than the identical pair (95% CI = $[73.6\%, 81.8\%]$).

Question 3 also can be addressed via a test for marginal homogeneity in the cross-classification of Comparison 2 and 3 choices regarding conflict. In Comparison 2 the first pair of intervals was chosen as more agreeing by 83.8% of respondents and in Comparison 3 87.6% chose the first pair. Because the widths of the nested pair in Comparison 2 differ by less than those in Comparison 3 we should indeed expect a higher percentage in Comparison 3. However, a 95% CI for the paired difference yields $[-0.11\%, 7.75\%]$ so we fail to reject the null hypothesis of no difference.

4.2 Overall Uncertainty and Question 4

Comparison 1 offers indirect corroboration of Smithson's (1999) conflict aversion hypothesis, because 60.0% of respondents chose the pointwise pair of disagreeing estimates as more uncertain than the pair of agreeing interval estimates (95% CI = $[54.1\%, 63.7\%]$). A similar percentage, 58.0%, chose the nested pair of intervals in Comparison 2 as more uncertain than the pair of agreeing interval estimates (95% CI = $[53.1\%, 62.7\%]$), and a substantially greater percentage, 76.7%, made the same choice in Comparison 3 (95% CI = $[72.3\%, 80.6\%]$). Finally, in Comparison 4 55.2% chose the nested pair of intervals as more uncertain than the overlapping pair (95% CI = $[50.3\%, 60.0\%]$). These latter three findings indicate that both perceived conflict and ambiguity may be independently contributing to overall perceived uncertainty.

A direct test of this (i.e., question 4) is a mixed logistic regression model utilizing the data from all four comparisons. This model included main-effects terms for comparisons, ambiguity and agreement (conflict), with random effects for the latter two covariates. A model with interaction terms did not improve fit significantly ($\chi^2(6) = 9.642, p = .141$). Both the agreement and ambiguity terms were significant in the expected directions ($z = -6.576, p < .0005$ and $z = 12.568, p < .0005$, respectively), the ambiguity effect being nearly twice as large.

4.3 Model Performance

The performance of the five models can be evaluated in two ways. First, we can simply assign each a "pass" or "fail" grade for every prediction made by each model regarding comparative conflict, ambiguity, or uncertainty. Second, for ambiguity and conflict we may use the differences between the scores each model assigns to every relevant pair of estimates to predict respondent choices via mixed logistic regressions.

Table 1 summarizes the "pass" or "fail" results. Only Comparisons 2-4 are shown because all models passed Comparison 1 on conflict and ambiguity and made no determinate predictions for uncertainty. "P" indicates that the model's prediction is in accordance with the empirical result; "F" indicates that the model's prediction is the opposite of the result; "N" that the model's prediction is equality whereas the result suggests a difference; and "U" that the status of the model's prediction is undetermined by the result because the null hypothesis could not be rejected.

Table 1: Model Pass-Fail Results

	Conflict			Ambiguity			Uncertainty		
	2	3	4	2	3	4	2	3	4
GV	P	P	N	U	N	P	F	P	F
D1	N	N	F	U	N	P	F	N	F
D2	P	P	N	U	N	P		P	N
VC1	N	N	F	U	P	U	F	P	F
VC2	P	P	N	U	P	U		P	P

Beginning with the conflict results, models D1 and VC1 fail three of the four comparisons because they are the models predicting that pairs of intervals with identical midpoints will not be considered to be conflicting. The other models pass three of the four comparisons. None of the models pass Comparison 4. The ambiguity results are equivocal, with all models passing two comparisons and none performing markedly better than the others. The uncertainty results also are mixed. No model with a determinate prediction passes Comparison 2, all but one pass Comparison 3, and only one (VC2) passes Comparison 4.

We now turn to the mixed logistic regressions. The D1 and VC1 models' conflict scores for the five distinct pairs of estimates used in the comparisons are proportional to one another, and the GV, D2 and VC2 models' conflict scores are proportional to one another. So there are two mixed logistic regressions to compare: The D1-VC1 and GV-D2-VC2 models. The log-likelihood of the GV- D2-VC2 model is -1073.39 and, as might be expected, markedly higher than the log-likelihood of the D1-VC1 model (-1083.40).

As with the conflict scores, the ambiguity scores for the D1 and VC1 models are proportional to one another and the ambiguity scores for the GV, D2 and VC2 models are proportional to one another. The The log-likelihoods of the GV- D2-VC2 and D1-VC1 models are fairly similar (-1336.76 and -1332.28 respectively).

5 Discussion

The results strongly indicate that the answer to question 1 is "no", at least for the rather de-contextualized comparisons used in this study. Even so, I urge caution regarding generalizability, having witnessed at least one applied context (a consultancy with a banking organization) in which stakeholders decided that nested estimates should *not* be considered as disagreeing. Other contextual factors could alter the answer to this question. Fior example, it is plausible that if one estimate is known to be based on a larger data set than the other, nested intervals might not be taken to indicate disagreement but instead attributed to the different sizes of the data sets.

Likewise, the conflict comparison results suggest the answer to question 2 is "no". However, the ambiguity comparisons are inconclusive regarding this question, and further investigations will be required to ascertain the conditions under which identical envelopes of intervals confer equal ambiguity. As indicated above, additional information about the basis for the estimates could alter this outcome as well.

Question 3 also has been answered in the negative, both in the failure to find a significant difference between choices in Comparisons 2 and 3, and in another unexpected fashion. None of the models or pooling rule considerations anticipated the finding in Comparison 4 that a nested pair of intervals would be regarded as more conflictive than a non-nested overlapping pair whose pairs of endpoints differed identically to the nested pair. This finding begs for interpretation, and that will be addressed shortly.

The mixed logistic regression demonstrated that both conflict and ambiguity choices made independent contributions to predicting uncertainty choices between pairs of estimates in the four comparisons. Moreover, the type of comparison did not significantly moderate the effect of either conflict or ambiguity. Thus, respondents generally behaved as though they perceived ambiguity and conflict as distinct contributors to overall uncertainty.

Nonetheless, this result suggests another question, namely whether ambiguity and conflict choices are associated. In this sample, judgments of ambiguity and agreement are strongly negatively related for all four comparisons (i.e., ambiguity and conflict are positively associated). That is, the odds of choosing $\{P_1, Q_1\}$ as the more ambiguous pair are higher if the respondent also chose $\{P_2, Q_2\}$ as the more agreeable (and vice versa). The odds-ratios for Comparisons 1, 2, 3, and 4 are 2.90, 6.79, 14.13, and 22.84, respectively. For Comparison 1 this finding is somewhat surprising because the pairs of estimates are constructed so that one pair is clearly ambiguous and the other clearly conflicting. It is not as surprising for Comparisons 2 and 4 because there is no definite majority view on which pair of estimates is the more ambiguous in either comparison. However, it is unsurprising for Comparison 3 because substantial majorities of respondents chose the second pair of estimates, $\{P_2, Q_2\}$, as more ambiguous and the first pair as showing more agreement.

The consistency of this relationship suggests that people may regard conflict and ambiguity as entailing one another: The greater the perceived conflict, the greater the perceived ambiguity, and vice-versa. This is not an irrational association to make, given that there are situations where ambiguity can generate conflict or conflict can generate ambiguity.

Table 2 displays the crosstabulations of the choices for all four comparisons. The negative association between the ambiguity and conflict choices is especially clear in Comparisons 2-4, where the majority of respondents who have chosen $\{P_1, Q_1\}$ as the more agreeing pair also have chosen $\{P_2, Q_2\}$ as the more ambiguous, while the majority who have chosen $\{P_2, Q_2\}$ as the more agreeing have chosen $\{P_1, Q_1\}$ as the more ambiguous.

Finally, let us consider the issue of modeling conflict and ambiguity jointly. Starting with ambiguity, as mentioned earlier, none of the models were clearly superior to the others in predicting ambiguity choices. The GV, D2 and VC2 models differ from the D1 and VC1 models in their predictions for Comparisons 3 and 4, so that the first three pass Comparison 4 while the latter two pass Comparison 3. Inspection of Table 2 reveals that D1 and VC1 pass both Comparisons 3 and 4 for those people who chose the first pair of intervals in each comparison as showing more agreement. As mentioned above, in Comparisons 2-4 the majority choice of which pair is more ambiguous switches depending on which pair is seen as showing more agreement. The clear suggestion is to build and test models of conflict and ambiguity assessment that take this positive relationship into account.

Table 2: Ambiguity-Agreement Association

	Agreement		
Ambig.	$\{P_1, Q_1\}$	$\{P_2, Q_2\}$	
$\{P_1, Q_1\}$	173	47	
$\{P_2, Q_2\}$	160	15	Comparison 1
	$\{P_1, Q_1\}$	$\{P_2, Q_2\}$	
$\{P_1, Q_1\}$	129	52	
$\{P_2, Q_2\}$	202	12	Comparison 2
	$\{P_1, Q_1\}$	$\{P_2, Q_2\}$	
$\{P_1, Q_1\}$	52	35	
$\{P_2, Q_2\}$	294	14	Comparison 3
	$\{P_1, Q_1\}$	$\{P_2, Q_2\}$	
$\{P_1, Q_1\}$	57	133	
$\{P_2, Q_2\}$	186	19	Comparison 4

Turning now to conflict, the GV, D2 and VC2 models perform markedly better than the D1 and VC1 models in predicting conflict choices because the latter two models consider nested interval estimates as having no conflict. However, none of the models passed Comparison 4.

One interpretation of the respondents' conflict choices in Comparisons 2-4 is that some people may perceive differences in interval widths as indicating disagreement. Thus, the second pair of estimates in Comparison 4 is doubly penalized for conflict because the endpoints differ and so do the interval widths, whereas in the first pair the endpoints differ by the same amounts but the interval widths agree (i.e., the experts are equally vague).

It is not difficult to amend the conflict models presented thus far to accommodate a penalty for differing vagueness. In the distance and variance component models it simply amounts to adding a distance measure and a variance component, respectively, that accounts for differences in interval widths. Doing so does not alter their predictions for any of the other comparisons, so they now pass Comparison 4. Moreover, their mixed logistic regression log-likelihoods are markedly better than their original counterparts (-1062.63 and -1064.75). The new models also present novel predictions regarding other comparisons and thus suggest specific tests of their validity. These will be undertaken in future experiments.

Readers will have noticed that the estimate scenarios in this study were considerably simplified, omitting any information about how the experts arrived at their estimates, the data on which the estimates were based, the experts' qualifications, and so on. As mentioned earlier in this section, such information can affect perceptions of conflict and ambiguity. For instance, two differing estimates based on separate analyses of the same data set would be likely to be percieved as a more striking conflict than the same two estimates based on separate (but, say, equal-sized) data sets. Likewise, knowledge of two experts' prior (dis)agreements with one another on similar issues could substantially influence perceptions of how strong their current disagreement is. Examples of factors potentially affecting perceptions of ambiguity are the amounts of evidence on which estimates are based and the level of relevant expertise possessed by the estimator. Finally, relevant perceiver characteristics include tolerance of uncertainty, agreeableness, need for closure, and prior alignment with one or another expert's position on issues relevant to the estimates. There is considerable scope, therefore, for experimentally investigating the effects of particular kinds of information and assessing the impacts of psychological covariates on perceptions of conflict and ambiguity.

References

[1] Baillon, A., Cabantous, L. & Wakker, P. (2012) Aggregating imprecise or conflicting beliefs: An experimental investigation using modern ambiguity theories. *Journal of Risk and Uncertainty*, 44, 115-147.

[2] Cabantous, L. (2007) Ambiguity aversion in the field of insurance: insurers attitude to imprecise and conflicting probability estimates. *Theory and Decision*, 62, 219240.

[3] Cabantous, L., Hilton, D., Kunreuther, H., & Michel-Kerjan, E. (2011) Is imprecise knowledge better than conflicting expertise? Evidence from insurers decisions in the United States. *Journal of Risk and Uncertainty*, 42, 211-232.

[4] Gajdos, T. & Vergnaud, J-C. (2012) Decisions with conflicting and imprecise information. *Social Choice and Welfare*, in press.

[5] Intergovernmental Panel on Climate Change, *Summary for policymakers: Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Retrieved May 2010 from http://www.ipcc.ch/pdf/assessment-report/ar4/wg1/ar4-wg1-spm.pdf(2007).

[6] Pushkarskaya, H., Smithson, M., Joseph, J.E., Corbly, C., & Levy, I. (2013) Decision making under ambiguity and conflict. *Unpublished manuscript.*

[7] Rode, C., Cosmides, L., Hell, W., & Tooby, J. (1999). When and why do people avoid unknown probabilities in decisions under uncertainty? Testing some predictions from optimal foraging theory. *Cognition*, 72, 269304.

[8] Smithson, M. (1999). Conflict aversion: preference for ambiguity vs. conflict in sources and evidence. *Organizational Behavior and Human Decision Processes*, 79, 179198.

A Robust Data Driven Approach to Quantifying Common-Cause Failure in Power Networks

Matthias C. M. Troffaes
Durham University, UK
matthias.troffaes@gmail.com

Simon Blake
Newcastle University, UK
simon.blake@ncl.ac.uk

Abstract

The standard alpha-factor model for common cause failure assumes symmetry, in that all components must have identical failure rates. In this paper, we generalise the alpha-factor model to deal with asymmetry, in order to apply the model to power networks, which are typically asymmetric. For parameter estimation, we propose a set of conjugate Dirichlet-Gamma priors, and we discuss how posterior bounds can be obtained. Finally, we demonstrate our methodology on a simple yet realistic example.

Keywords. robust, alpha-factor, failure, reliability, Gamma, Dirichlet

1 Introduction

When modelling power networks, typically, the basic event we are interested in are loss of so-called *security zones*. A security zone makes up a collection of components, so that if one component in the zone fails, power in the whole zone is lost. Security zones are typically bounded by circuit breakers, which allow isolating consequences of faults.

An interesting problem occurs when faults in different zones do not occur independently. For example, power lines in adjacent zones often share transmission towers. A landslide, for instance, can cause the tower to collapse, affecting both zones simultaneously. It is important that the frequency of such events is taken into account, as otherwise the actual risk to the network might be underestimated.

The standard literature for common cause failure modelling assumes symmetry [5, 9], however, clearly, for our purpose, security zones will typically not exhibit symmetry, due to differences in layout, composition, and age of constituents. In this paper, we adapt the approach of Troffaes et al. [9] to allow for asymmetry.

In doing so, as opposed to existing methods [3, 4], we enable a more data driven approach to network reliability analysis. Specifically, we allow actual failure data on the network—which is most informative, but typically also very sparse—to be combined with say national average failure rates—such data is typically far more abundant, but also not necessarily as applicable to the specific network at hand due to specific local conditions which may be hard to model, let alone to be quantified.

A key feature of our approach is built-in sensitivity analysis against ill-known parameters, following [10, 7, 8, 9]. Following recent work on prior-data conflict [11, 12, 9], in this paper, we will focus on sensitivity analysis in the so-called learning parameters of the model, which essentially tells us how much we should weigh network specific data against our prior expectations informed by say national averages.

The paper is structured as follows. Section 2 derives the mathematical model for dealing with common cause failures in asymmetric two component systems. Section 3 discusses the statistical problem of how to estimate the parameters of the model. We construct a likelihood for typical kinds of data available. We then propose a conjugate prior, which is an independent product of a Dirichlet (or beta) prior and two Gamma priors. Finally, we discuss how sensitivity analysis can be performed to obtain posterior bounds. Section 4 works through an actual example. Section 5 concludes the paper.

2 Modelling Common Cause Failure for Asymmetric Components

2.1 Two Component Model

In this discourse, a 'component' denotes any subsystem, which, for the purpose of common cause analysis, we do not subdivide any further. In particular, it does not need to denote a separate electrical component of

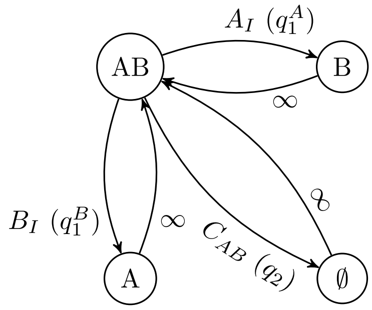


Figure 1: Markov chain for failure with instant repair. The nodes show non-faulty zones.

the power network. For example, if we are merely interested in the loss of security zones, a component could be taken to be such security zone.

Let us call these components A and B. Now, following the basic parameter model of Mosleh et al. [5] (also see [9]), one traditional way to model common cause failures is to attribute all failures to any of the following three events:

- A_I: independent failure of A
- B_I: independent failure of B
- C_{AB}: common cause failure of both A and B

These three events are assumed to be generated by independent Poisson processes. For simplicity, in this exposition, we assume that repair is immediate.[1] Figure 1 depicts the corresponding continuous time Markov chain, along with rates for all transitions.

Following standard notation in common cause failure modelling, by q_1^A we denote the rate of A_I, by q_1^B we denote the rate of B_I, and by q_2 we denote the rate of C_{AB}. The subscript of the q denotes the number of components involved (or is t for 'total', as in the next paragraph). The superscript denotes the particular component, and is required due to lack of symmetry. For comparison, in the standard basic parameter model, we would have $q_1^A = q_1^B = q_1$.

A key challenge is that we do not observe these events directly. Indeed, often, we have a good idea of the rate at which each component fails, that is, we know

$$q_t^A = q_1^A + q_2 \tag{1}$$
$$q_t^B = q_1^B + q_2 \tag{2}$$

Additionally, we may have a fairly good idea of what fraction α_2 of faults is due to a common cause. The fraction of faults not due to a common cause is $\alpha_1 := 1 - \alpha_2$.

For example, say that we have a sequence of 100 independent observations in which a fault occurs, and say that in exactly 18 of those observations, both components failed. Then, to a good approximation, α_2 would simply be 0.18. The parameters α_1 and α_2 are called *alpha-factors.*

So, we have three observable quantities: q_t^A, q_t^B, and α_1—note that $\alpha_2 = 1 - \alpha_1$. From these, we need to derive three model parameters: q_1^A, q_1^B, and q_2. Here, the only difference with the standard basic parameter model in the literature is that we do not assume $q_1^A = q_1^B$ (and whence, also not that $q_t^A = q_t^B$). This difference may seem only very subtle, particularly for the case where only two components are involved, however, the consequent mathematical treatment is notably different to merit a careful consideration, as follows.

We can easily express α_1 and α_2 in terms of the above parameters, once noted that a fraction of faults can be written as a ratio of fault rates:

$$\alpha_1 = \frac{q_1^A + q_1^B}{q_1^A + q_1^B + q_2} \tag{3}$$
$$\alpha_2 = \frac{q_2}{q_1^A + q_1^B + q_2} \tag{4}$$

Now, consider the combination $\alpha_1 + 2\alpha_2$:

$$\alpha_1 + 2\alpha_2 = \frac{q_1^A + q_1^B + 2q_2}{q_1^A + q_1^B + q_2} = \frac{q_t^A + q_t^B}{q_1^A + q_1^B + q_2} \tag{5}$$

Consequently,

$$q_1^A + q_1^B + q_2 = \frac{q_t^A + q_t^B}{\alpha_1 + 2\alpha_2} \tag{6}$$

where the right hand side now consists of observable quantities. Plugging this expression into our earlier expression for α_2, we find:

$$\alpha_2 = \frac{q_2(\alpha_1 + 2\alpha_2)}{q_t^A + q_t^B} \tag{7}$$

so, consequently:

$$q_2 = \frac{\alpha_2}{\alpha_1 + 2\alpha_2}(q_t^A + q_t^B) \tag{8}$$

We recovered one of the model parameters. For the other ones, simply use:

$$q_1^A = q_t^A - q_2 \tag{9}$$
$$q_1^B = q_t^B - q_2 \tag{10}$$

[1] Note that, consequently, simultaneous failures due to independent causes of the two components have probability zero.

2.2 Preliminary Example

To demonstrate the theory so far developed, we apply it on a simple example. Athough in the following, the probabilities are entirely made up, they are representative of typical power networks.

Suppose we have a collection of customers supplied from two security zones, named A and B, where loss of power in both zones will result in customer interruption. Suppose, for the sake of argument, that, on average, per year, we observe 3 faults in zone A, and 5 faults in zone B. We also know that, from historical data, 15% of all faults in these zones results in customer interruption. What is the rate at which we lose customers?

Following the above model, we have:

$$q_t^A = 3 \tag{11}$$
$$q_t^B = 5 \tag{12}$$

assuming rates are expressed per year, and

$$\alpha_1 = 0.85 \tag{13}$$
$$\alpha_2 = 0.15 \tag{14}$$

Then, following the earlier analysis, we find that:

$$q_2 = \frac{\alpha_2}{\alpha_1 + 2\alpha_2}(q_t^A + q_t^B) \tag{15}$$
$$= \frac{0.15}{0.85 + 2 \times 0.15}(3+5) = 1.043 \tag{16}$$

The rate at which customer interruption occurs is exactly $q_2 = 1.043$, or about one per year. Note that we can also derive the rate at which independent failures occur:

$$q_1^A = q_t^A - q_2 = 3 - 1.043 = 1.957 \tag{17}$$
$$q_1^B = q_t^B - q_2 = 5 - 1.043 = 3.957 \tag{18}$$

3 Parameter Estimation from Data

An obvious challenge with our statistical model is that we need to estimate the failure rates of each component (or, security zone), as well as the fraction of double failures. Information relating to these probabilities can come from a variety of sources.

Two options present themselves:

1. Use historical failure data of single and double failures in the network under study to estimate the parameters q_1^A, q_1^B, and q_2, directly, say using maximum likelihood. A problem here is that, typically, for one specific network, not very much data may be available.
2. Use average nationwide failure rates q_t^A and q_t^B, along with average nationwide double failure fraction α_2. The methodology of Section 2.1 then applies to find q_1^A, q_1^B, and q_2. As there is far more nationwide data available, one would hope that this leads to more accurate estimates for q_t^A, q_t^B, and α_2. A key problem here is that it is not clear to what extent nationwide averages will also apply to the specific network under study.

In this treatment, we use both sources of information: aggregated nationwide failure probabilities for components, obtained by averaging, as well as local data specific to the location of interest. As already mentioned, the latter sort of data is typically very sparse, but at the same time also more informative, as it can incorporate known information about individual asset condition, age, location (e.g. exposure to extreme weather, marine corrosion or industrial pollution), level of utilisation and actual fault history.

We now propose a conjugate Bayesian model for dealing with both types of data. Specifically, we use the aggregated data to construct a prior, and then use the likelihood of the local data to update this prior to a posterior. From a likelihood perspective, the prior simply represents pseudo counts, so effectively, we are really simply adding local data to the nationwide data to obtain a local prediction.

A key question is: how strong should the nationwide data be weighed in comparison to the local data? Or, phrased differently: how relevant do we believe is the nationwide data for making predictions about the local situation? In conjugate analysis [1], there is a natural parameter which represents this subjective judgement. What we will do is perform a sensitivity analysis against this parameter, very similar to what is done in for instance the imprecise Dirichlet model, or more generally, in the exponential family [10, 6, 11, 12].

Observe that the expression for q_2 in terms of the alpha-factors α_1, α_2 and total failure rates q_1^A, q_1^B (Eq. (8)) can be written as a function of just the alpha-factors, times a function of just the total failure rates. So, inspired by [9], for a joint prior, we use an independent product of two Gamma distributions, one on q_t^A and one on q_t^B, and a Dirichlet (or, beta) distribution jointly on α_1 and α_2. We now elaborate on this in the following sections.

3.1 Dirichlet Prior for Alpha-Factors

A natural way to estimate alpha-factors goes via a sequence of N observations, where n_1 of those involved single failures of either A or B (but not both), and the remaining n_2 involved double failures of both A and B. The corresponding likelihood is:

$$\Pr(n_1, n_2 \mid \alpha_1, \alpha_2) = \binom{N}{n_1} \alpha_1^{n_1} \alpha_2^{n_2}. \tag{19}$$

A conjugate prior for the above likelihood is the Dirichlet density (or, beta density, as we have only two categories):

$$f(\alpha_1, \alpha_2 \mid s, t_1, t_2) \propto \alpha_1^{st_1 - 1} \alpha_2^{st_2 - 1} \tag{20}$$

with hyperparameters $s > 0$ and t_1, $t_2 \in (0, 1)$ such that $t_1 + t_2 = 1$. The posterior density is simply:

$$f(\alpha_1, \alpha_2 \mid n_1, n_2, s, t_1, t_2) \propto \alpha_1^{st_1 + n_1 - 1} \alpha_2^{st_2 + n_2 - 1} \tag{21}$$

By Eq. (8), we will need to find the posterior expectation of

$$\frac{\alpha_2}{\alpha_1 + 2\alpha_2}, \tag{22}$$

where we remind the reader that $\alpha_1 + \alpha_2 = 1$, and typically, α_2 is expected to be small. As discussed in great detail in [9], we can do so via Taylor expansion. For example, with second order expansion:

$$E\left(\frac{\alpha_2}{\alpha_1 + 2\alpha_2} \middle| n_1, n_2, s, t_1, t_2\right) \tag{23}$$

$$\approx E(\alpha_2 - \alpha_2^2 | n_1, n_2, s, t_1, t_2) \tag{24}$$

$$= \frac{n_2 + st_2}{N + s}\left(1 - \frac{n_2 + st_2 + 1}{N + s + 1}\right) \tag{25}$$

using the well-known properties of the Dirichlet distribution (for example, see [9, Eq. (10)]); we remind the reader that $N = n_1 + n_2$. For this approximation, the absolute error is less than:

$$\frac{n_2 + st_2}{N + s} \frac{n_2 + st_2 + 1}{N + s + 1} \frac{n_2 + st_2 + 2}{N + s + 2}. \tag{26}$$

3.2 Gamma Prior for Total Failure Rates

To estimate total failure rates, assume we have observed a component (A or B) for time T, during which this component failed M times. The likelihood for the failure rate q_t of this component is then:

$$\Pr(M \mid q_t, T) = \frac{(q_t T)^M \exp(-q_t T)}{M!} \tag{27}$$

as we assumed a Poisson process. A conjugate prior for this likelihood is the Gamma density:

$$f(q_t \mid u, v) \propto q_t^{uv - 1} \exp(-q_t u) \tag{28}$$

with hyperparameters $u > 0$ and $v > 0$. The posterior density is:

$$f(q_t \mid M, T, u, v) \propto q_t^{uv + M - 1} \exp(-q_t(u + T)) \tag{29}$$

By Eq. (8), of interest is the posterior expectation of q_t, which is simply:

$$E(q_t \mid M, T, u, v) = \frac{T}{u + T}\frac{M}{T} + \frac{u}{u + T} v. \tag{30}$$

Considering this posterior expectation when $T = 0$, we see that v represents a prior expectation for q_t, and considering this posterior expectation when $T = u$, we see that u represents the time T needed before the posterior starts to move away from this prior [9, Sec. 3.2].

3.3 Full Analysis

Let us put everything together.

For the alpha-factors, suppose our prior expected fraction of single failures is t_1, and our prior fraction of double failures is t_2. Moreover, we observed n_1 single failures, and n_2 double failures. We are rather unsure about how much weight to assign to the prior, that is, we are unsure about the hyperparameter s. Remember, in a likelihood interpretation of Bayesian inference, s can be thought of the total pseudo count assigned to the prior. Say, $s \in [\underline{s}, \overline{s}]$; for example, with $\underline{s} = 0$ and $\overline{s} = 5$, we count the prior for no more than five observations. As discussed in [9], it seems quite sensible to perform a sensitivity analysis over s, to properly cope with prior-data conflict.

For the total failure rates, suppose our prior expected failure rates are v^A and v^B. Moreover, we observed M^A failures of component A during a time span of T, and M^B failures of component B during a time span of T. For simplicity we take the observed time spans for both components to be identical, as this is the case for our application, but it could be relaxed easily. Again, we are rather unsure about the hyperparameters u^A and u^B—for simplicity, we will also take these to be equal: $u := u^A = u^B$ (again this could be relaxed easily). Here, u can be thought of a pseudo observation time assigned to the prior. Say, $u \in [\underline{u}, \overline{u}]$; for example, with $\underline{u} = 0$ and $\overline{u} = 3$, we count the prior failure rates for no more than 3 years.

Consequently, by Eqs. (8), (25), and (30),

$$\underline{E}(q_2 \mid D) = \inf_{\substack{s \in [\underline{s}, \overline{s}] \\ u \in [\underline{u}, \overline{u}]}} E(q_2 \mid D, s, u), \tag{31}$$

$$\overline{E}(q_2 \mid D) = \sup_{\substack{s \in [\underline{s}, \overline{s}] \\ u \in [\underline{u}, \overline{u}]}} E(q_2 \mid D, s, u). \tag{32}$$

When we assume independence between the alpha-factors and the total failure rates, the expectation de-composes into a product:

$$\begin{aligned} E(q_2 \mid D, s, u) &= \frac{n_2 + st_2}{N + s}\left(1 - \frac{n_2 + st_2 + 1}{N + s + 1}\right) \\ &\quad \times \frac{u(v^A + v^B) + M^A + M^B}{u + T} \end{aligned} \tag{33}$$

and

$$D := (n_1, n_2, M^A, M^B, T, t_1, t_2, v^A, v^B). \tag{34}$$

Note that the optimization problem for s and u can be solved through two independent optimisation problems, one in just s, and one in just u. For the optimisation in u, due to the monotonicity of the objective function, it suffices look at just $\underline{u}$ and $\overline{u}$. The objective function in s is not always monotone (although it often will be), but nevertheless numerical optimisation is still quite easy. The example provides more detail.

Note that bounds for the lower and upper posterior expectations of q_1^A and q_1^B can be derived in a very similar way, through Eqs. (9) and (10)—we leave this to the reader.

4 Network Risk Example

4.1 Problem Description

Following is a generic double circuit reliability problem, based on an actual case study in the North-East of England.

There are two unequal circuits. Circuit A has an expected failure rate of 0.3856 per year, based on 2 transformers and 24.1 km of line and cable. Circuit B has an expected failure rate of 0.3279 per year, based on 1 transformer and 21.5 km of line and cable. No adjustments have been made for asset condition. In the past 12 years, circuit A has experienced 7 failures in 12 years, and circuit B has experienced 4 failures in 12 years. Of these failures, 3 were double failures. For a group of 11 neighbouring (and similar) circuits, there have been 38 failures, of which 24 were single failures, and 14 were double failures—these 38 include the circuit we are studying. On average, for a much larger group of circuits at that voltage, but not necessarily similar to the double circuit under study, about 18% of all failures are double failures.

4.2 Prior and Data

As global prior for the alpha-factors, we use the global average: $t_1 = 0.82$ and $t_2 = 0.18$. It seems reasonable to use neighbouring circuits to correct our prior information about the alpha-factors of our circuit: so $n_1 = 24$ and $n_2 = 14$.

For the total failure rates, an expert provided us with some prior expectations based on global averages of failures for the particular components that make up the circuits: $v^A = 0.3856$ and $v^B = 0.3279$. We have $M^A = 7$ failures during $T^A = 12$ years of circuit A, and $M^B = 4$ failures during $T^B = 12$ years of circuit B.

All we need in addition is some assessment about s (number of total failures needed before we start to move away from the prior in the direction of the data for alpha-factors) and u (time needed before starting to move away from prior in direction of the data for total failure rates). As discussed in Section 3.3, we will perform a sensitivity analysis over intervals for both s and u. Let us take $s = [0, 15]$ and $u = [0, 10]$, which seem conservative yet reasonable given their interpretation discussed earlier.

4.3 Posterior Bounds

We must solve the optimisation problems in Eqs. (31) and (32), using Eq. (33). Let

$$f(s) := \frac{n_2 + st_2}{N + s}\left(1 - \frac{n_2 + st_2 + 1}{N + s + 1}\right), \tag{35}$$

$$e(s) := \frac{n_2 + st_2}{N + s}\frac{n_2 + st_2 + 1}{N + s + 1}\frac{n_2 + st_2 + 2}{N + s + 2}, \tag{36}$$

$$g(u) := \frac{u(v^A + v^B) + M^A + M^B}{u + T}. \tag{37}$$

where $e(s)$ represents a bound on the absolute error, as $f(s)$ is only an approximation (see Eq. (26)). With

$$\underline{f} := \inf_{s \in [0,15]} f(s) \qquad \overline{f} := \sup_{s \in [0,15]} f(s) \tag{38}$$

$$\overline{e} := \sup_{s \in [0,15]} e(s) \tag{39}$$

and

$$\underline{g} := \inf_{u \in [0,10]} g(u) = \min_{u \in \{0,10\}} g(u) \tag{40}$$

$$\overline{g} := \sup_{u \in [0,10]} g(u) = \max_{u \in \{0,10\}} g(u), \tag{41}$$

where we used that g is a monotone function, we then have that,[2]

$$\underline{E}(q_2 \mid D) \geq (\underline{f} - \overline{e})\underline{g} \tag{42}$$

$$\overline{E}(q_2 \mid D) \leq (\overline{f} + \overline{e})\overline{g} \tag{43}$$

[2] Instead of $\overline{e}$, we could use $e(\arg\inf_{s\in[0,15]} f(s))$ and $e(\arg\sup_{s\in[0,15]} f(s))$ to arrive at slightly better error bounds, but in practice it makes little difference.

By numerical optimisation, we find

$$\underline{f} = 0.212 \qquad \overline{f} = 0.227 \qquad \overline{e} = 0.057 \tag{44}$$

$$\underline{g} = 0.824 \qquad \overline{g} = 0.917 \tag{45}$$

Concluding,

$$\underline{E}(q_2 \mid D) \geq 0.128 \tag{46}$$

$$\overline{E}(q_2 \mid D) \leq 0.260 \tag{47}$$

Note that the absolute error $\overline{e}$ is rather large in comparison to $\underline{f}$ and $\overline{f}$—this is due to the fact that the data reflects a rather high value for α_2, and low order approximations only work well when α_2 is less than 0.1. Using instead a sixth order approximation (the equations are very easy to compute, but rather long to write down, see [9] for details; also note that the approximation scheme is designed for ease of computation at the expense of requiring the use of higher order terms, and that more sophisticated techniques might achieve this accuracy with fewer terms), we find:

$$\underline{f} = 0.237 \qquad \overline{f} = 0.266 \tag{48}$$

$$\overline{e} = 0.002 \tag{49}$$

so, because $\overline{e}$ is quite small,

$$\underline{E}(q_2 \mid D) \approx 0.194 \tag{50}$$

$$\overline{E}(q_2 \mid D) \approx 0.245, \tag{51}$$

or in other words, we expect a double failure every four or five years.

5 Conclusions

We have explored a model for dealing with common cause failures in simple power networks, allowing data from various sources to be merged into a meaningful number, or range of numbers when robustness is at stake.

We assumed immediate repair, which is clearly not realistic. Non-immediate repair is typically modelled through continuous time Markov chains [2, Chapters 7–13], which have not yet received that much attention in the imprecise literature. The other unrealistic assumption is the Markov assumption itself, although that assumption seems still pervasive in the standard literature. In practice, failure rates are rarely independent of the history of the system, so the ability to build some level of non-stationarity into the model would be desirable. Moreover, it is not entirely clear how the typical simulation techniques that deal with these issues can be made to work to achieve a robust analysis over a range of parameters.

For more complex power networks, the model would need to be extended to handle multiple components. Although this is mathematically quite easy, difficulties are to be expected with estimating parameters that relate to common cause events, because there can be many more ways in which multiple failures occur when three or more components are involved. Some level of symmetry between common cause events would likely need to be accepted.

Another interesting question would be to investigate how the analysis impacts decisions, say on asset replacement.

Acknowledgements

The first author gratefully acknowledges funding from the Durham Energy Institute (EPSRC Grant EP/J501323/1). We thank all reviewers for their comments which have helped us improving various aspects of the paper. We are also indebted to Chris Dent for many useful discussions and comments.

References

[1] Jose M. Bernado and Adrian F. M. Smith. *Bayesian Theory.* John Wiley and Sons, 1994.

[2] R. Billinton and R. N. Allan. *Reliability Evaluation of Power Systems.* Plenum Press, 2nd edition, 1996.

[3] Simon Blake and Philip Taylor. *Handbook of Power Systems II*, chapter Aspects of Risk Assessment in Distribution System Asset Management: Case Studies, pages 449–480. Springer, 2010.

[4] Simon Blake, Philip Taylor, and David Miller. A composite methodology for evaluating network risk. In *CIRED 21st International Conference on Electricity Distribution*, Frankfurt, Germany, June 2011.

[5] A. Mosleh, K. N. Fleming, G. W. Parry, H. M. Paula, D. H. Worledge, and D. M. Rasmuson. Procedures for treating common cause failures in safety and reliability studies: Procedural framework and examples. Technical Report NUREG/CR-4780, PLG Inc., Newport Beach, CA (USA), January 1988.

[6] Erik Quaeghebeur and Gert de Cooman. Imprecise probability models for inference in exponential families. In Fabio G. Cozman, Robert Nau, and Teddy Seidenfeld, editors, *ISIPTA'05:*

Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, pages 287–296, Pittsburgh, USA, July 2005.

[7] Matthias C. M. Troffaes, Dana L. Kelly, and Gero Walter. Elicitation and inference for the imprecise Dirichlet model with arbitrary sets of hyperparameters. In *Programme and Abstracts: 5th CSDA International Conference on Computational and Financial Econometrics (CFE 2011) and 4th International Conference of the ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on Computing & Statistics (ERCIM 2011)*, page 38, 2011.

[8] Matthias C. M. Troffaes, Dana L. Kelly, and Gero Walter. Imprecise Dirichlet model for common-cause failure. In *Proceedings of PSAM 11 & ESREL 2012*, June 2012.

[9] Matthias C. M. Troffaes, Gero Walter, and Dana Kelly. A robust Bayesian approach to modelling epistemic uncertainty in common-cause failure models. Submitted. `arXiv:1301.0533`.

[10] Peter Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58(1):3–34, 1996.

[11] Gero Walter and Thomas Augustin. Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice*, 3:255–271, 2009.

[12] Gero Walter, Thomas Augustin, and Frank P. A. Coolen. On prior-data conflict in predictive Bernoulli inferences. In F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger, editors, *ISIPTA'11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, pages 391–400, Innsbruck, 2011. SIPTA.

A Note on the Temporal Sure Preference Principle and the Updating of Lower Previsions

Matthias C. M. Troffaes
Durham University, UK
matthias.troffaes@gmail.com

Michael Goldstein
Durham University, UK
michael.goldstein@durham.ac.uk

Abstract

This paper reviews the temporal sure preference principle as a basis for inference over time. We reformulate the principle in terms of desirability, and explore its implications for lower previsions. We report some initial results. Specifically, we present a simple condition for consistency of the temporal sure preference principle with any given collection of assessments, and we derive various bounds on the natural extension. We also discuss some of the technical difficulties encountered.

Keywords. updating, inference, temporal coherence, desirability, lower prevision

1 Introduction

Probabilistic inference has two components, one static and one dynamic. The static component is a description of probabilistic judgements now, where we are free to make any allocations of uncertainty that we consider to be appropriate, expressed, for example, through buying and selling prices on appropriate gambles, subject only to the constraints imposed by coherence over the collection of uncertainty judgements, precise or imprecise, that we choose now to make. The dynamic component describes how these uncertainty statements may change over time, as we receive further information, reflect further on the information that is currently available to us, and so forth.

Aspects of the dynamic component are expressed within the static component, for example through conditioning statements, which express our current buying and selling prices given various called-off bets which describe conditions under which the bets will or will not take place. Such conditioning is informative for our future judgements, but does not determine them, partly as our future experiences will not be summarisable as the observation of membership of a partition that we could specify in advance of our inferences, partly because we are always free to reflect further on the information that we have already received and change our judgements to those that we feel are in closer accord with the prior evidence, and partly because, in any case, there is nothing in the usual probabilistic formalism that forces an equivalence between current views on certain called-off bets, and actual future uncertainty assessments about the relevant quantities. This should not be seen as a failure of conditional reasoning itself—indeed, conditional reasoning is still a perfectly valid and extremely useful formalism for embedding the dynamic features of inference strictly within our current static judgements as to how such an inference might proceed.

At this point, perhaps we should note that one might indeed not care about modelling future beliefs, and take the stance that all future decisions are fully determined solely by current beliefs about those random variables that affect these decisions. For example, normal form decision making is precisely concerned with such scenario: if a subject makes all future decisions right now, only his current beliefs count, and his future beliefs are completely irrelevant. In practice however, beliefs are revised over time, and it is rarely the case that future beliefs, which will determine future decisions, are determined solely on the basis of called-off bets with respect to current beliefs, say through repeated application of Bayes theorem. Analyzing our current beliefs about our future beliefs, as in this paper, is thus important if we now wish to know how we will act in the future based on the actual, but now still uncertain, beliefs that we will hold in the future.

Temporal coherence is concerned with the careful description of the relationships between the static and dynamic features of probabilistic reasoning. We do not know what our future uncertainty judgements will be, but we may now express views about them. These views are, themselves, probabilistic. The basic questions that we must ask are:

(i) Are there any constraints that are reasonable to impose on our current judgements about our future judgements?
(ii) How may such constraints be exploited within the general approach to inference?
(iii) How does the conventional approach to probabilistic

reasoning, via conditioning, fit into the actual temporal evolution of beliefs?

This paper is a modest initial exploration of a particular development of such temporal reasoning, based on the work of [3, 4] and summarised in [5]. In particular, we discuss some of the implications of temporal reasoning for inference with coherent lower previsions. We will only explore questions (i) and (ii).

The key concept in studying temporal coherence is the so-called *temporal sure preference principle*, which establishes a link between certain future preferences and current preferences, thereby allowing us to say something now about our future beliefs.

In imprecise probability theory, preferences come about as a very natural way of modelling beliefs, and it has been argued that the concept of desirability, that is, which gambles we (possibly marginally) prefer to the zero gamble, forms one of the most elegant mathematical and philosophical foundations for imprecise probability [10, 11, 9].

The traditional way of looking at updating in the subjective approach to imprecise probability goes by means of conditioning, that is, looking at called-off gambles. For instance, very recently, Zaffalon and Miranda [12] provided a justification for conditioning and conglomerability, through temporal reasoning, in a setting where future beliefs are assumed to be fixed now.

However, in practice, future subjective beliefs rarely reflect past called-off gambles, and in fact there is no compelling reason for this to be so, simply because there is no compelling reason for them to be fixed now. Indeed, it seems far more natural to start out from the premise that future beliefs are inherently random, which leads to a more general theory, but of course we also risk it to be far less tractable—interestingly, in the precise case, the generality gained leads to updating rules which are far more efficient than computing with called-off gambles, particularly for large scale problems (for instance, see [1]). Having preference, in the form of desirability, at its foundations, imprecise probability is a natural candidate for temporal coherence. We hope it might lead us, as in the precise case, to say something meaningful now about future beliefs, in a way that updating is more flexible, more realistic, and potentially also numerically easier, than the traditional called-off gamble approach (i.e. the generalized Bayes rule [9, Sec. 6.4]).

This paper is organised as follows. Section 2 briefly summarises the main results that we need for lower previsions and desirability. Section 3 reviews temporal coherence and its main implications for previsions. Section 4 explores an approach to temporal coherence for lower previsions. We conclude in Section 5.

2 Lower Previsions and Desirability

Let Ω denote a possibility space. A *gamble* is simply a bounded random quantity, and is mathematically represented by a real-valued function on Ω. We will denote gambles by capital letters X, Y, ... The set of all gambles on Ω is denoted by $\mathcal{L}(\Omega)$. The set of all gambles on Ω that are constant on elements of some partition $\mathcal{A}$ is denoted by $\mathcal{L}(\mathcal{A})$.

As mentioned in the introduction, we will take desirability to be the basic concept, and will use it for studying the implications of temporal coherence on lower previsions. To keep the treatment as simple as possible, however, we will restrict ourselves to sets of almost-desirable gambles induced by lower previsions.

The following serves to fix the notation and conventions used in the paper. It is assumed that the reader is familiar with lower previsions and desirability. We refer to [9] for much more information on the topic. In particular, throughout the paper, we will use the properties of coherent lower previsions extensively [9, Sec. 2.6.1].

Specifically, let $\underline{\mathrm{E}}$ be a coherent lower prevision on $\mathcal{L}(\Omega)$ (without loss of generality, through natural extension [9, Sec. 3.1]), that is, $\underline{\mathrm{E}}$ satisfies:

C1 $\underline{\mathrm{E}}(X) \geq \inf X$

C2 $\underline{\mathrm{E}}(X+Y) \geq \underline{\mathrm{E}}(X) + \underline{\mathrm{E}}(Y)$

C3 $\underline{\mathrm{E}}(\lambda X) = \lambda \underline{\mathrm{E}}(X)$

for all X, $Y \in \mathcal{L}(\Omega)$ and all $\lambda \geq 0$. The upper prevision $\overline{\mathrm{E}}$ corresponding to $\underline{\mathrm{E}}$ is defined as:

$$\overline{\mathrm{E}}(X) = -\underline{\mathrm{E}}(-X). \tag{1}$$

By $\mathfrak{P}(\Omega)$ we denote the set of all coherent lower previsions on $\mathcal{L}(\Omega)$.

With $\underline{\mathrm{E}}$ we can then associate a *set of (almost) desirable gambles*:

$$\mathcal{D} := \{X \in \mathcal{L} \colon \underline{\mathrm{E}}(X) \geq 0\}. \tag{2}$$

For simplicity of exposition, when in the following we say desirable, we really mean almost-desirable. The following conditions are satisfied:

D1 if $X \geq 0$ then $X \in \mathcal{D}$,

D2 if $\sup X < 0$ then $X \notin \mathcal{D}$,

D3 if $X \in \mathcal{D}$ and $Y \in \mathcal{D}$ then $X + Y \in \mathcal{D}$,

D4 if $\lambda \geq 0$ and $X \in \mathcal{D}$ then $\lambda X \in \mathcal{D}$, and

D5 if $X + \epsilon \in \mathcal{D}$ for all $\epsilon > 0$, then $X \in \mathcal{D}$.

Note that we can recover $\underline{\mathrm{E}}$ from $\mathcal{D}$ through:

$$\underline{\mathrm{E}}(X) = \sup\{a \in \mathbb{R} \colon X - a \in \mathcal{D}\} \tag{3}$$

so in the following, we can use $\underline{\mathrm{E}}$ and $\mathcal{D}$ interchangeably.

A lower prevision is called a prevision when it is self-conjugate, that is, when $\underline{\mathrm{E}} = \overline{\mathrm{E}}$, in which case we simply denote it by E. It is well known that previsions correspond to expectation operators, and lower previsions correspond to lower envelopes of expectation operators.

We will consider lower previsions at different points in time—in fact, at just two points in time, 0 and $t > 0$.

By Ω we denote the possibility space at time 0: it represents our subjective judgement, now, about what events are possible. Because Ω will include events involving future beliefs, which we do not know, we emphasize that, in general, a full specification of Ω is not possible.

We assume that we are able to specify a partition $\mathcal{A}$ of Ω which generates all events relevant to the problem domain at hand. Unlike Ω, the partition $\mathcal{A}$ is explicitly modelled, and hence, represents the operational part of Ω. We assume that our current assessments about the problem domain only involve gambles that are constant on the elements of $\mathcal{A}$, and we need not make any further assessments about any other gambles, that is, we can specify a lower prevision $\underline{P}^{\mathcal{A}}$ defined on some subset of $\mathcal{L}(\mathcal{A})$. In particular, we need not make any direct assessments about our future beliefs—those will come in later through the temporal sure preference principle.

We also assume the existence of a partition $\mathcal{B}_t$ of Ω, such that exactly one of the elements of this partition will occur at time t. We will not make any assumption about $\mathcal{B}_t$, in fact, operationally, it is usually impossible identify now what $\mathcal{B}_t$ ought to be. One could consider an element of $\mathcal{B}_t$ to be a possible possibility space at time t, thus elements of $\mathcal{B}_t$ will be denoted by Ω_t. For any $\omega \in \Omega$, by $[\omega]_t$ we denote the unique element Ω_t of $\mathcal{B}_t$ that contains ω. Perhaps we need to emphasize that we do not assume any relationship between $\mathcal{A}$ and $\mathcal{B}_t$. In particular, we do not assume that, say, $\mathcal{B}_t$ refines $\mathcal{A}$: this would mean that, at time t, we would know which element A of $\mathcal{A}$ obtains, and generally, of course this will not be the case.

By $\underline{\mathrm{E}}_t$ we denote our coherent lower prevision at time t—its value is known to us at time t. So, $\underline{\mathrm{E}}_0$ is our current lower prevision, and embodies both our current assessments $\underline{P}^{\mathcal{A}}$ concerning the problem domain at hand, as well as any further principles taken into account, such as for instance the temporal sure preference principle, which we will discuss in detail later. However, $\underline{\mathrm{E}}_t$ is in fact a random lower prevision now:[1]

$$\underline{\mathrm{E}}_t(\Omega_t) \in \mathfrak{P}(\Omega_t) \text{ for any } \Omega_t \in \mathcal{B}_t, \tag{4}$$

whose value is only realised at time t.

When comparing gambles, as we will need to do further in the paper, it is convenient that those gambles are expressed with respect to the same possibility space. For this reason, it is more convenient to consider $\underline{\mathrm{E}}_t$ as a mapping from Ω to $\mathfrak{P}(\Omega)$:

$$\underline{\mathrm{E}}_t(\omega)(X) := \underline{\mathrm{E}}_t([\omega]_t)(X|_{[\omega]_t}), \tag{5}$$

for any $\omega \in \Omega$ and $X \in \mathcal{L}(\Omega)$. We will follow this convenient notation for the remainder of the paper. Note that one may think of $\mathcal{B}_t$ as the partition generated by $\underline{\mathrm{E}}_t$.

For any gamble $X \in \mathcal{L}(\Omega)$, by $\underline{\mathrm{E}}_t(X)$ we denote the random lower prevision of X at time t:

$$\underline{\mathrm{E}}_t(X)(\omega) := \underline{\mathrm{E}}_t(\omega)(X). \tag{6}$$

Clearly, $\underline{\mathrm{E}}_t(X) \in \mathcal{L}(\Omega)$, and it is constant on the elements of $\mathcal{B}_t$.

Similarly, we write $\mathcal{D}_t$ for the set of desirable gambles corresponding to $\underline{\mathrm{E}}_t$. So, $\mathcal{D}_t$ is a random set of gambles:

$$\mathcal{D}_t \colon \Omega \to \wp(\mathcal{L}(\Omega)) \tag{7}$$

where

$$\mathcal{D}_t(\omega) := \{X \in \mathcal{L}(\Omega) \colon \underline{\mathrm{E}}_t(\omega)(X) \geq 0\}, \tag{8}$$

and as with $\underline{\mathrm{E}}_t$, the value of $\mathcal{D}_t$ is only realised at time t.[2] Clearly, $\mathcal{D}_t$ is constant on the elements of $\mathcal{B}_t$.

3 Temporal Coherence for Previsions

In this section, we review the existing theory of temporal coherence for previsions.

3.1 Beliefs and Updating

By X, we denote a gamble whose value is unknown to us. Of course, we may have present beliefs about X. We assume that X is constant on the elements of the partition $\mathcal{A}$. Our present expectation for X is denoted by $\mathrm{E}_0(X)$, and our present variance for X is denoted by $\mathrm{var}_0(X)$. The subscript in E_0 and var_0 denotes time, where time 0 corresponds to the present.

As X is unknown, we may try to learn about X by observing another random quantity, which we denote by Y: say we actually observe the value of Y at time $t > 0$, whilst X remains unknown to us at time t. Again, we assume that Y is constant on the elements of the partition $\mathcal{A}$—because

[1] Remember that $\mathfrak{P}(\Omega_t)$ denotes the set of all coherent lower previsions on $\mathcal{L}(\Omega_t)$.

[2] We should note that $\mathcal{D}_t(\omega)$, when defined as a subset of $\mathcal{L}(\Omega)$ as in Eq. (8), may not satisfy D5, however of course $\mathcal{D}_t(\omega)$ will satisfy D5 as a subset of $\mathcal{L}([\omega]_t)$. Also note that $X \in \mathcal{D}_t(\omega)$ if and only if $I_{[\omega_t]}X \in \mathcal{D}_t(\omega)$.

we assumed that Y is known at time t, it will also be constant on the elements of $\mathcal{B}_t$. Here too, we may have present beliefs about Y, such as its present expectation $\mathrm{E}_0(Y)$ and present variance $\mathrm{var}_0(Y)$. In fact, we may hold present beliefs about X and Y jointly, such as for instance the present covariance between X and Y, which we denote by $\mathrm{cov}_0(X, Y)$.

As mentioned, whilst the value of Y will be known at time t, X remains unknown. Consequently, we may also consider, now, our future beliefs about X. However, because the future has yet to obtain, those future beliefs are uncertain in themselves. In other words, $\mathrm{E}_t(X)$ and $\mathrm{var}_t(X)$, the actual expectation and variance of X which represents our beliefs about X at time t, are gambles in themselves, whose values are only known to us at time t:

$$\mathrm{E}_t(X)\colon \Omega \to \mathbb{R}, \qquad \mathrm{var}_t(X)\colon \Omega \to \mathbb{R}. \tag{9}$$

For example, we can think about our current beliefs about our future beliefs, and could consider for instance our present expectation and variance of these gambles: $\mathrm{E}_0(\mathrm{E}_t(X))$, $\mathrm{E}_0(\mathrm{var}_t(X))$, $\mathrm{var}_0(\mathrm{E}_t(X))$, and $\mathrm{var}_0(\mathrm{var}_t(X))$.

The general problem of updating might then be concerned with answering the following questions. First, what should be the relationship between:

- our current beliefs about $\mathrm{E}_t(X)$,
- our current beliefs about X, and
- our current beliefs about Y?

More challengingly, what should be the relationship between:

- our actual beliefs $\mathrm{E}_t(X)$ about X at time t, and
- any updating rule for X as a function of Y?

3.2 The Temporal Sure Preference Principle

In order to establish relationships between current and future beliefs, we must impose conditions that go beyond coherence at a single time point. These conditions should be sufficiently weak and compelling to be widely applicable, while leading to a meaningful account of inference.

Any principle which asserts that beliefs now are compelling for beliefs in the future is, by its nature, unconvincing, as we cannot know what future information we may receive or what the outcome of our future reflections may be. The converse, however, is that we may often view our future beliefs as compelling for our current beliefs, as all such future reflections and information will be taken into account in such future judgements. In order for future judgements to influence our current judgements, we must know what such future judgements are. We therefore introduce the notion of a sure prefererence, at a future time, as one which we are now sure that we will hold at that time.

It may seem unreasonable, now, to think that we hold any such sure preferences. However, it so happens that we do indeed hold many such, and recognising them explicitly, and formalising their implications for our current judgements, provides a natural account of temporal reasoning. For this reason, Goldstein introduced the following principle (see [3], [4], [5, Sec. 3.5]):

Principle 1 (The Temporal Sure Preference Principle I)**.** *For any gambles $U \in \mathcal{L}(\Omega)$ and $W \in \mathcal{L}(\Omega)$, if you have a **sure** preference for U over W at future time t, then you should not have a strict preference for W over U now.*

It is useful to briefly reflect on what it means to have a sure preference for U over W at future time t. Remember, at future time t, an element Ω_t of $\mathcal{B}_t$ obtains, and we hold beliefs $\mathrm{E}_t(\Omega_t) \in \mathfrak{P}(\Omega_t)$—for now these beliefs are assumed to be precise. A *sure* preference means a preference regardless of the outcome Ω_t in $\mathcal{B}_t$. So in other words, we are sure to prefer U to W at time t whenever

$$\mathrm{E}_t(\Omega_t)(U|_{\Omega_t}) \geq \mathrm{E}_t(\Omega_t)(W|_{\Omega_t}) \text{ for all } \Omega_t \in \mathcal{B}_t, \tag{10}$$

or equivalently, whenever

$$\mathrm{E}_t(U)(\omega) \geq \mathrm{E}_t(W)(\omega) \text{ for all } \omega \in \Omega, \tag{11}$$

where we use the notation introduced earlier in Eqs. (5) and (6).

The temporal sure preference principle should be considered as a prescription for a particular domain of discourse, rather than as a fundamental condition for rationality. There are various reasons why, in a particular application, it might not hold. For example, we might consider that, at the future time, we could undergo personality changes which render our future judgements suspect to us now (the Doctor Jekyll and Mister Hyde scenario). More prosaically, we might just recognise situations where our future judgements are likely to be less reliable than our current judgements (for example, the problem of forgetting). Therefore, the intention of the temporal sure preference principle is that it should be viewed as a very weak, and widely applicable principle, whose relevance we should consider for the problem at hand. If we consider the temporal sure preference principle applicable in our problem, then we may draw on the strong implications of the principle to provide an account of temporal coherence for this situation. We know of no weaker alternative principle that allows a similar account of the inferential process, for the many applications where we will be willing to assert temporal sure preference.

The aim of this section is to study this principle in terms of desirability [10, 11, 9], whilst at the same time reviewing the main well-known consequences of the temporal sure preference principle for previsions, in order to provide a good understanding of the ideas and techniques involved before we move on to lower previsions in Section 4.

If we take preference $U \succeq W$ to mean that $U - W + \epsilon$ is desirable for all $\epsilon > 0$ [9, Sec. 3.7.5, first paragraph], and $U \succ W$ to mean that $U - W - \epsilon$ is desirable for some $\epsilon > 0$ [9, Sec. 3.7.7, second paragraph], then it is a trivial exercise to reformulate the above principle in terms of desirability:[3]

Principle 2 (The Temporal Sure Preference Principle II). *For any gamble $U \in \mathcal{L}(\Omega)$, if, for all $\epsilon > 0$, $U + \epsilon$ is sure to be desirable for us at future time t, then, for all $\epsilon > 0$, $-U - \epsilon$ should not be desirable for us now.*

Perhaps it is useful to note already here that many variations of Principle 2 are possible. We will consider some of those variations, which are all equivalent for previsions, but which are no longer equivalent for lower previsions.

We give a quick proof of equivalence, which holds generally—not just for sets of desirable gambles corresponding to previsions, but for arbitrary sets of desirable gambles; we do not even need to rely on coherence.

Proposition 3. *Principles 1 and 2 are equivalent.*

Proof. Suppose Principle 1 is satisfied. Suppose that, for all $\epsilon > 0$, $U + \epsilon$ is sure to be desirable to us at future time t. This means that, surely, $U \succeq_t 0$ at time t. Consequently, by Principle 1, $0 \not\succ_0 U$ now, or in other words, $0 - U - \epsilon$ is not desirable now for any $\epsilon > 0$. In other words, Principle 2 is satisfied.

Conversely, suppose that Principle 2 is satisfied. Suppose that, surely, $U \succeq_t W$ at time t. This means that, for all $\epsilon > 0$, $U - W + \epsilon$ is surely desirable at time t. Consequently, by Principle 2, for all ϵ, $-U + W - \epsilon$ is not desirable now. But this means precisely that $W \not\succ_0 U$, now. In other words, Principle 1 is satisfied. □

An obvious question at this point is: what kind of gambles can be surely desirable at some future time t? Obviously, any positive constant gamble would be, but that is hardly useful, as we already know that these are desirable to us now. For more interesting examples, consider cases where U is a function of E_t. For example, at time t, surely, the gamble $\mathrm{E}_t(X) - X + \epsilon$ is desirable for all $\epsilon > 0$ (note that at time t, $\mathrm{E}_t(X)$ is a constant, whilst X is still a gamble). The temporal sure preference principle then tells us that the gamble $-\mathrm{E}_t(X) + X - \epsilon$ is not desirable to us now.

3.3 Implications

The next proposition, due to Goldstein [4, Theorem 1], forms the basis for linking future beliefs about expectation and variance to current beliefs about expectation and variance. The proof is short, and provides an excellent example of how the temporal sure preference principle can be invoked to make non-trivial statements about $\mathrm{E}_t(X)$, so we reproduce it below.

Proposition 4. *If Principle 2 is satisfied, then it must hold that*

$$\mathrm{E}_0((X - \mathrm{E}_t(X))^2) \leq \mathrm{E}_0((X - Y)^2). \tag{12}$$

where Y is surely known by time t.

Proof. Note that, for previsions, $U \preceq_t W$ precisely when $\mathrm{E}_t(U)(\omega) \leq \mathrm{E}_t(W)(\omega)$ for all $\omega \in \Omega$, and $U \not\succ_0 W$ precisely when $\mathrm{E}_0(U) \leq \mathrm{E}_0(W)$. Also, note that $\mathrm{E}_t(U)(\omega) = U(\omega)$ for any gamble U that is constant on the elements of $\mathcal{B}_t$, such as $\mathrm{E}_t(X)$ and Y.

Consequently, for any $\omega \in \Omega$,

$$\mathrm{E}_t((X - Y)^2 - (X - \mathrm{E}_t(X))^2)(\omega) \tag{13}$$
$$= \mathrm{E}_t(-2XY + Y^2 + 2X\mathrm{E}_t(X) - \mathrm{E}_t(X)^2)(\omega) \tag{14}$$
$$= -2\mathrm{E}_t(X)(\omega)Y(\omega) + Y^2(\omega) + \mathrm{E}_t(X)^2(\omega) \tag{15}$$
$$= (\mathrm{E}_t(X)(\omega) - Y(\omega))^2 \geq 0 \tag{16}$$

where we have used the linearity of $\mathrm{E}_t(\omega)$. So, at time t,[4]

$$(X - \mathrm{E}_t(X))^2 \preceq_t (X - Y)^2. \tag{17}$$

Whence, by Principle 2, now,

$$(X - \mathrm{E}_t(X))^2 \not\succ_0 (X - Y)^2, \tag{18}$$

which yields the desired inequality. □

Those readers familiar with the usual called-off argument for conditional previsions may fear that we have, inadvertedly, relied on conglomerability of E_0 to complete the above argument. Perhaps, it is instructive to try follow this misinterpretation to put such fears at rest. Indeed, in the proof, we first show that, effectively,

$$(X - Y)^2 - (X - \mathrm{E}_t(X))^2 \tag{19}$$

is desirable at time t. One might correctly, but confusingly, understand that this means that the called-off gamble

$$I_{\Omega_t}\left((X - Y)^2 - (X - \mathrm{E}_t(X))^2\right) \tag{20}$$

is now desirable. In fact, it is sure to be desirable at time t—if Ω_t does not obtain, then it is zero and thus desirable, and if Ω_t does obtain, then the reasoning in the proof can be used to show that it is desirable as well—thus, by the temporal sure preference, indeed, the called-off gamble defined in Eq. (20) is desirable now. Then, assuming conglomerability, we can glue all these called-off gambles together to prove that the gamble in Eq. (19) is desirable now. We simply emphasize here that the actual proof works quite differently. In particular, the temporal sure preference principle is only applied once, namely on the gamble in Eq. (19): called-off gambles are never considered.

[3] The attentive reader will note that in Principle 2, we can actually take desirability to be actual desirability, rather than almost-desirability.

[4] It is interesting to compare Eq. (17) with the operational definition of expectation of de Finetti [2], in which Eq. (17) is the definition of $\mathrm{E}_t(X)$, rather than a derived property.

The proof of Proposition 4, and the above discussion, already hint at a slightly simpler version of the temporal sure preference principle:

Principle 5 (The Temporal Sure Preference Principle III)**.** *For any gamble $U \in \mathcal{L}(\Omega)$, if U is sure to be desirable for us at future time t, then U should be desirable for us now:*

$$\bigcap_{\Omega_t \in \mathcal{B}_t} \mathcal{D}_t(\Omega_t) = \bigcap_{\omega \in \Omega} \mathcal{D}_t(\omega) \subseteq \mathcal{D}_0. \tag{21}$$

Proposition 6. *Principle 5 implies Principle 2.*

Proof. Assume that Principle 5 holds. If $U + \epsilon$ is sure to be desirable at time t, for all $\epsilon > 0$, then consequently, $U + \epsilon$ is desirable now, for all $\epsilon > 0$. If $-U - \delta$ would be desirable for us now for some $\delta > 0$, then $U + \delta/2 - U - \delta = -\delta/2$ would be desirable as well, which would lead us to incur a sure loss, so $-U - \delta$ cannot be desirable now for any $\delta > 0$. In other words, Principle 2 holds. □

Proposition 7. *If our set of desirable gambles corresponds to a prevision, that is, if*

$$\mathcal{D}_0 = \{U \colon \mathrm{E}_0(U) \geq 0\} \tag{22}$$

for some prevision E_0, then Principle 5 is equivalent to Principle 2.

Proof. Assume Principle 2 holds. If U is sure to be desirable at time t, then obviously $U + \epsilon$ is also sure to be desirable at time t, for all $\epsilon > 0$. Consequently, $-U - \epsilon$ is not desirable now, for all $\epsilon > 0$, or in other words, $\mathrm{E}_0(-U) - \epsilon < 0$ for all $\epsilon > 0$. This means that $\mathrm{E}_0(U) \geq 0$, so U is desirable to us now. □

In other words, for the remainder of this section, where we are concerned with previsions only, we can assume Principle 5 without loss of generality. We will thus assume that desirability is as in Eq. (22).

Proposition 4 has a number of very interesting consequences:

Corollary 8. *If Principle 5 is satisfied, then*

$$\mathrm{E}_0(X - \mathrm{E}_t(X)) = 0. \tag{23}$$

Proof. In Proposition 4, let $Y := \mathrm{E}_t(X) + b$ where $b \in \mathbb{R}$, and take the minimum over b. □

Note that Eq. (23) is very similar to the usual definition of conglomerability as in for instance [9, p. 305, (C15)], so it is worth emphasizing that Eq. (23) is *not* your usual conglomerability, because $\mathrm{E}_t(X)$ is not necessarily obtained through conditioning.

We can also say something about the expected future variance, that is, $\mathrm{E}_0(\mathrm{var}_t(X))$.

Corollary 9 (Adjusted Variance)**.** *If Principle 5 is satisfied, then it holds that:*

$$\mathrm{var}_0(X - \mathrm{E}_t(X)) = \mathrm{E}_0(\mathrm{var}_t(X)) \leq \mathrm{var}_Y(X), \tag{24}$$

with

$$\mathrm{var}_Y(X) := \mathrm{var}_0(X) - \frac{\mathrm{cov}_0(X,Y)^2}{\mathrm{var}_0(Y)}, \tag{25}$$

where Y is surely known by time t.

Proof. To prove the inequality in Eq. (24), take $a + bY$ for Y in Proposition 4, and minimize over a and b.

Note that the usual formulation uses $\mathrm{var}_0(X - \mathrm{E}_t(X))$ only. It is easy to see that this is $\mathrm{E}_0(\mathrm{var}_t(X))$, which seems easier to interpret, and is also relevant for what comes later:

$$\begin{aligned}\mathrm{var}_0(X - \mathrm{E}_t(X)) = \mathrm{E}_0((X - \mathrm{E}_t(X) \\ - \mathrm{E}_0(X - \mathrm{E}_t(X)))^2)\end{aligned} \tag{26}$$

and by Eq. (23) $\mathrm{E}_0(X - \mathrm{E}_t(X)) = 0$, so

$$= \mathrm{E}_0((X - \mathrm{E}_t(X))^2) \tag{27}$$

and again by Eq. (23) $\mathrm{E}_0(\cdot) = \mathrm{E}_0(\mathrm{E}_t(\cdot))$, so

$$= \mathrm{E}_0(\mathrm{E}_t((X - \mathrm{E}_t(X))^2)) \tag{28}$$

$$= \mathrm{E}_0(\mathrm{var}_t(X)). \tag{29}$$

□

So, the temporal sure preference principle allows us to quantify uncertainty about future variance.

In the proof of Corollary 9, the value for $a + bY$ where the minimum is achieved is precisely the *adjusted expectation*:

Corollary 10 (Adjusted Expectation)**.** *If Principle 5 is satisfied, then*

$$\mathrm{E}_t(X) = \mathrm{E}_Y(X) + S_t(X), \tag{30}$$

where

$$\mathrm{E}_Y(X) := \mathrm{E}_0(X) + \frac{\mathrm{cov}_0(Y,X)}{\mathrm{var}_0(Y)}(Y - \mathrm{E}_0(Y)), \tag{31}$$

and

$$\mathrm{E}_0(S_t(X)) = 0, \quad \mathrm{cov}_0(S_t(X), \mathrm{E}_Y(X)) = 0. \tag{32}$$

Proof. Take $a + bY + c\mathrm{E}_t(X)$ for Y in Proposition 4, and do the usual magic. □

In other words, the temporal sure preference principle also allows us to quantify a linear connection between observations and future beliefs.

If Y is the indicator of some event E, then $\mathrm{E}_Y(X) = E(X|E)$, that is, adjusted expectation coincides with conditional expectation. So, Eq. (30) also provides an interpretation of the relation between conditioning and our actual posterior expectation.

The above results are only an initial tasting of the realm of possibilities. Of considerable interest is that the above treatment generalises almost trivially to the multivariate case.

4 Temporal Coherence for Lower Previsions

Let us now investigate the implications of the temporal sure preference principle for lower previsions.

4.1 The Temporal Sure Preference Principle for Lower Previsions

In the context of desirability, it makes sense to adopt Principle 5, for at least two reasons:

1. The principle seems reasonably compelling. Indeed, if U is sure to be desirable for us at time t, then it does not matter whether we accept it already now, or whether we accept it only at time t: the gamble has the same outcome either way.
2. We may use it as a production rule in natural extension.

By the second point, we mean the following. As mentioned in the introduction, we assume a partition $\mathcal{A}$ which represents what we could call the operational part of Ω. Specifically, all direct assessments of lower previsions $\underline{P}_0^{\mathcal{A}}(Y)$, which represent our beliefs now, concern gambles $Y \in \mathcal{L}(\mathcal{A})$. In other words, our initial assessments are embodied by a lower prevision $\underline{P}_0^{\mathcal{A}}$ which is defined on a subset of $\mathcal{L}(\mathcal{A})$. We can then consider the natural extension of $\underline{\mathrm{E}}_0^{\mathcal{A}}$ to all gambles $\mathcal{L}(\mathcal{A})$; let us denote that natural extension by $\underline{\mathrm{E}}_0^{\mathcal{A}}$. It is different from $\underline{\mathrm{E}}_0$, which embodies our beliefs about $\underline{P}_0^{\mathcal{A}}$ but also those implied by the temporal sure preference principle. Indeed, under Principle 5, all gambles V for which

$$\underline{\mathrm{E}}_t(V)(\omega) \geq 0 \text{ for all } \omega \in \Omega, \tag{33}$$

or briefly, for which $\underline{\mathrm{E}}_t(V) \geq 0$, are desirable now. Consequently,

$$\underline{\mathrm{E}}_0(U) = \sup_{\substack{\alpha\in\mathbb{R} \\ Y\in\mathcal{L}(\mathcal{A})\colon \underline{\mathrm{E}}_0^{\mathcal{A}}(Y)\geq 0 \\ V\in\mathcal{L}(\Omega)\colon \underline{\mathrm{E}}_t(V)\geq 0}} \{\alpha\colon U-\alpha \geq Y+V\} \tag{34}$$

for any gamble $U \in \mathcal{L}(\Omega)$.

Before we proceed investigating actual inferences from the above expression for natural extension, we need to address a few concerns. First, there is no guarantee that Principle 5 is consistent with our initial assessments $\underline{P}_0^{\mathcal{A}}$. Eq. (34) provides us with a means to verify this: we merely have to check that $\underline{\mathrm{E}}_0(0) < +\infty$ [9, p. 123, ll. 4–7]. Secondly, there is no guarantee that Principle 5 does not modify $\underline{\mathrm{E}}_0^{\mathcal{A}}$ on $\mathcal{L}(\mathcal{A})$. Thirdly, this form of natural extension is inherently non-constructive: it involves an operator $\underline{\mathrm{E}}_t$ about which we have not specified much at all. The next proposition answers the first two concerns. The last concern of course remains, but nevertheless, we will show that we still can derive something non-trivial about $\underline{\mathrm{E}}_t$, just as in the precise case discussed earlier.

Proposition 11. *If, for every $A \in \mathcal{A}$, there is an $\Omega_t^A \in \mathcal{B}_t$ such that*

$$\underline{\mathrm{E}}_t(\Omega_t^A)(A) = 1, \tag{35}$$

then Principle 5 is consistent with $\underline{P}_0^{\mathcal{A}}$, and, for all $X \in \mathcal{L}(\mathcal{A})$,

$$\underline{\mathrm{E}}_0(X) = \underline{\mathrm{E}}_0^{\mathcal{A}}(X). \tag{36}$$

Proof. If we can prove Eq. (36), then consistency follows immediately.

Consider any $X \in \mathcal{L}(\mathcal{A})$. Clearly, $\underline{\mathrm{E}}_0(X) \geq \underline{\mathrm{E}}_0^{\mathcal{A}}(X)$. We now prove the converse inequality. Indeed,

$$\underline{\mathrm{E}}_0(X) = \sup_{\substack{\alpha\in\mathbb{R} \\ Y\in\mathcal{L}(\mathcal{A})\colon \underline{\mathrm{E}}_0^{\mathcal{A}}(Y)\geq 0 \\ V\in\mathcal{L}(\Omega)\colon \underline{\mathrm{E}}_t(V)\geq 0}} \{\alpha\colon X-\alpha \geq Y+V\} \tag{37}$$

$$= \sup_{\substack{\alpha\in\mathbb{R} \\ Y\in\mathcal{L}(\mathcal{A})\colon \underline{\mathrm{E}}_0^{\mathcal{A}}(Y)\geq 0 \\ V\in\mathcal{L}(\Omega)\colon \underline{\mathrm{E}}_t(V)\geq 0}} \Bigg\{\alpha\colon (\forall A \in \mathcal{A}) \tag{38}$$

$$\left(X(A) - \alpha \geq Y(A) + \sup_{\omega\in A} V(\omega)\right)\Bigg\} \tag{39}$$

so, if we can show that $\sup_{\omega\in A} V(\omega) \geq 0$ whenever $\underline{\mathrm{E}}_t(V) \geq 0$, then

$$\leq \sup_{\substack{\alpha\in\mathbb{R} \\ Y\in\mathcal{L}(\mathcal{A})\colon \underline{\mathrm{E}}_0^{\mathcal{A}}(Y)\geq 0}} \{\alpha\colon X-\alpha \geq Y\} \tag{40}$$

$$= \underline{\mathrm{E}}_0^{\mathcal{A}}(X). \tag{41}$$

We are left to show that $\sup_{\omega\in A} V(\omega) \geq 0$ whenever $\underline{\mathrm{E}}_t(V) \geq 0$. In fact, we will show that $\sup_{\omega\in A\cap\Omega_t^A} V(\omega) \geq 0$, by contraposition. Note that Eq. (35) already implies that $A \cap \Omega_t^A$ is non-empty.

Suppose that $\sup_{\omega\in A\cap\Omega_t^A} V(\omega) < 0$, then there would be an $\epsilon > 0$ such that for all $\omega \in A \cap \Omega_t^A$,

$$V(\omega) < -\epsilon. \tag{42}$$

Therefore, necessarily, also

$$\underline{\mathrm{E}}_t(\Omega_t^A)(V) \leq \underline{\mathrm{E}}_t(\Omega_t^A)(I_{A^c}V) + \overline{\mathrm{E}}_t(\Omega_t^A)(-I_A\epsilon) = -\epsilon \tag{43}$$

because $\overline{\mathrm{E}}_t(\Omega_t^A)(A^c) = 0$, so $\underline{\mathrm{E}}_t(\Omega_t^A)(I_{A^c}V) = 0$, and $\underline{\mathrm{E}}_t(\Omega_t^A)(A) = 1$, so $\overline{\mathrm{E}}_t(\Omega_t^A)(-I_A\epsilon) = -\epsilon$. But Eq. (43) contradicts the assumption that $\underline{\mathrm{E}}_t(V) \geq 0$. □

The consistency condition in Eq. (35) has a simple interpretation: for every $A \in \mathcal{A}$, we must allow for the possibility that at time t, we will be certain that A has obtained. Note that we only must logically allow for this possibility—it may well have zero probability—so the condition is really very weak.

We also immediately have the following important result, which effectively reformulates Principle 5 in terms of lower previsions:[5]

Proposition 12. *Principle 5 holds if and only if, for every gamble $U \in \mathcal{L}(\Omega)$,*

$$\inf_{\omega\in\Omega} \underline{\mathrm{E}}_t(U)(\omega) \le \underline{\mathrm{E}}_0(U). \tag{44}$$

Proof. "only if". Suppose Principle 5 holds. We could rely on our expression for natural extension, Eq. (34), however it is instructive to use only Principle 5 in the proof.

For any $\epsilon > 0$, simply note that

$$U - \underline{\mathrm{E}}_t(U) + \epsilon \le U - \inf_{\omega\in\Omega} \underline{\mathrm{E}}_t(U)(\omega) + \epsilon \tag{45}$$

so $U - \inf_{\omega\in\Omega} \underline{\mathrm{E}}_t(U)(\omega) + \epsilon$ is sure to be desirable at time t, because $U - \underline{\mathrm{E}}_t(U) + \epsilon$ is. Consequently, we have that

$$\underline{\mathrm{E}}_0\left(U - \inf_{\omega\in\Omega} \underline{\mathrm{E}}_t(U)(\omega) + \epsilon\right) \ge 0 \tag{46}$$

and because this holds for all $\epsilon > 0$, we arrive at Eq. (44), after using the constant additivity of $\underline{\mathrm{E}}_0$.

"if". Suppose Eq. (44) holds. Consider any gamble $U \in \mathcal{L}(\Omega)$. If U is sure to be desirable at time t, then $\underline{\mathrm{E}}_t(U)(\omega) \ge 0$ for all $\omega \in \Omega$. Consequently, by Eq. (44),

$$\underline{\mathrm{E}}_0(U) \ge \inf_{\omega\in\Omega} \underline{\mathrm{E}}_t(U)(\omega) \ge 0 \tag{47}$$

so U is desirable now. Principle 5 follows. □

4.2 Implications

The treatment for previsions relied on the scoring definition of expectation, via Proposition 4. However, no proper scoring rules exist for lower previsions [7]. We try to generalise Proposition 4 anyway. We do so in two ways: first without scoring, and secondly using the relationship between expressions of the form $(X - a)^2$, and lower and upper variance—which is the closest notion to scoring we have for lower previsions. There are certainly more ways to go about it, but for this introductory paper, we will stick to these two.

First, we derive the following imprecise counterpart of Corollary 8.

Corollary 13. *If Principle 5 is satisfied, then*

$$\underline{\mathrm{E}}_0(X - \underline{\mathrm{E}}_t(X)) \ge 0. \tag{48}$$

Proof. By Eq. (44):

$$\inf_{\omega\in\Omega} \underline{\mathrm{E}}_t(X - \underline{\mathrm{E}}_t(X))(\omega) \le \underline{\mathrm{E}}_0(X - \underline{\mathrm{E}}_t(X)). \tag{49}$$

Now note that $\underline{\mathrm{E}}_t(X - \underline{\mathrm{E}}_t(X))(\omega) = 0$ for all $\omega \in \Omega$, by coherence of $\underline{\mathrm{E}}_t(\omega)$. □

Clearly, if we were to impose a conditioning interpretation, Eq. (48) corresponds to one of Walley's conditions for coherence [9, p. 303, (C11)].

Corollary 13 has a number of interesting immediate consequences:

Corollary 14. *If Principle 5 is satisfied, then*

$$\overline{\mathrm{E}}_0(X - \overline{\mathrm{E}}_t(X)) \le 0, \tag{50}$$

$$\underline{\mathrm{E}}_0(\underline{\mathrm{E}}_t(X)) \le \underline{\mathrm{E}}_0(X) \le \underline{\mathrm{E}}_0(\overline{\mathrm{E}}_t(X)), \tag{51}$$

$$\overline{\mathrm{E}}_0(\underline{\mathrm{E}}_t(X)) \le \overline{\mathrm{E}}_0(X) \le \overline{\mathrm{E}}_0(\overline{\mathrm{E}}_t(X)). \tag{52}$$

Proof. The first inequality holds by:

$$0 \le \underline{\mathrm{E}}_0(-X - \underline{\mathrm{E}}_t(-X)) = -\overline{\mathrm{E}}_0(X - \overline{\mathrm{E}}_t(X)). \tag{53}$$

The second inequality holds because

$$\underline{\mathrm{E}}_0(X - \underline{\mathrm{E}}_t(X)) \ge 0 \tag{54}$$

$$\Longrightarrow \underline{\mathrm{E}}_0(X) + \overline{\mathrm{E}}_0(-\underline{\mathrm{E}}_t(X)) \ge 0 \tag{55}$$

$$\Longrightarrow \underline{\mathrm{E}}_0(X) \ge \underline{\mathrm{E}}_0(\underline{\mathrm{E}}_t(X)) \tag{56}$$

and

$$\underline{\mathrm{E}}_0(-X - \underline{\mathrm{E}}_t(-X)) \ge 0 \tag{57}$$

$$\Longrightarrow \overline{\mathrm{E}}_0(-X) + \underline{\mathrm{E}}_0(\overline{\mathrm{E}}_t(X)) \ge 0 \tag{58}$$

$$\Longrightarrow \underline{\mathrm{E}}_0(\overline{\mathrm{E}}_t(X)) \ge \underline{\mathrm{E}}_0(X). \tag{59}$$

The thrid one is proved similarly. □

We can derive neither a lower bound on $\underline{\mathrm{E}}_0(\underline{\mathrm{E}}_t(X))$, nor an upper bound on $\overline{\mathrm{E}}_0(\overline{\mathrm{E}}_t(X))$, for example, due to the possibility of dilation [8].

Finally, let us see how far we can get with lower and upper variance. We need the following lemma [9, p. 618, G2]:

Lemma 15. *For every gamble X, there are previsions E_1 and E_2 in the credal set of $\underline{\mathrm{E}}$, such that for all $a \in \mathbb{R}$:*

$$\underline{\mathrm{var}}(X) := \underline{\mathrm{E}}((X - \mathrm{E}_1(X))^2) \le \underline{\mathrm{E}}((X - a)^2), \tag{60}$$

$$\overline{\mathrm{var}}(X) := \overline{\mathrm{E}}((X - \mathrm{E}_2(X))^2) \le \overline{\mathrm{E}}((X - a)^2). \tag{61}$$

In particular, for all $a \in \mathbb{R}$, $(X - a)^2 - \underline{\mathrm{var}}(X)$ is desirable. Note that $\overline{\mathrm{var}}(X) - (X - a)^2 - \epsilon$ is non-desirable, however this does not help us very much—in fact, this led us to investigate temporal sure preference also for non-desirability, yet the resulting principle seems not very compelling, and leads to serious issues.

[5] The attentive reader will note that the 'only if' part of the proof of Proposition 12 remains valid, if in Principle 5, we take desirability to be actual desirability, rather than almost-desirability.

Proposition 16. *If Principle 5 is satisfied, then*

$$\underline{\mathrm{E}}_0(\underline{\mathrm{var}}_t(X)) \leq \underline{\mathrm{E}}_0((X-Y)^2), \tag{62}$$

$$\overline{\mathrm{E}}_0(\underline{\mathrm{var}}_t(X)) \leq \overline{\mathrm{E}}_0((X-Y)^2). \tag{63}$$

Proof. By definition of variance,

$$\underline{\mathrm{E}}_t\left((X-Y)^2 - \underline{\mathrm{var}}_t(X)\right) \geq 0 \tag{64}$$

(remember that Y is a known constant at time t). Whence, by Eq. (44), also

$$\underline{\mathrm{E}}_0((X-Y)^2 - \underline{\mathrm{var}}_t(X)) \geq 0. \tag{65}$$

Concluding, by coherence,

$$\underline{\mathrm{E}}_0((X-Y)^2) \geq \underline{\mathrm{E}}_0(\underline{\mathrm{var}}_t(X)), \tag{66}$$

and

$$\overline{\mathrm{E}}_0((X-Y)^2) \geq \overline{\mathrm{E}}_0(\underline{\mathrm{var}}_t(X)). \tag{67}$$

□

Again, we cannot say anything about, say, $\overline{\mathrm{E}}_0(\overline{\mathrm{var}}_t(X))$.

As for adjusted lower expectation, if we are happy to bound, say, the upper expectation of the future lower variance, by Eq. (63), any function Y of observed quantities at time t which aims to minimize $\overline{\mathrm{E}}_0((X-Y)^2)$ could be a candidate. A good choice of function of course depends on the optimisation problem, and an obvious stumbling block is that even already for a simple linear form, say $a+bY$, $\overline{\mathrm{E}}_0((X-(a+bY))^2)$ cannot be written as a function of the imprecise expectation and imprecise variance of X and Y. In other words, at this point, we seem to get stuck, although there might be interesting and feasible solutions for specific cases, for instance, using techniques from imprecise regression.

5 Conclusion

We have discussed the temporal sure preference principle in the context of desirability and lower previsions. We found more than one way to generalise the temporal sure preference principle to lower previsions, so we used the simplest version, related directly to desirability.

We have identified an expression for natural extension under the suggested temporal sure preference principle. We then derived a simple condition, which guarantees consistency of the temporal sure preference principle with prior specifications, and which also guarantees that those prior specifications are not modified by adopting the temporal sure preference principle, so we can still use the usual (non-temporal) form of natural extension for gambles as far as our current beliefs are concerned.

We have also derived a host of bounds on lower and upper expectations of future lower and upper expecations and variances. In this initial investigation, a particular challenge which remains is to provide lower and upper bounds on *all* future lower and upper expectations and variances.

An obvious next step would be to investigate possible updating rules implied by the temporal sure preference principle, for example using ideas from imprecise regression. The optimisation problems involved do not appear to have nice closed solutions in general, essentially due to the nonlinearity of the lower and upper previsions. It would be very interesting to find non-trivial imprecise instances of lower previsions where such updating rules could be calculated explicitly. In this paper, we had an initial look at linear updating rules and lower and upper variance, but of course there might be many more ways to go about it.

Temporal reasoning without conditioning also raises interesting questions about the need for a possibility space. In fact, it is one of the premises of temporal reasoning that we cannot specify in advance what the possibility space ought to be. In the current paper, it serves only as a mathematical construct to establish a clear link with Walley's [9] approach to lower previsions and desirability. We might be better off simply ignoring the possibility space entirely, and instead working with random quantities directly, following the approach of de Finetti [2] and Williams [10, 11].

Finally, one might wonder, why not also introduce a principle for temporal coherence concerning non-desirability: say, if a gamble is surely non-desirable at a future time t, should it also be non-desirable now? One can show that, for previsions, this principle is equivalent to the usual temporal sure preference principle.

However, for lower previsions, this is no longer so, and it leads us to infer additional constraints. In fact, it leads to additional constraints that are usually not satisfied in the standard theory when updating is taken to be conditioning. We simply note here that temporal reasoning on non-desirability seems far less compelling, certainly so under the standard interpretation that non-desirability merely means that we do not say whether we accept a gamble or not. Here, a reject-accept approach to desirability [6] might lead to a better treatment.

Acknowledgements

The first author would like to thank Enrique Miranda and Gert de Cooman for interesting discussions on the topic of temporal coherence, which fed the desire to write this paper. We are particularly indebted to Gert de Cooman for making many useful suggestions which improved the clarity of the paper. We also thank all reviewers, whose many scrutinizing yet very constructive and supportive comments have helped us clarifying numerous technical aspects of the paper.

References

[1] Peter S. Craig, Michael Goldstein, Jonathan C. Rougier, and Allan H. Seheult. Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association*, 96(454):717–729, June 2001. URL: http://www.jstor.org/stable/2670309, doi:10.1198/016214501753168370.

[2] Bruno de Finetti. *Theory of Probability: A Critical Introductory Treatment*. Wiley, New York, 1974–5. Two volumes.

[3] Michael Goldstein. Temporal coherence. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 2*. 1985.

[4] Michael Goldstein. Prior inferences for posterior judgements. In M. C. D. Chiara et al., editors, *Structures and Norms in Science*, pages 55–71. Kluwer, 1997.

[5] Michael Goldstein and David A. Wooff. *Bayes Linear Statistics: Theory and Methods*. Wiley, Chichester, 2007.

[6] Erik Quaeghebeur, Gert de Cooman, and Filip Hermans. Accept & reject statement-based uncertainty models. Submitted. arXiv:1208.4462.

[7] Teddy Seidenfeld, Mark J. Schervish, and Joseph B. Kadane. Forecasting with imprecise probabilities. In F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger, editors, *ISIPTA'11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, pages 317–326, Innsbruck, 2011. SIPTA.

[8] Teddy Seidenfeld and Larry Wasserman. Dilation for sets of probabilities. *The Annals of Statistics*, 21(3):1139–1154, 1993. doi:10.1214/aos/1176349254.

[9] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[10] Peter M. Williams. Notes on conditional previsions. Technical report, School of Math. and Phys. Sci., Univ. of Sussex, 1975.

[11] Peter M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44(3):366–383, 2007. doi:10.1016/j.ijar.2006.07.019.

[12] Marco Zaffalon and Enrique Miranda. Probability and time. *Artificial Intelligence*, 198:1–51, May 2013. doi:10.1016/j.artint.2013.02.005.

Logistic Regression on Markov Chains for Crop Rotation Modelling

Matthias C. M. Troffaes
Durham University, UK
matthias.troffaes@gmail.com

Lewis Paton
Durham University, UK
l.w.paton@durham.ac.uk

Abstract

Often, in dynamical systems, such as farmer's crop choices, the dynamics is driven by external non-stationary factors, such as rainfall, temperature, and economy. Such dynamics can be modelled by a non-stationary Markov chain, where the transition probabilities are logistic functions of such external factors. We investigate the problem of estimating the parameters of the logistic model from data, using conjugate analysis with a fairly broad class of priors, to accommodate scarcity of data and lack of strong prior expert opinions. We show how maximum likelihood methods can be used to get bounds on the posterior mode of the parameters.

Keywords. logistic regression, Markov chain, robust Bayesian, conjugate, maximum likelihood, crop

1 Introduction

We wish to accurately model agricultural land use, that is, to predict what crop is grown in any particular field. Usually, farmers follow set patterns of successive yearly crop choices in order to preserve nutrients in the soil. For example, they may have a 3 year cycle, in which they, under normal circumstances, grow wheat for two years, and then leave the field empty for the third year. A very simple model for such crop choices on any particular field is a Markov chain (see for instance [4, 3]), where the state at time i is the crop choice at year i. Such a model makes a simplifying assumption, namely that crop choice in any given year only depends on crop choice in the previous year.

However, crop choices are not only affected by crop choices of the previous year(s): they are also affected by various environmental and economical conditions. In an earlier study, Luo [10] identified some of the most important factors as rainfall, temperature, profit margin, and soil type. To model the effect of these variables on crop choice, in this paper, we propose a logistic regression model for the crop choice transition probabilities. For simplicity, in this paper, we only investigate the impact of rainfall on a simple binary crop choice: wheat, or something else. Generalisation to more than one regressor and to more than two crop choices will be the subject of another paper.

A key challenge with any regression model is to estimate its parameters. First, following [5], we identify a class of conjugate priors for our model. Next, we follow a similar approach to that of the imprecise Dirichlet model [13]: we identify a reasonably vacuous set of conjugate priors, and calculate posterior bounds. A benefit of this approach is that it can also incorporate expert opinion, which will be very useful when studying crop types that are uncommon, such as oats. Our model is thus designed to handle situations in which data is scarce and in which prior expert opinion may be lacking.

The novel contributions of this paper are:

1. We present a first step at including imprecision in non-stationary Markov chains influenced by non-stationary random variables.
2. We propose a novel approach to imprecise logistic regression, based on conjugate analysis.
3. The use of maximum likelihood methods for approximate Bayesian inference in logistic regression, to arrive at fast algorithms when dealing with sets of priors, is new, even though relatively obvious.

The paper is structured as follows. Section 2 introduces the model. Section 3 describes the conjugate prior and posterior distributions, discusses the parameters of the model and their interpretation. Section 4 explains how we can use sets of distributions to obtain posterior bounds. Section 5 has an example. Section 6 concludes the paper, and details future areas of research.

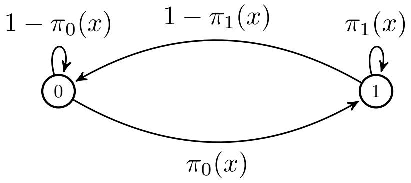


Figure 1: A Markov chain for crop rotations.

2 Logistic Model

We model crop rotations as a non-stationary Markov chain, as depicted in Figure 1.

The model has two states: either our current crop is wheat, which we denote as 1, or our current crop is not wheat, denoted as 0. The transition probabilities in the Markov chain only depend on the previous crop grown and the rainfall—which is where the non-stationarity comes from, as rainfall may change over years. We denote this by:

$$\pi_y(x) := P(Y_{i+1} = 1 | Y_i = y, X_i = x) \tag{1}$$

for all $y \in \{0, 1\}$ and $x \in \mathbb{R}$, where Y_i is the previous crop choice, and X_i is the rainfall recorded just before the planting of crop Y_{i+1}. Note that X_i is not assumed to be part of the state space of the Markov chain, and is simply a non-stationary random variable influencing the transition probabilities.

The impact of rainfall on these transition probabilities is typically either monotonically increasing, or monotonically decreasing. Therefore, a logistic regression model for $\pi_y(x)$ seems fairly reasonable:

$$\pi_y(x) = \frac{e^{\alpha_y + x\beta_y}}{1 + e^{\alpha_y + x\beta_y}}. \tag{2}$$

where α_y and β_y are parameters of the model.

For example, when it rains a lot, farmers are usually more likely to grow wheat, if the previous crop grown was also wheat; see Figure 2. To produce Figure 2, we used maximum likelihood to fit a logistic regression curve to some actual data when the previous crop grown was wheat—in fact, the data is shown in Table 3 for $y = 1$, and will be explained further in the paper. Note that the relationship is actually reversed if the previous crop is not wheat ($y = 0$).

Also note that the data used here is quite limited, as we used only 10 observations. In reality, wheat versus non-wheat will not be an issue as wheat is a very common crop. However, some crop types, such as oats, are very rare, and will suffer from scarcity of data. For actual applications, our model will be appropriate to handle such crop types specifically. Here, we chose wheat versus non-wheat because that data was readily available, but of course other crop types will be investigated in the future, including rare ones.

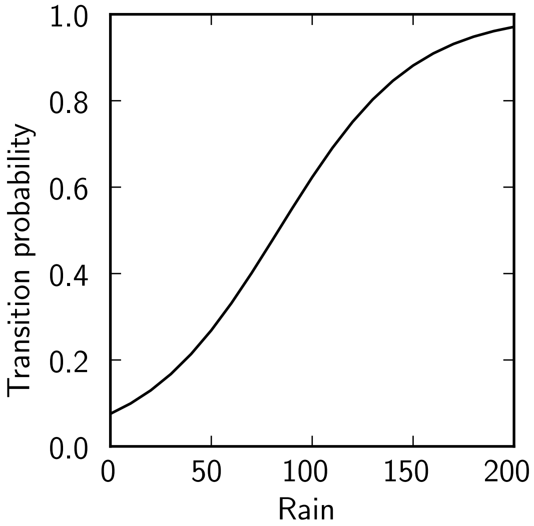


Figure 2: Logistic regression of the probability of growing wheat against rainfall, when the previous crop grown was wheat.

We also assume that we have some model for the regressor X, say a probability density $f_\gamma(x)$ with parameter γ.

For further details about logistic regression, see for instance [1].

3 Parameter Estimation

3.1 Data

We now wish to estimate the parameters of the model, given some data. We have recorded crop transitions and rainfall of a number of fields over a number of years. Specifically, we have $n_y(x)$ observations where the previous crop choice was y and rainfall was x—obviously, $n_y(x)$ will be zero at all but a finite number of $x \in \mathbb{R}$. Of these $n_y(x)$ observations, the crop choice was 1 in $k_y(x)$ cases.

Because we effectively have two separate logistic regression models—one for $y = 0$ and one for $y = 1$—it makes sense to split our data into two sets accordingly. Table 1 tabulates the full data set. Table 2 tabulates the same data, but split according to the value of y.

previous crop y	rain x	current crop total $n_y(x)$	current crop count $k_y(x)$
1	46	1	0
0	52	1	0
0	38	1	1
1	30	1	1
1	37	1	0
⋮	⋮	⋮	⋮

Table 1: Crop rotation data.

previous $y = 0$		
rain x	current crop total $n(x)$	current crop count $k(x)$
52	1	0
38	1	1
⋮	⋮	⋮

previous $y = 1$		
rain x	current crop total $n(x)$	current crop count $k(x)$
46	1	0
30	1	1
37	1	0
⋮	⋮	⋮

Table 2: Crop rotation data split by y.

3.2 Likelihood

Our inspiration is the work by Chen and Ibrahim [5], who propose a conjugate prior distribution of the form:

$$\exp\left(\sum_{i=1}^{m} s\left[t_i(\alpha + x_i\beta) - \ln(1 + e^{\alpha + x_i\beta})\right]\right) \tag{3}$$

where $\vec{x} = (x_1, \dots, x_m)$ are the observed locations of the regressor, α and β are parameters of the logistic model (as in Eq. (2)), and s and $\vec{t}$ are hyperparameters. However, our notation is simpler if we work directly with the count functions $n_y(x)$ and $k_y(x)$ which are defined for all $x \in \mathbb{R}$, rather than having to enumerate over observed locations explicitly.

Specifically, in terms of $n_y(x)$ and $k_y(x)$, our likelihood is:

$$L_y(\alpha_y, \beta_y, \gamma_y | n_y, k_y) = p_y(k_y | n_y, \alpha_y, \beta_y) f(n_y | \gamma) \tag{4}$$

where

$$f(n_y|\gamma) = \prod_{x\in\mathbb{R}} f_\gamma(x)^{n_y(x)} \tag{5}$$

and

$$\begin{aligned} &p_y(k_y | n_y, \alpha_y, \beta_y) \\ &= \prod_{x\in\mathbb{R}} \binom{n_y(x)}{k_y(x)} \pi_y(x)^{k_y(x)} (1 - \pi_y(x))^{n_y(x) - k_y(x)}. \end{aligned} \tag{6}$$

The above products over $x \in \mathbb{R}$ are well defined: because $k_y(x)$ and $n_y(x)$ are zero at all but a finite number of x, all but a finite number of factors are equal to one.

Because the likelihood is a product of a function of γ and a function of (α_y, β_y), we can separate our inference procedure accordingly. In the following, we will concern ourselves with inference about (α_y, β_y) only, and leave inference about γ to another paper.

Note that we have subscript y everywhere. To keep notation readable, we will drop it in the remainder of this section. So, we can write:

$$\begin{aligned} &p(k|n, \alpha, \beta) \\ &= \prod_{x\in\mathbb{R}} \binom{n(x)}{k(x)} \pi(x)^{k(x)} (1 - \pi(x))^{n(x) - k(x)} \end{aligned} \tag{7}$$

For conjugate analysis later, we rewrite this in canonical form [2, p. 202, Definition 4.12], which, after some manipulations, yields:

$$\propto \exp\left(\sum_{x\in\mathbb{R}} k(x)(\alpha + x\beta) - n(x)\ln\left(1 + e^{\alpha + x\beta}\right)\right) \tag{8}$$

up to a normalisation constant that is a function of x only. The above sum over $x \in \mathbb{R}$ is well defined, because $k(x)$ and $n(x)$ are zero at all but a finite number of x.

3.3 Conjugate Prior and Posterior

Following [5, p. 470, Eq. (6.1)], we can now simply define a conjugate prior [2, p. 266, Proposition 5.4] for logistic regression:

$$\begin{aligned} &f_0(\alpha, \beta | s, t) \\ &\propto \exp\left(\sum_{x\in\mathbb{R}} s(x)\left[t(x)(\alpha + x\beta) - \ln\left(1 + e^{\alpha + x\beta}\right)\right]\right), \end{aligned} \tag{9}$$

where s and t are non-negative functions on $\mathbb{R}$ such that $s(x) = t(x) = 0$ for all but a finite number of $x \in \mathbb{R}$, and $0 \le t(x) \le 1$ for all $x \in \mathbb{R}$.

Writing the posterior distribution down is a simple task [2, p. 269, Proposition 5.5]. We simply multiply

Eq. (8) and Eq. (9), to obtain:

$$
\begin{aligned}
&f(\alpha,\beta|k,n,s,t) \\
&\quad\propto f_0(\alpha,\beta|s,t)p(k,n|\alpha,\beta) && (10)\\
&\quad\propto \exp\Bigg(\sum_{x\in\mathbb{R}}(s(x)t(x)+k(x))(\alpha+x\beta) \\
&\qquad\qquad -(n(x)+s(x))\ln\left(1+e^{\alpha+x\beta}\right)\Bigg) && (11)
\end{aligned}
$$

It is clear the prior distribution and posterior distribution are of the same family:

$$f(\alpha,\beta|k,n,s,t)=f_0(\alpha,\beta|\sigma,\tau) \tag{12}$$

where

$$\sigma(x):=s(x)+n(x),\ \text{and} \tag{13}$$

$$\tau(x):=\frac{s(x)t(x)+k(x)}{s(x)+n(x)}. \tag{14}$$

We now study this family in a bit more detail.

3.4 Interpretation of Hyperparameters

A key problem we are faced with is the choice of prior hyperparameters $s(x)$ and $t(x)$. Ideally we want a direct interpretation of the parameters. Eqs. (13) and (14) show that, as usual, the hyperparameters can be interpreted as a prior virtual sample, with $s(x)$ observations at $X=x$, $s(x)t(x)$ of which are wheat ($Y_{i+1}=1$). The implications of such specification may however not be entirely clear to an expert, and therefore it seems more appealing, at least to us, to relate the hyperparameters to the prior predictive instead, as is commonly done for the regular exponential family through a famous result by Diaconis and Ylvisaker [6, Theorem 2].

To apply [6, Theorem 2], the number of parameters must be equal to the dimension d of the space $\mathbb{R}^d$ in which the hyperparameter t lives. Therefore, if we relax the model by replacing $\alpha+x\beta$ with an arbitrary function $\theta(x)$—i.e. if we were to drop the assumption that $\pi(x)$ has a logistic form—then [6, Theorem 2] applies, and $t(x)$ is precisely the prior prediction for $\pi(x)$ (see [5, Eqs. (2.4) and (2.5)]).

For our actual model, however, there are only two parameters to estimate (α and β), but unfortunately, the hyperparameter t effectively lives in $\mathbb{R}^d$, where d is the number of x where $s(x)$ is non-zero. Specifically, although we have conjugacy, it is very easy to see that, in general, the prior predictive $\hat{\pi}_0(x)$ is not equal to the hyperparameter $t(x)$, i.e. $t(x)$ is not a prior expectation for $\pi(x)$, unless $d=2$.

We can still arrive at some sort of interpretation for $t(x)$ as follows. Inspired by [6, Theorem 2], by the usual properties of integration and densities:

$$\iint_{\mathbb{R}^2}\frac{\partial}{\partial\alpha}f_0(\alpha,\beta|s,t)\,\mathrm{d}\alpha\,\mathrm{d}\beta=0 \tag{15}$$

$$\iint_{\mathbb{R}^2}\frac{\partial}{\partial\beta}f_0(\alpha,\beta|s,t)\,\mathrm{d}\alpha\,\mathrm{d}\beta=0 \tag{16}$$

These equations yield:

$$\sum_{x\in\mathbb{R}}s(x)t(x)=\sum_{x\in\mathbb{R}}s(x)\hat{\pi}_0(x) \tag{17}$$

$$\sum_{x\in\mathbb{R}}xs(x)t(x)=\sum_{x\in\mathbb{R}}xs(x)\hat{\pi}_0(x) \tag{18}$$

where

$$
\begin{aligned}
\hat{\pi}_0(x)&:=P(Y_{i+1}=1|Y_i=y,X_i=x,s,t) && (19)\\
&=\iint_{\mathbb{R}^2}\pi(x)f_0(\alpha,\beta|s,t)\,\mathrm{d}\alpha\,\mathrm{d}\beta && (20)
\end{aligned}
$$

Note that we should write $\hat{\pi}_{0y}(x)$ but we omit the subscript y for ease of notation as usual.

These equations show that $t(x)$ in some sense 'matches' $\hat{\pi}_0(x)$, the more so for values of x where $s(x)$ is larger. Of course, for any given prior specification of the function $\hat{\pi}_0$, even for fixed s, there will be many different functions t that satisfy Eqs. (17) and (18), so the choice of $t(x)$ is not uniquely determined by our prior expectation about $\pi(x)$.

As mentioned, there is however a special case where the conditions of [6, Theorem 2] are satisfied, and so where we do get a direct interpretation of $t(x)$. This occurs when there are only two points $\{x_1,x_2\}$ where $s(x)$ is non-zero. In this case, Eqs. (17) and (18) do have a unique solution, namely:

$$t(x_1)=\hat{\pi}_0(x_1)\ \text{and}\ t(x_2)=\hat{\pi}_0(x_2) \tag{21}$$

regardless of $s(x_1)$ and $s(x_2)$ (of course, this also follows directly from [6, Theorem 2]). Whence, for simplicity and interpretability, this is the case that we will consider in practical examples later. In this case, as we shall see, $s(x_1)$ and $s(x_2)$ also carry their usual interpretation, in determining the speed by which our posterior will move away from our prior.

4 Inference

4.1 Posterior Transition Probability

For inference, we are mostly interested in the posterior transition probability:

$$
\begin{aligned}
\hat{\pi}(x)&:=P(Y_{i+1}=1|Y_i=y,X_i=x,k,n,s,t) && (22)\\
&=\iint_{\mathbb{R}^2}\pi(x)f(\alpha,\beta|k,n,s,t)\,\mathrm{d}\alpha\,\mathrm{d}\beta && (23)
\end{aligned}
$$

where it is worth recalling that $\pi(x)$ is a non-linear (logistic) function of α and β. Specifically, taking into account our uncertainty about α and β as given by the posterior, we are interested in evaluating Eq. (23). The challenge now is the evaluation of the integral. One option is to directly numerically integrate. However, as eventually, we want to use sets of distributions, this may not necessarily be the most sensible route to take.

Therefore, we may prefer to rely on faster approximations of the integral. A first crude idea would be to approximate the prior (Eq. (9)) by a multivariate normal distribution; Chen and Ibrahim [5] mention that for large sample sizes this approximation yields the exact solution. Whilst the mean can be easily approximated through the mode, obtaining the covariance structure is somewhat more difficult (a starting point would be [5, Theorem 2.3]). Interestingly, there are variational techniques for direct updating of the mean and covariance structure [9], which means that we would need to perform the multivariate normal approximation only once, on the initial prior.

However, this approach still requires numerical integration. As just mentioned, when we move to sets of priors, this might easily become computationally intractable, as we will have to update, approximate, and integrate, for every prior in the set. A more crude but also much faster approximation would be to simply pretend that all probability mass is concentrated at the mode of the posterior. It is relatively straightforward to show that the mode can be obtained by solving the following system of non-linear equations for α and β:

$$\sum_{x\in\mathbb{R}} \sigma(x)\tau(x) = \sum_{x\in\mathbb{R}} \sigma(x)\pi(x) \tag{24}$$

$$\sum_{x\in\mathbb{R}} x\sigma(x)\tau(x) = \sum_{x\in\mathbb{R}} x\sigma(x)\pi(x) \tag{25}$$

where it is again worth recalling that $\pi(x)$ is a non-linear (logistic) function of α and β. To obtain an approximate value for $\hat{\pi}(x)$, we simply plug in the solution (α^*, β^*) into the expression for $\pi(x)$ (see Eq. (2)):

$$\hat{\pi}(x) \approx \frac{e^{\alpha^*+x\beta^*}}{1+e^{\alpha^*+x\beta^*}}. \tag{26}$$

Although this approximation is obviously horribly crude, we note that in fact it corresponds to the maximum likelihood estimate, where the data has been augmented with pseudo counts. Hence, it reflects current practice quite well, and arguably even improves it, by allowing for additional prior information to be taken into account.

Solving a system of non-linear equations is non-trivial. However, Green [8] provides a Newton Raphson algorithm specifically for the maximum likelihood estimate of logistic regression. We can essentially recycle algorithms like these to find the mode, simply by adding some pseudo counts to the data to reflect our prior.

4.2 Sets of Prior Distributions

We now want to propose sets of prior distributions, in a similar vein to Walley's imprecise Dirichlet Model [13]. In this section, we study the inferences resulting from an arbitrary but fixed prior function for $s(x)$, namely:

$$s(x) := \begin{cases} s & \text{if } x \in \mathcal{X}, \\ 0 & \text{otherwise,} \end{cases} \tag{27}$$

for some finite set $\mathcal{X} \subseteq \mathbb{R}$, and an arbitrary set of prior functions $\mathfrak{T}$ for $t(x)$. We explain how to calculate posterior bounds based on this set of priors, and the observed data. Practical choices for reasonably vacuous sets of prior distributions will be discussed further in Section 5.

4.3 Posterior Transition Probability Bounds

For the above choice of $s(x)$, Eqs. (24) and (25) can be written as:

$$s\sum_{x\in\mathcal{X}}(\pi(x)-t(x)) + \sum_{x\in\mathbb{R}}(n(x)\pi(x)-k(x)) = 0, \tag{28}$$

$$s\sum_{x\in\mathcal{X}} x(\pi(x)-t(x)) + \sum_{x\in\mathbb{R}} x(n(x)\pi(x)-k(x)) = 0. \tag{29}$$

If we can solve the above equations for all $t \in \mathfrak{T}$, then we obtain a set Θ^* of solutions (α^*, β^*), one solution for each $t \in \mathfrak{T}$. Each member of Θ^* corresponds to an estimate of the posterior transition probability as in Eq. (26). Whence,

$$\underline{\pi}(x) \approx \inf_{(\alpha^*,\beta^*)\in\Theta^*} \frac{e^{\alpha^*+x\beta^*}}{1+e^{\alpha^*+x\beta^*}}, \tag{30}$$

$$\overline{\pi}(x) \approx \sup_{(\alpha^*,\beta^*)\in\Theta^*} \frac{e^{\alpha^*+x\beta^*}}{1+e^{\alpha^*+x\beta^*}}, \tag{31}$$

are the desired lower and upper posterior approximations of the transition probability.

5 Example

As discussed in Section 3.4, there is a direct interpretation of $t(x)$ when $\mathcal{X} = \{x_1, x_2\}$. We will explore this case here.

previous $y = 0$		
rain x	current crop total $n(x)$	current crop count $k(x)$
18	1	1
68	1	1
24	1	1
19	1	1
99	1	0
16	1	0
20	1	0
119	1	0
102	1	0
87	1	1
17	1	0
29	1	0

previous $y = 1$		
rain x	current crop total $n(x)$	current crop count $k(x)$
72	1	1
105	1	1
6	1	0
104	1	1
77	1	0
69	1	0
15	1	0
63	1	0
35	1	1
25	1	0

Table 3: Actual crop rotation data split by y.

We take a set of functions for $t(x)$ and a constant s. The most vacuous choice would be:

$$\mathfrak{T}_v = \{t \in \mathbb{R}^{\mathbb{R}} : t(x) = 0 \text{ when } x \notin \mathcal{X}, \\ 0 < t(x) < 1 \text{ when } x \in \mathcal{X}\} \qquad (32)$$

Solving the optimisation problem (Eqs. (30) and (31)) over $\mathfrak{T}_v$ is rather involved. For a simple quick analysis, we restrict ourselves to the extreme points of $\mathfrak{T}_v$, namely:

$$\mathfrak{T}'_v = \{t \in \mathbb{R}^{\mathbb{R}} : t(x) = 0 \text{ when } x \notin \mathcal{X}, \\ (t(x_1), t(x_2)) \in \{(0,0),(0,1),(1,0),(1,1)\}\} \qquad (33)$$

We will use data collected from actual fields to illustrate the ideas we have talked about. The data is shown in Table 3. It consists of 22 observations of crop transitions [12], and the corresponding rainfall recorded in the month of planting for each crop [11].

Figure 3 shows $\hat{\pi}(x)$ for each element of $\mathfrak{T}'_v$, where we have specified $s = 2$, $x_1 = 30$, and $x_2 = 80$, and we are looking at the model for $y = 1$. Each line corresponds to one estimate. The grey region represents the posterior estimates from the most vacuous set $\mathfrak{T}_v$ (we actually used a 21×21 grid over the unit square). As can be seen $\underline{\pi}$ and $\overline{\pi}$ are very closely matched in both cases (almost shockingly so!), so it seems very reasonable to use only $\mathfrak{T}'_v$ instead of the full set $\mathfrak{T}_v$, for ease of computation.

Note that the case $t_1 = 1$, $t_2 = 0$, goes against the data, and corresponds to a non-natural shape for π in this problem. Thus, Figure 3 also highlights the importance of including constraints on π which follow from prior expert opinion, for instance by removing those values from $\mathfrak{T}_v$ for which π violates those constraints. A less vacuous prediction for larger values of x would result. Of course, other techniques for learning under order constraints, which have been studied for instance in the context of Bayesian network learning [7], could also have some potential here.

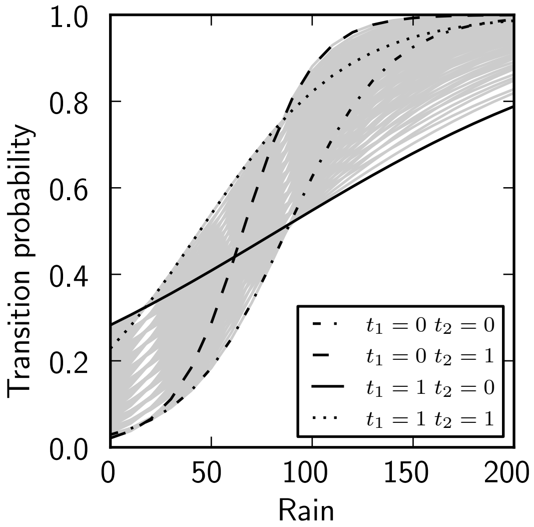


Figure 3: $\hat{\pi}(x)$ based on $\mathfrak{T}_v$ and $\mathfrak{T}'_v$ for $y = 1$.

Our choice for x_1 and x_2 is also important. In Figure. 4, we use $x_1 = 10$ and $x_2 = 100$, which lean more towards the extremes of the range of observations in x. This changes our inferences quite substantially. The largest impact is observed for the case $t_1 = 1$, $t_2 = 0$, and as we just saw, removing such unnatural values for t from $\mathfrak{T}_v$ might be reasonable. In any case, this also demonstrates the importance of choosing x_1 and x_2 sensibly, particularly under the vacuous model $\mathfrak{T}_v$. For example, a sensible choice would be to take for x_1 the first quartile and for x_2 the third quartile, of the observations in x (or of our prior distribution for x).

The inference also depends on the the value of s. As in the imprecise Dirichlet model, smaller values of s produce tighter bounds, as seen in Figure. 5.

6 Conclusion

In this paper, we proposed a new model for land use, which aims to properly capture epistemic uncertainty about crop rotations, in an interpretable, robust, and efficient way. Thereby, we presented a first step at including imprecision in non-stationary Markov chains influenced by non-stationary random variables.

In a nutshell, starting from earlier work by Chen and

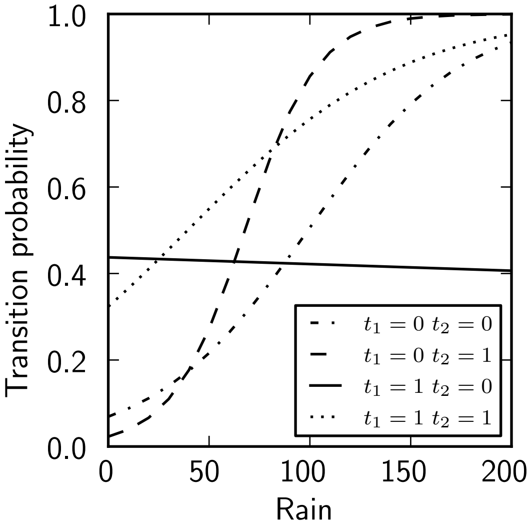


Figure 4: The impact of changing x_1 and x_2.

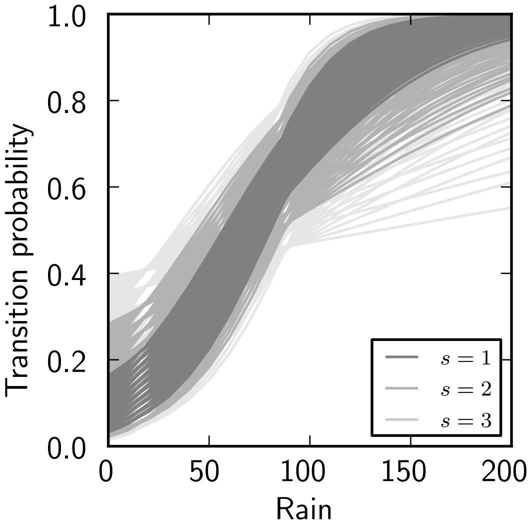


Figure 5: 3 different sets of $\hat{\pi}(x)$ for different values of s.

Ibrahim [5], we proposed a new model for imprecise logistic regression, using sets of conjugate prior distributions for a generalised linear model with logistic link function, to get bounds on the probability of growing wheat as a function of rainfall.

We investigated the interpretation of the hyperparameters of the model, which turns out to be somewhat non-trivial, unless the model is constrained in a very specific way.

We care about robustness, because typically, for certain rare crop types, only a small amount of data is available. This results in posterior probabilities which are highly sensitive to prior specifications. By using sets of priors, our approach allows us to draw accurate robust inferences even from near-vacuous prior knowledge about crop rotations.

Due to the non-linearity of our model, one might fear that the updating process is highly complicated. We proposed the use of maximum likelihood methods for approximate Bayesian inference, effectively via data augmentation, to arrive at fast algorithms when dealing with sets of priors. Much to our surprise, it turns out that using the set of extreme points in our prior specifications still captures the posterior bounds extremely well. We suspect that this is due to the monotonicity of the link function.

An obvious weakness of our analysis is the use of the posterior mode as a very crude approximation to the actual posterior expectation. However, the other options for evaluating the posterior expectation are computationally far more complex, making a robust analysis over sets of parameter values infeasible, at least in our initial attempts. Nevertheless, the use of the posterior mode corresponds quite well to current practice: a standard technique for estimating the parameters in logistic regression goes by maximum likelihood estimation, and the posterior mode can be interpreted as such.

Concerning the actual crop modelling, this work is still in its infancy. We have yet to judge the effects of the simplifying assumptions we have made, and we still need to assess the validity of our model. We plan to use the posterior bounds in conjunction with a predictive model for rainfall, to make predictions about future crop distributions. We also plan to extend the model to deal with multiple crop choices (i.e. more than just wheat) and multiple regressors (i.e. not just rainfall, but also economic factors).

Acknowledgements

The second author is supported by the Food and Environmental Research Agency (FERA), and EPSRC.

We thank Andy Hart and Nigel Boatman from FERA for technical advice. Finally, we are indebted to all three reviewers for their comments which have improved the paper considerably.

References

[1] Alan Agresti. *Categorical Data Analysis.* John Wiley and Sons, second edition, 2002.

[2] Jose M. Bernado and Adrian F. M. Smith. *Bayesian Theory.* John Wiley and Sons, 1994.

[3] M. S. Castellazzi, J. Matthews, F. Angevin, C. Sausse, G. A. Wood, P. J. Burgess, I. Brown, K. F. Conrad, and J. N. Perry. Simulation scenarios of spatio-temporal arrangement of crops at the landscape scale. *Environmental Modelling and Software*, 25(12):1881–1889, 2010.

[4] M. S. Castellazzi, G. A. Wood, P. J. Burgess, J. Morris, K. F. Conrad, and J. N. Perry. A systematic representation of crop rotations. *Agricultural Systems*, 97:26–33, 2008.

[5] Ming-Hui Chen and Joseph G. Ibrahim. Conjugate priors for generalized linear models. *Statistica Sinica*, 13:461–476, 2003.

[6] Persi Diaconas and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281, 1979.

[7] Ad Feelders and Linda C. van der Gaag. Learning Bayesian network parameters under order constraints. *International Journal of Approximate Reasoning*, 42:37–53, May 2006.

[8] P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society Series B*, 46(2):149–192, 1984.

[9] Tommi S. Jaakkola and Michael I. Jordan. A variational approach to Bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, 1997.

[10] Weiqi Luo. Land use modelling. Internal Report, Food and Envirnoment Research Agency, 2010.

[11] Data collected by the Met Office. `http://www.metoffice.gov.uk/climate/uk/stationdata/`. Accessed: 11/02/2013.

[12] Data collected by the Rural Payments Agency under the integrated administration and control system for the administration of subsidies under the common agricultural policy.

[13] Peter Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58(1):3–34, 1996.

Model checking for imprecise Markov chains

Matthias C. M. Troffaes
Durham University, UK
matthias.troffaes@gmail.com

Damjan Škulj
University of Ljubljana, Slovenia
damjan.skulj@fdv.uni-lj.si

Abstract

We extend probabilistic computational tree logic for expressing properties of Markov chains to imprecise Markov chains, and provide an efficient algorithm for model checking of imprecise Markov chains. Thereby, we provide a formal framework to answer a very wide range of questions about imprecise Markov chains, in a systematic and computationally efficient way, whilst at the same time improving and simplifying model checking for a fairly broad class of Markov decision processes.

Keywords. imprecise Markov chain, model checking, parse tree, logic, computation

1 Introduction

In a nutshell, model checking [2] is a model-based technique which automates the verification of the reliability of a system. To do so, firstly we need a specification of the *system model*, and secondly a specification of *system properties*. The system is then deemed reliable if the model satisfies those properties. Model checking can be performed manually, for instance through peer review, however, for larger models, this can take a lot of time and can cost a lot of money. Moreover, peer review typically does not catch all errors. Therefore, there is a strong need for automating the model checking process.

Interestingly, in model checking literature, a lot of work has been done for so-called *Markov decision processes*—these are Markov chains where the transition probabilities depend on non-deterministic choices, about which we are completely vacuous, and so in fact they are quite closely related to imprecise probability [4, 13, 14, 11, 12].[1] Indeed, so-called *imprecise Markov chains* have been studied by many authors. Hartfiel [7, 6] proposed Markov set-chains models, where transition matrices form matrix intervals. He does not connect his work with imprecise probability formally, but uses several methods similar to those developed in imprecise probability theory. Another more recent case where interval probabilities are involved in the study of Markov chains, outside the formal theory of imprecise probability, is described in [8]. In that work, the interval probabilities are formed by abstraction of a precise Markov chain, that is by merging states. Also this model could be seen as an imprecise Markov chain. A more formal connection between Markov chains and interval probabilities was established by Kozine and Utkin [9], and generalised by Škulj [10] and De Cooman et al. [5]. In [5], lower and upper expectation operators are used instead of sets of probabilities, leading to simpler calculations and more elegant proofs. We follow their approach as well.

In this paper, we investigate model checking techniques for imprecise Markov chains. Although, theoretically, these models can already be checked using existing techniques for Markov decision processes [2, Sec. 10.6] [1], in this paper we follow [5] and restrict ourselves to a very special type of Markov decision process, namely those imprecise Markov chains for which bounds on transition probabilities can be calculated simply by means of linear programming.

Indeed, imprecise Markov chains can be formally connected to Markov decision processes as follows: the extreme points of the set of transition probabilities in the imprecise Markov chain correspond to the set of actions in the Markov decision process. Existing techniques for Markov decision processes require the set of actions to be finite. Interestingly, in our approach, the set of extreme points is not required to be finite, as our model checking algorithm does not depend on the cardinality of the set of extreme points. A second advantage of our approach is that we express our algorithm directly in terms of constraints (lower and upper expectations), rather than in terms of extreme

[1] In operations research, the term Markov decision process has an entirely different meaning.

points or, actions, if you like. This is computationally much more efficient than model checking algorithms which work with actions directly, because it is well known that the number of extreme points is usually much greater than the number of constraints required to describe the same set (for example, see [3]). Even in those cases where the number of constraints is larger, one can still use the extreme points to calculate lower and upper expectations, whence our approach never does worse than existing algorithms which use actions directly.

From the model checking perspective, this paper contributes algorithms that are potentially much faster than traditional methods for Markov decision processes, in essence because we can circumvent sets of probabilities, and instead focus on the constraints. The algorithm is also simpler, and resembles much more closely the one for precise Markov chains. From the imprecise probability perspective, the contribution of this paper is the development of a formal framework to answer a very wide range of questions about Markov chains in a computationally efficient way. In doing so, we put various existing results from the literature about imprecise Markov chains into a common framework.

This paper is structured as follows. Section 2 reviews the existing theory of model checking for Markov chains. Section 3 explains how the logic and model checking algorithm can be adapted to suit imprecise Markov chains. Section 4 has an example. We conclude in Section 5.

2 Model Checking for Markov Chains

Before we move on to imprecise Markov chains, in this section, we briefly review the standard model checking framework for Markov chains [2, Chapter 10]. For simplicity, in this paper, we will restrict ourselves to finite state discrete time Markov chains.

2.1 Model Specification: Transitions, Labelling, Paths, Probabilities

Definition 1 (Markov Chain) *A (finite state, discrete time)* Markov chain *consists of:*

- *a finite set of* states S*,*
- *an* initial probability $\mathrm{P}_0(s)$ *for all* $s \in S$*, and*
- transition probabilities $\mathbf{P}(s,t)$ *for all* $(s,t) \in S^2$*.*

For specifying properties of Markov chains, it is useful to introduce labels as well:

Definition 2 (Labelling) *Consider a finite set of* atomic propositions AP*. A* labelling *of states is then simply a mapping* $L\colon S \to \wp(AP)$ *which associates a set of atomic propositions with every state.*

An atomic proposition is just a convenient way to specify a subset of states. For example, in a reliability problem, we could have

$$AP = \{\text{system working}, \text{system broken}\}, \tag{1}$$

with states of the Markov chain labelled accordingly. In more advanced problems, it is sometimes convenient to allow each state to carry more than one atomic proposition. A trivial labelling is $L(s) = \{s\}$ for every $s \in S$; this is what we will usually assume, unless otherwise stated.

The *digraph* of a Markov chain is a graph where each state is represented by a vertex, and each possible transition ($\mathbf{P}(s,t) > 0$) is an edge—this is the picture we generally draw for a Markov chain.

A *path* is then simply an infinite sequence of states on the digraph:

$$\pi = s_0 s_1 s_2 \cdots \in S^{\mathbb{N}}. \tag{2}$$

The *trace* of a path is its induced sequence of labels:

$$\operatorname{trace}(s_0 s_1 s_2 \cdots) = L(s_0)L(s_1)L(s_2)\cdots \in \wp(AP)^{\mathbb{N}}. \tag{3}$$

A *cylinder set* is a set of paths with a common prefix:

$$\operatorname{cyl}(s_0 s_1 \cdots s_n) = \{s_0 s_1 \cdots s_n s_{n+1} s_{n+2} \cdots : s_{n+1} s_{n+2} \cdots \in S^{\mathbb{N}}\}. \tag{4}$$

For example, the set of paths starting from state s is the cylinder set

$$\operatorname{cyl}(s) = \{s_0 s_1 s_2 \cdots \in S^{\mathbb{N}} \colon s_0 = s\}. \tag{5}$$

Cylinder sets play a central role in Markov chains because these are the sets for which we can very easily calculate their probability:

$$\Pr(\operatorname{cyl}(s_0 s_1 \cdots s_n)) = \mathrm{P}_0(s_0) \prod_{i=0}^{n-1} \mathbf{P}(s_i, s_{i+1}). \tag{6}$$

Also of interest is the probability of a cylinder set conditional on knowing the initial state s_0:

$$\Pr(\operatorname{cyl}(s_0 s_1 \cdots s_n) \mid s_0) = \prod_{i=0}^{n-1} \mathbf{P}(s_i, s_{i+1}). \tag{7}$$

state formula	meaning
$s \models \text{true}$	always satisfied
$s \models a$	$a \in L(s)$
$s \models \neg\Phi$	not $s \models \Phi$
$s \models \Phi \wedge \Psi$	$s \models \Phi$ and $s \models \Psi$
$s \models \mathbb{P}_J(\phi)$	$\Pr(s \models \phi) \in J$
path formula	**meaning**
$\pi \models \bigcirc\Phi$	$\pi[1] \models \Phi$
$\pi \models \Phi \mathbin{\mathsf{U}} \Psi$	$\exists j \geq 0$: $\big((\forall 0 \leq k < j\colon \pi[k] \models \Phi)$ and $\pi[j] \models \Psi\big)$
$\pi \models \Phi \mathbin{\mathsf{U}}^{\leq n} \Psi$	$\exists 0 \leq j \leq n$: $\big((\forall 0 \leq k < j\colon \pi[k] \models \Phi)$ and $\pi[j] \models \Psi\big)$

Table 1: Semantics of state and path formulas.

2.2 Property Specification: Probabilistic Computation Tree Logic

A formal and very useful way of specifying properties of Markov chains goes via *probabilistic computation tree logic*, where we distinguish between two types of properties:

1. Properties of *states* of the system:
$$s \models \Phi \text{ if state } s \text{ satisfies state formula } \Phi. \quad (8)$$
2. Properties of *paths* of the system:
$$\pi \models \phi \text{ if path } \pi \text{ satisfies path formula } \phi. \quad (9)$$

State formulas are denoted by upper case greek letters Φ, Ψ, and so on. Path formulas are denoted by lower case greek letters ϕ, ψ, and so on.

State and path formulas Φ and ϕ are taken from a grammar with the following syntax:

$$\Phi ::= \text{true} \quad | \quad a \quad | \quad \Phi_1 \wedge \Phi_2 \quad | \quad \neg\Phi \quad | \quad \mathbb{P}_J(\phi) \quad (10)$$

$$\phi ::= \bigcirc\Phi \quad | \quad \Phi \mathbin{\mathsf{U}} \Psi \quad | \quad \Phi \mathbin{\mathsf{U}}^{\leq n} \Psi \quad (11)$$

where $\bigcirc$ ("next") and U ("until") operators must be directly preceded by a $\mathbb{P}_J$ operator. The semantics, or meaning, of these formulas is summarized in Table 1, where

$$\Pr(s \models \phi) = \Pr(\{\pi \in \text{cyl}(s)\colon \pi \models \phi\} \mid s) \quad (12)$$

and $\pi[i] = s_i$ for $\pi = s_0 s_1 \cdots$. The usual operators can be derived from the above ones:

$$\Phi \vee \Psi := \neg((\neg\Phi) \wedge (\neg\Psi)) \qquad \Phi \text{ or } \Psi \quad (13)$$
$$\Diamond\Phi := \text{true} \mathbin{\mathsf{U}} \Phi \qquad \text{eventually } \Phi \quad (14)$$
$$\Box\Phi := \neg(\Diamond(\neg\Phi)) \qquad \text{always } \Phi \quad (15)$$

as well as bounded versions $\Diamond^{\leq n}$ and $\Box^{\leq n}$ of $\Diamond$ and $\Box$, implication, exclusive or, equivalence, and so on.

For a non-trivial example of a state formula, consider a system modelled by a Markov chain whose states are labelled with atomic propositions taken from $AP = \{\text{working}, \text{broken}\} = \{w, b\}$. The property

$$s \models \mathbb{P}_{[0.99,1]}\Big(\Diamond^{\leq 20}\big(w \wedge \mathbb{P}_{[0,0.01]}(w \mathbin{\mathsf{U}}^{\leq 10} b)\big)\Big) \quad (16)$$

is then satisfied when, starting from state s, with probability at least 0.99, within 20 steps, the system reaches a working state, from which the probability that the system breaks down within the next 10 steps is less than 0.01. Model checking provides an automated method for verifying any such formula.

Before we move on to the algorithm, one technical issue that arises is is whether $\mathbb{P}_J(\phi)$ is well defined, or more specifically, whether the probability $\Pr(s \models \phi)$ (see Eq. (12)) exists. The key observation is:

Theorem 1 *For every state s and every path formula ϕ,*

$$\{\pi \in \text{cyl}(s)\colon \pi \models \phi\} \quad (17)$$

is a countable union of cylinder sets.

The above theorem, along with the fact that the probability specification in Eqs. (6) and (7) can be extended to a σ-field containing all countable unions of cylinder sets, imply that $\Pr(s \models \phi)$ exists; see for instance [2, Lemma 10.39] for a proof of Theorem 1.

2.3 Model Checking: Automated Algorithm

The central question we aim to answer is: given a state s and a state formula Φ, does s satisfy Φ? In a nutshell, the algorithm works as follows:

- traverse the *parse tree* of Φ, visiting all subformulas, starting at the leaves of the tree and working back to its root,
- at each subformula Ψ, calculate set of states which *satisfy* Ψ
$$\text{Sat}(\Psi) = \{s' \in S\colon s' \models \Psi\}, \quad (18)$$
- check that $s \in \text{Sat}(\Phi)$.

Figure 1 shows the parse tree of the formula on the right hand side of Eq. 16, and Figure 2 demonstrates how we evaluate Sat through the parse tree.

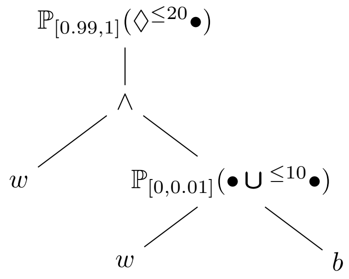

$\mathbb{P}_{[0.99,1]}(\Diamond^{\leq 20}\bullet)$
$\wedge$
w
$\mathbb{P}_{[0,0.01]}(\bullet \mathsf{U}^{\leq 10}\bullet)$
w
b

Figure 1: Parse tree of the formula on the right hand side of Eq. 16.

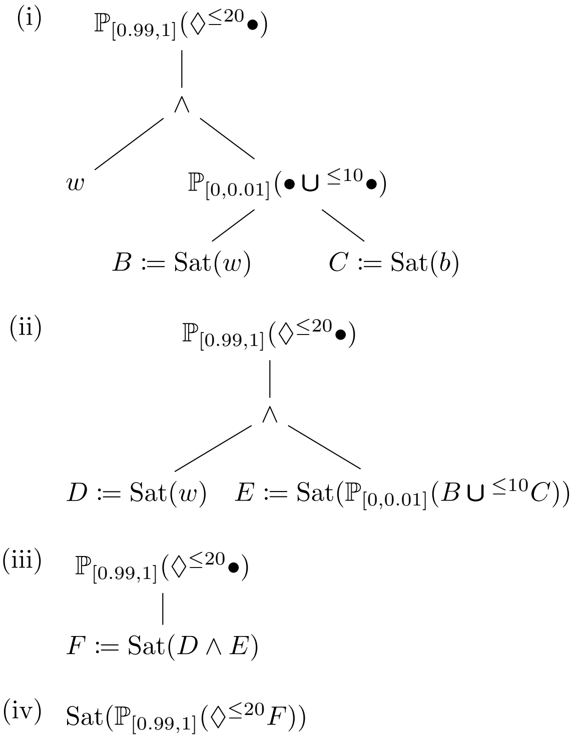

(i)
$\mathbb{P}_{[0.99,1]}(\Diamond^{\leq 20}\bullet)$
$\wedge$
w
$\mathbb{P}_{[0,0.01]}(\bullet \mathsf{U}^{\leq 10}\bullet)$
$B := \mathrm{Sat}(w)$
$C := \mathrm{Sat}(b)$
(ii)
$\mathbb{P}_{[0.99,1]}(\Diamond^{\leq 20}\bullet)$
$\wedge$
$D := \mathrm{Sat}(w)$
$E := \mathrm{Sat}(\mathbb{P}_{[0,0.01]}(B \mathsf{U}^{\leq 10}C))$
(iii)
$\mathbb{P}_{[0.99,1]}(\Diamond^{\leq 20}\bullet)$
$F := \mathrm{Sat}(D \wedge E)$
(iv)
$\mathrm{Sat}(\mathbb{P}_{[0.99,1]}(\Diamond^{\leq 20}F))$

Figure 2: Evaluating Sat through the parse tree.

Effectively, we calculate $\mathrm{Sat}(\Psi)$ by applying the following formulas recursively:

$$\mathrm{Sat}(\mathrm{true}) = S \tag{19}$$

$$\mathrm{Sat}(a) = \{s \in S\colon a \in L(s)\} \tag{20}$$

$$\mathrm{Sat}(\Phi \wedge \Psi) = \mathrm{Sat}(\Phi) \cap \mathrm{Sat}(\Psi) \tag{21}$$

$$\mathrm{Sat}(\neg\Phi) = S \setminus \mathrm{Sat}(\Phi) \tag{22}$$

$$\mathrm{Sat}(\mathbb{P}_J(\phi)) = \{s \in S\colon \Pr(s \vDash \phi) \in J\} \tag{23}$$

where

$$\Pr(s \vDash \bigcirc\Phi) = \mathbf{P}(s, \mathrm{Sat}(\Phi)) = \sum_{t \in \mathrm{Sat}(\Phi)} \mathbf{P}(s,t) \tag{24}$$

and, under certain regularity conditions,

$$\Pr(s \vDash \Phi \mathsf{U} \Psi) = \lim_{n\to\infty} \Pr(s \vDash \Phi \mathsf{U}^{\leq n}\Psi). \tag{25}$$

Note that there are efficient ways to determine $\Pr(s \vDash \Phi \mathsf{U}^{\leq n}\Psi)$ for large n. There are also methods for evaluating $\Pr(s \vDash \Phi \mathsf{U} \Psi)$ directly, under arbitrary conditions; for details we refer to [2, pp. 761–762].

Thus, the only probability we yet have to evaluate is $\Pr(s \vDash \Phi \mathsf{U}^{\leq n}\Psi)$. Let $C := \mathrm{Sat}(\Phi)$ and $B := \mathrm{Sat}(\Psi)$, and for simplicity assume a trivial labelling $L(s) = \{s\}$, so we can write $\Pr(s \vDash C \mathsf{U}^{\leq n}B)$ for $\Pr(s \vDash \Phi \mathsf{U}^{\leq n}\Psi)$.

- If $s \in B$ then $\Pr(s \vDash C \mathsf{U}^{\leq n}B) = 1$.
- Otherwise, if $s \notin C$ then $\Pr(s \vDash C \mathsf{U}^{\leq n}B) = 0$.
- Otherwise, $s \in C \setminus B$, and

$$\begin{aligned}\Pr(s \vDash C \mathsf{U}^{\leq n}B) &= \Pr_*(s \vDash \bigcirc^n B) && (26)\\ &= \mathbf{P}_*^n(s,B) && (27)\\ &= \sum_{t\in B}\mathbf{P}_*^n(s,t) && (28)\end{aligned}$$

where $\Pr_*$ and $\mathbf{P}_*$ denotes the probabilities associated with the modified Markov chain where all states, except those in $C \setminus B$, have been made absorbing. There are many efficient ways to evaluate the n-step transition probability matrix $\mathbf{P}_*^n(s,t)$.

3 Model Checking for Imprecise Markov Chains

3.1 Model Specification: Credal Sets and Upper Transition Operator

Definition 3 (Imprecise Markov Chain) *A (finite state, discrete time)* imprecise Markov chain *consists of:*

- *a finite set of* states S*,*
- *an* initial credal set $\mathcal{P}_0$ *on* S*, specified through linear constraints on* $\mathrm{P}_0(\cdot)$*, and*
- *a* transition credal set $\boldsymbol{\mathcal{P}}(s)$ *for each* $s \in S$*, specified through linear constraints on* $\mathbf{P}(s, \cdot)$*.*

The sensitivity interpretation is that, at each step, the transition is described by $\mathbf{P}(s, \cdot) \in \boldsymbol{\mathcal{P}}(s)$ but we do not know which element.

Clearly, this model is not the most general one. Firstly, it has *separately specified rows* [10], that is, a separate model for transitions from each state. Secondly, it features *non-stationarity* (in the sensitivity interpretation), as the actual transition probabilities may change from step to step, and are only constrained to belong to their credal set at each step. These assumptions make the model computationally tractable, and almost as easy to work with as precise Markov chains. Indeed, it turns out that for most calculations of typical interest, we can entirely ignore the credal sets, and instead work with a single operator that can be evaluated through linear programming [5, 10].

Indeed, for typical calculations, we are interested in lower and upper probabilities of events, or more generally, lower and upper expectations of random quantities. Let $\mathcal{L}(S)$ denote the set of all random quantities, also called *gambles*, on S. Gambles are denoted by lower case letters f, g, and so on.

For instance, we could be interested in the lower and upper expectation of a gamble f on the next state, given the current state s:

Definition 4 (Transition Operators) *The operators* $\underline{\mathrm{T}}\colon \mathcal{L}(S) \to \mathcal{L}(S)$ *and* $\overline{\mathrm{T}}\colon \mathcal{L}(S) \to \mathcal{L}(S)$ *defined by*

$$(\underline{\mathrm{T}}(f))(s) := \min_{\mathbf{P}(s,\cdot)\in\boldsymbol{\mathcal{P}}(s)} \sum_{t\in S} \mathbf{P}(s,t) f(t), \tag{29}$$

$$(\overline{\mathrm{T}}(f))(s) := \max_{\mathbf{P}(s,\cdot)\in\boldsymbol{\mathcal{P}}(s)} \sum_{t\in S} \mathbf{P}(s,t) f(t), \tag{30}$$

are called the lower and upper transition operators*.*

A key point is that calculation of $\underline{\mathrm{T}}(f)$ and $\overline{\mathrm{T}}(f)$ is efficient. In fact, once we have specified the linear constraints that determine the credal sets $\boldsymbol{\mathcal{P}}(s)$ for each $s \in S$, Eqs. (29) and (30) can be evaluated via linear programming [11, Chapter 3]. Many interesting characteristics of the imprecise Markov chain can be derived just from $\overline{\mathrm{T}}$ [5]. To ease notation, we will often write $\overline{\mathrm{T}}f$ for $\overline{\mathrm{T}}(f)$.

More generally, we might be interested in the lower and upper expectations of gambles on the state after exactly n steps, given the current state s. For instance, what is the upper probability of ending up in $B \subseteq S$ after exactly n steps? Let us use the usual notation for the indicator gamble of a set B:

$$I_B(s) := \begin{cases} 1 & \text{if } s \in B, \\ 0 & \text{if } s \notin B. \end{cases} \tag{31}$$

Clearly, $(\overline{\mathrm{T}} I_B)(s)$ is the desired upper probability for $n = 1$. By marginal extension [11, Sec. 6.7.2], it follows that $\overline{\mathrm{T}}(\overline{\mathrm{T}} I_B))(s)$ is the desired upper probability for $n = 2$; we will use the notation $\overline{\mathrm{T}}^2 I_B$ for $\overline{\mathrm{T}}(\overline{\mathrm{T}} I_B)$. Similarly, by $(\overline{\mathrm{T}}^n I_B)(s)$, we denote the desired upper probability for arbitrary n, found by repeated application of $\overline{\mathrm{T}}$.

Definition 5 (n-Step Transition Operators) *The operators* $\underline{\mathrm{T}}^n\colon \mathcal{L}(S) \to \mathcal{L}(S)$ *defined by*

$$(\underline{\mathrm{T}}^n f)(s) = \begin{cases} (\underline{\mathrm{T}}(\underline{\mathrm{T}}^{n-1} f))(s) & \text{if } n > 1 \\ (\underline{\mathrm{T}} f)(s) & \text{if } n = 1 \end{cases} \tag{32}$$

is called the n-step lower transition operator*, and*

$$(\overline{\mathrm{T}}^n f)(s) = \begin{cases} (\overline{\mathrm{T}}(\overline{\mathrm{T}}^{n-1} f))(s) & \text{if } n > 1 \\ (\overline{\mathrm{T}} f)(s) & \text{if } n = 1 \end{cases} \tag{33}$$

is called the n-step upper transition operator*.*

For model checking, we will use the notation

$$\underline{\mathrm{T}}^n(s, B) := \underline{\mathrm{T}}^n(I_B)(s) \tag{34}$$

$$\overline{\mathrm{T}}^n(s, B) := \overline{\mathrm{T}}^n(I_B)(s) \tag{35}$$

to denote the lower and upper probability of ending up in B given that we started in s.

3.2 Property Specification: Imprecise Probabilistic Computation Tree Logic

The syntax and semantics of state and path formulas is as before, with only two differences.

First, for simplicity, here, we will exclude U from our logic, to avoid technical issues with countable unions of cylinder sets resulting from the U operator. The $\bigcirc$ and $\mathsf{U}^{\leq n}$ operators pose no problem. Consequently, we can impose an upper bound on the number of steps, after which we are no longer interested in the Markov chain, so all sets and partitions involved can be assumed to have finite cardinality. In this way, we avoid a host of technical problems. Problems involving infinite horizon require additional considerations and will therefore be tackled elsewhere. In any case, for practical applications, time-bounded properties will often be sufficient.

Secondly, and more crucially, we need to use a *different semantics for* $\mathbb{P}$:

$$s \vDash \mathbb{P}_J(\phi) \tag{36}$$

will mean that

$$[\underline{\Pr}(s \vDash \phi), \overline{\Pr}(s \vDash \phi)] \subseteq J \tag{37}$$

where

$$\underline{\Pr}(s \vDash \phi) := \underline{\Pr}(\{\pi \in \mathrm{cyl}(s) \colon \pi \vDash \phi\} \mid s) \tag{38}$$
$$\overline{\Pr}(s \vDash \phi) := \overline{\Pr}(\{\pi \in \mathrm{cyl}(s) \colon \pi \vDash \phi\} \mid s) \tag{39}$$

and the right hand sides denote the lower and upper expectations corresponding to *natural extension* [13, 14] [11, Sec. 8.1], where the Markov condition is expressed through epistemic irrelevance [5].

3.3 Model Checking: Automated Algorithm

Again, the central question is: given a state s and a state formula Φ, does s satisfy Φ? It is easy to see that we can implement an algorithm exactly as before, namely by *evaluating* Sat *throughout the parse tree.*

The non-trivial differences are:

$$\underline{\Pr}(s \vDash \bigcirc\Phi) = \underline{\mathrm{T}}(s, \mathrm{Sat}(\Phi)) \tag{40}$$

and

$$\underline{\Pr}(s \vDash \Phi \mathbin{\mathsf{U}}^{\leq n} \Psi) = \begin{cases} 1 & \text{if } s \in \mathrm{Sat}(\Psi) \\ 0 & \text{if } s \notin \mathrm{Sat}(\Phi) \cup \mathrm{Sat}(\Psi) \\ \underline{\mathrm{T}}_*^n(s, \mathrm{Sat}(\Psi)) & \text{if } s \in \mathrm{Sat}(\Phi) \setminus \mathrm{Sat}(\Psi) \end{cases} \tag{41}$$

where the $*$ denotes the imprecise Markov chain with all states outside $\mathrm{Sat}(\Phi) \setminus \mathrm{Sat}(\Psi)$ being made absorbing:

$$(\underline{\mathrm{T}}_* f)(s) = \begin{cases} (\underline{\mathrm{T}} f)(s) & \text{if } s \in \mathrm{Sat}(\Phi) \setminus \mathrm{Sat}(\Psi), \\ f(s) & \text{otherwise.} \end{cases} \tag{42}$$

The formulas for upper expectations are trivially similar.

4 Example

Consider a Markov chain with the set of states $S = \{s_1, s_2, s_3, s_4\}$ and the lower and upper transition probabilities:

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{4} & 0 \\ 0 & \frac{1}{4} & \frac{1}{6} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{1}{2} \end{pmatrix}, \tag{43}$$

$$U = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{7}{12} & \frac{5}{12} & \frac{1}{2} & 0 \\ 0 & \frac{7}{12} & \frac{1}{2} & \frac{7}{12} \\ 0 & 0 & \frac{1}{2} & \frac{3}{4} \end{pmatrix}, \tag{44}$$

that is,

$$\boldsymbol{\mathcal{P}}(s_i) = \{\mathbf{P}(s_i) \colon L(s_i) \leq \mathbf{P}(s_i) \leq U(s_i)\} \tag{45}$$

where $L(s_i)$ is the ith row of L, and $U(s_i)$ is the ith row of U. As mentioned, the corresponding transition operators $\underline{\mathrm{T}}$ and $\overline{\mathrm{T}}$ can be efficiently evaluated through linear programming.

We are interested in verifying the property:

$$s_2 \vDash \mathbb{P}_{[0.9,1]}\Big(\Diamond^{\leq 2}\Big(\mathbb{P}_{[0.4,1]}\big((s_2 \vee s_3) \mathbin{\mathsf{U}}^{\leq 6} s_1\big)\Big)\Big) \tag{46}$$

that is, starting from s_2, with probability at least 0.9, in at most two steps we end up in a state from which, with probability at least 0.4, we end up in s_1 in at most 6 steps without visiting s_4.

To answer the above question, we start with evaluating:

$$\mathrm{Sat}(s_2 \vee s_3) = \{s_2, s_3\} \text{ and } \mathrm{Sat}(s_1) = \{s_1\}. \tag{47}$$

Next, we need:

$$\mathrm{Sat}\Big(\mathbb{P}_{[0.4,1]}\big(\{s_2, s_3\} \mathbin{\mathsf{U}}^{\leq 6} \{s_1\}\big)\Big) \tag{48}$$

Clearly, s_1 belongs to the set because:

$$\underline{\Pr}(\{s_2, s_3\} \mathbin{\mathsf{U}}^{\leq 6} \{s_1\} \mid s_1) = \overline{\Pr}(\{s_2, s_3\} \mathbin{\mathsf{U}}^{\leq 6} \{s_1\} \mid s_1) = 1, \tag{49}$$

and s_4 does not belong to the set because:

$$\underline{\Pr}(\{s_2, s_3\} \mathbin{\mathsf{U}}^{\leq 6} \{s_1\} \mid s_4) = \overline{\Pr}(\{s_2, s_3\} \mathbin{\mathsf{U}}^{\leq 6} \{s_1\} \mid s_4) = 0, \tag{50}$$

For s_2,

$$\underline{\Pr}(\{s_2, s_3\} \mathbin{\mathsf{U}}^{\leq 6} \{s_1\} \mid s_2) = \underline{\mathrm{T}}_*^6(s_2, \{s_1\}) \tag{51}$$
$$= 0.4809 \tag{52}$$
$$\overline{\Pr}(\{s_2, s_3\} \mathbin{\mathsf{U}}^{\leq 6} \{s_1\} \mid s_2) = \overline{\mathrm{T}}_*^6(s_2, \{s_1\}) \tag{53}$$
$$= 0.8685 \tag{54}$$

so s_2 belongs to the set, as $[0.4809, 0.8685] \subseteq [0.4, 1]$. For s_3,

$$\underline{\Pr}(\{s_2, s_3\} \mathbin{\mathsf{U}}^{\leq 6} \{s_1\} \mid s_3) = \underline{\mathrm{T}}_*^6(s_2, \{s_1\}) \tag{55}$$
$$= 0.1415 \tag{56}$$
$$\overline{\Pr}(\{s_2, s_3\} \mathbin{\mathsf{U}}^{\leq 6} \{s_1\} \mid s_3) = \overline{\mathrm{T}}_*^6(s_2, \{s_1\}) \tag{57}$$
$$= 0.5934 \tag{58}$$

so s_3 does not belong to the set, as $[0.1415, 0.5934] \not\subseteq [0.4, 1]$. Concluding,

$$\mathrm{Sat}\Big(\mathbb{P}_{[0.4,1]}\big(\{s_2, s_3\} \cup^{\leq 6} \{s_1\}\big)\Big) = \{s_1, s_2\}. \tag{59}$$

Whence, finally, we need to calculate

$$\mathrm{Sat}\big(\mathbb{P}_{[0.9,1]}(\mathrm{true} \cup^{\leq 2} \{s_1, s_2\})\big). \tag{60}$$

In fact, to verify Eq. (46), we only need to determine whether s_2 belongs to this set. Clearly it does, because

$$\begin{aligned}\underline{\mathrm{Pr}}(&\mathrm{true} \cup^{\leq 2} \{s_1, s_2\} \mid s_2) \\ &= \overline{\mathrm{Pr}}(\mathrm{true} \cup^{\leq 2} \{s_1, s_2\} \mid s_2) = 1,\end{aligned} \tag{61}$$

and obviously $\{1\} \subseteq [0.9, 1]$.

5 Conclusion

We have provided a model checking algorithm for imprecise Markov chains that is equally easy as the corresponding algorithm for precise Markov chains. Rather surprisingly, the same bag of tricks from the precise case can be used for the imprecise case.

An interesting open problem is the evaluation of $\Phi \cup \Psi$, where there is no bound on number of steps. Intuitively, it seems obvious that we can do this by evaluating $\Phi \cup^{\leq n} \Psi$ for large enough n. How large should n be? When is convergence guaranteed? Are there also generally applicable direct methods as in the precise case? We may also need to deal with the technical issue of dealing with a countable number of partitions to express the Markov condition, a case which is not handled by Walley's theory [11, Sec. 8.1], but covered by Williams's approach [13, 14].

Finally, there are additional optimisation tricks possible for precise Markov chains (see for instance [2, Sec. 10.1.1, Remark 10.17]). Even though these are somewhat technical, it would be interesting to see whether we can recycle such tricks for imprecise Markov chains as well.

Acknowledgements

We thank all reviewers for their constructive comments, which have led to some improvements of the paper.

References

[1] Christel Baier, Holger Hermanns, Joost-Pieter Katoen, and Boudewijn R. Haverkort. Efficient computation of time-bounded reachability probabilities in uniform continuous-time Markov decision processes. *Theoretical Computer Science*, 345(1):2–26, November 2005. doi:10.1016/j.tcs.2005.07.022.

[2] Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking.* The MIT Press, 2008.

[3] Sancho E. Berenguer and Robert L. Smith. The expected number of extreme points of a random linear program. *Mathematical Programming*, 35(2):129–134, 1986. doi:10.1007/BF01580643.

[4] George Boole. *An investigation of the laws of thought on which are founded the mathematical theories of logic and probabilities.* Walton and Maberly, London, 1854.

[5] Gert de Cooman, Filip Hermans, and Erik Quaeghebeur. Imprecise Markov chains and their limit behavior. *Probability in the Engineering and Informational Sciences*, 23(4):597–635, October 2009. arXiv:0801.0980, doi:10.1017/S0269964809990039.

[6] D. J. Hartfiel. *Markov Set-Chains.* Springer-Verlag, Berlin, 1998.

[7] D. J. Hartfiel and E. Seneta. On the theory of Markov set-chains. *Advances in Applied Probability*, 26(4):947–964, 1994. doi:10.2307/1427899.

[8] Joost-Pieter Katoen, Daniel Klink, Martin Leucker, and Verena Wolf. Three-valued abstraction for probabilistic systems. *The Journal of Logic and Algebraic Programming*, 81(4):356–389, 2012. doi:10.1016/j.jlap.2012.03.007.

[9] Igor O. Kozine and Lev V. Utkin. Interval-valued finite Markov chains. *Reliable Computing*, 8:97–113, 2002. doi:10.1023/A:1014745904458.

[10] Damjan Škulj. Discrete time Markov chains with interval probabilities. *International Journal of Approximate Reasoning*, 50(8):1314–1329, 2009. doi:10.1016/j.ijar.2009.06.007.

[11] Peter Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.

[12] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I — Intervallwahrscheinlichkeit als umfassendes Konzept.* Physica, Heidelberg, 2001. In cooperation with Thomas Augustin and Anton Wallner.

[13] Peter M. Williams. Notes on conditional previsions. Technical report, School of Math. and Phys. Sci., Univ. of Sussex, 1975.

[14] Peter M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44(3):366–383, 2007. `doi:10.1016/j.ijar.2006.07.019`.

An imprecise boosting-like approach to regression

Lev V. Utkin
Department of Control, Automation, and
System Analysis, Saint Petersburg State
Forest Technical University
lev.utkin@mail.ru

Andrea Wiencierz
Department of Statistics, LMU Munich
andrea.wiencierz@stat.uni-muenchen.de

Abstract

This paper is about a generalization of ensemble methods for regression which are based on variants of the basic AdaBoost algorithm. The generalization of these regression methods consists in restricting the unit simplex for the weights of the instances to a smaller set of weighting probabilities. The proposed algorithms cover the standard AdaBoost-based regression algorithms and standard regression as special cases. Various imprecise statistical models can be used to obtain the restricted set of probabilities. One advantage of the proposed algorithms compared to the basic AdaBoost-based regression methods is that they have less tendency to over-fitting, because the weights of the hard instances are restricted. Finally, some simulations and applications also indicate a better performance of the proposed generalized methods.

Keywords. Regression, AdaBoost, algorithm, linear-vacuous mixture model, Kolmogorov–Smirnov bounds.

1 Introduction

Regression modeling is one of the main problems in applied statistics. Roughly speaking, the aim is to estimate a function $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^m$ with $m \in \mathbb{N}$ and $\mathcal{Y} \subset \mathbb{R}$, from a finite set of noisy samples $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ for some $n \in \mathbb{N}$. A large number of regression methods were developed in the last decades, many of which are based on the minimization of a risk functional defined by a certain loss function and by the probability distribution of the data (see, e.g., [9, 18, 21]). In practice, the estimated function is obtained by minimizing the so-called empirical risk (possibly regularized), the sum of the loss values for the given data points divided by n, which can be interpreted as the risk functional associated with the empirical distribution of the data. The empirical distribution can be represented as the point $\hat{p} = (n^{-1}, \ldots, n^{-1})$ in the unit simplex with n vertices denoted by $S(1, n)$. In this paper, we focus on this kind of regression methods within the proposed algorithms, because it is very easy to incorporate individual weights for the instances, which is a core element of the algorithms we want to generalize. The weighted estimates can simply be interpreted as minimizers of the risk functional associated with another discrete probability distribution $p = (p_1, \ldots, p_n)$ of the data than the empirical distribution $\hat{p}$.

A very popular approach to regression is the ensemble methodology. The popularity of ensemble methods for regression stems from success of boosting methods for classification, in particular, of the well-known AdaBoost (Adaptive Boosting) algorithm proposed by [5]. AdaBoost is a general purpose boosting algorithm that can be used in conjunction with many different learning algorithms to improve their performance. The basic scheme of the AdaBoost algorithm for classification is the following: Initially, a standard classifier is estimated, assigning identical weights to all examples, then, in each of a previously fixed number of iterations, the weights of all misclassified examples are increased, while the weights of correctly classified examples are decreased, before again computing a classifier accounting for the unequal weights of the instances. Thus, with each step, the classifier focuses more and more on the difficult examples of the training data set, thereby improving the classification accuracy. The final result obtained by AdaBoost is a weighted majority vote of the classifiers of each iteration, which has a better prediction performance than each of the individual classifiers alone. Detailed reviews of boosting methods can be found, e.g., in [1, 3, 12, 14].

One of the first boosting algorithms for regression is the so-called AdaBoost.R2 proposed in [2], where real-valued residuals replace the 0–1 misclassification errors in the evaluation of the estimates. However, the base regression estimates are evaluated by the weighted average of the absolute values of the residuals scaled to $[0, 1]$, which is a similar error measure to the misclassification rate. Up to the recent years, many more boosting methods for regression have been developed, a recent survey is provided in [13]. In contrast to most of the ensemble-based algorithms using the weighted average of base regression estimates as their final regression functions, [11] analyzed the choice of the weighted median and proposed the corresponding algorithm called MedBoost. The author proved boosting-type convergence of the algorithm and gave clear conditions for the convergence of the robust training error. Another interesting boosting scheme for regression problems is proposed in [17], where a threshold value for the residuals is introduced to transform the real-valued errors back to the 0–1 errors, which directly fit into the AdaBoost algorithm for classification. This adaptation of the AdaBoost algorithm is called AdaBoost.RT and its properties were further investigated in [15].

A common feature of these boosting algorithms is that they iteratively search for a discrete probability distribution of the training data such that the regression error is minimized. The adapted weighting probabilities may be arbitrary points in the unit simplex. This can lead to over-fitting, when too large weights are assigned to a few hard-to-learn examples. There are different approaches to deal with this problem. One way of overcoming the problem of over-fitting in the context of regression is the so-called shrinkage regularization, where the weights of the base regression estimates are reduced, and thus, the learning rate of the boosting algorithm (see, e.g., [7]). Another interesting approach is based on restricting the weights, e.g., by fixing a maximum size of the weights a priori. In this paper, we follow this idea but we propose to use imprecise statistical models like the linear-vacuous mixture model or the Kolmogorov–Smirnov bounds to restrict the set of weighting probabilities. To modify the boosting algorithms accordingly, we replace the adaption of the instances' weighting probabilities with the updating of weights in the convex linear combination of the extreme points of the restricted set. Thus, we here present a general tool for modifying available boosting algorithms and for constructing a number of new ensemble-based methods which avoid the problem of over-fitting.

In the following two sections, we propose the corresponding modifications of two popular boosting algorithms: AdaBoost.R2 introduced in [2] and AdaBoost.RT proposed in [17]. Section 4 reviews suitable imprecise probability models to obtain the restricted set of weighting probabilities. Finally, we present the results of simulations based on synthetic data and on real data.

2 AdaBoost.R2 and its modification

At first, we modify the AdaBoost.R2 algorithm proposed in [2]. The scheme of this boosting algorithm for regression is presented as Algorithm 1. Given a training data set $Z = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ and a regression method which is suitable for weighted estimation, the algorithm requires a maximum number of iterations $T \in \mathbb{N}$ to be chosen a priori. Then, the iteration index t is set to one and the weighting probabilities $p_i^{(1)}$ are set to n^{-1} for all $i \in \{1, \ldots, n\}$. (Alternatively, the vector $p^{(1)}$ could be randomly selected from the unit simplex $S(1, n)$.) In each iteration step $t \in \{1, \ldots, T\}$, a regression function $\hat{f}^{(t)}$ is estimated using the weights $p^{(t)}$. In contrast to AdaBoost for classification, where the estimated classifiers are evaluated by their average misclassification error, the regression estimates are evaluated on the basis of the absolute residuals $|y_i - \hat{f}^{(t)}(x_i)|$ with $i \in \{1, \ldots, n\}$. Yet, to obtain an overall error measure similar to the misclassification rate, the absolute residuals are divided by the maximum value $D^{(t)}$ such that the weighted sum $\epsilon^{(t)}$ of the normalized residuals $\hat{e}_1^{(t)}, \ldots, \hat{e}_n^{(t)}$ lies in the interval $[0, 1]$. If $\epsilon^{(t)} > 0.5$, we exit the loop and use only the first $t - 1$ regression estimates to determine the final result. In the context of classification this is a sensible stopping criterion, because it

Algorithm 1 AdaBoost.R2

```
Require: Maximum number of iterations $T$ and training data set $Z$.
Ensure: $\alpha^{(t)}$ and $\hat{f}^{(t)}$ for all $t \in \{1, \ldots, T\}$;
  set $t \leftarrow 1$ and $p^{(t)} \leftarrow (n^{-1}, \ldots, n^{-1})$;
  repeat
    estimate $\hat{f}^{(t)}$ using weighting probabilities $p^{(t)}$;
    compute $D^{(t)} \leftarrow \max_{j \in \{1,\ldots,n\}} |y_j - \hat{f}^{(t)}(x_j)|$;
    compute normalized errors for all $i \in \{1, \ldots, n\}$:
    $\hat{e}_i^{(t)} \leftarrow \frac{|y_i - \hat{f}^{(t)}(x_i)|}{D^{(t)}}$;
    calculate the overall error of $\hat{f}^{(t)}$:
    $\epsilon^{(t)} \leftarrow \sum_{i=1}^{n} p_i^{(t)} \hat{e}_i^{(t)}$;
    if $\epsilon^{(t)} > 0.5$ then
      $T \leftarrow t - 1$;
    end if
    compute contribution of $\hat{f}^{(t)}$ to the final result:
    $\alpha^{(t)} \leftarrow \ln\left(\frac{1-\epsilon^{(t)}}{\epsilon^{(t)}}\right)$;
    adapt weights for all $i \in \{1, \ldots, n\}$:
    $p_i^{(t+1)} \leftarrow p_i^{(t)} \exp\left(-\alpha^{(t)}(1 - \hat{e}_i^{(t)})\right)$;
    normalize $p^{(t+1)}$ to be a proper distribution;
    $t++$
  until $t > T$
  normalize $\alpha^{(1)}, \ldots, \alpha^{(T)}$ such that $\sum_{t=1}^{T} \alpha^{(t)} = 1$;
  compute regression function $\hat{f} \leftarrow \sum_{t=1}^{T} \alpha^{(t)} \hat{f}^{(t)}$.
```

means that classifiers with an error rate higher than 50% may not contribute to the combined result. However, in the regression context the usefulness of this stopping criterion is less clear. Here, it corresponds to stopping the iterations when the situation arises, where the average normalized residual is larger than 50% of the maximum absolute residual. If $\epsilon^{(t)} \leq 0.5$, the overall error is used to determine the contribution $\alpha^{(t)}$ of the estimated function $\hat{f}^{(t)}$ in the combined result $\hat{f}$. Furthermore, the weighting probabilities of the instances are adapted by the formula:

$$p_i^{(t+1)} = p_i^{(t)} \exp\left(-\alpha^{(t)}(1 - \hat{e}_i^{(t)})\right)$$

for all $i \in \{1, \ldots, n\}$. Thus, the weights of examples with relatively large residuals are increased, while the others are decreased. As the last step within each iteration, we normalize $(p_1^{(t+1)}, \ldots, p_n^{(t+1)})$ to obtain a proper weighting distribution where $\sum_{i=1}^n p_i^{(t+1)} = 1$. Finally, when the loop is ended, the $\alpha^{(1)}, \ldots, \alpha^{(T)}$ are adjusted such that $\sum_{t=1}^T \alpha^{(t)} = 1$ and the combined result $\hat{f} = \sum_{t=1}^T \alpha^{(t)} \hat{f}^{(t)}$ is determined.

According to the adaption rule, the distribution of weighting probabilities p can be an arbitrary point in $S(1, n)$ including its vertices. Indeed, as already shown for the basic AdaBoost algorithm in [5], the weighting probabilities of the examples tend to concentrate on instances which have large residuals compared with the other data points and may be outliers. Hence, the regression function will be estimated by taking mainly these hard-to-learn examples into account. The obtained estimated function will perform well on these extreme data points but may perform rather poor on the other examples. This property is called over-fitting, because the out-of-sample prediction performance of such a regression estimate may be very bad, as the actual functional relation is better reflected by the neglected examples.

Let us now consider a set $\mathcal{P}$ of probability distributions, which is a subset of the unit simplex, i.e., $\mathcal{P} \subset S(1, n)$. We assume that $\mathcal{P}$ is convex, i.e., it is produced by finitely many linear constraints. This implies that it is totally defined by its extreme points $q^{(k)} = (q_1^{(k)}, \ldots, q_n^{(k)})$ for all $k \in \{1, \ldots, r\}$ with $r \in \mathbb{N}$. Thus, every probability distribution $p \in \mathcal{P}$ can be represented as

$$p = \sum_{k=1}^{r} \lambda_k q^{(k)},$$

where $\lambda = (\lambda_1, \ldots, \lambda_r)$ is a vector of weights such that $\sum_{k=1}^r \lambda_k = 1$.

The core idea of the modification of AdaBoost.R2 we propose here is to adapt the weights in λ instead of updating directly p. This does not mean that the weighting distribution p is not updated in the iterations, but it changes only within the set $\mathcal{P}$ and through adaption of λ. For the weights $\lambda_1, \ldots, \lambda_r$ there are no additional restrictions, they move freely in the unit simplex having r vertices denoted by $S(1, r)$. Thus, in the scheme of the modified algorithm presented as Algorithm 2, we replace $p^{(t)}$ with $\sum_{k=1}^r \lambda_k^{(t)} q^{(k)}$. Instead of initializing $p^{(1)}$ with the empirical distribution, we set $\lambda^{(1)} = (r^{-1}, \ldots, r^{-1})$. (Alternatively, the vector $\lambda^{(1)}$ could be randomly selected from $S(1, r)$). When we substitute $p^{(t)}$ in the formula of the overall error measure of the t-th regression estimate, we obtain the following representation:

Algorithm 2 Imprecise AdaBoost.R2

```
Require: Maximum number of iterations $T$, training
    data set $Z$ and extreme points $q^{(1)}, \ldots, q^{(r)}$ of $\mathcal{P}$.
Ensure: $\alpha^{(t)}$ and $\hat{f}^{(t)}$ for all $t \in \{1, \ldots, T\}$;
  set $t \leftarrow 1$ and $\lambda^{(1)} \leftarrow (r^{-1}, \ldots, r^{-1})$;
  repeat
    compute the vector of weighting probabilities:
      $p^{(t)} \leftarrow \sum_{k=1}^r \lambda_k^{(t)} q^{(k)}$;
    estimate $\hat{f}^{(t)}$ using weighting probabilities $p^{(t)}$;
    compute $D^{(t)} \leftarrow \max_{j \in \{1,\ldots,n\}} |y_j - \hat{f}^{(t)}(x_j)|$;
    compute normalized errors for all $i \in \{1, \ldots, n\}$:
      $\hat{e}_i^{(t)} \leftarrow \frac{|y_i - \hat{f}^{(t)}(x_i)|}{D^{(t)}}$;
    compute error portions for all $k \in \{1, \ldots, r\}$:
      $\hat{\varepsilon}_k^{(t)} \leftarrow \sum_{i=1}^n \hat{e}_i^{(t)} q_i^{(k)}$;
    calculate the overall error of $\hat{f}^{(t)}$:
      $\epsilon^{(t)} \leftarrow \sum_{k=1}^r \lambda_k^{(t)} \hat{\varepsilon}_k^{(t)}$;
    if $\epsilon^{(t)} > 0.5$ then
      $T \leftarrow t - 1$;
    end if
    compute contribution of $\hat{f}^{(t)}$ to the final result:
      $\alpha^{(t)} \leftarrow \ln\left(\frac{1-\epsilon^{(t)}}{\epsilon^{(t)}}\right)$;
    adapt weights for all $k \in \{1, \ldots, r\}$:
      $\lambda_k^{(t+1)} \leftarrow \lambda_k^{(t)} \exp\left(-\alpha^{(t)}(1 - \hat{\varepsilon}_k^{(t)})\right)$;
    normalize $\lambda^{(t+1)}$ such that $\sum_{k=1}^r \lambda_k^{(t+1)} = 1$;
    $t{+}{+}$
  until $t > T$
  normalize $\alpha^{(1)}, \ldots, \alpha^{(T)}$ such that $\sum_{t=1}^T \alpha^{(t)} = 1$;
  compute regression function $\hat{f} \leftarrow \sum_{t=1}^T \alpha^{(t)} \hat{f}^{(t)}$.
```

$$\epsilon^{(t)} = \sum_{i=1}^{n} \hat{e}_i^{(t)} p_i^{(t)} = \sum_{i=1}^{n} \hat{e}_i^{(t)} \sum_{k=1}^{r} \lambda_k^{(t)} q_i^{(k)} = \sum_{k=1}^{r} \lambda_k^{(t)} \hat{\varepsilon}_k^{(t)},$$

where $\hat{\varepsilon}_k^{(t)} = \sum_{i=1}^n \hat{e}_i^{(t)} q_i^{(k)}$ can be interpreted as the contribution of the k-th extreme point to the average normalized residual. It corresponds to the mean value of the normalized residuals with respect to the discrete distribution $q^{(k)} \in \mathcal{P}$. Moreover, the above representation unveils a nice characteristic of the proposed modification of the algorithm. In fact, it implies that the n examples are transformed to $r \geq n$ virtual data points (i.e., the extreme points) with associated residuals $\hat{\varepsilon}_k^{(t)}$ and weights $\lambda_k^{(t)}$ for all $k \in \{1,\ldots,r\}$.

From this interpretation, it is straightforward to derive the updating rule to obtain the weights $\lambda_1^{(t+1)},\ldots,\lambda_r^{(t+1)}$. In the same way as the weighting probabilities of the data are adapted in Algorithm 1, we increase the weights of those extreme points with large errors $\hat{\varepsilon}_k^{(t)}$ and vice versa. Hence, we simply adapt the updating rule given in AdaBoost.R2 and update the weights of the extreme points by

$$\lambda_k^{(t+1)} = \lambda_k^{(t)} \exp\left(-\alpha^{(t)}(1-\hat{\varepsilon}_k^{(t)})\right)$$

for all $k \in \{1,\ldots,r\}$. The $\lambda_1^{(t+1)},\ldots,\lambda_r^{(t+1)}$ are also normalized to fulfill the condition $\sum_{k=1}^{r} \lambda_k^{(t+1)} = 1$. Note that the obtained weighting probability distribution $p^{(t+1)}$ again belongs to the set $\mathcal{P}$ because it is a convex linear combination of the corresponding extreme points.

Let us now consider the special case where we do not have additional information, and thus, $\mathcal{P} = S(1,n)$. In this case, there are $r = n$ extreme points corresponding to the vertices of the unit simplex, e.g., for $k = 1$ we have $q^{(1)} = (1,0,\ldots,0)$. Then, $p^{(t)} = (\lambda_1^{(t)},\ldots,\lambda_n^{(t)})$ and the k-th extreme point mean error $\hat{\varepsilon}_k^{(t)} = \hat{e}_k^{(t)}$ for all $t \in \{1,\ldots,T\}$. Hence, we get the following updated weights for all $k \in \{1,\ldots,n\}$:

$$\lambda_k^{(t+1)} = p_k^{(t)} \exp\left(-\alpha^{(t)}(1-\hat{e}_k^{(t)})\right) = p_k^{(t+1)},$$

which coincide with those obtained in AdaBoost.R2. This implies that the proposed algorithm is a generalization of the standard AdaBoost.R2 and covers it as the special case where the set of weighting probabilities is not restricted.

The proposed Imprecise AdaBoost.R2 algorithm has several positive features in comparison with the standard AdaBoost.R2. As the number of extreme points of $\mathcal{P}$ is always larger than or equal to the number of examples, the modified algorithm can have a larger number of parameters to adjust. In this case, the weighting probabilities can be adapted in finer steps within the set $\mathcal{P}$. Furthermore, when we have only a few examples, the overall errors $\epsilon^{(t)}$ of the $\hat{f}^{(t)}$ with $t \in \{1,\ldots,T\}$ can only be determined with much uncertainty due to the high variance of the estimates. As a result, the weights may change very quickly and the algorithm may become unstable. The proposed modification of the AdaBoost.R2 algorithm is less affected by this problem if $\mathcal{P}$ is a proper subset of $S(1,n)$, because in this case the weights cannot be too large and hence neither the differences between the weighting probabilities of an instance in two subsequent iteration steps. Finally, any set of discrete probabilities defined by linear constraints can be used in the algorithm. This allows to introduce any prior information of this kind about the training data. In Section 4, we discuss a selection of imprecise statistical models to derive $\mathcal{P}$, but in principle it can be any convex subset of $S(1,n)$. Moreover, it is possible to further generalize the proposed Imprecise AdaBoost.R2 algorithm and allow the set $\mathcal{P}$ to be changed in every iteration step according to some rule, for instance, by means of Bayesian updating.

3 Threshold AdaBoost algorithm and its modification

In this section, we consider the AdaBoost.RT algorithm introduced in [15]. This algorithm is based on the idea that the training examples can be classified into two classes by comparing the accuracy of the predicted values with a predefined relative error threshold. Then, the evaluation of the regression estimates $\hat{f}^{(t)}$ within the iterated loop of the algorithm can be done on the basis of the average misclassification error like in the basic AdaBoost algorithm for binary classification. Algorithm 3 outlines the scheme of the AdaBoost.RT algorithm.

In contrast to the normalized absolute residuals of AdaBoost.R2, here the regression errors $\hat{e}_1^{(t)},\ldots,\hat{e}_n^{(t)}$ are given by the absolute values of the relative residuals, for each $t \in \{1,\ldots,T\}$. These residuals are compared to a threshold value $\tau \in \mathbb{R}_{\geq 0}$. The corresponding examples are considered as misclassified if their residual exceeds τ and as correctly classified otherwise. Thus, as in AdaBoost for classification, each estimated function $\hat{f}^{(t)}$ is evaluated by its overall misclassification rate $\epsilon^{(t)} = \sum_{\{i:\hat{e}_i^{(t)}>\tau\}} p_i^{(t)}$. Furthermore, the weights are updated according to a rule depending on τ. The weights associated with examples with small relative residuals are decreased, while those of the examples considered as misclassified remain constant. By normalizing $p_1^{(t+1)},\ldots,p_n^{(t+1)}$ to obtain a probability distribution, the weighting probabilities of the misclassified examples are, in fact, increased.

An important feature of the algorithm is that it does not stop when the overall error rate $\epsilon^{(t)}$ is greater than 0.5. In AdaBoost.RT it is not necessary to explicitly state a stopping criterion, because the computation scheme for the weights $\alpha^{(t)}$ of the regression estimates in the combined result implies that poor estimates are almost neglected and vice versa. That is, if $\epsilon^{(t)}$ is high, so is $\beta^{(t)} = \left(\epsilon^{(t)}\right)^l$ for some $l \in \mathbb{N}$, and thus, $\alpha^{(t)} = -\ln\left(\beta^{(t)}\right)$ will be very small com-

Algorithm 3 AdaBoost.RT

```
Require: Maximum number of iterations $T$, training
    data set $Z$, threshold $\tau$ and power coefficient $l$.
Ensure: $\alpha^{(t)}, \beta^{(t)}$ and $\hat{f}^{(t)}$ for all $t \in \{1, \dots, T\}$;
  set $t \leftarrow 1$ and $p^{(t)} \leftarrow (n^{-1}, \dots, n^{-1})$;
  repeat
    estimate $\hat{f}^{(t)}$ using weighting probabilities $p^{(t)}$;
    compute relative errors for all $i \in \{1, \dots, n\}$:
     $\hat{e}_i^{(t)} \leftarrow \left| \frac{y_i - \hat{f}^{(t)}(x_i)}{y_i} \right|$;
    calculate the overall error rate of $\hat{f}^{(t)}$:
     $\epsilon^{(t)} \leftarrow \sum_{\{i:\hat{e}_i^{(t)} > \tau\}} p_i^{(t)}$;
    compute $\beta^{(t)} \leftarrow \left(\epsilon^{(t)}\right)^l$;
    compute contribution of $\hat{f}^{(t)}$ to the final result:
     $\alpha^{(t)} \leftarrow -\ln\left(\beta^{(t)}\right)$;
    adapt weights for all $i \in \{1, \dots, n\}$ by:
    if $\hat{e}_i^{(t)} \leq \tau$ then
      $p_i^{(t+1)} \leftarrow p_i^{(t)} \beta^{(t)}$;
    else
      $p_i^{(t+1)} \leftarrow p_i^{(t)}$;
    end if
    normalize $p^{(t+1)}$ to be a proper distribution;
    $t{+}{+}$
  until $t > T$
  normalize $\alpha^{(1)}, \dots, \alpha^{(T)}$ such that $\sum_{t=1}^{T} \alpha^{(t)} = 1$;
  compute regression function $\hat{f} \leftarrow \sum_{t=1}^{T} \alpha^{(t)} \hat{f}^{(t)}$.
```

pared to better estimates of other iterations. In [15] it is also argued that even if $\epsilon^{(t)} > 0.5$ for some of the estimates in the ensemble, the final output of the ensemble-based algorithm is better than that of a single regression estimate. That is why AdaBoost.RT does not have a stopping rule like the AdaBoost.R2 algorithm, although it would fit the framework of this algorithm very well, as the evaluation is based on a pseudo misclassification error rate.

In spite of the virtues of AdaBoost.RT, it has the shortcoming that the threshold must be selected a priori, because the performance of the algorithm is sensitive to τ. If τ is too low, then it is generally very difficult to get a sufficient number of correctly predicted examples. Furthermore, the standard AdaBoost.RT algorithm has the same tendency to overfitting as the AdaBoost.R2 algorithm due to the too large set of weighting probabilities. In order to overcome this disadvantage, we propose a modified version of AdaBoost.RT where the set of weighting probabilities is restricted to the convex set $\mathcal{P}$ with extreme points $q^{(k)} = (q_1^{(k)}, \dots, q_n^{(k)})$ with $k \in \{1, \dots, r\}$. The scheme of the modified AdaBoost.RT is presented as Algorithm 4. Again, we can interpret the proposed modification as replacing the n training data with r virtual examples (i.e., the extreme points of $\mathcal{P}$) with residuals $\hat{\varepsilon}_k^{(t)}$ and weights λ_k for all $k \in \{1, \dots, r\}$. Then, the overall error rates $\epsilon^{(t)}$ are obtained as $\sum_{\{k:\hat{\varepsilon}_k^{(t)} > \tau\}} \lambda_k^{(t)}$.

Algorithm 4 Imprecise AdaBoost.RT

```
Require: Maximum number of iterations $T$, train-
    ing data set $Z$, threshold $\tau$, power coefficient $l$ and
    extreme points $q^{(1)}, \dots, q^{(r)}$ of $\mathcal{P}$.
Ensure: $\alpha^{(t)}, \beta^{(t)}$ and $\hat{f}^{(t)}$ for all $t \in \{1, \dots, T\}$;
  set $t \leftarrow 1$ and $\lambda^{(1)} \leftarrow (r^{-1}, \dots, r^{-1})$;
  repeat
    compute the vector of weighting probabilities:
     $p^{(t)} \leftarrow \sum_{k=1}^{r} \lambda_k^{(t)} q^{(k)}$;
    estimate $\hat{f}^{(t)}$ using weighting probabilities $p^{(t)}$;
    compute relative errors for all $i \in \{1, \dots, n\}$:
     $\hat{e}_i^{(t)} \leftarrow \left| \frac{y_i - \hat{f}^{(t)}(x_i)}{y_i} \right|$;
    compute error portions for all $k \in \{1, \dots, r\}$:
     $\hat{\varepsilon}_k^{(t)} \leftarrow \sum_{i=1}^{n} \hat{e}_i^{(t)} q_i^{(k)}$;
    calculate the overall error rate of $\hat{f}^{(t)}$:
     $\epsilon^{(t)} \leftarrow \sum_{\{k:\hat{\varepsilon}_k^{(t)} > \tau\}} \lambda_k^{(t)}$;
    compute $\beta^{(t)} \leftarrow \left(\epsilon^{(t)}\right)^l$;
    compute contribution of $\hat{f}^{(t)}$ to the final result:
     $\alpha^{(t)} \leftarrow -\ln\left(\beta^{(t)}\right)$;
    adapt weights for all $k \in \{1, \dots, r\}$ by:
    if $\hat{\varepsilon}_k^{(t)} \leq \tau$ then
      $\lambda_k^{(t+1)} \leftarrow \lambda_k^{(t)} \beta^{(t)}$;
    else
      $\lambda_k^{(t+1)} \leftarrow \lambda_k^{(t)}$;
    end if
    normalize $\lambda^{(t+1)}$ such that $\sum_{k=1}^{r} \lambda_k^{(t+1)} = 1$;
    $t{+}{+}$
  until $t > T$
  normalize $\alpha^{(1)}, \dots, \alpha^{(T)}$ such that $\sum_{t=1}^{T} \alpha^{(t)} = 1$;
  compute regression function $\hat{f} \leftarrow \sum_{t=1}^{T} \alpha^{(t)} \hat{f}^{(t)}$.
```

Let us again consider the special case without additional information about the weighting probabilities, and thus, $\mathcal{P} = S(1, n)$ with $r = n$ vertices. Then, for all $t \in \{1, \dots, T\}$ we obtain $p^{(t)} = (\lambda_1^{(t)}, \dots, \lambda_n^{(t)})$ and $\lambda_k^{(t+1)} = p_k^{(t+1)}$ for all $k \in \{1, \dots, n\}$, while the k-th extreme point mean error is given by $\hat{\varepsilon}_k^{(t)} = \hat{e}_k^{(t)}$ and $\epsilon^{(t)} = \sum_{k:\hat{\varepsilon}_k^{(t)} > \tau} p_k^{(t)}$. Hence, also the standard AdaBoost.RT algorithm is a special case of its proposed modification.

4 Imprecise statistical models

In this section, we briefly review a selection of imprecise statistical models which can be used to determine the set of weighting probabilities $\mathcal{P} \subset S(1, n)$. In particular, we consider two different imprecise neighborhood models around the empirical distribution $\hat{p}$ of the training data and one statistical approach derived from the Kolmogorov–Smirnov test. Every imprecise neighborhood model is characterized by a common parameter ν, which in some cases can be interpreted as the (subjective) probability that the elicited probability distribution p is incorrect. Further interpretations of the parameter ν are, for example, as the size of possible errors in p or the amount of information on which the model is based.

4.1 The linear-vacuous mixture model

The linear-vacuous mixture or imprecise ν-contaminated models produce the set $\mathcal{P}(\nu, p)$ of probabilities $\pi = (\pi_1, \dots, \pi_n)$ such that $\pi_i = (1-\nu)p_i + \nu b_i$ for some $\nu \in [0,1]$ and for all $i \in \{1, \dots, n\}$, where $p = (p_1, \dots, p_n)$ is the elicited probability distribution and $(b_1, \dots, b_n)$ can be any probability distribution in $S(1, n)$. The set $\mathcal{P}(\nu, p)$ is a convex subset of the unit simplex; it coincides with $S(1, n)$ when $\nu = 1$, while $\mathcal{P}(\nu, p) = \{p\}$ if $\nu = 0$. For $p = \hat{p} = (n^{-1}, \dots, n^{-1})$, the set $\mathcal{P}(\nu, \hat{p})$ has $r = n$ extreme points $q_k \in S(1, n)$ with $k \in \{1, \dots, n\}$, which are all of the same form: the k-th element is given by $(1-\nu)n^{-1} + \nu$ and the other $n-1$ elements are equal to $(1-\nu)n^{-1}$. For example, the extreme point q_2 is given by the vector

$$q_2 = \left(\frac{1-\nu}{n}, \frac{1-\nu}{n} + \nu, \dots, \frac{1-\nu}{n} \right).$$

4.2 The pari-mutuel model

Another imprecise neighborhood model is the imprecise pari-mutuel model [22, Subsection 3.3.5], for which the set of probability distributions is defined as

$$\mathcal{P}_P(\nu, p) = \{\pi \in S(1, n) : \pi_i \le (1+\nu)p_i \; \forall \, i \in \{1, \dots, n\}\},$$

where $\nu \in [0, +\infty)$ and $p = (p_1, \dots, p_n)$ is the elicited distribution. The set $\mathcal{P}(\nu, p)$ consists of all probability distributions π such the weighting probabilities of the points do not exceed a constant multiple of the probabilities given by the distribution p. The set $\mathcal{P}(\nu, p)$ can also be obtained from $\mathcal{P}_P(\nu, p)$ by reflecting $\mathcal{P}(\nu, p)$ about the point p. Lower and upper probabilities of the i-th point are given by $\max\{0, (1+\nu)p_i - \nu\}$ and $\min\{(1+\nu)p_i, 1\}$, respectively. The difference between the upper and lower probabilities is ν, as long as p_i is far enough from 0 or 1.

When we consider the empirical distribution $\hat{p} = (n^{-1}, \dots, n^{-1})$ as the elicited distribution, the extreme points of the set $\mathcal{P}_P(\nu, \hat{p})$ depend on the chosen parameter ν as expressed in the following proposition.

Proposition 1 *Let $z_1, \dots, z_n$ with $n \in \mathbb{N}$ be a set of univariate data and let $\mathcal{P}_P(\nu, \hat{p})$ be the set of probabilities according to a pari-mutuel neighborhood model around the empirical distribution $\hat{p} = (n^{-1}, \dots, n^{-1})$ of the data for some $\nu \in [0, +\infty)$.*

(1) If $\nu \le (n-1)^{-1}$, then the set $\mathcal{P}_P(\nu, \hat{p})$ has $r = n$ extreme points $q_k \in S(1, n)$ with $k \in \{1, \dots, n\}$, which are of the following form: the k-th element is given by $(1+\nu)n^{-1} - \nu$ and the other $n-1$ elements are equal to $(1+\nu)n^{-1}$.

(2) If $(n-1)^{-1} < \nu < (n-1)$, then the set $\mathcal{P}_P(\nu, \hat{p})$ has $r = s\binom{n}{s}$ extreme points, where $s \in \mathbb{N}$ and it is defined by the inequality

$$\frac{1}{n-s+1} \le \frac{1+\nu}{n} \le \frac{1}{n-s}.$$

The extreme points have the same form: $n-s$ elements have value $(1+\nu)n^{-1}$, there is one element given by $1-(n-s)(1+\nu)n^{-1}$, and the other $s-1$ elements are equal to zero.

(3) If $\nu \ge (n-1)$, then $\mathcal{P}_P(\nu, \hat{p}) = S(1, n)$.

The third part of this proposition is obvious, because in this case the upper probabilities of all points are equal to one. The proofs of parts (1) and (2) can be found in [19, Propositions 1 and 3].

4.3 Kolmogorov–Smirnov bounds

A statistical approach to constructing bounds for the set of weighting probabilities can be derived from confidence bounds for the probability distribution of the data. Such confidence bounds can be obtained by inverting the so-called Kolmogorov–Smirnov test.

Let F denote the cumulative distribution function associated with the unknown probability measure P of some univariate data $z_1, \dots, z_n$ with $n \in \mathbb{N}$ and $\hat{F}_n$ their empirical cumulative distribution function. The Kolmogorov–Smirnov test is a nonparametric test for the null hypothesis that $z_1, \dots, z_n$ have been generated by some particular distribution F_0. As test statistic the supremum vertical distance of $\hat{F}_n$ and F_0 is considered. It can be shown that the distribution of this test statistic under the null hypothesis is independent of F_0. The quantiles $k_{n,1-\gamma}$ of the test distribution are available in tables for a certain range of sample sizes and some different test levels $\gamma \in (0,1)$. For large n, the quantiles can be approximated by a simple formula. The null hypothesis is rejected at level $\gamma \ge P_{F_0}(||\hat{F}_n - F_0||_\infty > k_{n,1-\gamma})$ if the observed supremum distance given by

$$\max_{1,\dots,n} \max \left\{ \frac{i}{n} - F_0(z_{(i)}), \, F_0(z_{(i)}) - \frac{i-1}{n} \right\},$$

where $z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(n)}$, is larger than $k_{n,1-\gamma}$. By considering all distribution functions such that the test does not reject the null hypothesis at the chosen γ, we obtain a $1-\gamma$ confidence band for F which is of the form

$$\mathcal{C}_{1-\gamma} = \{F' : \underline{F}_n(z) \leq F'(z) \leq \overline{F}_n(z) \ \forall z\},$$

with

$$\underline{F}_n(z) = \max\{\hat{F}_n(z) - k_{n,1-\gamma}, 0\} \text{ and}$$
$$\overline{F}_n(z) = \min\{\hat{F}_n(z) + k_{n,1-\gamma}, 1\}.$$

As the quantiles of the exact test distribution are not available for all sample sizes and all test levels, several modifications of the above confidence band have been suggested, replacing $k_{n,1-\gamma}$ by an upper approximation $d_{n,1-\gamma}$, e.g., the upper bounds provided by the so-called Dvoretzky–Kiefer–Wolfowitz inequality, resulting in more conservative but easy-to-compute confidence bands for F. Therefore, we use $d_{n,1-\gamma}$ as the general notation in the following, but refer to the limiting functions $\underline{F}_n(z)$ and $\overline{F}_n(z)$ of all confidence bands of the above type (with $d_{n,1-\gamma}$ instead of $k_{n,1-\gamma}$) as Kolmogorov–Smirnov bounds. See [23, Section 2] and [10, Subsection 8.9.3] for more details.

It has been shown in [20] that it is possible to derive a set of probability mass functions $\mathcal{P}_K(\gamma)$ corresponding to the confidence band for the cumulative distribution function of the type $\mathcal{C}_{1-\gamma}$. The set $\mathcal{P}_K(\gamma)$ is a convex subset of $S(1,n)$ with $s\binom{n}{s}$ extreme points, where $s \in \mathbb{N}$ is determined from the condition

$$nd_{n,1-\gamma} < s \leq 1 + nd_{n,1-\gamma}.$$

Every extreme point has $s-1$ elements of size 0, one element with the value $(s - nd_{n,1-\gamma})(n(1-d_{n,1-\gamma}))^{-1}$, and $n-s$ elements of size $(n(1-d_{n,1-\gamma}))^{-1}$.

5 Numerical experiments

To study how well the proposed algorithms may solve practical problems, we conduct several numerical experiments. Thereby, we use weighted Support Vector Regression (SVR, see, e.g., [16, 18]) with absolute loss function and Gaussian kernel as regression estimator within the algorithms and we determine the set of weighting probabilities by means of the linear-vacuous mixture model. We make different simulations to study the impact of the choice of ν on the performance of the proposed regression methods, before we apply them to analyze two data sets from the UCI Machine Learning Repository [4].

From each of the (synthetic or real) data sets we randomly select two distinct subsets: a training data set of n examples to learn the model, and a test data set of n_{test} instances to evaluate the performance of the algorithms. For the synthetic data, the performance is assessed by two measures: the square roots of the Mean Square Prediction Error (RMSPE) and of the average Residual Sum of Squares (RRSS), which are defined by

$$RMSPE = \sqrt{\frac{\sum_{i=1}^{n_{test}} (f(x_i) - \hat{f}(x_i))^2}{n_{test}}} \quad \text{and}$$
$$RRSS = \sqrt{\frac{\sum_{i=1}^{n_{test}} (y_i - \hat{f}(x_i))^2}{n_{test}}},$$

respectively, where f denotes the true function, $\hat{f}$ is the function estimated by one of the proposed algorithms, and $\hat{f}(x_i)$ is the predicted value of y_i for each $i \in \{1, \ldots, n_{test}\}$. Both measures are computed on the basis of the test data set in each simulation run. As usual, the RMSPE and RRSS values are finally averaged over the simulation runs. The smaller the values of these average error measures are, the better the corresponding regression method. Regarding the numerical examples analyzing real data the RMSPE cannot be computed, because the true function f is unknown. Hence, in this case, we only compare the overall RRSS obtained from repeatedly drawing training and test data sets. Furthermore, since the main purpose of the numerical examples is to show the application of the methods to simple and illustrative problems, the hyperparameters are not optimized.

5.1 Synthetic data

The aim of analyzing synthetic data is to investigate how the parameter ν of the linear-vacuous mixture model introduced in the previous section influences the performance of the regression methods based on the modified boosting algorithms. Therefore, we conduct the simulations for five different choices of ν, namely $\nu \in \{0, 0.25, 0.5, 0.75, 1\}$. Recall that, when $\nu = 1$ then $\mathcal{P} = S(1,n)$, and thus, we have the standard basic boosting algorithm on the one extreme of the ν range, whereas $\mathcal{P} = \{\hat{p}\}$ for $\nu = 0$, which reduces the modified boosting algorithms to the standard SVR with identical weights of examples.

In our simulations, we consider two different kinds of data sets. The first is generated according to the following setup. In each of 40 runs, we generate 200 examples $(x_j, y_j) \in \mathbb{R}^2$ for all $j \in \{1, \ldots, 200\}$ by

$$x_j = 0.02(j-1) - 2 \quad \text{and}$$
$$y_j = \exp(-x_j^2) + 0.5\eta_j,$$

where η_j is a random number drawn from the uniform distribution on the interval $[-1, 1]$. Similar data

Table 1: Performance of the modification of AdaBoost.R2 by different ν

ν	$RRSS$	$RMSPE$
0.0	0.456	0.344
0.25	0.428	0.311
0.5	0.427	0.31
0.75	0.435	0.32
1	0.443	0.329

Table 2: Performance of the modification of AdaBoost.R2 by different ν adding the asymmetric noise

ν	$RRSS$	$RMSPE$
0.0	0.835	0.437
0.25	0.704	0.373
0.5	0.717	0.378
0.75	0.726	0.374
1	0.749	0.39

sets have been used, e.g., in [8]. From these 200 data points, we randomly draw a training data set and a distinct test data set. The number of training examples is $n = 10$ and the number of testing examples $n_{test} = 60$. We then apply one of the proposed regression methods to the training data and obtain an estimate $\hat{f}$ of the function f. Here, we set the number of iterations in the boosting algorithms to $T = 20$ and consider $\nu \in \{0, 0.25, 0.5, 0.75, 1\}$. Finally, we compute the performance measures.

As a variant of this synthetic data set, we consider also an asymmetric noise. In contrast to the above case, we here generate the random errors according to the following rule: for all $j \in \{1, \dots, 200\}$, we draw a random number $a_j \in [0, 1]$, then η_j is drawn from $[-1, 1]$ if $a_j < 0.7$ and from the interval $[0, 4]$ if $a_j \geq 0.7$. All random numbers are generated according to uniform distributions on the corresponding intervals.

Table 1 shows the performance measures $RRSS$ and $RMSPE$ for the modified AdaBoost.R2 algorithm by different values of the parameter ν. We observe that the proposed regression method achieves the best results when the linear-vacuous model with $\nu = 0.5$ is used to restrict the set of weighting probabilities. Table 2 shows the results for the data set with asymmetric errors. Here, the best results are achieved for $\nu = 0.25$. The additional asymmetric noise produces some kind of outliers which the standard AdaBoost.R2 tends to fit too well, because these points are assigned high weights. As the proposed modification of the algorithm restricts these weights, the problem of over-fitting is avoided.

Table 3: Performance of the modification of AdaBoost.RT by different ν

ν	$RRSS$	$RMSPE$
0.0	0.465	0.364
0.25	0.409	0.295
0.5	0.408	0.288
0.75	0.411	0.293
1	0.424	0.311

Table 4: Performance of the modification of AdaBoost.RT by different ν adding the asymmetric noise

ν	$RRSS$	$RMSPE$
0.0	0.837	0.458
0.25	0.71	0.345
0.5	0.723	0.336
0.75	0.723	0.356
1	0.727	0.377

Let us now analyze the simulation results for the modification of AdaBoost.RT algorithm. By using the same initial data as for the modification of AdaBoost.R2, we get the performance measures for the same scenarios. The case of the symmetric errors is shown in Table 3 and the situation with the additional asymmetric noise is presented in Table 4. For the analyses, we set the threshold to $\tau = 0.5$ and $l = 2$ (see Algorithm 4). Also for the modified AdaBoost.RT, the values of ν implying the best performance of the algorithm in the first and second error scenarios are $\nu = 0.5$ and $\nu = 0.25$, respectively. Furthermore, we observe that the modification of AdaBoost.RT slightly outperforms the modification of AdaBoost.R2 in both data settings.

Hence, in the analyses of the first synthetic data set, the proposed modifications of the AdaBoost-based algorithms perform better than the original ones, corresponding to $\nu = 1$, and better than the standard SVR, corresponding to $\nu = 0$. In addition to the above analyses, we will consider another synthetic data set, which is the well-known data set Friedman#1 [6]. In each of 40 simulation runs we generate 200 examples of 10 independent variables, which are uniformly distributed in the interval $[0, 1]$. Only 5 of these 10 variables are selected randomly and then used to produce the values of the output variable for all $j \in \{1, \dots, 200\}$ in the following way:

$$y_j = 10\sin(\pi x_{j,1}x_{j,2}) + 20(x_{j,3} - 0.5)^2 + 10x_{j,4} + 5x_{j,5} + \eta_j,$$

where η_j is a random variable drawn from a standard normal distribution. Here, we used 20 training examples and 40 test examples.

The RRSS and RMSPE measures obtained by using the modification of AdaBoost.R2 are shown in

Table 5: Performance of the modification of AdaBoost.R2 for Friedman#1 data by different ν

ν	$RRSS$	$RMSPE$
0.0	3.12	2.8
0.25	3.08	2.74
0.5	3.06	2.72
0.75	3.07	2.723
1	3.09	2.75

Table 6: Performance of the modification of AdaBoost.RT for Friedman#1 data by different ν

ν	$RRSS$	$RMSPE$
0.0	3.12	2.8
0.25	3.08	2.75
0.5	2.96	2.63
0.75	2.98	2.63
1	3.07	2.71

Table 5 and the results for the modification of AdaBoost.RT with $\tau = 0.1$ and $l = 2$ are given in Table 6. Here again, we find that the modification of the AdaBoost.RT algorithm attains slightly lower average errors than the regression method based on the modification of AdaBoost.R2. The value $\nu = 0.5$ implies the best performance of both generalized algorithms, but here we observe that they outperform standard boosting only with higher values of ν.

5.2 Real data

In addition to the simulations, we evaluate the performance of the proposed regression methods by analyzing two publicly available data sets from the UCI Machine Learning Repository [4]: Slump Test [25], and Concrete [24]. The Slump Test data set contains 103 data points. There are seven input variables characterizing the slump flow of concrete and three output variables in the data set. Here, we use only the third output variable: 28-day compressive strength. In the Concrete Data Set there are 1 030 data points characterizing the concrete compressive strength as a highly nonlinear function of age and ingredients which include cement, blast furnace slag, fly ash, water, etc. There are eight input variables and one output variable, namely the concrete compressive strength.

We analyze both data sets with the proposed algorithms, again for different choices of $\nu \in \{0, 0.25, 0.5, 0.75, 1\}$ and with $T = 20$. To evaluate the average residual error measure RRSS we perform a cross-validation with 40 repetitions, where in each run, we randomly select $n = 40$ training data and $n_{test} = 40$ test data. The results of the computations are given in Table 7 for the modified AdaBoost.R2 method and in Table 8 those for the generalized AdaBoost.RT algorithm with $l = 2$. The obtained figures confirm the results of the simulations and indicate a superior fit of the proposed regression methods for $\nu \in \{0.25, 0.5, 0.75\}$. Thus, if the mixture probability ν is neither too small nor too big both modified algorithms perform better in terms of RRSS than their basic counterparts, which correspond to $\nu = 1$.

Table 7: RRSS for the UCI data sets by using AdaBoost.R2 with different ν

ν	0	0.25	0.5	0.75	1
Slump Test	9.08	8.79	8.75	8.85	9.08
Concrete	27	16.7	16.6	16.8	17

Table 8: RRSS for the UCI data sets by using AdaBoost.RT with different ν

ν	τ	0	0.25	0.5	0.75	1
Slump Test	0.08	9.14	8.74	8.51	8.48	8.75
Concrete	0.3	32.1	16.8	16.9	17.1	17.1

The results of the numerical examples indicate that the value of ν does indeed affect the performance of the proposed algorithms. Hence, in a practical setting, the choice of this parameter should be made very carefully, e.g., on the basis of a cross-validation scheme.

6 Conclusion

We proposed the generalizations of two ensemble-based boosting algorithms for regression where the unit simplex for the weights of the instances is restricted to a smaller set of weighting probabilities. This smaller set is obtained by imprecise statistical models like, e.g. the linear-vacuous mixture model. The modified algorithms are more flexible and tend less to over-fitting. Numerical experiments further indicate that among the extreme cases (recall that for $\nu = 0$ it corresponds to standard SVR and for $\nu = 1$ to the basic boosting algorithm), the standard AdaBoost algorithms are always at least as good as the standard SVR and often much better. Moreover, we found that both modified algorithms perform better than their standard counterparts, if the mixture probability ν is neither too small nor too big. The core idea behind the presented modifications could be adapted to deal with imprecise data, too, as the imprecise data induce a set of compatible probability distributions.

Acknowledgments

This work was conducted while the second author was visiting the Saint Petersburg State Forest Technical University on a fellowship from the German Academic Exchange Service (DAAD).

References

[1] P. Bühlmann and T. Hothorn. Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, 22(4):477–505, 2007.

[2] H. Drucker. Improving regressors using boosting techniques. In *Proc. of the 14th Intenational Conferences on Machine Learning*, pages 107–115, San Francisco, CA, USA, 1997. Morgan Kaufmann.

[3] A.J. Ferreira and M.A.T. Figueiredo. Boosting algorithms: A review of methods, theory, and applications. In C. Zhang and Y. Ma, editors, *Ensemble Machine Learning: Methods and Applications*, pages 35–85. Springer, New York, 2012.

[4] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[5] Y. Freund and R.E. Schapire. A decision theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[6] J.H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–82, 1991.

[7] J.H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[8] P.-Y. Hao. Interval regression analysis using support vector networks. *Fuzzy Sets and Systems*, 60:2466–2485, 2009.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, New York, 2001.

[10] N.L. Johnson and F. Leone. *Statistics and experimental design in engineering and the physical sciences*, volume 1. Wiley, New York, 1964.

[11] B. Kegl. Robust regression by boosting the median. In *Learning Theory and Kernel Machines. Lecture Notes in Computer Science*, volume 2777, pages 258–272. Springer, Berlin Heidelberg, 2003.

[12] L.I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms.* Wiley-Interscience, New Jersey, 2004.

[13] J.M. Moreira, C. Soares, A.M. Jorge, and J.F. de Sousa. Ensemble approaches for regression: A survey. *ACM Computing Surveys*, 45(1):1–40, 2012.

[14] L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

[15] D.L. Shrestha and D.P. Solomatine. Experiments with AdaBoost.RT, an improved boosting scheme for regression. *Neural Computation*, 18(7):1678–1710, 2006.

[16] A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004.

[17] D.P. Solomatine and D.L. Shrestha. AdaBoost.RT: a boosting algorithm for regression problems. In *Proc. of the International Joint Conference on Neural Networks*, pages 1163–1168, Budapest, Hungary, 2004.

[18] I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, 2008.

[19] L.V. Utkin. A framework for imprecise robust one-class classification models. *International Journal of Machine Learning and Cybernetics*, 2012. To appear.

[20] L.V. Utkin and F.P.A. Coolen. On reliability growth models using Kolmogorov-Smirnov bounds. *International Journal of Performability Engineering*, 7(1):5–19, 2011.

[21] V. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[22] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.

[23] L. Wasserman. *All of Nonparametric Statistics.* Springer, New York, 2006.

[24] I-Cheng Yeh. Modeling of strength of high performance concrete using artificial neural networks. *Cement and Concrete Composites*, 28(12):1797–1808, 1998.

[25] I-Cheng Yeh. Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and Concrete Composites*, 29(6):474–480, 2007.

Operator of Composition for Credal Sets

Jiřina Vejnarová
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
vejnar@utia.cas.cz

Abstract

This paper is the first attempt to introduce the operator of composition, already known from probability, possibility and evidence theories, also for credal sets. We prove that the proposed definition preserves all the necessary properties of the operator enabling us to define compositional models as an efficient tool for multidimensional models representation. Theoretical results are accompanied by numerous illustrative examples.

Keywords. Credal sets, graphical models, conditional independence.

1 Introduction

In the second half of 1990's a new approach to efficient representation of multidimensional probability distributions was introduced with the aim to be alternative to Graphical Markov Modeling. This approach is based on the following idea: multidimensional distribution is *composed* from a system of low-dimensional distributions by repetitive application of a special operator of composition, which is also the reason why the models are called *compositional models.* In several papers, in which the properties of the operator and models were studied [4, 5, 6], it was shown (among others) that these models are, in a way, equivalent to Bayesian networks. Roughly speaking, *any multidimensional distribution representable by a Bayesian network can also be represented in the form of a compositional model, and vice versa.*

Later, this compositional models were introduced also in possibility theory [12, 13] (here the models are parameterized by a continuous t-norm) and a few years ago also in evidence theory [8, 9]. In all these frameworks the original idea is kept, but there exist some slight differences among them, as we shall see later.

Although Bayesian networks and compositional models represent the same class of distributions, they do not make it in the same way. Bayesian networks use *conditional distributions* whereas compositional models consist of *unconditional distributions.* Naturally, both types of models contain the same information but while some marginal distributions are explicitly expressed in compositional models, it may happen that their computation from a corresponding Bayesian network is rather computationally expensive. Therefore it appears that some of computational procedures designed for compositional models are (algorithmically) simpler than their Bayesian network counterparts.

Furthermore, the research concerning relationship between compositional models in evidence theory and evidential networks [14] revealed probably a more important thing. Even though any evidential network (with proper conditioning rule and conditional independence concept) can be expressed as a compositional model, if we do it in the opposite way and transform a compositional model into an evidential network, we realize, that the model is more imprecise than the original one. It is caused by the fact that conditioning increases imprecision, and as it is typical not only for evidence theory, but also for other imprecise probability frameworks, compositional models in more general frameworks than evidence theory (e.g. for credal sets) seem to be worth-studying.

The goal of this paper is to show that the operator of composition can also be introduced for credal sets. Moreover, we will show that it keeps the basic properties of its counterparts in other frameworks, and therefore it will enable us to introduce compositional models for multidimensional credal sets.

The contribution is organized as follows. In Section 2 we summarize basic concepts and notation. Definition of the operator of composition is introduced in Section 3, where also its basic properties can be found, while Section 4 is devoted to more advanced properties. Finally, in Section 5 we introduce the concept of so-called *perfect sequences* and demonstrate their importance.

2 Basic Concepts and Notation

In this section we will recall basic concepts and notation necessary for understanding the contribution.

2.1 Variables and Distributions

For an index set $N = \{1, 2, \ldots, n\}$ let $\{X_i\}_{i\in N}$ be a system of variables, each X_i having its values in a finite set $\mathbf{X}_i$ and $\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n$ be the Cartesian product of these sets.

In this paper we will deal with groups of variables on its subspaces. Let us note that X_K will denote a group of variables $\{X_i\}_{i\in K}$ with values in

$$\mathbf{X}_K = \times_{i\in K} \mathbf{X}_i$$

throughout the paper.

Having two probability distributions P_1 and P_2 of X_K we say that P_1 is *absolutely continuous* with respect to P_2 (and denote $P_1 \ll P_2$) if for any $x_K \in \mathbf{X}_K$

$$P_2(x_K) = 0 \Longrightarrow P_1(x_K) = 0.$$

This concept plays an important role in the definition of the operator of composition.

2.2 Credal Sets

A *credal set* $\mathcal{M}(X_K)$ about a group of variables X_K is defined as a closed convex set of probability measures about the values of this variable.

In order to simplify the expression of operations with credal sets, it is often considered [10] that a credal set is the set of probability distributions associated to the probability measures in it. Under such consideration a credal set can be expressed as a *convex hull* of its extreme distributions

$$\mathcal{M}(X_K) = \mathrm{CH}\{\mathrm{ext}(\mathcal{M}(X_K))\}.$$

Consider a credal set about X_K, i.e. $\mathcal{M}(X_K)$. For each $L \subset K$ its *marginal credal set* $\mathcal{M}(X_L)$ is obtained by element-wise marginalization, i.e.

$$\mathcal{M}(X_L) = \mathrm{CH}\{P^{\downarrow L} : P \in \mathrm{ext}(\mathcal{M}(X_K))\}, \qquad (1)$$

where $P^{\downarrow L}$ denotes the marginal distribution of P on $\mathbf{X}_L$. If the above introduced notation (1) cannot be used (e.g. to avoid misunderstandings), then we use $\mathcal{M}(X_K)^{\downarrow L}$, or simply $\mathcal{M}^{\downarrow L}$, instead.

Having two credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ about X_K and X_L, respectively (assuming that $K, L \subseteq N$), we say that these credal sets are *projective* if their marginals about common variables coincide, i.e. if

$$\mathcal{M}_1(X_{K\cap L}) = \mathcal{M}_2(X_{K\cap L}).$$

Let us note that if K and L are disjoint, then $\mathcal{M}_1$ and $\mathcal{M}_2$ are projective, as $\mathcal{M}(X_\emptyset) = 1$.

Besides marginalization we will also need the opposite operation, called vacuous extension. *Vacuous extension* of a credal set $\mathcal{M}(X_L)$ about X_L to a credal set

$$\mathcal{M}(X_K) = \mathcal{M}(X_L)^{\uparrow K}$$

($L \subset K$) is the maximal credal set about X_K such that $\mathcal{M}(X_K)^{\downarrow L} = \mathcal{M}(X_L)$.

Example 1 Let

$$\mathcal{M}(X_1) = \mathrm{CH}(\{[0.2, 0.8], [0.4, 0.6]\})$$

be a credal set about variable X_1. Its vacuous extension $\mathcal{M}(X_1X_2)$ is then

$$\begin{aligned}\mathcal{M}(X_1X_2) \quad = \quad \mathrm{CH}(\{&[0.2, 0, 0.8, 0], [0, 0.2, 0.8, 0],\\ &[0.2, 0, 0, 0.8], [0, 0.2, 0, 0.8],\\ &[0.4, 0, 0.6, 0], [0, 0.4, 0.6, 0],\\ &[0.4, 0, 0, 0.6], [0, 0.4, 0, 0.6]\}),\end{aligned}$$

since evidently

$$\mathcal{M}(X_1X_2)^{\downarrow\{1\}} = \mathrm{CH}(\{[0.2, 0.8], [0.4, 0.6]\}),$$

as desired.

To show, that it is also maximal let us suppose, that there exists a credal set $\mathcal{M}'(X_1X_2)$ containing $\mathcal{M}(X_1X_2)$ and $\mathcal{M}(X_1) = \mathcal{M}'(X_1)$. Then $\mathcal{M}'(X_1X_2)$ must contain at least one $p = (p_1, p_2, p_3, p_4) \notin \mathcal{M}(X_1X_2)$. Nevertheless, it means, that either $p_1 + p_2 < 0.2$ or $p_1 + p_2 > 0.4$ (from which analogous inequalities for $p_3 + p_4$ follow). Therefore, $p^{\downarrow\{1\}} \notin \mathcal{M}(X_1)$ and $\mathcal{M}(X_1X_2)$ is maximal. ♢

The concept of absolute continuity of probability distributions can be generalized for credal sets in the following way. $\mathcal{M}_1(X_K)$ is absolutely continuous with respect to $\mathcal{M}_2(X_K)$, if $P_1 \ll P_2$ for any $P_1 \in \mathcal{M}_1(X_K)$ and $P_2 \in \mathcal{M}_2(X_K)$.

Evidently, it is not the only way how to generalize the concept of absolute continuity to credal sets. It can be done e.g. using lower previsions (but the definitions are not equivalent), nevertheless, the above-presented definition is more suitable for our purpose, as we shall see in the next section.

2.3 Strong Independence

Among numerous definitions of independence for credal sets [2] we have chosen strong independence, as it seems to be most appropriate for multidimensional models.

We say that (groups of) variables X_K and X_L (K and L disjoint) are *strongly independent* with respect to $\mathcal{M}(X_{K\cup L})$ iff (in terms of probability distributions)

$$\mathcal{M}(X_{K\cup L}) = \{P_1 \cdot P_2 : P_1 \in \mathcal{M}(X_K), P_2 \in \mathcal{M}(X_L)\}. \qquad (2)$$

Again, there exist several generalizations of this notion to conditional independence, see e.g. [10], but since the following definition is suggested by the authors as the most appropriate for the marginal problem, it seems to be a suitable concept also in our case, since the operator of composition can also be used as a tool for solution of a marginal problem, as shown (in the framework of possibility theory) e.g. in [13].

Given three groups of variables X_K, X_L and X_M (K, L, M be mutually disjoint subsets of N, such that K and L are nonempty), we say that X_K and X_L are *independent on the distribution* [10] given X_M under global set $\mathcal{M}(X_{K\cup L\cup M})$ (in symbols $K \perp\!\!\!\perp L|M[\mathcal{M}]$[1] iff

$$\mathcal{M}(X_{K\cup L\cup M}) = \{(P_1 \cdot P_2)/P_1^{\downarrow M} : P_1 \in \mathcal{M}(X_{K\cup M}), P_2 \in \mathcal{M}(X_{L\cup M}), P_1^{\downarrow M} = P_2^{\downarrow M}\}.$$

This definition is a generalization of stochastic conditional independence: if $\mathcal{M}(X_{K\cup L\cup M})$ is a singleton, then also $\mathcal{M}(X_{K\cup M})$ and $\mathcal{M}(X_{L\cup M})$ are (projective) singletons and the definition collapses into definition of stochastic conditional independence.

3 Operator of Composition and Its Properties

Now, let us start considering how to define composition of two credal sets. Consider two index sets $K, L \subset N$. At this moment we do not put any restrictions on K and L; they may be but need not be disjoint, one may be subset of the other. We even admit that one or both of them are empty.

In order to enable the reader the understanding of this concept, let us first present the definition of composition for precise probabilities [4]. Let P and Q be two probability distributions of (groups of) variables X_K and X_L. Then

$$(P \triangleright Q)(X_{K\cup L}) = \frac{P(X_K) \cdot Q(X_L)}{Q(X_{K\cap L})},$$

whenever $P(X_{K\cap L}) \ll Q((X_{K\cap L})$. Otherwise, it remains undefined.

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be credal sets about X_K and X_L, respectively. Our goal is to define a new credal set, denoted by $\mathcal{M}_1 \triangleright \mathcal{M}_2$, which will be about $X_{K\cup L}$ and will contain all of the information contained in $\mathcal{M}_1$ and as much as possible of information of $\mathcal{M}_2$ (for the exact meaning see properties *(ii)* and *(iii)* of Lemma 1). The required properties are met by the following definition.

Definition 1 For two arbitrary credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ about X_K and X_L, a *composition* $\mathcal{M}_1 \triangleright \mathcal{M}_2$ is defined by one of the following expressions:

[a] if $\mathcal{M}_1(X_{K\cap L}) = \mathcal{M}_2(X_{K\cap L})$, then

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L}) = \{(P_1 \cdot P_2)/P_2^{\downarrow K\cap L} : P_1 \in \mathcal{M}_1(X_K), P_2 \in \mathcal{M}_2(X_L), P_1^{\downarrow K\cap L} = P_2^{\downarrow K\cap L}\},$$

[b] if $\mathcal{M}_1(X_{K\cap L}) \neq \mathcal{M}_2(X_{K\cap L})$, and $\mathcal{M}_1(X_{K\cap L}) \ll \mathcal{M}_2(X_{K\cap L})$, then

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L}) = \{(P_1 \cdot P_2)/P_2^{\downarrow K\cap L} : P_1 \in \mathcal{M}_1(X_K), P_2 \in \mathcal{M}(X_L)\}),$$

[c] otherwise

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L}) = \mathcal{M}_1(X_K)^{\uparrow K\cup L}.$$

From point [b] of the definition one can see the importance of the definition of absolute continuity in the way presented in et the end of Section 2.2. Exactly this definition enables us to define the composition of two credal sets correctly.

The following lemma presents basic properties possessed by this operator of composition.

Lemma 1 *For arbitrary two credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ about X_K and X_L, respectively, the following properties hold true:*

(i) $\mathcal{M}_1 \triangleright \mathcal{M}_2$ *is a credal set about* $X_{K\cup L}$.

(ii) $(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_K) = \mathcal{M}_1(X_K)$.

(iii) $\mathcal{M}_1 \triangleright \mathcal{M}_2 = \mathcal{M}_2 \triangleright \mathcal{M}_1 \iff \mathcal{M}_1(X_{K\cap L}) = \mathcal{M}_2(X_{K\cap L})$.

[1] If there is no doubt, we will omit $[\mathcal{M}]$.

Proof.

(i) To prove that $\mathcal{M}_1 \triangleright \mathcal{M}_2$ is a credal set about $X_{K\cup L}$ we have to distinguish cases [a] and [b] from [c]. In cases [a] and [b] it is enough to show that any $P \in \mathcal{M}_1 \triangleright \mathcal{M}_2$ is a probability distribution on $\mathbf{X}_{K\cup L}$. But it is obvious, as any $P \in (\mathcal{M}_1 \triangleright \mathcal{M}_2)$ is obtained via formula for composition of probability distributions (cf. e.g. [4]). In case [c] it is obvious too, as $\mathcal{M}_1 \triangleright \mathcal{M}_2$ is a vacuous extension of an credal set about X_K to a credal set about $X_{K\cup L}$.

(ii) Again, we have to make the proof separately. If $(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L})$ is obtained via [c], then the equality follows directly from the definition of vacuous extension. In cases [a] and [b] marginalization of a credal set is element-wise (as mentioned in the preceding section), therefore, analogous to the proof of *(i)* it is enough to prove that $\left((P_1 \cdot P_2)/P_2^{\downarrow K\cap L}\right)^{\downarrow K} = P_1$ for any $P_1 \in \mathcal{M}_1(X_K)$ and $P_2 \in \mathcal{M}_2(X_L)$. But it immediately follows from the results obtained for precise probabilities (see e.g. [4]).

(iii) First, let us assume that

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L}) = (\mathcal{M}_2 \triangleright \mathcal{M}_1)(X_{K\cup L}).$$

Then also its marginals must be identical, particularly

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cap L}) = (\mathcal{M}_2 \triangleright \mathcal{M}_1)(X_{K\cap L}).$$

Taking into account *(ii)* of this lemma we have

$$\begin{aligned}(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cap L}) &= \left(((\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L}))^{\downarrow K}\right)^{\downarrow K\cap L}\\ &= ((\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_K))^{\downarrow K\cap L}\\ &= (\mathcal{M}_1(X_K))^{\downarrow K\cap L} = \mathcal{M}_1(X_{K\cap L})\end{aligned}$$

and similarly

$$(\mathcal{M}_2 \triangleright \mathcal{M}_1)(X_{K\cap L}) = \mathcal{M}_2(X_{K\cap L}),$$

from which the desired equality immediately follows.

Let, on the other hand, $\mathcal{M}_1(X_{K\cap L}) = \mathcal{M}_2(X_{K\cap L})$. In this case [a] of Definition 1 is applied and for any distribution P of $(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L})$ there exist $P_1 \in \mathcal{M}_1(X_K)$ and $P_2 \in \mathcal{M}_2(X_L)$ such that $P_1^{\downarrow K\cap L} = P_2^{\downarrow K\cap L}$ and $P = (P_1 \cdot P_2)/P_2^{\downarrow K\cap L}$. But simultaneously (due to projectivity) $P = (P_1 \cdot P_2)/P_1^{\downarrow K\cap L}$, which is an element of $(\mathcal{M}_2 \triangleright \mathcal{M}_1)(X_{K\cup L})$. Hence

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L}) = (\mathcal{M}_2 \triangleright \mathcal{M}_1)(X_{K\cup L}),$$

as desired. □

Let us now illustrate the application of the operator of composition and its properties by three examples. The first shows what happens when $K \cap L = \emptyset$.

Example 2 Let

$$\mathcal{M}_1(X_1) = \mathrm{CH}\{[0.2, 0.8], [0.7, 0.3]\}$$

and

$$\mathcal{M}_2(X_2) = \mathrm{CH}\{[0.6, 0.4], [1, 0]\}$$

be two credal sets about X_1 and X_2, respectively. Then, as mentioned above, $\mathcal{M}_1(X_1)$ and $\mathcal{M}_2(X_2)$ are projective, and therefore $\mathcal{M}_1 \triangleright \mathcal{M}_2$ is obtained via [a] in Definition 1:

$$\begin{aligned}&(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1X_2) \\ &\quad = \{[0.7 - 0.5\alpha - 0.28\beta + 0.2\alpha\beta,\\ &\qquad 0.28\beta - 0.2\alpha\beta,\\ &\qquad 0.3 + 0.5\alpha - 0.12\beta - 0.2\alpha\beta,\\ &\qquad 0.12\beta + 0.2\alpha\beta], \alpha, \beta \in [0,1]\},\end{aligned} \tag{3}$$

which is nothing else than strong independence product of $\mathcal{M}_1(X_1)$ and $\mathcal{M}_2(X_2)$. The extreme points of $\mathcal{M}_1 \triangleright \mathcal{M}_2$ are

$$\begin{aligned}&[0.12, 0.08, 0.48, 0.32], [0.2, 0, 0.8, 0],\\ &[0.42, 0.28, 0.18, 0.12], [0.7, 0, 0.3, 0],\end{aligned} \tag{4}$$

nevertheless

$$\begin{aligned}&(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1X_2)\\ &\quad \neq \mathrm{CH}\{[0.12, 0.08, 0.48, 0.32], [0.2, 0, 0.8, 0],\\ &\qquad [0.42, 0.28, 0.18, 0.12], [0.7, 0, 0.3, 0]\},\end{aligned}$$

as e.g.

$$\begin{aligned}&[0.41, 0.04, 0.39, 0.16]\\ &\quad \in \mathrm{CH}\{[0.12, 0.08, 0.48, 0.32], [0.2, 0, 0.8, 0],\\ &\qquad [0.42, 0.28, 0.18, 0.12], [0.7, 0, 0.3, 0]\},\end{aligned}$$

but $[0.41, 0.04, 0.39, 0.16] \notin \mathcal{M}_1 \triangleright \mathcal{M}_2$. ♢

It is evident, that one would obtain the same result by application of the formula in [b] (if he/she omits the fact that the condition $\mathcal{M}_1(X_{K\cap L}) \neq \mathcal{M}_2(X_{K\cap L})$ is not fulfilled), as trivially $\mathcal{M}_1(X_{K\cap L})) \ll \mathcal{M}_2(X_{K\cap L})$. Nevertheless, these two cases must be distinguished in general case, as can be seen from the following two examples.

Let us note that in the examples that follow we will prefer to use extreme points of credal sets (4) to their general form (3), as it seems to be more convenient if we want to compare e.g. the resulting credal sets (or their marginals).

Example 3 Let

$$\begin{aligned}&\mathcal{M}_1(X_1X_2)\\ &\quad = \mathrm{CH}\{[0.2, 0.8, 0, 0], [0.1, 0.4, 0.1, 0.4],\\ &\qquad [0.25, 0.25, 0.25, 0.25], [0, 0, 0.5, 0.5]\},\end{aligned}$$

and

$$\begin{aligned}&\mathcal{M}_2(X_2X_3)\\ &\quad = \mathrm{CH}\{[0.5, 0, 0.5, 0], [0.2, 0.3, 0.2, 0.3],\\ &\qquad [0.3, 0.3, 0.2, 0.2], [0, 0.6, 0, 0.4]\},\end{aligned}$$

be two credal sets about variables X_1X_2 and X_2X_3, respectively. These two credal sets are not projective, as $\mathcal{M}_1(X_2) = \mathrm{CH}\{[0.2, 0.8], [0.5, 0.5]\}$, while $\mathcal{M}_2(X_2) = \mathrm{CH}\{[0.5, 0.5], [0.6, 0.4]\}$. Therefore [b] of Definition 1 should be applied:

$$\begin{aligned}&(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1X_2X_3)\\ &\quad \subseteq \mathrm{CH}\{[0.2, 0, 0.8, 0, 0, 0, 0, 0],\\ &\qquad [0.08, 0.12, 0.32, 0.48, 0, 0, 0, 0],\\ &\qquad [0.1, 0, 0.4, 0, 0.1, 0, 0.4, 0],\\ &\qquad [0.04, 0.06, 0.16, 0.24, 0.04, 0.06, 0.16, 0.24],\\ &\qquad [0.1, 0.1, 0.4, 0.4, 0, 0, 0, 0],\\ &\qquad [0, 0.2, 0, 0.8, 0, 0, 0, 0],\\ &\qquad [0.05, 0.05, 0.2, 0.2, 0.05, 0.05, 0.2, 0.2],\\ &\qquad [0, 0.1, 0, 0.4, 0, 0.1, 0, 0.4],\\ &\qquad [0.25, 0, 0.25, 0, 0.25, 0, 0.25, 0],\\ &\qquad [0.1, 0.15, 0.1, 0.15, 0.1, 0.15, 0.1, 0.15],\\ &\qquad [0, 0, 0, 0, 0.5, 0, 0.5, 0],\\ &\qquad [0, 0, 0, 0, 0.2, 0.3, 0.2, 0.3]\\ &\qquad [0.125, 0.125, 0.125, 0.125,\\ &\qquad\quad 0.125, 0.125, 0.125, 0.125],\\ &\qquad [0, 0.25, 0, 0.25, 0, 0.25, 0, 0.25],\\ &\qquad [0, 0, 0, 0, 0.25, 0.25, 0.25, 0.25],\\ &\qquad [0, 0, 0, 0, 0, 0.5, 0, 0.5]\}.\end{aligned}$$

If we, despite this fact, try to apply [a] of Definition 1, we will realize that only probability distributions P_1 and P_2 from $\mathcal{M}_1(X_1X_2)$ and $\mathcal{M}_2(X_2X_3)$, respectively, with marginal $P_i^{\downarrow\{2\}} = [0.5, 0.5]$ are projective, and therefore we obtain only a subset of $(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1X_2X_3)$, namely a subset of

$$\begin{aligned}\mathrm{CH}\{&[0.25, 0, 0.25, 0, 0.25, 0, 0.25, 0],\\ &[0.1, 0.15, 0.1, 0.15, 0.1, 0.15, 0.1, 0.15],\\ &[0, 0, 0, 0, 0.5, 0, 0.5, 0],\\ &[0, 0, 0, 0, 0.2, 0.3, 0.2, 0.3]\},\end{aligned}$$

which does not keep the first marginal in contrary to $(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1X_2X_3)$, as can easily be checked. ♢

Example 4 Let $\mathcal{M}_1(X_1X_2)$ be as in previous example and

$$\begin{aligned}\mathcal{M}_2(X_2X_3) = \mathrm{CH}\{&[0.2, 0, 0.3, 0.5], [0, 0.2, 0, 0.8],\\ &[0.5, 0, 0.5, 0], [0.2, 0.3, 0.2, 0.3]\},\end{aligned}$$

be a credal set about variables X_2X_3. Contrary to the previous example these two credal sets are projective, as

$$\mathcal{M}_1(X_2) = \mathrm{CH}\{[0.2, 0.8], [0.5, 0.5]\} = \mathcal{M}_2(\mathrm{X}_2),$$

therefore [a] of Definition 1 should be applied:

$$\begin{aligned}&(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1X_2X_3)\\ &\quad \subseteq \mathrm{CH}\{[0.2, 0, 0.3, 0.5, 0, 0, 0, 0],\\ &\qquad [0, 0.2, 0, 0.8, 0, 0, 0, 0],\\ &\qquad [0.1, 0, 0.15, 0.25, 0.1, 0, 0.15, 0.25],\\ &\qquad [0, 0.1, 0, 0.4, 0, 0.1, 0, 0.4],\\ &\qquad [0.25, 0, 0.25, 0, 0.25, 0, 0.25, 0],\\ &\qquad [0.1, 0.15, 0.1, 0.15, 0.1, 0.15, 0.1, 0.15],\\ &\qquad [0, 0, 0, 0, 0.5, 0, 0.5, 0],\\ &\qquad [0, 0, 0, 0, 0.2, 0.3, 0.2, 0.3]\},\end{aligned}$$

If, instead of it, one used [b] of the same definition, he/she would arrive to the credal set containing in addition the following extreme points

$$\begin{aligned}&[0.2, 0, 0.8, 0, 0, 0, 0, 0],\\ &[0.08, 0.12, 0.32, 0.48, 0, 0, 0, 0],\\ &[0.1, 0, 0.4, 0, 0.1, 0, 0.4, 0],\\ &[0.04, 0.06, 0.16, 0.24, 0.04, 0.06, 0.16, 0.24],\\ &[0.25, 0, 0.09375, 0.15625, 0.25, 0, 0.09375, 0.15625],\\ &[0, 0.25, 0, 0.25, 0, 0.25, 0, 0.25],\\ &[0, 0, 0, 0, 0.5, 0, 0.1875, 0.3125],\\ &[0, 0, 0, 0, 0, 0.5, 0, 0.5].\end{aligned}$$

Although both of these composed credal sets keep the first marginal, as can easily be checked, they differ form each other as concerns the second marginal: the correctly composed credal set keeps it, while the other has much bigger marginal, containing in addition the following extreme points:

$$\begin{aligned}&[0.2, 0, 0.8, 0], [0.08, 0.12, 0.32, 0.48],\\ &[0.5, 0, 0.1875, 0.3125], [0, 0.5, 0, 0.5].\end{aligned}$$

♢

Unfortunately, the definition is not elegant, nevertheless, its basic properties are satisfied and, as we shall see later, it holds also for other properties necessary for the introduction of compositional models.

4 Further Properties

As said in the Introduction, the operator of composition was originally introduced in (precise) probability theory. Nevertheless, any probability distribution may be viewed also as a singleton credal set (i.e. credal set containing a single point). One would expect that the operator of composition we have introduced in this contribution coincides with the probabilistic one if applied to singleton credal sets. And it is the case, as can be seen from the following lemma.

Lemma 2 *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two singleton credal sets about X_K and X_L, respectively, where $\mathcal{M}_1(X_{K\cap L})$ is absolutely continuous with respect to $\mathcal{M}_2(X_{K\cap L})$. Then $(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K\cup L})$ is also a singleton.*

Proof. Let us suppose that $\mathcal{M}_1 \triangleright \mathcal{M}_2$ is not a singleton, i.e. it contains at least two different points. Due to the condition of absolute continuity both these points can be expressed in the form

$$P^i = P_1^i \cdot P_2^i/(P_2^i)^{\downarrow K\cap L}.$$

As $P^1 \neq P^2$, it is evident that either $P_1^1 \neq P_1^2$ or $P_2^1/(P_2^1)^{\downarrow K\cap L} \neq P_2^2/(P_2^2)^{\downarrow K\cap L}$ (and therefore also $P_2^1 \neq P_2^2$), or both. In any case, it is a contradiction as both credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ are singletons. □

The reader should however realize that the definition of the operator of composition for singleton credal sets is not completely equivalent to the definition of composition for probabilistic distributions. They equal each other only in case that the probabilistic version is defined. This is ensured in Lemma 2 by assuming the absolute continuity. In case it does not hold, the probabilistic operator is not defined while its credal version introduced in this paper is always defined (analogous to evidential operator of composition). Nevertheless, in this case, the result is not a singleton credal set. We shall illustrate it by a simple example.

Example 5 Let

$$\mathcal{M}_1(X_1X_2) = \{[0.25, 0.25, 0.25, 0.25]\},$$

and

$$\mathcal{M}_2(X_2X_3) = \{[0.5, 0.5, 0, 0]\},$$

be two singleton credal sets about variables X_1X_2 and X_2X_3, respectively. Let us compute $\mathcal{M}_1 \triangleright \mathcal{M}_2$. As $\mathcal{M}_1(X_2) = \{[0.5, 0.5]\}$, while $\mathcal{M}_2(X_2) = \{[1, 0]\}$, it is evident, that $\mathcal{M}_1$ is not absolutely continuous with respect to $\mathcal{M}_2$. Therefore we have via [c] of Definition 1:

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1X_2X_3) = \mathcal{M}_1(X_1X_2)^{\uparrow\{1,2,3\}}$$

which is evidently not a singleton any more.

Let us remark that $(\mathcal{M}_2 \triangleright \mathcal{M}_1)(X_1X_2X_3)$, in contrast to $(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1X_2X_3)$, is a singleton credal set

$$(\mathcal{M}_2 \triangleright \mathcal{M}_1)(X_1X_2X_3) = \{[0.25, 0.25, 0, 0, 0.25, 0.25, 0, 0]\},$$

because $\mathcal{M}_2(X_2)$ is absolutely continuous with respect to $\mathcal{M}_1(X_2)$. ♢

From this example one can see that the operator of composition is not commutative. The following example demonstrates that this operator is neither associative.

Example 6 Let

$$\mathcal{M}_1(X_1) = \mathrm{CH}\{[0.2, 0, 8], [0.7, 0.3]\}$$

and

$$\mathcal{M}_2(X_2) = \{[0.5, 0.5]\}$$

be two credal sets about X_1 and X_2, respectively, and

$$\mathcal{M}_3(X_1X_2) = \mathrm{CH}\{[1, 0, 0, 0], [0, 1, 0, 0]\ [0, 0, 1, 0], [0, 0, 0, 1]\}.$$

Then $\mathcal{M}_1 \triangleright \mathcal{M}_2$ is obtained via [a] in Definition 1:

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1X_2) = \mathrm{CH}\{[0.1, 0.1, 0.4, 0.4], [0.35, 0.35, 0.15, 0.15]\},$$

due to Definition 1 and $(\mathcal{M}_1 \triangleright \mathcal{M}_2) \triangleright \mathcal{M}_3 = \mathcal{M}_1 \triangleright \mathcal{M}_2$ according to property (2) of Lemma 1. On the other hand

$$(\mathcal{M}_2 \triangleright \mathcal{M}_3)(X_1X_2) = \mathrm{CH}\{[0.5, 0.5, 0, 0], [0.5, 0, 0, 0.5]\ [0, 0.5, 0.5, 0], [0, 0, 0.5, 0.5]\},$$

via [c] of Definition 1. Now, computing $\mathcal{M}_1 \triangleright (\mathcal{M}_2 \triangleright \mathcal{M}_3)$ we obtain again via [c] of Definition 1

$$\begin{aligned}(\mathcal{M}_1 \triangleright (\mathcal{M}_2 \triangleright \mathcal{M}_3))(X_1X_2) = \mathrm{CH}\{&[0.2, 0, 0.8, 0], [0.2, 0, 0, 0.8],\\ &[0, 0.2, 0.8, 0], [0, 0.2, 0, 0.8],\\ &[0.7, 0, 0.3, 0], [0.7, 0, 0, 0.3]\\ &[0, 0.7, 0.3, 0], [0, 0.7, 0, 0.3]\},\end{aligned}$$

which evidently differs from $(\mathcal{M}_1 \triangleright \mathcal{M}_2) \triangleright \mathcal{M}_3$. ♢

The following theorem reveals the relationship between strong independence and the operator of composition. It is, together with Lemma 1, the most important assertion enabling us to introduce multidimensional models.

Theorem 1 *Let $\mathcal{M}$ be a credal set about $X_{K\cup L}$ with marginals $\mathcal{M}(X_K)$ and $\mathcal{M}(X_L)$. Then*

$$\mathcal{M}(X_{K\cup L}) = (\mathcal{M}^{\downarrow K} \triangleright \mathcal{M}^{\downarrow L})(X_{K\cup L}) \qquad (5)$$

iff

$$(K \setminus L) \perp\!\!\!\perp (L \setminus K)|(K \cap L). \qquad (6)$$

Proof. Let us suppose that (5) holds. Since $\mathcal{M}_1(X_K)$ and $\mathcal{M}_2(X_L)$ are projective, [a] of Definition 1 is applied and therefore

$$\begin{aligned}&\mathcal{M}(X_{K\cup L})\\ &\quad= \{(P_1 \cdot P_2)/P_2^{\downarrow K\cap L} : P_1 \in \mathcal{M}(X_K),\\ &\qquad P_2 \in \mathcal{M}(X_L), P_1^{\downarrow K\cap L} = P_2^{\downarrow K\cap L}\}).\end{aligned}$$

To prove (6) means to find for any P from $\mathcal{M}(X_{K\cup L})$ a pair of projective distributions P_1 and P_2 from $\mathcal{M}(X_K)$ and $\mathcal{M}(X_L)$, respectively, such that $P = (P_1 \cdot P_2)/P_1{}^{\downarrow K\cap L}$. But due to condition of projectivity, $\mathcal{M}(X_{K\cup L})$ consists of exactly this type of distributions.

Let on the other hand (6) be satisfied. Then any P from $\mathcal{M}(X_{K\cup L})$ can be expressed as conditional product of its marginals, namely

$$P = (P^{\downarrow K} \cdot P^{\downarrow K})/P^{\downarrow K\cap L},$$

$P^{\downarrow K} \in \mathcal{M}(X_K)$ and $P^{\downarrow L} \in \mathcal{M}(X_L)$. Therefore,

$$\begin{aligned}&\mathcal{M}(X_{K\cup L})\\ &= \quad \{(P^{\downarrow K} \cdot P^{\downarrow K})/P^{\downarrow K\cap L} : P^{\downarrow K} \in \mathcal{M}_1(X_K),\\ &\qquad P^{\downarrow L} \in \mathcal{M}_2(X_L))\},\end{aligned}$$

which concludes the proof. □

5 Compositional models

In this section we will consider repetitive application of the operator of composition with the goal to create a multidimensional model. Since the operator is neither commutative nor associative we have to specify in which order the low-dimensional credal sets are composed together. To make the formulae more transparent we will omit parentheses in case that the operator is to be applied from left to right, i.e., in what follows

$$\begin{aligned}&\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \mathcal{M}_3 \triangleright \ldots \triangleright \mathcal{M}_{n-1} \triangleright \mathcal{M}_n\\ &\quad= (\ldots((\mathcal{M}_1 \triangleright \mathcal{M}_2) \triangleright \mathcal{M}_3) \triangleright \ldots \triangleright \mathcal{M}_{n-1}) \triangleright \mathcal{M}_n.\end{aligned}$$

Furthermore, we will always assume $\mathcal{M}_i$ be a credal set about X_{K_i}.

The reader familiar with some papers on probabilistic, possibilistic or evidential compositional models knows that one of the most important notions of this theory is that of a so-called *perfect sequence*, which will be now introduced also for credal sets.

Definition 2 A generating sequence of credal sets $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$ is called *perfect* if

$$\begin{aligned}\mathcal{M}_1 \triangleright \mathcal{M}_2 &= \mathcal{M}_2 \triangleright \mathcal{M}_1,\\ \mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \mathcal{M}_3 &= \mathcal{M}_3 \triangleright (\mathcal{M}_1 \triangleright \mathcal{M}_2),\\ &\vdots\\ \mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \ldots \triangleright \mathcal{M}_n &= \mathcal{M}_n \triangleright (\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_{n-1}).\end{aligned}$$

It is evident that the necessary condition for perfectness is pairwise projectivity of low-dimensional credal sets. However, the following example demonstrates the fact that it need not be sufficient.

Example 7 Let $\mathcal{M}_1(X_1)$ and $\mathcal{M}_2(X_2)$ as in Example 2 and let $\mathcal{M}_3(X_1, X_2)$ be defined as follows:

$$\begin{aligned}&\mathcal{M}_3(X_1, X_2)\\ &\quad= \mathrm{CH}\{[0.1, 0.1, 0.5, 0.3], [0.2, 0, 0.8, 0],\\ &\qquad [0.4, 0.3, 0.2, 0.1], [0.7, 0, 0.3, 0]\}.\end{aligned}$$

It is evident, that $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ are pairwise projective, as

$$\mathcal{M}_3(X_1) = \mathrm{CH}\{[0.2, 0.8,], [0.7, 0.3]\} = \mathcal{M}_1(\mathrm{X}_1)$$

and

$$\mathcal{M}_3(X_2) = \mathrm{CH}\{[0.6, 0.4,], [1, 0]\} = \mathcal{M}_2(\mathrm{X}_2)$$

and $\mathcal{M}_1$ and $\mathcal{M}_2$ are trivially projective, as already mentioned above. But they do not form a perfect sequence, as

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \mathcal{M}_3)(X_1X_2) = (\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_1X_2),$$

whose extreme points are in (4), while

$$(\mathcal{M}_3 \triangleright (\mathcal{M}_1 \triangleright \mathcal{M}_2))(X_1X_2) = \mathcal{M}_3(X_1X_3),$$

which is different. ♢

Therefore a stronger condition, expressed by the following assertion, must be fulfilled.

Lemma 3 *A generating sequence $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$ is perfect* iff *the pairs of credal sets $\mathcal{M}_j$ and $(\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_{j-1})$ are projective, i.e. if*

$$\begin{aligned}&\mathcal{M}_j(X_{K_j\cap(K_1\cup\ldots\cup K_{j-1})})\\ &\qquad= (\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_{j-1})(X_{K_j\cap(K_1\cup\ldots\cup K_{j-1})}),\end{aligned}$$

for all $j = 2, 3, \ldots, n$.

Proof. This assertion is proved just by a multiple application of assertion (3) of Lemma 1:

$$\begin{aligned}
&\mathcal{M}_1 \triangleright \mathcal{M}_2 = \mathcal{M}_2 \triangleright \mathcal{M}_1 \\
&\qquad \iff \mathcal{M}_1(X_{K_2 \cap K_1}) = \mathcal{M}_2(X_{K_2 \cap K_1}), \\
&\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \mathcal{M}_3 = \mathcal{M}_3 \triangleright (\mathcal{M}_1 \triangleright \mathcal{M}_2) \\
&\qquad \iff (\mathcal{M}_1 \triangleright \mathcal{M}_2)(X_{K_3 \cap (K_1 \cup K_2)}) \\
&\qquad\qquad = \mathcal{M}_3(X_{K_3 \cap (K_1 \cup K_2)}), \\
&\qquad \vdots \\
&\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \ldots \triangleright \mathcal{M}_n = \mathcal{M}_n \triangleright (\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}n-1) \\
&\qquad \iff (\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_{n-1})(X_{K_n \cap (K_1 \cup \ldots \cup K_{n-1})}) \\
&\qquad\qquad = \mathcal{M}_n(X_{K_n \cap (K_1 \cup \ldots \cup K_{n-1})}). \qquad \square
\end{aligned}$$

From Definition 2 one can hardly see what are the properties of the perfect sequences; the main one is expressed by the following characterization theorem, which, hopefully, also reveals why we call these sequences perfect.

Theorem 2 *A generating sequence of credal sets $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$ is perfect* iff *all the credal sets from this sequence are marginal to the composed credal set $\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \ldots \triangleright \mathcal{M}_n$:*

$$(\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \ldots \triangleright \mathcal{M}_n)(X_{K_j}) = \mathcal{M}_j(X_{K_j}),$$

for all $j = 1, \ldots, m$.

Proof. The fact that all credal sets $\mathcal{M}_j$ from perfect sequence $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$ are marginals of $(\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \ldots \triangleright \mathcal{M}_n)$ follows from the fact that $(\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_j)$ is marginal to $(\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_n)$ (due to *(ii)* of Lemma 1) and $\mathcal{M}_j$ is marginal to

$$\mathcal{M}_j \triangleright (\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_{j-1}) = \mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_j.$$

Suppose now that for all $j = 1, \ldots, n$, $\mathcal{M}_j$ are marginal assignments to $\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_n$. It means that all the assignments from the sequence are pairwise projective, and that each $\mathcal{M}_j$ is projective with any marginal assignment of $\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_n$, and consequently also with $\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_{j-1}$. So we get that

$$\begin{aligned}
&\mathcal{M}_j(X_{K_j \cap (K_1 \cup \ldots \cup K_{j-1})}) \\
&\qquad = (\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_{j-1})(X_{K_j \cap (K_1 \cup \ldots \cup K_{j-1})})
\end{aligned}$$

for all $j = 2, \ldots, n$, which is equivalent, due to Lemma 3, to the fact that $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$ is perfect. □

The following (almost trivial) assertion, which brings the sufficient condition for perfectness, resembles assertions concerning decomposable models.

Theorem 3 *Let a generating sequence of pairwise projective credal sets $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$ be such that $K_1, K_2, \ldots, K_n$ meets the well-known running intersection property:*

$$\begin{aligned}
&\forall j = 2, 3, \ldots, n \;\; \exists \ell (1 \le \ell < j) \\
&\qquad \textit{such that} \quad K_j \cap (K_1 \cup \ldots \cup K_{j-1}) \subseteq K_\ell.
\end{aligned}$$

Then $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$ is perfect.

Proof. Due to Lemma 3 it is enough to show that for each $j = 2, \ldots, n$ credal set $\mathcal{M}_j$ and the composed credal set $\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_{j-1}$ are projective. Let us prove it by induction.

For $j = 2$ the required projectivity is guaranteed by the fact that we assume pairwise projectivity of all $\mathcal{M}_1, \ldots, \mathcal{M}_n$. So we have to prove it for general $j > 2$ under the assumption that the assertion holds for $j - 1$, which means (due to Theorem 2) that all $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_{j-1}$ are marginal to $\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_{j-1}$. Since we assume that $K_1, \ldots, K_n$ meets the running intersection property, there exists $\ell \in \{1, 2, \ldots j-1\}$ such that $K_j \cap (K_1 \cup \ldots \cup K_{j-1}) \subseteq K_\ell$. Therefore $(\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_{j-1})(X_{K_j \cap (K_1 \cup \ldots \cup K_{j-1})})$ and $\mathcal{M}_\ell(X_{K_j \cap (K_1 \cup \ldots \cup K_{j-1})})$ are the same marginals of $\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_{j-1}$ and therefore they have to equal to each other:

$$\begin{aligned}
&(\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_{j-1})(X_{K_j \cap (K_1 \cup \ldots \cup K_{j-1})}) \\
&\qquad = \mathcal{M}_\ell(X_{K_j \cap (K_1 \cup \ldots \cup K_{j-1})}).
\end{aligned}$$

However we assume that $\mathcal{M}_j$ and $\mathcal{M}_\ell$ are projective and therefore also

$$\begin{aligned}
&(\mathcal{M}_1 \triangleright \ldots \triangleright \mathcal{M}_{j-1})(X_{K_j \cap (K_1 \cup \ldots \cup K_{j-1})}) \\
&\qquad = \mathcal{M}_j(X_{K_j \cap (K_1 \cup \ldots \cup K_{j-1})}),
\end{aligned}$$

as desired. □

It should be stressed at this moment that running intersection property of $K_1, K_2, \ldots, K_n$ is a sufficient condition which guarantees perfectness of a generating sequence of pairwise projective assignments. By no means this condition is necessary as it will be shown in the following example.

Example 8 Simple example is given by two credal sets $\mathcal{M}_1$ and $\mathcal{M}_2$ from Example 7 about X_1 and X_2, respectively, and the third credal set $\tilde{\mathcal{M}}_3 = \mathcal{M}_1 \triangleright \mathcal{M}_2$. Considering sequence $\mathcal{M}_1, \mathcal{M}_2, \tilde{\mathcal{M}}_3$, it is evident that $K_1 = \{1\}, K_2 = \{2\}, K_3 = \{1, 2\}$ do not meet the running intersection property. And yet the sequence $\mathcal{M}_1, \mathcal{M}_2, \tilde{\mathcal{M}}_3$ is perfect because all the credal sets are marginal (or equal) to $\mathcal{M}_1 \triangleright \mathcal{M}_2 \triangleright \tilde{\mathcal{M}}_3$. Let us note that if we chose instead of $\tilde{\mathcal{M}}_3$ any other credal set $\mathcal{M}_3$ about $X_1 X_2$ different from $\tilde{\mathcal{M}}_3 = \mathcal{M}_1 \triangleright \mathcal{M}_2$, e.g. that from Example 7 the generating sequence $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ would not be perfect any more. ♢

Therefore we can see that perfectness of a sequence is not only a structural property connected with the properties of $K_1, K_2, \ldots, K_n$ but depends also on specific values of the respective basic assignments.

As said already in the introduction, in precise probability framework any multidimensional distribution representable by a Bayesian network can also be represented in the form of a perfect sequence, and vice versa. For more details the reader is referred to [7], where also an algorithm for transformation of a perfect sequence of probability distributions into a Bayesian network can be found.

Recently we have found out, that in evidence theory transformation from evidential network to a compositional model is exactly the same as in precise probability framework, but the opposite process is a bit different — it may happen that resulting model expressed by evidential network is less precise than that the compositional model [14].

At present we do not know too much about the relationship between compositional models of multidimensional credal sets and credal networks. We conjecture it will be similar to the evidential framework. But it is only a conjecture, the research is just at the beginning. Nevertheless, to clarify this relationship is our first goal.

6 Conclusions

Graphical Markov Models were designed to enable description of real-life problems by probabilistic models. This is because problems of practice lead to multidimensional models, where the number of dimensions is expressed rather in hundreds than in tens. Inspired by the original probabilistic approach the paper is the first attempt to build up compositional models of multidimensional credal sets as an alternative to Graphical Markov Models with imprecision.

We have defined credal set operator of composition manifesting all the main characteristics of its probabilistic pre-image. Even more, there is one point in which the credal set operator of composition is superior to the probabilistic one (similarly to the operator in the evidential framework): thanks to the ability of credal sets to model total ignorance, the operator of composition is for credal sets always defined, which is not the case in the (precise) probabilistic framework.

In the paper we have proved the basic properties of the operator (including the relationship to strong independence) necessary for the introduction of compositional models and their most important special case, *perfect sequence models.* Naturally, there are still many open problems to be solved. The most important one is a design of efficient computational procedures for this type of models. At this moment we know very little about similarities and differences between the described compositional models and other multidimensional models such as [1, 3, 11], as well as about the relation between the compositional models developed for credal sets and those introduced in possibility [12, 13] and evidence [8, 9] theories.

Acknowledgements

The support of Grant GAČR P402/11/0378 is gratefully acknowledged.

References

[1] A. Benavoli and A. Antonucci, Aggregating imprecise probabilistic knowledge. In Augustin, T., Coolen, F., Moral, S., Troffaes, M.C.M. (Eds.), ISIPTA '09: *Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications.* SIPTA, pp. 31-40.

[2] I. Couso, S. Moral and P. Walley, Examples of independence for imprecise probabilities, *Proceedings of ISIPTA'99,* eds. G. de Cooman, F. G. Cozman, S. Moral, P. Walley, Ghent, June 29 – July 2, 1999, pp. 121–130.

[3] F. G. Cozman, Credal networks, *Artificial Intelligence Journal*, **120** (2000), pp. 199-233.

[4] R. Jiroušek. Composition of probability measures on finite spaces. *Proc. of UAI'97*, (D. Geiger and P. P. Shenoy, eds.). Morgan Kaufmann Publ., San Francisco, California, pp. 274–281, 1997.

[5] R. Jiroušek. Graph modelling without graphs. *Proc. of IPMU'98*, (B. Bouchon-Meunier, R.R. Yager, eds.). Editions E.D.K. Paris, pp. 809–816, 1988.

[6] R. Jiroušek. Marginalization in composed probabilistic models. *Proc. of UAI'00* (C. Boutilier and M. Goldszmidt eds.), Morgan Kaufmann Publ., San Francisco, California, pp. 301–308, 2000.

[7] R. Jiroušek and J. Vejnarová, Construction of multidimensional models by operators of composition: current state of art. *Soft Computing,* **7** (2003), pp. 328–335.

[8] R. Jiroušek, J. Vejnarová and M. Daniel, Compositional models for belief functions. *Proceedings of 5th International Symposium on Imprecise Probability: Theories and Applications*

ISIPTA'07, eds. G. De Cooman, J. Vejnarová, M. Zaffalon, Praha, 2007, pp. 243-252.

[9] R. Jiroušek and J. Vejnarová, Compositional models and conditional independence in Evidence Theory, *Int. J. Approx. Reasoning,* **52** (2011), 316-334.

[10] S. Moral and A. Cano, Strong conditional independence for credal sets, *Ann. of Math. and Artif. Intell.*, **35** (2002), 295–321.

[11] S. Moral and J. Sagrado, Aggregation of imprecise probabilities. *Aggregation and fusion of imperfect information,* 1997, pp. 162–188.

[12] J. Vejnarová, Composition of possibility measures on finite spaces: preliminary results. In: *Proc. of 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU'98,* (B. Bouchon-Meunier, R.R. Yager, eds.). Editions E.D.K. Paris, 1998, pp. 25–30.

[13] J. Vejnarová, On possibilistic marginal problem, *Kybernetika* **43**, 5 (2007), pp. 657–674.

[14] J. Vejnarová, Evidential networks from a different perspective. In: *Synergies of Soft Computing and Statistics for Intelligent Data Analysis,* Soft Methods In Probability and Statistics, (2012). pp. 429–436.

Modelling practical certainty and its link with classical propositional logic

Arthur Van Camp
Ghent University
SYSTeMS Research Group
Arthur.VanCamp@UGent.be

Gert de Cooman
Ghent University
SYSTeMS Research Group
Gert.deCooman@UGent.be

Abstract

We model practical certainty in the language of accept & reject statement-based uncertainty models. We present three different ways, each time using a different nature of assessment: we study coherent models following from (i) favourability assessments, (ii) acceptability assessments, and (iii) indifference assessments. We argue that a statement of favourability, when used with an appropriate background model, essentially boils down to stating a belief of practical certainty using acceptability assessments. We show that the corresponding models do not form an intersection structure, in contradistinction with the coherent models following from an indifferenc assessment. We construct embeddings of classical propositional logic into each of our models for practical certainty.

Keywords. Imprecise probabilities, accept & reject statement-based uncertainty models, classical propositional logic, strong belief structure.

1 Introduction

In classical propositional logic, a subject who is certain of the truth of some propositions, or equivalently, of the occurrence of the corresponding events, models this by giving his set of certain events—or true propositions. In this paper, we investigate to what extent classical propositional logic can be embedded in accept and reject statement-based uncertainty models [9]. The embedding is not perfect, therefore we speak of *practical uncertainty*. The language of the uncertainty models used is rich enough to encompass the different approaches of Walley and de Finetti. In order to obtain more insight in these approaches, we study three different types of assessments.

The first type of assessments fits well into Walley's approach to defining lower previsions, and focusses on strict preference. The second type appears to be weaker, as it focusses on weak preference, but we show that the difference essentially does not matter: the derived coherent models from both types of assessments are the same. The third and last type of assessments is more in line with de Finetti's approach to defining previsions, and focusses on indifference.

Because strong belief structures [2] have nice properties, we investigate whether the derived coherent models constitute such structures. It turns out that only the coherent models derived from the third type of assessments do.

The basic concepts are introduced in Section 2. In the subsequent three sections, we study three different ways of modelling practical certainty. We start with favourability assessments in Section 3, and study the consequences of the rationality requirements of No Confusion, Deductive Closure and No Limbo. We proceed with acceptability assessments in Section 4, where we also investigate the connection with the models of the previous section. The last type of assessments—those based on indifference—are discussed in Section 5. In Section 6, we find the corresponding coherent lower prevision models, and we investigate when they are coherent. We make the link with (strong) belief structures in Section 7. Finally, in Section 8, we embeds classical propositional logic into the models introduced in this paper.

2 Notations and concepts

We consider a subject who is uncertain about the value of a variable X that takes values in the—not necessarily finite but non-empty—possibility space $\mathscr{X}$. We want to model this subject's beliefs about the value that X assumes.

2.1 Events and sets of events

An *event* is a subset of $\mathscr{X}$, or equivalently, an element of the *power set* $\mathscr{P} := \{A\colon A \subseteq \mathscr{X}\}$: the collection of all events. A non-empty subset $\mathscr{C}$ of $\mathscr{P}$ is called a *filter base* if it is *closed under finite intersections* (closed under *conjunction*): if both A and B are elements of $\mathscr{C}$, then also $A \cap B \in \mathscr{C}$. A filter base $\mathscr{C}$ is called *proper* if additionally $\emptyset \notin \mathscr{C}$. A non-empty subset $\mathscr{F}$ of $\mathscr{P}$ is called a *filter* if: (i) $\mathscr{F}$ is closed under finite intersections, and (ii) $\mathscr{F}$

is *increasing* (closed under *modus ponens*): if $A \in \mathscr{F}$ and $A \subseteq B$, then also $B \in \mathscr{F}$. A filter $\mathscr{F}$ is called *proper* if additionally $\emptyset \notin \mathscr{F}$, or equivalently, $\mathscr{F} \neq \mathscr{P}$. We denote the set of all proper filters by $\mathbb{F}$. A proper filter $\mathscr{U}$ is called an *ultrafilter* if either $A \in \mathscr{U}$ or $A^c \in \mathscr{U}$ for all events A. We denote the set of all ultrafilters by $\mathbb{U}$.

As an example, consider a filter base $\mathscr{C}$, then the set $\{A \in \mathscr{P} \colon (\exists B \in \mathscr{C}) B \subseteq A\}$ is a filter—the filter generated by the filter base $\mathscr{C}$. It is proper if and only if the filter base $\mathscr{C}$ is.

2.2 Gambles and sets of gambles

A *gamble* f is a bounded real-valued function on the possibility space $\mathscr{X}$. It is interpreted as an uncertain reward $f(X)$. If the value of the variable X turns out to be x, it results in a—positive or negative—payoff $f(x)$, expressed in units of some predetermined linear utility scale. The set of all gambles on $\mathscr{X}$ is denoted by $\mathscr{L}$.

We can compare two gambles f and g in $\mathscr{L}$. We write $f \geq g$ if $f(x) \geq g(x)$ for all x in $\mathscr{X}$. For example, $f \geq 0$ if f is nowhere negative, and we then say that f is *non-negative*. The subset $\mathscr{L}_{\geq 0}$ of $\mathscr{L}$ is the set of all non-negative gambles. We write $f > g$ if $f \geq g$ and $f \neq g$. For example, $f < 0$ if f is nowhere positive—so $f \leq 0$—and $f(x) < 0$ for at least one x in $\mathscr{X}$, and we then say that f is *negative*. The subset $\mathscr{L}_{<0}$ of $\mathscr{L}$ is the set of all negative gambles. We write $f \gtrdot g$ if $\inf(f-g) > 0$. For example, $f \lessdot 0$ if $\sup f < 0$, meaning that the gamble f is negative and bounded away from zero. The subset $\mathscr{L}_{\lessdot 0}$ of $\mathscr{L}$ is the set of all such gambles. We write $f \rhd g$ if $f(x) > g(x)$ for all x in $\mathscr{X}$. For example, $f \rhd 0$ if f is everywhere (strictly) positive, and we then say that f is *point-wise positive*. The subset $\mathscr{L}_{\rhd 0}$ of $\mathscr{L}$ is the set of all such gambles.

We also introduce a number of operations on sets of gambles $\mathscr{K}, \mathscr{K}' \subseteq \mathscr{L}$. The first is the *Minkowski sum* $\mathscr{K} + \mathscr{K}' := \{f + g \colon f \in \mathscr{K}, g \in \mathscr{K}'\}$. The *positive scalar hull* $\overline{\mathscr{K}} := \{\lambda f \colon \lambda \in \mathbb{R}_{>0}, f \in \mathscr{K}\}$ is the collection of all positive multiples of gambles in $\mathscr{K}$, where we use the notation $\mathbb{R}_{>a}$ for the set of real numbers (strictly) greater than the real number a. The *positive linear hull* $\operatorname{posi}\mathscr{K}$ is the collection of all positive linear combinations of gambles in $\mathscr{K}$:

$$\operatorname{posi}\mathscr{K} := \left\{ \sum_{k=1}^{n} \lambda_k f_k \colon n \in \mathbb{N}, \lambda_k \in \mathbb{R}_{>0}, f_k \in \mathscr{K} \right\},$$

where we use the notation $\mathbb{N}$ for the set of natural numbers (positive integers). Observe that

$$\begin{aligned}\operatorname{posi}(\mathscr{K} \cup \mathscr{K}') = {} & \operatorname{posi}\mathscr{K} \\ & \cup \operatorname{posi}\mathscr{K}' \cup (\operatorname{posi}\mathscr{K} + \operatorname{posi}\mathscr{K}'). \quad (1)\end{aligned}$$

We call a set $\mathscr{K}$ a *cone* if $\operatorname{posi}\mathscr{K} = \mathscr{K}$.

2.3 Accept & reject statement-based uncertainty models

In order to have greater flexibility in expressing beliefs, we use the framework and language of *accept & reject statement-based uncertainty models*, as introduced and described in detail by Quaeghebeur et al. [9]. In contrast with the slightly older and more common framework of *sets of desirable gambles* [1, 4, 13], where the subject gives only one set of gambles, in this framework a subject is supposed to give two sets: a set of *acceptable* gambles $\mathscr{A}_{\succeq} \subseteq \mathscr{L}$, and a set of *rejected* gambles $\mathscr{A}_{\prec} \subseteq \mathscr{L}$. An assessment is then represented by $\mathscr{A} = \langle \mathscr{A}_{\succeq}; \mathscr{A}_{\prec} \rangle$. Following the discussion in Ref. [9], a subject's accepting a gamble f implies a commitment for him to engage in the following transaction: (i) the actual value x of the variable X is determined, and (ii) the subject gets the—possibly negative—payoff $f(x)$. On the other hand, the subject's rejecting a gamble implies that he excludes it from being accepted.

From an assessment, one can derive other types of statements. For any gamble $f \in \mathscr{A}_{\simeq} := \mathscr{A}_{\succeq} \cap -\mathscr{A}_{\succeq}$, the subject accepts both f and its negation $-f$. We say that he is *indifferent* about f, and $\mathscr{A}_{\simeq}$ is his set of indifferent gambles. For any $f \in \mathscr{A}_{\rhd} := \mathscr{A}_{\succeq} \cap -\mathscr{A}_{\prec}$, the subject accepts f and rejects its negation $-f$. We say that he finds f *favourable*, and $\mathscr{A}_{\rhd}$ is his set of favourable gambles. These are the gambles that the subject strictly prefers to 0, which is the interpretation that is usually given to desirable gambles [1, 4]. Finally, gambles in the set $\mathscr{A}_{\smile} := \mathscr{L} \setminus (\mathscr{A}_{\succeq} \cup \mathscr{A}_{\prec})$ are called *unresolved*. For unresolved gambles no accept or reject statement has been made.

3 Modelling practical certainty using favourability

3.1 Assessment

If a subject is *practically* certain that a proposition is true, or that the corresponding event A occurs, we will first take this to mean that he finds any gamble of the form $\mathbb{I}_A - 1 + \varepsilon$, with $\varepsilon \in \mathbb{R}_{>0}$, favourable.[1] Here $\mathbb{I}_A$ is the *indicator* of the event A, which assumes the value 1 on A (if the proposition is true) and 0 elsewhere. Finding $\mathbb{I}_A - 1 + \varepsilon$ favourable means: (i) the transaction that yields ε if A occurs and $\varepsilon - 1$ otherwise, is accepted, and (ii) the transaction that yields $-\varepsilon$ if A occurs, and $1 - \varepsilon$ otherwise, is rejected (excluded from being accepted). The first assessment means that the subject accepts to bet on A at odds $\varepsilon/(1-\varepsilon)$, and the second that he excludes accepting a bet against A at odds $(1-\varepsilon)/\varepsilon$. So our subject accepts to bet on A at all odds, and rejects betting against A at any odds.

A subject can be practically certain about a number of

[1] Actually, it is enough to look at $\varepsilon \in (0,1)$, because for $\varepsilon \geq 1$, $\mathbb{I}_A - 1 + \varepsilon \in \mathscr{L}_{\geq 0}$ already belongs to the background model; see further on.

events. We collect the events he is practically certain about in the set $\mathscr{T} \subseteq \mathscr{P}$. So his initial assessment is:

$$\mathscr{A} = \langle \mathscr{A}_{\rhd}; -\mathscr{A}_{\rhd} \rangle \\ \text{with } \mathscr{A}_{\rhd} = \{\mathbb{I}_A - 1 + \varepsilon \colon A \in \mathscr{T}, \varepsilon \in \mathbb{R}_{>0}\}. \quad (2)$$

Even before an assessment is given, some gambles can be presumed to be accepted and others to be rejected. Such *a priori* assumptions can be captured by positing a *background model* $\mathscr{S}$. In the context of favourability assessments, it follows from the discussion in Ref. [9, Section 5] that it is convenient to use the following background model:

$$\mathscr{S} = \langle \mathscr{L}_{\geq 0}; \mathscr{L}_{<0} \rangle,$$

so we take for granted that all non-negative gambles should be accepted, and all negative gambles should be rejected—be excluded from being accepted. The background model $\mathscr{S}$ is an instance of a *favour-indifference model* [9, Section 4.3], meaning that it fulfils the two conditions $-\mathscr{S}_{\prec} \subseteq \mathscr{S}_{\succeq}$ and $\mathscr{S}_{\succeq} = \mathscr{S}_{\rhd} \cup \mathscr{S}_{\simeq}$.

We use $\mathscr{B} := \mathscr{A} \cup \mathscr{S} = \langle \mathscr{A}_{\rhd} \cup \mathscr{L}_{\geq 0}; -\mathscr{A}_{\rhd} \cup \mathscr{L}_{<0} \rangle$ to denote the smallest assessment that includes both the subject's assessment $\mathscr{A}$ and the background model $\mathscr{S}$.

Clearly, we will have to impose conditions on the set $\mathscr{T}$ of events that the subject is practically certain to occur. To give just one example, suppose $\mathscr{T} = \mathscr{P}$, then the subject is practically certain about the occurrence of every event and of its complement, which—as we shall see—is not a rational belief. The conditions we impose on the set $\mathscr{T}$ follow from three rationality criteria, described in full detail in Ref. [9]. In the next three sections, we discuss these rationality criteria and the resulting requirements on $\mathscr{T}$.

3.2 Deductive closure

That we are working with a linear utility scale for rewards has certain consequences. If the gambles f and g are acceptable, then so should be $f + g$, and λf, with $\lambda \in \mathbb{R}_{>0}$. These two observations are summarised in the *deductive extension* $\mathrm{ext}_{\mathbf{D}}$:

$$\mathrm{ext}_{\mathbf{D}}\, \mathscr{B} := \langle \mathrm{posi}\, \mathscr{B}_{\succeq}; \mathscr{B}_{\prec} \rangle,$$

and we call an assessment $\mathscr{D}$ deductively closed if $\mathrm{ext}_{\mathbf{D}}\, \mathscr{D} = \mathscr{D}$. This leads us to the first rationality criterion: assessments should be deductively closed.

Proposition 1. *The positive linear hull of* $\mathscr{B}_{\succeq}$ *is given by* $\mathrm{posi}\, \mathscr{B}_{\succeq} = \mathscr{L}_{\geq 0} \cup \mathscr{L}^{\gtrdot}_{\mathscr{T}}$, *where we use the notations* $\mathscr{L}^{\gtrdot}_{\mathscr{T}} := \{f \in \mathscr{L} \colon (\exists B \in \mathscr{C}_{\mathscr{T}}) \inf(f|B) > 0\}$,[2] $\inf(f|B) := \inf\{f(x) \colon x \in B\}$ *and*

$$\mathscr{C}_{\mathscr{T}} := \left\{ \bigcap_{k=1}^{n} A_k \colon n \in \mathbb{N}, A_k \in \mathscr{T} \right\} \quad (3)$$

is the collection of all finite intersections of elements of $\mathscr{T}$. *Note that* $\mathrm{posi}\, \mathscr{B}_{\succeq} \neq \mathscr{L}$ *if and only if* $\emptyset \notin \mathscr{C}_{\mathscr{T}}$, *meaning that* $\mathscr{T}$ *has the intuitively appealing* finite intersection property.

Proof. We infer from Eq. (1) that $\mathrm{posi}(\mathscr{A}_{\rhd} \cup \mathscr{L}_{\geq 0}) = \mathrm{posi}\, \mathscr{A}_{\rhd} \cup \mathscr{L}_{\geq 0} \cup (\mathrm{posi}\, \mathscr{A}_{\rhd} + \mathscr{L}_{\geq 0})$. Since $0 \in \mathscr{L}_{\geq 0}$, we see that $\mathrm{posi}\, \mathscr{A}_{\rhd} + \mathscr{L}_{\geq 0} \supseteq \mathrm{posi}\, \mathscr{A}_{\rhd}$, and therefore $\mathrm{posi}(\mathscr{A}_{\rhd} \cup \mathscr{L}_{\geq 0}) = \mathscr{L}_{\geq 0} \cup (\mathrm{posi}\, \mathscr{A}_{\rhd} + \mathscr{L}_{\geq 0})$. A gamble f belongs to $\mathrm{posi}\, \mathscr{A}_{\rhd} + \mathscr{L}_{\geq 0}$ if and only if there are $n \in \mathbb{N}$, $\lambda_1, \dots, \lambda_n \in \mathbb{R}_{>0}$, $A_1, \dots, A_n \in \mathscr{T}$ and $\varepsilon_1, \dots, \varepsilon_n \in \mathbb{R}_{>0}$ such that

$$f \geq \sum_{k=1}^{n} \lambda_k (\varepsilon_k - \mathbb{I}_{A_k^c}).$$

By an appropriate choice of the $\lambda_k > 0$ and the $\varepsilon_k \in (0,1)$, the lower bound in the inequality above can be made arbitrarily low (negative) provided that $\bigcap_{k=1}^{n} A_k = \emptyset$, and only then. This shows that $\emptyset \in \mathscr{C}_{\mathscr{T}} \Leftrightarrow \mathrm{posi}\, \mathscr{A}_{\rhd} + \mathscr{L}_{\geq 0} = \mathscr{L}$. So let us assume that $\emptyset \notin \mathscr{C}_{\mathscr{T}}$.

Consider any gamble f in $\mathrm{posi}\, \mathscr{A}_{\rhd} + \mathscr{L}_{\geq 0}$, then there are $n \in \mathbb{N}$, $\lambda_1, \dots, \lambda_n \in \mathbb{R}_{>0}$, $A_1, \dots, A_n \in \mathscr{T}$ and $\varepsilon_1, \dots, \varepsilon_n \in \mathbb{R}_{>0}$ such that $f \geq \sum_{k=1}^{n} \lambda_k (\varepsilon_k - \mathbb{I}_{A_k^c})$, and therefore $\inf(f | \bigcap_{k=1}^{n} A_k) \geq \sum_{k=1}^{n} \lambda_k \varepsilon_k > 0$.

Conversely, if $\inf(f | \bigcap_{k=1}^{n} A_k) =: \delta > 0$ for some $n \in \mathbb{N}$ and $A_1, \dots, A_n \in \mathscr{T}$, then let all $\lambda_k > \lambda := \delta - \inf f \geq 0$ and all $\varepsilon_k := \frac{\delta}{n\lambda_k} > 0$, so

$$\begin{aligned} \sum_{k=1}^{n} \lambda_k (\varepsilon_k - \mathbb{I}_{A_k^c}) &\leq \mathbb{I}_{\bigcap_{k=1}^{n} A_k} \delta + \mathbb{I}_{\bigcup_{k=1}^{n} A_k^c} (\delta - \lambda) \\ &= \mathbb{I}_{\bigcap_{k=1}^{n} A_k} \delta + \mathbb{I}_{\bigcup_{k=1}^{n} A_k^c} \inf f \leq f, \end{aligned}$$

meaning that $f \in \mathrm{posi}\, \mathscr{A}_{\rhd} + \mathscr{L}_{\geq 0}$. □

For notational convenience, we define $\mathscr{L}^{\lessdot}_{\mathscr{T}} := -\mathscr{L}^{\gtrdot}_{\mathscr{T}}$.

The set $\mathscr{C}_{\mathscr{T}}$, as defined in Eq. (3), satisfies all the requirements for a filter base. It is called the *filter base generated by* the set $\mathscr{T}$.

The deductively closed $\mathrm{ext}_{\mathbf{D}}\, \mathscr{B}$ is not yet "perfect enough": for it to be a so-called *model*, we need to further impose the criteria of No Confusion and No Limbo.

3.3 No Confusion

Given the interpretation attached to an accept and to a reject statement, there should be no gambles in the set $(\mathrm{ext}_{\mathbf{D}}\, \mathscr{B})_{\between} := (\mathrm{ext}_{\mathbf{D}}\, \mathscr{B})_{\succeq} \cap (\mathrm{ext}_{\mathbf{D}}\, \mathscr{B})_{\prec}$: a gamble cannot be accepted and rejected at the same time. This observation leads us to the second rationality criterion: deductively closed assessments need to have

$$\text{No Confusion:} \quad (\mathrm{ext}_{\mathbf{D}}\, \mathscr{B})_{\between} = \emptyset.$$

The following proposition gives the conditions to be imposed on $\mathscr{T}$ in order to have No Confusion.

[2] We let $\inf(f|\emptyset)$ be $+\infty$ everywhere.

Proposition 2. *The deductively closed assessment* $\operatorname{ext}_{\mathbf{D}}\mathscr{B}$ *has No Confusion if and only if* $\mathscr{T}$ *satisfies the* finite intersection property*:* $\bigcap_{k=1}^{n} A_k \neq \emptyset$ *for all* $n \in \mathbb{N}$ *and all* $A_1,\dots,A_n \in \mathscr{T}$*, or equivalently,* $\emptyset \notin \mathscr{C}_{\mathscr{T}}$*.*

Proof. $\operatorname{ext}_{\mathbf{D}}\mathscr{B}$ has No Confusion if and only if $\mathscr{L}_{\mathscr{T}}^{\gtrdot} \cap -\mathscr{A}_{\rhd} = \emptyset$, $\mathscr{L}_{\geq 0} \cap -\mathscr{A}_{\rhd} = \emptyset$, $\mathscr{L}_{\mathscr{T}}^{\gtrdot} \cap \mathscr{L}_{<0} = \emptyset$ and $\mathscr{L}_{\geq 0} \cap \mathscr{L}_{<0} = \emptyset$. The last intersection is obviously empty, and the condition for the third one to be empty is clearly that $\emptyset \notin \mathscr{C}_{\mathscr{T}}$, taking into account Prop. 1.

The second intersection $\mathscr{L}_{\geq 0} \cap -\mathscr{A}_{\rhd}$ is empty if and only if $\mathbb{I}_{A^c} - \varepsilon \not\geq 0$ for all events A in $\mathscr{T}$ and all $\varepsilon \in \mathbb{R}_{>0}$, which is equivalent with $\emptyset \notin \mathscr{T}$.

The first intersection is non-empty if and only if there are $A \in \mathscr{T}$ and $B \in \mathscr{C}_{\mathscr{T}}$ such that $\inf(\mathbb{I}_{A^c} - \varepsilon | B) > 0$ for some $\varepsilon \in \mathbb{R}_{>0}$, or equivalently, such that $B \cap A = \emptyset$. This tells us that the first intersection is empty if and only if

$$(\forall A \in \mathscr{T})(\forall B \in \mathscr{C}_{\mathscr{T}}) B \cap A \neq \emptyset,$$

which is equivalent with $\emptyset \notin \mathscr{C}_{\mathscr{T}}$. □

Because of its form, $\mathscr{C}_{\mathscr{T}}$ is a filter base. Moreover, No Confusion is equivalent to $\mathscr{C}_{\mathscr{T}}$ being a *proper* filter base: in addition to $\mathscr{C}_{\mathscr{T}}$ being closed under finite intersections, it cannot contain the empty set. From now on, we consider only proper filter bases $\mathscr{C}_{\mathscr{T}}$, or equivalently, sets $\mathscr{T}$ that satisfy the finite intersection property.

3.4 No Limbo

For $\operatorname{ext}_{\mathbf{D}}\mathscr{B}$ to be a *model*, besides being deductively closed and having No Confusion, it also needs to satisfy a third and last rationality criterion: it must have No Limbo.

To see what this means, consider any deductively closed assessment $\mathscr{D} = \langle \mathscr{D}_{\succeq}; \mathscr{D}_{\prec} \rangle$ with No Confusion. At this point, all the gambles in $\mathscr{D}_{\smile} = \mathscr{L} \setminus (\mathscr{D}_{\succeq} \cup \mathscr{D}_{\prec})$ are unresolved, and can therefore in principle still be accepted or rejected. But it is proved in Ref. [9, Corollary 6] that the gambles in the so-called *limbo*

$$\big(\overline{\mathscr{D}_{\prec}} - (\mathscr{D}_{\succeq} \cup \{0\})\big) \setminus \mathscr{D}_{\prec}, \tag{4}$$

which is a subset of $\mathscr{D}_{\smile}$, cannot be made acceptable without creating Confusion. In other words, these are the unresolved gambles that have exactly the same effect as gambles in $\mathscr{D}_{\prec}$: if we considered them as acceptable too, the resulting assessment would have Confusion. So they are still in an unresolved state, but if we want to avoid Confusion, there is nothing for it: we must also reject them.

Starting from the deductively closed assessment $\mathscr{D}$ with No Confusion, additionally rejecting the gambles that are in its limbo results in its *reckoning extension* $\operatorname{ext}_{\mathbf{M}}$:

$$\operatorname{ext}_{\mathbf{M}}\mathscr{D} := \big\langle \mathscr{D}_{\succeq}; \overline{\mathscr{D}_{\prec}} \cup \big(\overline{\mathscr{D}_{\prec}} - \mathscr{D}_{\succeq}\big) \big\rangle, \tag{5}$$

and we say that a deductively closed assessment $\mathscr{D}$ without Confusion has No Limbo if and only if $\operatorname{ext}_{\mathbf{M}}\mathscr{D} = \mathscr{D}$, or equivalently, if and only if the set in Eq. (4) is empty.

We end up with $\mathscr{M} := \operatorname{ext}_{\mathbf{M}}\operatorname{ext}_{\mathbf{D}}\mathscr{B}$, a model that is deductively closed and has No Limbo and No Confusion; see Ref. [9, Prop. 7] for details. We call it a *coherent model.* The next proposition characterises $\mathscr{M}$, where the notation emphasises the set of favourable gambles.

Proposition 3. *The coherent model* $\mathscr{M} = \operatorname{ext}_{\mathbf{M}}\operatorname{ext}_{\mathbf{D}}\mathscr{B}$ *is given by* $\mathscr{M} = \langle \mathscr{M}_{\rhd} \cup \{0\}; -\mathscr{M}_{\rhd} \rangle$*, with*

$$\mathscr{M}_{\rhd} := \mathscr{L}_{\mathscr{T}}^{\gtrdot} \cup \mathscr{L}_{>0}.$$

Proof. The proof for the set of acceptable gambles $\mathscr{M}_{\succeq}$ follows from Prop. 1, $(\operatorname{ext}_{\mathbf{M}}\mathscr{D})_{\succeq} = \mathscr{D}_{\succeq}$ and $\mathscr{L}_{\geq 0} = \mathscr{L}_{>0} \cup \{0\}$. Taking into account Eq. (5) and Prop. 1, the set of rejected gambles is given by $\mathscr{M}_{\prec} = \overline{\mathscr{B}_{\prec}} \cup \big(\overline{\mathscr{B}_{\prec}} - (\mathscr{L}_{\geq 0} \cup \mathscr{L}_{\mathscr{T}}^{\gtrdot})\big) = \overline{\mathscr{B}_{\prec}} - (\mathscr{L}_{\geq 0} \cup \mathscr{L}_{\mathscr{T}}^{\gtrdot})$, where we used the fact that $0 \in \mathscr{L}_{\geq 0}$. Because $\overline{\mathscr{A}_{\prec} \cup \mathscr{L}_{<0}} = \overline{\mathscr{A}_{\prec}} \cup \mathscr{L}_{<0}$, it follows that

$$\begin{aligned} \mathscr{M}_{\prec} &= (\overline{\mathscr{A}_{\prec}} \cup \mathscr{L}_{<0}) - (\mathscr{L}_{\geq 0} \cup \mathscr{L}_{\mathscr{T}}^{\gtrdot}) \\ &= \big(\overline{\mathscr{A}_{\prec}} - \mathscr{L}_{\geq 0}\big) \cup \big(\overline{\mathscr{A}_{\prec}} - \mathscr{L}_{\mathscr{T}}^{\gtrdot}\big) \\ &\qquad \cup (\mathscr{L}_{<0} - \mathscr{L}_{\geq 0}) \cup \big(\mathscr{L}_{<0} - \mathscr{L}_{\mathscr{T}}^{\gtrdot}\big). \end{aligned} \tag{6}$$

We first prove that $\mathscr{L}_{<0} \cup \mathscr{L}_{\mathscr{T}}^{\lessdot} \subseteq \mathscr{M}_{\prec}$. Observe that $\mathscr{L}_{<0} \subseteq \overline{\mathscr{A}_{\prec}} \cup \mathscr{L}_{<0} \subseteq (\overline{\mathscr{A}_{\prec}} \cup \mathscr{L}_{<0}) - (\mathscr{L}_{\geq 0} \cup \mathscr{L}_{\mathscr{T}}^{\gtrdot}) = \mathscr{M}_{\prec}$, where the last inclusion holds because $0 \in \mathscr{L}_{\geq 0}$. To show that also $-\mathscr{L}_{\mathscr{T}}^{\gtrdot} \subseteq \mathscr{M}_{\prec}$, use the next Lem. 1 to see that $-\mathscr{L}_{\mathscr{T}}^{\gtrdot} = (-\mathscr{L}_{\mathscr{T}}^{\gtrdot}) + \mathscr{L}_{<0}$, and by Eq. (6), this is a subset of $\mathscr{M}_{\prec}$.

Next, we prove that $\mathscr{L}_{<0} \cup \mathscr{L}_{\mathscr{T}}^{\lessdot} \supseteq \mathscr{M}_{\prec}$. We prove that each of the four terms of the union of Eq. (6) is a subset of $\mathscr{L}_{<0} \cup \mathscr{L}_{\mathscr{T}}^{\lessdot}$. To do so, it is useful to remark that

$$\mathscr{L}_{\mathscr{T}}^{\lessdot} = \operatorname{posi}(\mathscr{A}_{\prec} + \mathscr{L}_{\leq 0}) \supseteq \operatorname{posi}\mathscr{A}_{\prec} \supseteq \overline{\mathscr{A}_{\prec}}. \tag{7}$$

For $\overline{\mathscr{A}_{\prec}} - \mathscr{L}_{\geq 0}$, use Eq. (7) to infer that $\overline{\mathscr{A}_{\prec}} \subseteq -\mathscr{L}_{\mathscr{T}}^{\gtrdot}$, so $\overline{\mathscr{A}_{\prec}} - \mathscr{L}_{\geq 0} \subseteq -\mathscr{L}_{\mathscr{T}}^{\gtrdot} - \mathscr{L}_{\geq 0} = -\mathscr{L}_{\mathscr{T}}^{\gtrdot}$, where the equality follows from Lem. 1. For $\overline{\mathscr{A}_{\prec}} - \mathscr{L}_{\mathscr{T}}^{\gtrdot}$, use Eq. (7) to obtain $\overline{\mathscr{A}_{\prec}} - \mathscr{L}_{\mathscr{T}}^{\gtrdot} \subseteq -\mathscr{L}_{\mathscr{T}}^{\gtrdot} - \mathscr{L}_{\mathscr{T}}^{\gtrdot} = -\mathscr{L}_{\mathscr{T}}^{\gtrdot}$, where the equality follows from the fact that $-\mathscr{L}_{\mathscr{T}}^{\gtrdot}$ is a cone. Since $\mathscr{L}_{<0} - \mathscr{L}_{\geq 0} = \mathscr{L}_{<0}$, it only remains to consider the last term: use Lem. 1 to find that $\mathscr{L}_{<0} - \mathscr{L}_{\mathscr{T}}^{\gtrdot} = -\mathscr{L}_{\mathscr{T}}^{\gtrdot}$. □

Lemma 1. *For any collection of events* $\mathscr{T} \subseteq \mathscr{P}$ *that satisfies the finite intersection property,* $\mathscr{L}_{\mathscr{T}}^{\gtrdot} = \mathscr{L}_{\mathscr{T}}^{\gtrdot} + \mathscr{L}_{>0} = \mathscr{L}_{\mathscr{T}}^{\gtrdot} + \mathscr{L}_{\geq 0}$.

Proof. Since $0 \in \mathscr{L}_{\geq 0}$, we have $\mathscr{L}_{\mathscr{T}}^{\gtrdot} \subseteq \mathscr{L}_{\mathscr{T}}^{\gtrdot} + \mathscr{L}_{\geq 0}$, and since $\mathscr{L}_{>0} \subseteq \mathscr{L}_{\geq 0}$, also $\mathscr{L}_{\mathscr{T}}^{\gtrdot} + \mathscr{L}_{>0} \subseteq \mathscr{L}_{\mathscr{T}}^{\gtrdot} + \mathscr{L}_{\geq 0}$. The proof is complete if we can prove that $\mathscr{L}_{\mathscr{T}}^{\gtrdot} + \mathscr{L}_{\geq 0}$ is also included in both $\mathscr{L}_{\mathscr{T}}^{\gtrdot}$ and $\mathscr{L}_{\mathscr{T}}^{\gtrdot} + \mathscr{L}_{>0}$. Consider any gamble $f \in \mathscr{L}_{\mathscr{T}}^{\gtrdot} + \mathscr{L}_{\geq 0}$, so there are $g \in \mathscr{L}$ and $B \in \mathscr{C}_{\mathscr{T}}$ such that $\delta := \inf(g|B) > 0$ and $f \geq g$. This means that also $\inf(f|B) > 0$, and therefore $f \in \mathscr{L}_{\mathscr{T}}^{\gtrdot}$. Also, consider the gamble $h := \delta/2\,\mathbb{I}_B + g\mathbb{I}_{B^c} < f$. Because $\inf(h|B) = \delta/2 > 0$, it follows that $h \in \mathscr{L}_{\mathscr{T}}^{\gtrdot}$ and therefore $f = h + (f - h) \in \mathscr{L}_{\mathscr{T}}^{\gtrdot} + \mathscr{L}_{>0}$. □

To summarise, we started out with the assessment $\mathscr{A}$ of a subject who is practically certain of the occurrence of all events in $\mathscr{T}$, and added the background model $\mathscr{S}$, leading to a larger assessment $\mathscr{B} = \mathscr{A} \cup \mathscr{S}$. Using deductive and reckoning extension, and by imposing restrictions on $\mathscr{T}$, namely that $\mathscr{T}$ has the finite intersection property, we added acceptable as well as rejected gambles to end up with the coherent model $\mathscr{M} = \mathrm{ext}_{\mathbf{M}}\,\mathrm{ext}_{\mathbf{D}}\,\mathscr{B}$. Prop. 3 guarantees that, as was the case for the initial assessment of Eq. (2), the coherent model $\mathscr{M}$ is fully determined by the set of favourable gambles

$$\mathscr{M}_{\triangleright} = \{f \in \mathscr{L} \colon (\exists B \in \mathscr{C}_{\mathscr{T}}) \inf(f|B) > 0\} \cup \mathscr{L}_{>0},$$

leaving aside the always indifferent zero gamble. Because $\mathscr{M}$ is a coherent model, we call this set $\mathscr{M}_{\triangleright}$ a *coherent set of favourable gambles*. In this model $\mathscr{M}$, 0 is the only indifferent gamble: $\mathscr{M}_{\simeq} = \mathscr{M}_{\succeq} \cap -\mathscr{M}_{\succeq} = \{0\}$, and $\mathscr{M}$ is an instance of a favour-indifference model, because $-\mathscr{M}_{\prec} \subseteq \mathscr{M}_{\succeq}$ and $\mathscr{M}_{\succeq} = \mathscr{M}_{\triangleright} \cup \mathscr{M}_{\simeq}$.

3.5 Finding all practically certain events

We now ask ourselves whether the inference procedure described above, which allowed us to infer from the set of favourable gambles $\mathscr{A}_{\triangleright}$ the larger set of favourable gambles $\mathscr{M}_{\triangleright}$, bears any relationship to inference in classical propositional logic? Which are the other events, besides the ones in $\mathscr{T}$, that the inference procedure tells us our subject, if he is rational, should also be practically certain of?

As we have suggested above, a subject who is certain about an event A expresses this as finding favourable the gambles of the form $-\mathbb{I}_{A^c} + \varepsilon$, with $\varepsilon \in \mathbb{R}_{>0}$. We denote the corresponding set of favourable gambles by $\mathscr{A}_{\triangleright}^{A} := \{-\mathbb{I}_{A^c} + \varepsilon \colon \varepsilon \in \mathbb{R}_{>0}\}$. The question therefore becomes: *for which events A is the set $\mathscr{A}_{\triangleright}^{A}$ a subset of $\mathscr{M}_{\triangleright}$?* As the gambles in $\mathscr{A}_{\triangleright}^{A}$ are, for ε small enough, positive only on A, the answer to this question is immediate:

$$\mathscr{A}_{\triangleright}^{A} \subseteq \mathscr{M}_{\triangleright} \Leftrightarrow (\exists B \in \mathscr{C}_{\mathscr{T}}) B \subseteq A.$$

This tells us that the subject should be practically certain of all events in the filter generated by $\mathscr{T}$:

$$\mathscr{F}_{\mathscr{T}} := \{A \in \mathscr{P} \colon (\exists B \in \mathscr{C}_{\mathscr{T}})(B \subseteq A)\}.$$

This is a proper filter provided that $\emptyset \notin \mathscr{C}_{\mathscr{T}}$. Also observe that $\mathscr{L}_{\mathscr{T}}^{\gtrdot} = \{f \in \mathscr{L} \colon (\exists B \in \mathscr{F}_{\mathscr{T}}) \inf(f|B) > 0\}$.

Any filter is a set-theoretic counterpart of a collection of propositions that is deductively closed (closed under conjunction and modus ponens), and the generated filter corresponds to the deductive closure of a set of propositions, in classical propositional logic. We see that on our specific interpretation of it—or semantics for it—the logic of practical certainty has the same basic machinery as classical propositional logic. In simple terms: if someone is practically certain that both the events A and B occur, it is reasonable to be practically certain of $A \cap B$; and if someone is practically certain that the event A occurs, then it is reasonable to be practically certain of every event $B \supseteq A$.

Our argument goes further than that, because it also allows us to infer which gambles a subject should find favourable if he is practically certain that all events in $\mathscr{T}$ occur: all gambles in $\mathscr{M}_{\triangleright}$, which are the gambles that are strictly positive, or that have a strictly positive return, bounded away from zero, on some practically certain event.

4 Modelling practical certainty using acceptability

When a subject is practically certain that an event A occurs, we have taken this to mean, in Section 3, that he finds favourable every gamble of the form $-\mathbb{I}_{A^c} + \varepsilon$, with $\varepsilon \in \mathbb{R}_{>0}$. Here, we repeat the same reasoning with a weaker assessment of acceptability, rather than favourability: if a subject is practically certain that an event A occurs, we now take this to mean that he finds every gamble of the form $-\mathbb{I}_{A^c} + \varepsilon$, with $\varepsilon \in \mathbb{R}_{>0}$, acceptable. With $\mathscr{T}$ the collection of events he is practically certain of, his assessment is therefore:

$$\mathscr{A}^{-} := \langle \{-\mathbb{I}_{A^c} + \varepsilon : A \in \mathscr{T}, \varepsilon \in \mathbb{R}_{>0}\}; \emptyset \rangle.$$

We make the same a priori assumptions summarised in the background model $\mathscr{S} = \langle \mathscr{L}_{\geq 0}; \mathscr{L}_{<0} \rangle$, leading to the smallest background-including assessment $\mathscr{B}^{-} = \mathscr{A}^{-} \cup \mathscr{S}$.

In the next proposition, we determine the relationship between $\mathscr{M}^{-} = \mathrm{ext}_{\mathbf{M}}\,\mathrm{ext}_{\mathbf{D}}\,\mathscr{B}^{-}$ and $\mathscr{M}$, and show that the (apparently) weaker acceptability assessment leads to the same conclusions.[3]

Proposition 4. *Using deductive extension we obtain* $\mathrm{ext}_{\mathbf{D}}\,\mathscr{B}^{-} = \left\langle \mathscr{L}_{\geq 0} \cup \mathscr{L}_{\mathscr{T}}^{\gtrdot}; \mathscr{L}_{<0} \right\rangle$. *This deductively closed assessment has No Confusion if and only if $\mathscr{T}$ satisfies the finite intersection property. The corresponding coherent model is* $\mathscr{M}^{-} = \mathrm{ext}_{\mathbf{M}}\,\mathrm{ext}_{\mathbf{D}}\,\mathscr{B}^{-} = \mathscr{M}$.

Proof. The argument is analogous to, but less involved than, that in the proofs of Props. 1–3. □

5 An alternative way of modelling practical certainty using indifference

Williams [15] and Walley [12] define a lower prevision $\underline{p}$ for a gamble f as a supremum acceptable buying price:

[3] The equivalence between the implications of favourability and acceptability assessments does not hold in more general cases. As an example, consider the background model $\langle \mathscr{L}_{\geq 0}; \emptyset \rangle$. Then the conclusions from every non-empty favourability assessment differs from the corresponding acceptability assessment.

the highest price $\underline{p}$ such that $f-\underline{p}+\varepsilon$ is acceptable—or equivalently as it turns out, favourable—for all $\varepsilon>0$. In the previous sections, we have used an approach with a very similar flavour to account for practical certainty: the supremum acceptable buying price for (the indicator of) a practically certain event is 1.

The approach that Bruno de Finetti [5, 7] follows in defining the (precise) prevision p for a gamble f, is rather different:[4] it is the unique number p such that the subject is indifferent between the uncertain f and the fixed p, or equivalently, between $f-p$ and 0.

We therefore also propose an alternative way of modelling practical certainty, more along the lines of de Finetti's approach to previsions: we model a subject's practical certainty of the occurrence of an event A by an assessment of indifference between $\mathbb{I}_A$ and 1, or equivalently, between $\mathbb{I}_{A^c}$ and 0. This amounts to a statement of acceptability for both $\mathbb{I}_{A^c}$ and its negation $-\mathbb{I}_{A^c}$. But, since $\mathbb{I}_{A^c}\geq 0$, and since we will use $\mathscr{L}_{\geq 0}$ as a background model for acceptability, meaning that all non-negative gambles are a priori assumed to be acceptable (see further on), we need only explicitly state the acceptability of $-\mathbb{I}_{A^c}$. This assessment is stronger than the corresponding one in the previous sections: here the subject actually accepts the gamble $-\mathbb{I}_{A^c}$, whereas before he only accepted gambles of the form $-\mathbb{I}_{A^c}+\varepsilon$, with $\varepsilon\in\mathbb{R}_{>0}$.

If our subject is practically certain of every event in the collection $\mathscr{T}\subseteq\mathscr{P}$, this leads to the (indifference) assessment:

$$\mathscr{A}':=\langle\{-\mathbb{I}_{A^c}\colon A\in\mathscr{T}\};\emptyset\rangle\,.$$

Before, we used the background model $\mathscr{S}=\langle\mathscr{L}_{\geq 0};\mathscr{L}_{<0}\rangle$. The nature of an indifference assessment no longer allows us to use $\mathscr{S}$ as background model, as this would lead to difficulties: since $-\mathbb{I}_{A^c}\in\mathscr{L}_{<0}$ if $A^c\neq\emptyset$, in order to avoid No Confusion, the set $\mathscr{T}$ can only contain the trivial certain event $\mathscr{X}$.[5] For this reason, we propose a slightly more conservative background model:

$$\mathscr{S}'=\langle\mathscr{L}_{\geq 0};\mathscr{L}_{\lhd 0}\rangle\,,$$

where we take for granted that all non-negative gambles should be accepted, and all gambles that are point-wise (strictly) negative should be rejected:

$$\mathscr{L}_{\lhd 0}:=\{f\in\mathscr{L}\colon(\forall x\in\mathscr{X})f(x)<0\}\,.$$

We use $\mathscr{B}':=\mathscr{A}'\cup\mathscr{S}'=\big\langle\mathscr{A}'_{\succeq}\cup\mathscr{L}_{\geq 0};\mathscr{L}_{\lhd 0}\big\rangle$ to denote the smallest assessment that includes both the subject's indifference assessment $\mathscr{A}'$ and the background model $\mathscr{S}'$.

In this section, due to page limitations, and because the reasoning uses similar arguments to the ones in Section 3, we will omit the proofs.

As before, in order to obtain a coherent model, we have to impose rationality conditions on the set $\mathscr{T}$ of practically certain events, which we explore next.

5.1 Deductive closure

The first rationality criterion states that we have to accept every gamble that can be deduced from $\mathscr{B}'_{\succeq}$: the deductive closure is $\operatorname{ext}_{\mathbf{D}}\mathscr{B}'=\big\langle\operatorname{posi}\mathscr{B}'_{\succeq};\mathscr{B}'_{\prec}\big\rangle$.

Proposition 5. *The positive linear hull of $\mathscr{B}'_{\succeq}$ is*

$$\operatorname{posi}\mathscr{B}'_{\succeq}=\mathscr{L}^{\geqq}_{\mathscr{T}}:=\{f\in\mathscr{L}\colon(\exists B\in\mathscr{C}_{\mathscr{T}})\mathbb{I}_Bf\geq 0\}\,,$$

with $\mathscr{C}_{\mathscr{T}}$ defined as in Eq. (3). *Note that* $\operatorname{posi}\mathscr{B}'_{\succeq}\neq\mathscr{L}$ *if and only if $\emptyset\notin\mathscr{C}_{\mathscr{T}}$.*

Compared with $\operatorname{posi}\mathscr{B}_{\succeq}$, $\operatorname{posi}\mathscr{B}'_{\succeq}$ contains more gambles: those gambles f that are non-negative on an event B in $\mathscr{C}_{\mathscr{T}}$, but for which $\inf(f|B)$ is zero.

5.2 No Confusion

The second rationality criterion requires that the deductively closed assessment $\operatorname{ext}_{\mathbf{D}}\mathscr{B}'$ should have No Confusion. This leads to the same condition on $\mathscr{T}$ as before in Section 3:

Proposition 6. *The deductively closed assessment* $\operatorname{ext}_{\mathbf{D}}\mathscr{B}'$ *has No Confusion if and only if $\mathscr{T}$ satisfies the finite intersection property, or equivalently, if $\emptyset\notin\mathscr{C}_{\mathscr{T}}$.*

As in Section 3, the second rationality criterion turns $\mathscr{C}_{\mathscr{T}}$ into a proper filter base. From now on, we will assume $\mathscr{C}_{\mathscr{T}}$ to be proper.

5.3 No Limbo

The final rationality criterion of No Limbo leads us to apply the reckoning extension $\operatorname{ext}_{\mathbf{M}}$ to the deductive extension $\operatorname{ext}_{\mathbf{D}}\mathscr{B}'$ with No Confusion, leading to a coherent model $\mathscr{M}':=\operatorname{ext}_{\mathbf{M}}\operatorname{ext}_{\mathbf{D}}\mathscr{B}'$.

Proposition 7. *The coherent model $\mathscr{M}'$ is given by $\mathscr{M}'=\big\langle\mathscr{M}'_{\succeq};\mathscr{M}'_{\prec}\big\rangle$, with $\mathscr{M}'_{\succeq}=\mathscr{L}^{\geqq}_{\mathscr{T}}$ and $\mathscr{M}'_{\prec}=-\mathscr{L}^{\trianglerighteq}_{\mathscr{T}}=\mathscr{L}^{\trianglelefteq}_{\mathscr{T}}:=$ $\{f\in\mathscr{L}\colon(\exists B\in\mathscr{C}_{\mathscr{T}})(\forall x\in B)f(x)<0\}$.*

The corresponding set of favourable gambles $\mathscr{M}'_{\rhd}$ is:

$$\begin{aligned}\mathscr{M}'_{\rhd}&=\mathscr{M}'_{\succeq}\cap-\mathscr{M}'_{\prec}=\mathscr{L}^{\geqq}_{\mathscr{T}}\cap\mathscr{L}^{\trianglerighteq}_{\mathscr{T}}=\mathscr{L}^{\trianglerighteq}_{\mathscr{T}}\\&=\{f\in\mathscr{L}\colon(\exists B\in\mathscr{C}_{\mathscr{T}})(\forall x\in B)f(x)>0\}\,.\end{aligned}$$

5.4 Finding all practically certain events

As in Section 3.5, we ask ourselves whether, in addition to the events in $\mathscr{T}$, the criteria of rationality allow the subject to infer the practical certainty of more events. Since, here, we are modelling practical certainty via indifference, we

[4] For an extensive discussion of the difference between the two approaches, we refer to Refs. [9] and [11].

[5] See also the discussion in Ref. [9, Section 5].

look at the indifferent gambles $\mathscr{M}'_{\simeq}$ in the coherent model $\mathscr{M}'$, and our subject is practically certain about an event A precisely when he is indifferent about $\mathbb{I}_{A^c}$, meaning that $-\mathbb{I}_{A^c}$ (in addition to $\mathbb{I}_{A_c}$) belongs to his inferred set of indifferent gambles $\mathscr{M}'_{\simeq} = \mathscr{M}'_{\succeq} \cap -\mathscr{M}'_{\succeq}$.

So let us look for an expression for $\mathscr{M}'_{\simeq}$. This set contains the gambles for which there are B and B' in $\mathscr{C}_{\mathscr{T}}$ such that both $\mathbb{I}_B f \geq 0$ and $\mathbb{I}_{B'} f \leq 0$. Since $\mathscr{C}_{\mathscr{T}}$ is closed under finite intersections, we find that

$$\begin{aligned}\mathscr{M}'_{\simeq} &= \{f \in \mathscr{L} \colon (\exists B \in \mathscr{C}_{\mathscr{T}}) \mathbb{I}_B f = 0\} \\ &= \{f \in \mathscr{L} \colon (\exists B \in \mathscr{F}_{\mathscr{T}}) \mathbb{I}_B f = 0\},\end{aligned}$$

and therefore also

$$-\mathbb{I}_{A^c} \in \mathscr{M}'_{\simeq} \Leftrightarrow (\exists B \in \mathscr{C}_{\mathscr{T}}) A^c \cap B = \emptyset \Leftrightarrow A \in \mathscr{F}_{\mathscr{T}}.$$

This tells us that the subject can be practically certain of all events A in the proper filter $\mathscr{F}_{\mathscr{T}}$ generated by $\mathscr{T}$, as in Section 3.5. Here too, our approach allows us to say even more: the subject should regard as favourable all gambles that are (strictly) positive on some practically certain event, and be indifferent about any gamble that is zero on some practically certain event.

6 Coherent lower prevision and coherent lower probability

6.1 Coherent lower prevision

With every set of favourable gambles we can associate a *lower prevision* $\underline{P}$ and an *upper prevision* $\overline{P}$. Lower previsions (or lower expectation functionals) $\underline{P}$ as wel as upper previsions (or upper expectation functionals) $\overline{P}$ are real-valued functionals defined on $\mathscr{L}$. Given any set of favourable gambles $\mathscr{D}$, then the corresponding lower prevision $\underline{P}$ and upper prevision $\overline{P}$ are defined by:

$$\begin{aligned}&\underline{P}(f) := \sup\{\mu \in \mathbb{R} \colon f - \mu \in \mathscr{D}\} \text{ and} \\ &\overline{P}(f) := \inf\{\mu \in \mathbb{R} \colon \mu - f \in \mathscr{D}\} \text{ for every } f \text{ in } \mathscr{L}.\end{aligned}$$

If the defining set of favourable gambles is coherent, then we call $\underline{P}$ and $\overline{P}$ coherent. Since $\underline{P}(f) = -\overline{P}(-f)$ for every $f \in \mathscr{L}$, lower and upper previsions contain the same information, and we focus on lower previsions.

Let us calculate the coherent lower prevision corresponding with $\mathscr{M}_{\triangleright}$. For any gamble f, $\underline{P}(f)$ is the supremum μ such that $f - \mu$ is an element of $\mathscr{M}_{\triangleright}$, or equivalently, it is the supremum μ such that

$$\mu < f \text{ or } (\exists B \in \mathscr{C}_{\mathscr{T}}) \mu < \inf(f|B).$$

This tells us that $\underline{P}(f)$ is the maximum of $\inf f$ and $\sup_{B \in \mathscr{C}_{\mathscr{T}}} \inf(f|B)$. Since the latter number is never smaller than the former, we conclude:[6]

$$\underline{P}(f) = \sup_{B \in \mathscr{C}_{\mathscr{T}}} \inf(f|B) = \sup_{B \in \mathscr{F}_{\mathscr{T}}} \inf(f|B).$$

To make explicit the proper filter of events $\mathscr{C}_{\mathscr{T}}$ we are using, we denote this lower prevision also as $\underline{P}_{\mathscr{F}_{\mathscr{T}}}$. Observe that $\underline{P}_{\mathscr{F}_{\mathscr{T}}}$ is coherent if and only if $\mathscr{C}_{\mathscr{T}}$ is a proper filter base.

Using a similar argument as above, it follows that the lower prevision $\underline{P}'$ corresponding with the set of favourable gambles $\mathscr{M}'_{\triangleright}$ is the supremum μ such that $(\exists B \in \mathscr{C}_{\mathscr{T}}) \inf(f|B) > \mu$, whence $\underline{P}'(f) = \sup_{B \in \mathscr{C}_{\mathscr{T}}} \inf(f|B) = \underline{P}(f)$ for every gamble $f \in \mathscr{L}$. This tells us that, regardless of whether we formulate practical certainty using favourability or indifference assessments, we end up with the same corresponding coherent lower prevision.

6.2 Coherent lower probability

With every lower prevision $\underline{P}$, we can associate a lower probability $\underline{Q}$. A lower probability $\underline{Q}$ is a real-valued set function defined on $\mathscr{P}$. Given a lower prevision $\underline{P}$, then the corresponding lower probability $\underline{Q}$ is defined by:

$$\underline{Q}(A) := \underline{P}(\mathbb{I}_A) \text{ for each event } A \text{ in } \mathscr{P}.$$

If the defining lower prevision is coherent, then the corresponding lower probability is called coherent as well.

We look at the lower probability $\underline{R}_{\mathscr{F}_{\mathscr{T}}}$ corresponding with the lower prevision $\underline{P}_{\mathscr{F}_{\mathscr{T}}}$. For any event A, the lower probability $\underline{R}_{\mathscr{F}_{\mathscr{T}}}(A)$ equals $\sup_{B \in \mathscr{C}_{\mathscr{T}}} \inf(\mathbb{I}_A|B)$. Since $\inf(\mathbb{I}_A|B)$ is 1 if $B \subseteq A$ and 0 otherwise, we have:

$$\underline{R}_{\mathscr{F}_{\mathscr{T}}}(A) = \begin{cases} 1 & \text{if } (\exists B \in \mathscr{C}_{\mathscr{T}}) B \subseteq A \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} 1 & \text{if } A \in \mathscr{F}_{\mathscr{T}} \\ 0 & \text{otherwise.} \end{cases}$$

This lower probability is coherent because $\underline{P}_{\mathscr{F}_{\mathscr{T}}}$ is. This tells us that the subject is willing to bet at all odds on the occurrence of every event $A \in \mathscr{F}_{\mathscr{T}}$. For all other events, he has no commitment whatsoever: he is only willing to bet on these other evens at zero odds. Compare this with the discussion in Sections 3.5 and 5.4.

Conversely, an event A for which the upper probability is zero—which means that the subject is willing to bet at all odds against the occurrence of A—reflects practical certainty that A does not occur. For an event A, the upper probability $\overline{R}_{\mathscr{F}_{\mathscr{T}}}(A)$ equals $\inf_{B \in \mathscr{C}_{\mathscr{T}}} \sup(\mathbb{I}_A|B)$, which is zero iff $A \in \mathscr{I} := \{B^c \colon B \in \mathscr{T}\}$. This ideal[7] of subsets $\mathscr{I}$ is the set of events the agent is practically certain will not occur.

[6] This lower prevision constitute a particular case of the so-called filter maps, see [3].

[7] The notion of an ideal is the dual notion of a filter: an ideal $\mathscr{I}$ is a subset of $\mathscr{P}$ that is *closed under finite unions* ($A \cup B \in \mathscr{I}$ when $A, B \in \mathscr{I}$) and *decreasing* (if $A \in \mathscr{I}$ and $B \subseteq A$, then also $B \in \mathscr{I}$).

7 Connection with strong belief structures

7.1 Strong belief structures

For this section, we will need some extra notation. We call $\mathbf{A}$ the collection of all the assessments—with or without Confusion: $\mathbf{A} = \{\langle \mathscr{D}_{\succeq};\mathscr{D}_{\prec}\rangle : \mathscr{D}_{\succeq},\mathscr{D}_{\prec} \subseteq \mathscr{L}\}$. Assessments in $\mathbf{A}$ can be partially ordered by set inclusion $\subseteq$: with two assessments $\mathscr{D}$ and $\mathscr{D}'$ in $\mathbf{A}$, we write $\mathscr{D} \subseteq \mathscr{D}'$ if and only if $\mathscr{D}_{\succeq} \subseteq \mathscr{D}'_{\succeq}$ and $\mathscr{D}_{\prec} \subseteq \mathscr{D}'_{\prec}$. The corresponding partially ordered set is denoted by $(\mathbf{A},\subseteq)$.

Not all assessments in $\mathbf{A}$ are of interest; we can restrict our attention to some generic subclass of models $\mathbb{M} \subseteq \mathbf{A}$. This $\mathbb{M}$ inherits the partial order $\subseteq$ from $\mathbf{A}$. We call $\hat{\mathbb{M}}$ the set of *maximal*, or undominated, models in $\mathbb{M}$: $\hat{\mathbb{M}} := \{\mathscr{D} \in \mathbb{M}\colon (\forall \mathscr{D}' \in \mathbb{M})(\mathscr{D} \subseteq \mathscr{D}' \Rightarrow \mathscr{D} = \mathscr{D}')\}$. In contradistinction with $\mathbf{A}$, where $\hat{\mathbf{A}} = \{\langle \mathscr{L};\mathscr{L}\rangle\}$ is its top (and unique maximal element), the family of models $\mathbb{M}$ may have no, one or multiple maximal elements.

We are interested in whether the structure $(\mathbf{A},\mathbb{M},\subseteq)$ is a *strong belief structure* [2], meaning that it satisfies the following four criteria:

S1. $(\mathbf{A},\subseteq)$ is a complete lattice: for any subset $\mathbf{B}$ of $\mathbf{A}$, its supremum $\sup\mathbf{B}$ and its infimum $\inf\mathbf{B}$ with respect to the order $\subseteq$ exist. Here the component-wise union operator $\bigcup$ plays the role of supremum and the component-wise intersection operator $\bigcap$ that of infimum.

S2. $(\mathbb{M},\subseteq)$ is a (component-wise) *intersection structure*, meaning that $\mathbb{M}$ is closed under arbitrary non-empty infima: for any non-empty subset $\mathbf{B}$ of $\mathbb{M}$, $\inf\mathbf{B} \in \mathbb{M}$.

S3. The partially ordered set $(\mathbb{M},\subseteq)$ has no top.

S4. The partially ordered set $(\mathbb{M},\subseteq)$ is *dually atomic*: $\hat{\mathbb{M}} \neq \emptyset$ and $\mathscr{D} = \inf\{\mathscr{D}' \in \hat{\mathbb{M}}\colon \mathscr{D} \subseteq \mathscr{D}'\}$ if $\mathscr{D} \in \mathbb{M}$.

A structure $(\mathbf{A},\mathbb{M},\subseteq)$ that satisfies requirements S1–S3 is called a *belief structure*. The relevance of the additional requirement S4 is that the maximal coherent models can be used to construct any coherent model. We want to investigate whether the coherent models encountered in Sections 3 and 5 constitute strong belief structures.

7.2 Favourability of acceptability assessments

We consider the family of models for practical certainty following from favourability or acceptability assessments, as described in Sections 3 and 4:

$$\mathbb{C} := \left\{\left\langle \mathscr{L}_{\mathscr{F}}^{\gtrless} \cup \mathscr{L}_{\geq 0};\mathscr{L}_{\mathscr{F}}^{\lessgtr} \cup \mathscr{L}_{<0}\right\rangle : \mathscr{F} \in \mathbb{F}\right\}.$$

For this family $\mathbb{C}$, it is not difficult to show by means of a counterexample that $(\mathbf{A},\mathbb{C},\subseteq)$ does not constitute a strong belief structure: it is not even a belief structure as it violates requirement S2.

7.3 Indifference assessments

We consider the family of models for practical certainty following from indifference assessments, as described in Section 5:

$$\mathbb{C}' := \left\{\left\langle \mathscr{L}_{\mathscr{F}}^{\gtrless};\mathscr{L}_{\mathscr{F}}^{\lhd}\right\rangle : \mathscr{F} \in \mathbb{F}\right\}.$$

The elements of $\mathbb{C}'$ are the coherent models identified in Prop. 7, and to make explicit which filter we are using, we denote them by $\mathscr{M}'(\mathscr{F}) = \left\langle \mathscr{M}'_{\succeq}(\mathscr{F});\mathscr{M}'_{\prec}(\mathscr{F})\right\rangle := \left\langle \mathscr{L}_{\mathscr{F}}^{\gtrless};\mathscr{L}_{\mathscr{F}}^{\lhd}\right\rangle$. In contrast with the structure considered in Section 7.2, $(\mathbf{A},\mathbb{C}',\subseteq)$ is a strong belief structure.

Proposition 8. *$(\mathbf{A},\mathbb{C}',\subseteq)$ is a strong belief structure.*

Proof. We have to prove that $(\mathbf{A},\mathbb{C}',\subseteq)$ fulfils the requirements S1–S4. S1 is fulfilled thanks to [9, Section 2.6]. S2 is fulfilled thanks to the next Lem. 2. For S3, consider Lem. 4 and take into account that the set of maximal elements of $\mathbb{F}$ is the set of ultrafilters $\mathbb{U}$, so $\mathbb{C}'$ has no top. S4 is fulfilled thanks to Lem. 4 and the *Ultrafilter Theorem* [10]. □

Lemma 2. *$(\mathbb{C}',\subseteq)$ is an intersection structure.*

Proof. Consider an arbitrary non-empty subset $\mathbf{B} \subseteq \mathbb{C}'$. We can describe $\mathbf{B}$ using a family of filters $\mathscr{F}_i$, $i \in I$ with a non-empty index set $I \neq \emptyset$: $\mathbf{B} = \{\mathscr{M}'(\mathscr{F}_i) : i \in I\}$. We now have to prove that $\inf\mathbf{B} \in \mathbb{C}'$, or equivalently, that $\bigcap_{i\in I}\mathscr{M}'(\mathscr{F}_i) \in \mathbb{C}'$, since taking infima corresponds to taking component-wise intersections. Consider any gamble f, then:

$$\begin{aligned} f \in \bigcap_{i\in I}\mathscr{M}'_{\succeq}(\mathscr{F}_i) &\Leftrightarrow (\forall i \in I)(\exists B_i \in \mathscr{F}_i)(\forall x \in B_i)f(x) \geq 0 \\ &\Leftrightarrow (\exists B \in \bigcap_{i\in I}\mathscr{F}_i)(\forall x \in B)f(x) \geq 0. \end{aligned}$$

For the second equivalence the converse implication is trivial. The direct implication holds because it follows that $(\forall x \in \bigcup_{i\in I}B_i)f(x) \geq 0$ and $\bigcup_{i\in I}B_i$ belongs to all $\mathscr{F}_j$, $j \in I$. By the next Lem. 3, $\bigcap_{i\in I}\mathscr{F}_i$ is a proper filter. Using a completely similar argument leads to a similar conclusion for the rejected gambles $\bigcap_{i\in I}\mathscr{M}'_{\prec}(\mathscr{F}_i)$. □

The proof of Lem. 2 tells us more than that $\mathbb{C}'$ is closed under arbitrary non-empty intersections; it also tells us how to find the filter that is associated with this intersection:

$$\bigcap_{i\in I}\mathscr{M}'(\mathscr{F}_i) = \mathscr{M}'\left(\bigcap_{i\in I}\mathscr{F}_i\right). \tag{8}$$

Lemma 3. *Given a non-empty family of proper filters $\mathscr{F}_i$, $i \in I$, $\mathscr{F} := \bigcap_{i\in I}\mathscr{F}_i \in \mathbb{F}$.*

Proof. Since $\emptyset \notin \mathscr{F}_i$, also $\emptyset \notin \mathscr{F}$. Because $\mathscr{X} \in \mathscr{F}_i$ for every $i \in I$, also $\mathscr{F} \neq \emptyset$. Furthermore, let $A,B \in \mathscr{F}$, meaning that $A,B \in \mathscr{F}_i$ for every $i \in I$. Then also also $A \cap B \in \mathscr{F}_i$ for every $i \in I$, what tells us that $A \cap B \in \mathscr{F}$, meaning that $\mathscr{F}$ is closed under conjunction. Finally, let $A \in \mathscr{F}$ and $B \supseteq A$. Then $B \in \mathscr{F}_i$ for every $i \in I$, whence $B \in \mathscr{F}$, meaning that $\mathscr{F}$ is increasing. □

Lemma 4. *The partially ordered sets* $(\mathbb{C}', \subseteq)$ *and* $(\mathbb{F}, \subseteq)$ *are* order isomorphic*, meaning that there is a bijection* ϕ *from* $\mathbb{C}'$ *to* $\mathbb{F}$ *such that* $\mathscr{M}'(\mathscr{F}_k) \subseteq \mathscr{M}'(\mathscr{F}_\ell)$ *if and only if* $\phi(\mathscr{M}'(\mathscr{F}_k)) \subseteq \phi(\mathscr{M}'(\mathscr{F}_\ell))$ *for all* $\mathscr{M}'(\mathscr{F}_k), \mathscr{M}'(\mathscr{F}_\ell) \in \mathbb{C}'$*.*

Proof. Consider the map ϕ from $\mathbb{C}'$ to $\mathbb{F}$ defined by $\phi(\mathscr{M}'(\mathscr{F})) := \mathscr{F}$. ϕ is clearly injective and surjective, and therefore a bijection. We then have to prove for all $\mathscr{F}_k, \mathscr{F}_\ell \in \mathbb{F}$ that $\mathscr{M}'(\mathscr{F}_k) \subseteq \mathscr{M}'(\mathscr{F}_\ell)$ if and only if $\mathscr{F}_k \subseteq \mathscr{F}_\ell$. The 'if' is immediate from the definition of the $\mathscr{F}_i$. For the 'only if', start with $\mathscr{M}'(\mathscr{F}_k) \subseteq \mathscr{M}'(\mathscr{F}_\ell)$, and focuss on the accepted gambles. It follows that $\mathscr{M}'_{\succeq}(\mathscr{F}_k) \subseteq \mathscr{M}'_{\succeq}(\mathscr{F}_\ell)$. This is equivalent with $(\forall B_k \in \mathscr{F}_k)(\exists B_\ell \in \mathscr{F}_\ell) B_\ell \subseteq B_k$. Since $\mathscr{F}_\ell$ is increasing, it follows that $\mathscr{F}_k \subseteq \mathscr{F}_\ell$. □

8 Embedding classical propositional logic into models for practical certainty

We want to formally embed classical propositional logic into our framework. Since, in contradistinction with the models following from favourability assessments, the models that follow from indifference assessments constitute an intersection structure, this embedding is easier for the latter models.

8.1 Indifference assessments

Eq. (8) and Lem. 4 tell us that that language of proper filters is interchangeable with the language of models following from indifference assessments as far as modelling practical certainty is concerned.

8.2 Favourability assessments

Since the partially ordered set $(\mathbb{C}, \subseteq)$ is no intersection structure, there is no counterpart to Eq. (8):

$$\bigcap_{i\in I} \mathscr{M}(\mathscr{F}_i) \supseteq \mathscr{M}\left(\bigcap_{i\in I} \mathscr{F}_i\right),$$

where $\mathscr{M}(\mathscr{F}) := \left\langle \mathscr{L}^{\gtrdot}_{\mathscr{F}} \cup \mathscr{L}_{\geq 0}; \mathscr{L}^{\lessdot}_{\mathscr{F}} \cup \mathscr{L}_{<0} \right\rangle \in \mathbb{C}$, but the converse inclusion does not generally hold. Despite of this observation, Prop. 9 guarantees that we can still find an embedding of the set of filters $\mathbb{F}$ into $\mathbb{C}$.

Proposition 9. *Consider a coherent set of favourable gambles* $\mathscr{D}_\triangleright$ *derived from a coherent model that includes the background model* $\mathscr{S}$ *and take any collection of events* $\mathscr{A} \subseteq \mathscr{P}$ *such that* $\mathscr{M}_\triangleright(\mathscr{A}) \subseteq \mathscr{D}_\triangleright$*. Let* $\mathscr{F} := \{B \in \mathscr{P} \colon (\forall \varepsilon \in \mathbb{R}_{>0}) - \mathbb{I}_{B^c} + \varepsilon \in \mathscr{D}_\triangleright\}$*, then*

(i) $\mathscr{F} \in \mathbb{F}$; *(ii)* $\mathscr{M}_\triangleright(\mathscr{F}) \subseteq \mathscr{D}_\triangleright$; *(iii)* $\mathscr{A} \subseteq \mathscr{F}$.

Proof. Due to [9, Prop. 8 (iii)], $\mathscr{D}_\triangleright$ is a cone, and $\mathscr{S}_\triangleright \subseteq \mathscr{D}_\triangleright$. This guarantees, by the way, that we can always find such $\mathscr{A}$: if $\mathscr{D}_\triangleright = \mathscr{S}_\triangleright$, use $\mathscr{A} = \{\mathscr{X}\}$. No Confusion guarantees that $\mathscr{L}_{\leq 0}$ and $\mathscr{D}_\triangleright$ are disjoint, ensuring that $\emptyset \notin \mathscr{F}$. Since $\varepsilon \in \mathscr{D}_\triangleright$ for all $\varepsilon \in \mathbb{R}_{>0}$, we see that $\mathscr{X} \in \mathscr{F}$, ensuring that $\mathscr{F} \neq \emptyset$. Consider two events $A, B \in \mathscr{F}$, then both $-\mathbb{I}_{A^c} + \varepsilon_1$ and $-\mathbb{I}_{B^c} + \varepsilon_2 \in \mathscr{D}_\triangleright$ for all $\varepsilon_1, \varepsilon_2 \in \mathbb{R}_{>0}$, so also $\varepsilon_1 + \varepsilon_2 - \mathbb{I}_{A^c} - \mathbb{I}_{B^c} \in \mathscr{D}_\triangleright$. From this, we infer $\varepsilon_1 + \varepsilon_2 - \mathbb{I}_{A^c} - \mathbb{I}_{B^c} \leq \varepsilon_1 + \varepsilon_2 - \mathbb{I}_{(A\cap B)^c} \in \mathscr{D}_\triangleright$ for all $\varepsilon_1, \varepsilon_2 \in \mathbb{R}_{>0}$, so $A \cap B \in \mathscr{F}$, meaning that $\mathscr{F}$ is closed under finite intersections. Consider an event $A \in \mathscr{F}$ and $B \supseteq A$, then $-\mathbb{I}_{A^c} + \varepsilon \in \mathscr{D}_\triangleright$ for all $\varepsilon \in \mathbb{R}_{>0}$. Because $-\mathbb{I}_{A^c} \leq -\mathbb{I}_{B^c}$, also $-\mathbb{I}_{B^c} + \varepsilon \in \mathscr{D}_\triangleright$, so $B \in \mathscr{F}$, meaning that $\mathscr{F}$ is increasing. This proves (i).

For (ii), consider any gamble $f \in \mathscr{M}_\triangleright(\mathscr{F})$. Then there is some $B \in \mathscr{F}$ such that $\inf(f|B) =: \delta > 0$, and it follows that $f \geq \mathbb{I}_{B^c} \inf f + \mathbb{I}_B \delta = \mathbb{I}_{B^c}\gamma + \delta$, where $\gamma := \inf(f) - \delta \leq 0$. Because of the definition of $\mathscr{F}$ and taking into account that $\mathscr{D}_\triangleright$ is a cone that includes $\mathbb{R}_{>0}$, $\{-\lambda \mathbb{I}_{B^c} + \varepsilon \colon \lambda \in \mathbb{R}_{\geq 0}, \varepsilon \in \mathbb{R}_{>0}\} \subseteq \mathscr{D}_\triangleright$, hence $f \in \mathscr{D}_\triangleright$.

For (iii), consider any event $B \in \mathscr{A}$. Then $-\mathbb{I}_{B^c} + \varepsilon \in \mathscr{M}_\triangleright(\mathscr{A})$ because $\inf(-\mathbb{I}_{B^c} + \varepsilon | B) > 0$. Since $\mathscr{M}_\triangleright(\mathscr{A}) \subseteq \mathscr{D}_\triangleright$ by assumption, then also $-\mathbb{I}_{B^c} + \varepsilon \in \mathscr{D}_\triangleright$, which tells us that $B \in \mathscr{F}$. □

9 Conclusions

We have shown that the language of accept & reject statement-based uncertainty models is well-suited for describing practical certainty about the validity of some propositions, or the occurrence of the corresponding events. We have studied three different ways of translating such beliefs of practical certainty into this language, each time modelled by a different type of assessment. All three types formulations lead to the same logical inferences: a collection of events the subject is practically certain of must be closed under conjunction and modus ponens. This conclusion can be drawn as well by calculating the corresponding coherent lower probability: it is formulated in terms of a filter. We concluded with the result that the collection of coherent models following from the latter type of assessments constitute a strong belief structure, and we found a belief embedding of classical propositional logic into all our models for practical certainty.

Future goals include deriving belief expansion and belief revision operators in the language of sets of favourable gambles, inspired by the ideas in [2].

Acknowledgements

The authors wish to thank the three anonymous referees for their helpful comments.

References

[1] Inés Couso and Serafín Moral. Sets of desirable gambles: conditioning, representation, and precise

probabilities. *International Journal of Approximate Reasoning*, 52(7):1034–1055, 2011.

[2] Gert de Cooman. Belief models: an order-theoretic investigation. *Annals of Mathematics and Artificial Intelligence*, 45:5–34, 2005.

[3] Gert de Cooman and Enrique Miranda. Lower previsions induced by filter maps. In *Proceedings of IPMU'12*, 2012.

[4] Gert de Cooman and Erik Quaeghebeur. Exchangeability and sets of desirable gambles. *International Journal of Approximate Reasoning*, 53(3):363–395, 2012. Special issue in honour of Henry E. Kyburg, Jr.

[5] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937. English translation in [8].

[6] B. de Finetti. *Teoria delle Probabilità*. Einaudi, Turin, 1970.

[7] B. de Finetti. *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons, Chichester, 1974–1975. English translation of [6], two volumes.

[8] H. E. Kyburg Jr. and H. E. Smokler, editors. *Studies in Subjective Probability*. Wiley, New York, 1964. Second edition (with new material) 1980.

[9] Erik Quaeghebeur, Gert de Cooman, and Filip Hermans. Accept & reject statement-based uncertainty models. 2013. Submitted for publication, arXiv:1208.4462v2 [math.PR].

[10] Matthias C. M. Troffaes and Gert de Cooman. *Lower Previsions*. Wiley, 2013.

[11] Carl G. Wagner. The Smith–Walley interpretation of subjective probability: An appreciation. *Studia Logica*, 86:343–350, 2007.

[12] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[13] Peter Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24:125–148, 2000.

[14] Peter M. Williams. Notes on conditional previsions. Technical report, School of Mathematical and Physical Science, University of Sussex, UK, 1975. Revised journal version: [15].

[15] Peter M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44:366–383, 2007. Revised journal version of [14].

Interval-Valued Linear Model

Xun Wang
Beijing University of Technology
Beijing, China
wangxun@emails.bjut.edu.cn

Shoumei Li
Beijing University of Technology
Beijing, China
lisma@bjut.edu.cn

Thierry Denoeux
Universitié de Technologie de
Compiègne, Heudiasyc, CNRS
Compiègne, France
thierry.denoeux@hds.utc.fr

Abstract

This paper introduces a new type of statistical model: the interval-valued linear model, which describes the linear relationship between an interval-valued output random variable and real-valued input variables. Firstly, we discuss the notions of variance and covariance of set-valued and interval-valued random variables. Then, we give the definition of the interval-valued linear model and its least square estimation, as well as some properties of the least square estimation. Thirdly, we show that, whereas the best linear unbiased estimation does not exist, the best binary linear unbiased estimator exists and it is just the least square estimator. Finally, we present simulation experiments and an application example regarding temperature of cities affected by their latitude, which illustrates the application of our model.

Keywords. Interval-valued linear model, least square estimation, best binary linear unbiased estimation, D_p metric.

1 Introduction

Traditional statistical models have played a significant role in a wide range of areas. However, in real life situations, many problems cannot be handled by traditional statistical models due to imperfectness of data. Therefore, specialized statistical techniques are needed. In many practical cases, we often have to face a particular kind of imperfect data: interval-valued data (e.g., [8], [9] and [13]).

Interval-valued data may represent uncertainty or variability. In the former case, the interval data represent incomplete observations, i.e., we just know the true data belong to a range (an interval), rather than the precise values. For example, assume that researchers test the service life of a group of products, such as light bulbs. Since testing time is very long, they cannot stay in the laboratory at any time. They could come to the laboratory to see how many bulbs are burnt out every two or three hours. Then, the data regarding service life of bulbs they get are interval-valued. In contrast, in the variability case, an interval is not interpreted as a set containing a single true value, but the observation themselves are interval-valued. For instance, a weather forecast typically provides the highest and lowest temperature of the next day, which is an interval including almost all the useful information about tomorrow's temperature. This interval reflects variability of temperature of one day.

The linear model is probably the simplest and most common statistical model. It describes a random output variable determined by a few input variables and an error term in a linear way. In this paper, we consider the situation in which observations are interval-valued, i.e., the random variable is an interval-valued random variable, which is determined by real-valued variables in a linear way. This interval-valued linear model could play a significant role in dealing with imperfect data, e.g., to investigate how (interval-valued) temperature is impacted by (point-valued) intensity of solar radiation, air pressure, latitude of location , or the statistical relationship between interval-valued service life of light bulbs and point-valued properties of materials used in making bulbs.

Interval-valued random variables are a special kind of set-valued random variables, whose values are compact convex subsets of the real line $\mathbb{R}^1$. Since we have at our disposal many results on the theory of set-valued random variables (e.g., [16], [17] and [26]), this is a suitable framework to tackle the problem addressed in this paper. For a long time, however, there has been only a few works to discuss the variance and covariance of set-valued random variables, since the difference between two sets is difficult to define and the hyperspace (e.g., the space of all intervals) is not linear with respect to addition and multiplication. Vital [21] studied the metric for compact convex sets via the support functions. In 2005, Yang and Li [24], Yang [25] investigated the d_p metric for sets and the D_p metric in the space of set-valued random variables; they proposed to use the D_p metric to define the variance and covariance of set-valued and interval-valued random variables, which proved to be a good approach to deal with this problem. In Chapter 5 of [25], Yang also built a linear regression model with interval-valued regression coefficients. The underlying space in [24] and [25] is $\mathbb{R}^d$. In 2008, Blanco et al. [4] defined the d_K-variance for interval-valued random variables with underlying space $\mathbb{R}^1$, which is a special case of [24] and [25].

Some other works about interval-valued and set-valued statistical models are as follows. Tanaka and Lee [19] introduced the interval linear regression model, which is not based on the interval-valued random variable framework, and estimated the coefficients using a quadratic optimization method. Blanco-Fernandez et al. [5] and Sinova et al. [18] investigated the linear relationship between two interval-valued random variables, considering the input variable as two real-valued random variables (center and radius of the interval). They gave the least square estimation of the coefficients under the d_2 metric of intervals. Blanco-Fernandez et al. [6] studied the strong consistency and asymptotic distributions of the least square estimator. Beresteanu and Molinari [3] investigated inference for partially observed models via the asymptotic approach; they supposed the observations to be uncertain and proposed an estimation method for the real-valued parameters. Hsu and Wu [14] investigated interval-valued time series and gave three evaluation criteria of estimation and forecast efficiency for interval-valued time series. Wang and Li [22] introduced a new type of interval-valued time series (the interval autoregressive time series model) and proposed methods for parameter estimation and forecasting based on the evaluation criteria in [14]. Wang and Li [23] investigated set-valued and interval-valued stationary time series, based on the definition of variance and covariance of set-valued and interval-valued random variables introduced in [24] and [25].

In this paper, we start with the set-valued framework and consider the interval-valued random variable as a special case of set-valued random variable. We then introduce the interval-valued linear model and its least square estimation, prove its unbiasedness and discuss the best binary unbiased estimation. Treating an interval-valued random variable as two separate point-valued random variables (the left- and right-endpoints of the interval, or the center and radius of the interval) is deemed to be unreasonable. One reason is that it is quite easy to obtain estimation or forecast results such that the left-endpoint is larger than the right-endpoint or the center is negative, because these two linear models are unrelated. In this paper, we also show the limitation of using two separate linear models in terms of forecast efficiency via a simulation experiment.

The organization of this paper is as follows. In Section 2, we define the variance and covariance of set-valued random variables based on the d_p metric for sets and the D_p metric for interval-valued random variables. In Section 3, we introduce the interval-valued linear model and its least square estimator (LSE), prove the unbiasedness of this LSE and give the covariance matrix of this estimator. In Section 4, we show that the best linear unbiased estimation does not exist in general, but the best binary linear unbiased estimation (BBLUE) exists and is unique, and the BBLUE is just the LSE. In Section 5, we present a simulation study to show the methodology, and illustrate the efficiency of estimations introduced in Sections 3 and 4. We then present another simulation experiment to compare our model with using two separate linear models. Finally, in Section 6, we use the interval-valued linear model to investigate the relationship between city temperature and latitude. This example also shows how this model can be used to deal with some practical problems.

Due to page limitation, we have to omit all the proofs of theorems in Sections 3 and 4 in this paper.

2 Variance and Covariance of Set-Valued Random Variables

2.1 d_p Metric between Sets

In this section, we assume that $(\Omega, \mathcal{A}, P)$ is a probability space, $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ is a Banach space, $\mathbf{K}(\mathcal{X})$ is the family of all nonempty closed subsets of $\mathcal{X}$ and $\mathbf{K}_{kc}(\mathcal{X})$ is the family of all nonempty compact convex subsets of $\mathcal{X}$.

For any $A, B \in \mathbf{K}(\mathcal{X}), \lambda \in \mathbb{R}$, define

$$A + B = \{a + b : a \in A, b \in B\},$$

$$\lambda A = \{\lambda a : a \in A\}.$$

If $A, B \in \mathbf{K}_{kc}(\mathcal{X})$, then $A + B \in \mathbf{K}_{kc}(\mathcal{X})$.

For each $A \in \mathbf{K}_{kc}(\mathcal{X})$, the support function is defined by

$$s(x^*, A) = \sup\{x^*(a) : a \in A\}, \ x^* \in \mathcal{X}^*,$$

where $\mathcal{X}^*$ is the dual space of $\mathcal{X}$, i.e., the set of all bounded linear functionals on $\mathcal{X}$. For example, if $\mathcal{X} = \mathbb{R}^1$, $\mathcal{X}^* = \mathbb{R}^1$. Take an interval $[a, b]$ with

$0 \leq a < b$, $x \in \mathbb{R}^1$, then $s(x, [a, b]) = \begin{cases} bx, & x \geq 0 \\ ax, & x < 0 \end{cases}$.
Regarding the support function, we have the following properties:

$$s(x^*, A + B) = s(x^*, A) + s(x^*, B),$$

$$s(x^*, \lambda A) = \lambda s(x^*, A), \quad \lambda \geq 0.$$

For $1 \leq p < \infty$, take $A, B \in \mathbf{K}_{kc}(\mathcal{X})$. We define the metric d_p on $\mathbf{K}_{kc}(\mathcal{X})$ ([1], [16], [24]) by

$$d_p(A, B) = \left[\int\limits_{S^*} |s(x^*, A) - s(x^*, B)|^p d\mu \right]^{1/p},$$

where S^* is the unit sphere of $\mathcal{X}^*$, i.e., $S^* = \{x^* \in \mathcal{X}^* : \|x^*\|_{\mathcal{X}^*} = 1\}$, μ is a measure on $(\mathcal{X}^*, \mathcal{B}(\mathcal{X}^*))$.

Remark 2.1. If $\mathcal{X} = \mathbb{R}^1$, then $\mathbf{K}_{kc}(\mathbb{R}^1) = \{[a, b] : -\infty < a \leq b < \infty\}$ is the family of all intervals on $\mathbb{R}^1$. If $A_1 = [a_1, b_1] = (c_1; r_1)$, $A_2 = [a_2, b_2] = (c_2; r_2)$, where $c_i = (a_i + b_i)/2$ and $r_i = (b_i - a_i)/2$ for $i = 1, 2$, then

$$A_1 + A_2 = [a_1 + a_2, b_1 + b_2] = (c_1 + c_2; r_1 + r_2)$$

$$kA_1 = (kc_1; |k|r_1)$$

and

$$\begin{aligned} d_p(A_1, A_2) &= [|a_2 - a_1|^p + |b_2 - b_1|^p]^{1/p} \\ &= [|(c_2 - c_1) - (r_2 - r_1)|^p \\ &\quad + |(c_2 - c_1) + (r_2 - r_1)|^p]^{1/p}. \end{aligned}$$

2.2 D_p Metric Space of Set-Valued Random Variables

A set-valued mapping $F : \Omega \to \mathbf{K}(\mathcal{X})$ is called a set-valued random variable (e.g., [11], [16]) if, for each open subset O of $\mathcal{X}$, $F^{-1}(O) \in \mathcal{A}$, where $F^{-1}(O) = \{\omega \in \Omega : F(\omega) \cap O \neq \emptyset\}$ and $\emptyset$ is the empty set. Any two set-valued random variables are considered *identical* if $F_1(\omega) = F_2(\omega)$ for almost every $\omega \in \Omega$ (for short, denoted by "$a.s.(P)$").

Let $\mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathcal{X})]$ denote the family of set-valued random variables taking values in $\mathbf{K}_{kc}(\mathcal{X})$.

The D_p metric with respect to set-valued random variables is defined by

$$D_p(F_1, F_2) = [E(d_p^p(F_1(\omega), F_2(\omega)))]^{1/p},$$

where $F_1, F_2 \in \mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathcal{X})]$ ([24]).

Remark 2.2. If $\mathcal{X} = \mathbb{R}^1$, $\mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathbb{R}^1)]$ is the family of all interval-valued random variables. For $F_i \in \mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathbb{R}^1)]$, $F_i(\omega) = [f_i(\omega), g_i(\omega)] = (c_i(\omega); r_i(\omega))$, where $f_i(\omega), g_i(\omega)$ are random variables and $f_i(\omega) \leq g_i(\omega)$, and $c_i(\omega) = (f_i(\omega) + g_i(\omega))/2, r_i(\omega) = (g_i(\omega) - f_i(\omega))/2$, $i = 1, 2$. By the definition of D_p, we have

$$\begin{aligned} &D_p(F_1(\omega), F_2(\omega)) \\ =\ &[E|f_2(\omega) - f_1(\omega)|^p + E|g_2(\omega) - g_1(\omega)|^p]^{1/p} \\ =\ &[E|(c_2(\omega) - c_1(\omega)) - (r_2(\omega) - r_1(\omega))|^p \\ &+ E|(c_2(\omega) - c_1(\omega)) + (r_2(\omega) - r_1(\omega))|^p]^{1/p}. \end{aligned}$$

Let $\mathcal{L}^p[\Omega, \mathbf{K}_{kc}(\mathcal{X})] = \{F : E[\|F\|_{d_p}^p] < +\infty, F \in \mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathcal{X})]\}$. Then we have the following theorem:

Theorem 2.1. *$(\mathcal{L}^p[\Omega, \boldsymbol{K}_{kc}(\mathbb{R}^d)], D_p)$ is a complete metric space for each $1 \leq p < \infty$.* [24]

2.3 Variance and Covariance of Set-Valued Random Variables

The expectation of set-valued random variable F was introduced by Aumann [2].

Definition 2.1. *For each integrable bounded set-valued random variable F, which means* $\sup\{\|f\| : f \in F\}$ *has finite expectation, the Aumann integral of F, denoted by $E[F]$, is defined by*

$$E[F] = \left\{ \int_\Omega f dP : f \in S_F \right\},$$

where $S_F = \{f : f(\omega) \in F(\omega)\ a.s.(P), and\ f\ is\ integrable\}$ is called the selection of set-valued random variable F, $\int_\Omega f dP$ is the usual Bochner integral.

The properties of the expectation of set-valued random variables have been discussed in [11] and [16].

However, since the space of subsets of $\mathcal{X}$ is not a linear space with respect to the addition and multiplication, the minus between two sets is difficult to define. Thus, extending the important notions of variance and the covariance to set-valued random variables is not a trivial task. Yang and Li [24] proposed to define variance and covariance using the D_p metric on $\mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathbb{R}^d)]$, based on the fact that the support function of sets is subtractive. Later, Wang and Li [23] extended these definitions to the more general space $\mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathcal{X})]$.

Definition 2.2. *For each set-valued random variable $F \in \mathcal{U}[\Omega, \boldsymbol{K}_{kc}(\mathcal{X})]$, the variance of F, denoted by* $\mathrm{Var}(F)$, *is defined as*

$$\begin{aligned} \mathrm{Var}(F) &= [D_2(F, E(F))]^2 \\ &= E\left\{ \int\limits_{S^*} [s(x^*, F(\omega)) - s(x^*, E(F(\omega)))]^2 d\mu \right\}. \end{aligned}$$

For two set-valued random variables $F_1, F_2 \in \mathcal{U}[\Omega, \boldsymbol{K}_{kc}(\mathcal{X})]$, the covariance of F_1 and F_2, denoted *by* $\mathrm{Cov}(F_1, F_2)$*, is defined as*

$$\begin{aligned}
&\mathrm{Cov}(F_1, F_2)\\
=\ & E\Bigg\{ \int_{S^*} [s(x^*, F_1(\omega)) - s(x^*, E(F_1))]\\
& [s(x^*, F_2(\omega)) - s(x^*, E(F_2))] d\mu \Bigg\}.
\end{aligned}$$

The correlation coefficient of F_1 *and* F_2*, denoted by* $\rho(\mathrm{F}_1, \mathrm{F}_2)$*, is defined as*

$$\rho(F_1, F_2) = \frac{\mathrm{Cov}(F_1, F_2)}{\sqrt{\mathrm{Var}(F_1) \cdot \mathrm{Var}(F_2)}}.$$

The variance, covariance and correlation coefficient of set-valued random variables have the following properties. The proofs of Theorem 2.3-2.6 can be found in [23].

Theorem 2.2. *The variance* $\mathrm{Var}(F)$ *of* $F \in \mathcal{U}[\Omega, \boldsymbol{K}_{kc}(\mathcal{X})]$ *has the following properties:*

(1) $\mathrm{Var}(C) = 0$ *for any constant* $C \in \boldsymbol{K}_k(\mathcal{X})$*.*

(2) $\mathrm{Var}(aF) = a^2\mathrm{Var}(F)$ *for any* $a \geq 0$*.*

(3) $\mathrm{Var}(F_1 + F_2) = \mathrm{Var}(F_1) + 2\mathrm{Cov}(F_1, F_2) + \mathrm{Var}(F_2)$*.*

(4) (Chebyshev Inequality) $P(d_2(F, E(F)) \geq \varepsilon)) \leq \mathrm{Var}(F)/\varepsilon^2$*, for any* $\varepsilon > 0$*.*

Theorem 2.3. *The covariance* $\mathrm{Cov}(F_1, F_2)$ *of* $F_1, F_2 \in \mathcal{U}[\Omega, \boldsymbol{K}_{kc}(\mathcal{X})]$ *has the following properties:*

(1) $\mathrm{Cov}(aF_1, F_2) = \mathrm{Cov}(F_1, aF_2) = a\mathrm{Cov}(F_1, F_2)$ *for any* $a \geq 0$*.*

(2) $\mathrm{Cov}(F_1 + F_2, F_3) = \mathrm{Cov}(F_1, F_3) + \mathrm{Cov}(F_2, F_3)$*,* $\mathrm{Cov}(F_1, F_2 + F_3) = \mathrm{Cov}(F_1, F_2) + \mathrm{Cov}(F_1, F_3)$*.*

Theorem 2.4. *For any two interval-valued random variables* $X_1(\omega) = [a_1(\omega), b_1(\omega)] = (c_1(\omega); r_1(\omega))$ *and* $X_2(\omega) = [a_2(\omega), b_2(\omega)] = (c_2(\omega); r_2(\omega))$*, where* $c_i(\omega) = (a_i(\omega) + b_i(\omega))/2$ *is the center and* $r_i(\omega) = (b_i(\omega) - a_i(\omega))/2$ *is the radius of* $X_i(\omega)$*,* $i = 1, 2$*, the following equalities hold:*

$$\begin{aligned}
&\mathrm{Cov}(X_1(\omega), X_2(\omega))\\
=\ & \mathrm{Cov}(a_1(\omega), a_2(\omega)) + \mathrm{Cov}(b_1(\omega), b_2(\omega))\\
=\ & 2\mathrm{Cov}(c_1(\omega), c_2(\omega)) + 2\mathrm{Cov}(r_1(\omega), r_2(\omega)).
\end{aligned}$$

Theorem 2.5. *The correlation coefficient* ρ *of* $F_1, F_2 \in \mathcal{U}[\Omega, \boldsymbol{K}_{kc}(\mathcal{X})]$ *has the following properties:*

(1) $|\rho| \leq 1$*.*

(2) If F_1 *and* F_2 *are independent, then* $\rho = 0$*.*

(3) $\rho(F_1, F_2) = 1$ *if and only if* $F_2 + \lambda E(F_1) = E(F_2) + \lambda F_1$*, a.s.(P),* $\rho(F_1, F_2) = -1$ *if and only if* $F_2 + \lambda F_1 = E(F_2) + E(\lambda F_1)$*, a.s.(P), where* $\lambda = \sqrt{\mathrm{Var}(F_2)/\mathrm{Var}(F_1)}$*.*

Remark 2.3. For an interval-valued random variable $F \in \mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathbb{R}^1)]$, denoted as $F(\omega) = [f(\omega), g(\omega)] = (c(\omega); r(\omega))$, where $f(\omega), g(\omega)$ are real-valued random variables and $f(\omega) \leq g(\omega)$, $c(\omega) = (f(\omega) + g(\omega))/2, r(\omega) = (g(\omega) - f(\omega))/2$, by the definition of Aumann integral and variance of set-valued random variables, we have

$$E(F(\omega)) = [E(f(\omega)), E(g(\omega))] = (E(c(\omega)); E(r(\omega)))$$

and

$$\begin{aligned}
&\mathrm{Var(F(\omega))}\\
=\ & E(|f(\omega) - E(f)|^2) + E(|g(\omega) - E(g)|^2)\\
=\ & E(|c(\omega) - E(c) - (r(\omega) - E(r))|^2)\\
& +E(|c(\omega) - E(c) + (r(\omega) - E(r))|^2).
\end{aligned}$$

For interval-valued random variables $F_1, F_2 \in \mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathbb{R}^1)]$,

$$\begin{aligned}
&\mathrm{Cov(F_1(\omega), F_2(\omega))}\\
=\ & E(|f_1(\omega) - E(f_1)||f_2(\omega) - E(f_2)|)\\
& +E(|g_1(\omega) - E(g_1)||g_2(\omega) - E(g_2)|)\\
=\ & E(|c_1(\omega) - E(c_1) - (r_1(\omega) - E(r_1))|\\
& |c_2(\omega) - E(c_2) - (r_2(\omega) - E(r_2))|)\\
& +E(|c_1(\omega) - E(c_1) + (r_1(\omega) - E(r_1))|\\
& |c_2(\omega) - E(c_2) + (r_2(\omega) - E(r_2))|).
\end{aligned}$$

3 Interval-Valued Linear Model and Least Square Estimation

In this section, we consider an interval-valued linear model with the following general form

$$E(y) = X\beta, \tag{1}$$

where $y = (y_1, y_2, \cdots, y_n)^T$ is an $n \times 1$ vector of interval-valued observations, $X = (x_{ij})_{i=1,j=1}^{n,p}$ is an $n \times p$ design matrix, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is a $p \times 1$ interval-valued parameter vector.

Definition 3.1. *If* $(y_i; x_{i1}, x_{i2}, \cdots, x_{ip})$*,* $i = 1, 2, \cdots, n$ *is a sample of interval-valued linear model (1), the least square estimator of unknown parameters* β *is the estimator which minimizes* $d_2(y, X\beta)$*.*

By the definition of the d_p metric, we have

$$\begin{aligned}
&d_2^2(y, X\beta)\\
=\ & \sum_{i=1}^{n} d_2^2(y_i, x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots, +x_{ip}\beta_p)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^{n}\Big[(c_{y_i}-x_{i1}c_{\beta_1}-\cdots-x_{ip}c_{\beta_p})\\
&\quad -(r_{y_i}-|x_{i1}|r_{\beta_1}-\cdots-|x_{ip}|r_{\beta_p})\Big]^2\\
&\quad +\sum_{i=1}^{n}\Big[(c_{y_i}-x_{i1}c_{\beta_1}-\cdots-x_{ip}c_{\beta_p})\\
&\quad +(r_{y_i}-|x_{i1}|r_{\beta_1}-\cdots-|x_{ip}|r_{\beta_p})\Big]^2\\
&= 2\sum_{i=1}^{n}\Big[(c_{y_i}-x_{i1}c_{\beta_1}-\cdots-x_{ip}c_{\beta_p})^2\\
&\quad +(r_{y_i}-|x_{i1}|r_{\beta_1}-\cdots-|x_{ip}|r_{\beta_p})^2\Big],
\end{aligned}$$

where c_A, r_A represent the center and radius of interval A, respectively. This is a quadratic function of $c_{\beta_1},\cdots,c_{\beta_p},r_{\beta_1},\cdots,r_{\beta_p}$ and $d_2^2(y,X\beta)\geq 0$, so there exists a minimum value, which satisfies

$$\frac{\partial d_2^2(y,X\beta)}{\partial c_{\beta_j}}=0,\ \frac{\partial d_2^2(y,X\beta)}{\partial r_{\beta_j}}=0,\ j=1,2,\cdots,p,$$

that is

$$\begin{cases}\sum\limits_{i=1}^{n}(c_{y_i}-x_{i1}c_{\beta_1}-\cdots-x_{ip}c_{\beta_p})(-x_{ij})=0\\ \sum\limits_{i=1}^{n}(r_{y_i}-|x_{i1}|r_{\beta_1}-\cdots-|x_{ip}|r_{\beta_p})(-x_{ij})=0,\end{cases}$$

$j=1,2,\cdots,p$. Rewriting these equations in matrix form, we get:

$$\begin{cases}X^Tc_y=X^TXc_\beta\\ |X|^Tr_y=|X|^T|X|r_\beta,\end{cases}\tag{2}$$

where $|X|=(|x_{ij}|)_{i=1,j=1}^{n,p}$.

From the above discussions, we have the following theorem.

Theorem 3.1. *If $rank(X)=rank(|X|)=p$, the least square estimator for the interval-valued linear model (1), denoted as $\hat{\beta}_{LS}$, is unique, and*

$$\hat{\beta}_{LS}=((X^TX)^{-1}X^Tc_y;(|X|^T|X|)^{-1}|X|^Tr_y).\tag{3}$$

Furthermore, we can obtain the following theorems.

Theorem 3.2. *The LSE $\hat{\beta}_{LS}$ is an unbiased estimator of β.*

Theorem 3.3. *If $E(y)=X\beta$, $rank(X)=rank(|X|)=p$ and $\mathrm{Cov(c_y)}=\sigma_1^2\mathrm{I_n}$, $\mathrm{Cov(r_y)}=\sigma_2^2\mathrm{I_n}$, then the covariance matrix of $\hat{\beta}_{LS}$ is*

$$\mathrm{Cov}(\hat{\beta}_{\mathrm{LS}})=2\sigma_1^2(\mathrm{X^TX})^{-1}+2\sigma_2^2(\mathrm{|X|^T|X|})^{-1}.$$

4 Best Linear Unbiased and Binary Linear Unbiased Estimation

4.1 Best Linear Unbiased Estimation

Given n interval-valued data from the interval-valued linear model (1), $y_i=[a_{y_i},b_{y_i}]=(c_{y_i};r_{y_i}), i=1,2,\cdots,n$, the best linear unbiased estimator is a linear combination of $y_1,y_2,\cdots,y_n$

$$\hat{\beta}_j=\lambda_{j1}y_1+\lambda_{j2}y_2+\cdots+\lambda_{jn}y_n\doteq\lambda_j^Ty,\tag{4}$$

$j=1,2,\cdots,p$, and the estimation is unbiased, that is,

$$E(\hat{\beta}_j)=\beta_j.$$

Assume $\beta_j=[a_{\beta_j},b_{\beta_j}]=(c_{\beta_j};r_{\beta_j})$. By (1) and (4), we have

$$\begin{aligned}E(\hat{\beta}_j)&=\lambda_j^TE(y)\\&=\lambda_j^T(Xc_\beta;|X|r_\beta)=(\lambda_j^TXc_\beta;|\lambda_j|^T|X|r_\beta),\end{aligned}$$

where $|\lambda_j|=(|\lambda_{j1}|,|\lambda_{j2}|,\cdots,|\lambda_{jn}|)^T$. Therefore we obtain

$$E(\hat{\beta})=(\Lambda Xc_\beta;|\Lambda||X|r_\beta),\tag{5}$$

where $\Lambda=\begin{pmatrix}\lambda_1^T\\ \lambda_2^T\\ \vdots\\ \lambda_p^T\end{pmatrix}=\begin{pmatrix}\lambda_{11}&\lambda_{12}&\cdots&\lambda_{1n}\\ \lambda_{21}&\lambda_{22}&\cdots&\lambda_{2n}\\ \cdots&\cdots&\cdots&\cdots\\ \lambda_{p1}&\lambda_{p2}&\cdots&\lambda_{pn}\end{pmatrix}$

and $|\Lambda|=\begin{pmatrix}|\lambda_{11}|&|\lambda_{12}|&\cdots&|\lambda_{1n}|\\ |\lambda_{21}|&|\lambda_{22}|&\cdots&|\lambda_{2n}|\\ \cdots&\cdots&\cdots&\cdots\\ |\lambda_{p1}|&|\lambda_{p2}|&\cdots&|\lambda_{pn}|\end{pmatrix}$.

On the other hand, since $\hat{\beta}$ is unbiased, we get

$$E(\hat{\beta})=(c_{\beta_j};r_{\beta_j}).\tag{6}$$

Therefore, by (5) and (6), we have

$$\Lambda X=I_p,\quad |\Lambda||X|=I_p.\tag{7}$$

Unfortunately, the solution of (7) does not exist in general. For the case $p>1$, consider the interval-valued linear regression model as an example:

$$E(y)=\beta_1+\beta_2X_2,$$

where $X_2=(x_{12},x_{22},\cdots,x_{n2})$.

Let $\Lambda=\begin{pmatrix}\lambda_{11}&\lambda_{12}&\cdots&\lambda_{1n}\\ \lambda_{21}&\lambda_{22}&\cdots&\lambda_{2n}\end{pmatrix}$ and $X=\begin{pmatrix}1&1&\cdots&1\\ x_{21}&x_{22}&\cdots&x_{2n}\end{pmatrix}^T$, then the second equation of (7) is

$$\sum_{i=1}^{n}|\lambda_{1i}|=1,\ \sum_{i=1}^{n}|\lambda_{1i}||x_{2i}|=0,$$

$$\sum_{i=1}^{n} |\lambda_{2i}| = 0, \quad \sum_{i=1}^{n} |\lambda_{2i}||x_{2i}| = 1.$$

It is obvious that these equations are contradictory.

For the case $p = 1$, $E(y) = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} \beta_1$, then (7) becomes

$$\sum_{i=1}^{n} \lambda_{1i} x_{i1} = 1, \quad \sum_{i=1}^{n} |\lambda_{1i}||x_{i1}| = 0.$$

Therefore, a linear unbiased estimator exists if and only if $x_{i1} \geq 0, i = 1, 2, \cdots, n$.

4.2 Best Binary Linear Unbiased Estimation

From the above discussions, we know that, for the interval-valued linear model (1), the best linear unbiased estimation does not exist in general, which is a major difference with the traditional linear model. However, for the interval-valued linear model, we could introduce another notion: the binary best linear unbiased estimation, which has some interesting statistical properties.

Definition 4.1. *The binary linear combination of interval-valued data $y_i = [a_{y_i}, b_{y_i}] = (c_{y_i}; r_{y_i}), i = 1, 2, \cdots, n$ with coefficients k_i, l_i $(l_i \geq 0)$ is defined as*

$$\sum_{i=1}^{n} (k_i c_{y_i}; l_i r_{y_i}) = \left(\sum_{i=1}^{n} k_i c_{y_i}; \sum_{i=1}^{n} l_i r_{y_i} \right).$$

Definition 4.2. *An estimator of an interval-valued parameter is called binary linear estimator, if it is a binary linear combination of interval-valued observations. Assume $\hat{\theta}$ is a binary linear estimator of interval-valued parameter θ, if $\hat{\theta}$ is unbiased and for any binary linear unbiased estimator θ^* of θ,*

$$\mathrm{Var}(\theta^*) \geq \mathrm{Var}(\hat{\theta}),$$

$\hat{\theta}$ is called best binary linear unbiased estimator of θ, denoted as BBLUE.

If θ is a $p \times 1$ vector of interval-valued parameter, $\mathrm{Var}(\theta^*) \geq \mathrm{Var}(\hat{\theta})$ in this definition means that $\mathrm{Cov}(\theta^*) - \mathrm{Cov}(\hat{\theta})$ is a nonnegative definite matrix.

Theorem 4.1. *If $E(y) = X\beta$, $rank(X) = rank(|X|) = p$ and $\mathrm{Cov}(\mathrm{c_y}) = \sigma_1^2 \mathrm{I_n}$, $\mathrm{Cov}(\mathrm{r_y}) = \sigma_2^2 \mathrm{I_n}$, then the least square estimator $\hat{\beta}_{LS}$ is the unique BBLUE.*

Theorem 4.2. *If $E(y) = X\beta$, $rank(X) = rank(|X|) = p$ and $\mathrm{Cov}(\mathrm{c_y}) = \sigma_1^2 \mathrm{I_n}$, $\mathrm{Cov}(\mathrm{r_y}) = \sigma_2^2 \mathrm{I_n}$, then for for all $\alpha \in \mathbb{R}^p$, $\alpha^T \hat{\beta}_{LS}$ is the unique BBLUE of $\alpha^T \beta$.*

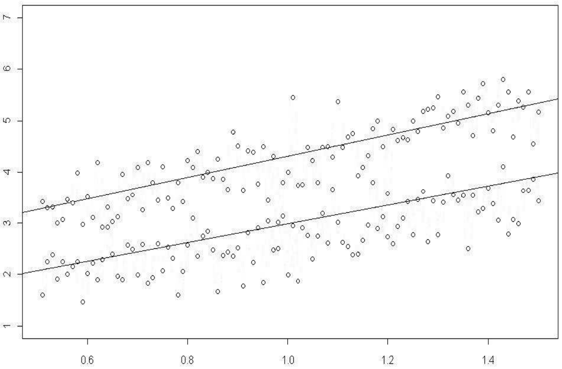


Figure 1: Points indicate 100 observations and the two lines represent the interval-valued linear regression function: $y = [1.06, 2.02] + [1.66, 2.32]x$.

5 Simulation Results

5.1 Test of Estimation Efficiency

In this section, we illustrate the interval-valued linear regression model by simulation. Let $\beta_1 = [1, 2] = (1.5; 0.5)$, $\beta_2 = [1.7, 2.3] = (2; 0.3)$ and

$$\begin{aligned} y_i &= \beta_1 + x_i \beta_2 + \varepsilon_i \\ &= (1.5 + 2x_i + c_{\varepsilon_i}; 0.5 + 0.3x_i + r_{\varepsilon_i}), \end{aligned}$$

$i = 1, 2, \cdots, n$, where $c_{\varepsilon_i}, r_{\varepsilon_i}$ are $N(0, 0.3^2)$ normal independent random variables, so that $E(y_i) = \beta_1 + E(x_i)\beta_2$. Therefore, we have

$$Ey = E \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = X \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Firstly, we let the quantity of observations n be 100, $x_i = 0.5 + 0.01i$, $i = 1, 2, \cdots, 100$. In one experiment, we get a least square estimator $\hat{\beta}_{LS}$ of β_1, β_2. Figure 1 shows the simulation experiment, in which $\hat{\beta}_{LS} = ([1.06, 2.02], [1.66, 2.32])^T$. In Figure 1, the points show the simulated data $y_i(x_i) = [1, 2] + [1.7, 2.3]x_i + \varepsilon_i$, $x_i = 0.5 + 0.01i$, $i = 1, 2, \cdots, 100$ and the two lines represent the interval-valued linear regression function computed by LSE (3): $y = [1.06, 2.02] + [1.66, 2.32]x$.

We repeated this experiment 1000 times, average value of $\hat{\beta}_{LS}^{(1)}$ was $[0.9959131, 1.996367] = (1.49614; 0.5002269)$, with a sample mean square error (sample MSE) equal to 0.0442. The average value of 1000 $\hat{\beta}_{LS}^{(2)}$ was $[1.706118, 2.300196] =$

Table 1: Average value and sample MSE of $\hat{\beta}^{(1)}_{LS}$.

	mean value of $\hat{\beta}^{(1)}_{LS}$	sample MSE of $\hat{\beta}^{(1)}_{LS}$
n=100	[0.9959131,1.996367]	0.0442
n=200	[1.002874,1.995194]	0.0236
n=300	[1.002542,2.006844]	0.0154

Table 2: Average value and sample MSE of $\hat{\beta}^{(2)}_{LS}$.

	mean value of $\hat{\beta}^{(2)}_{LS}$	sample MSE of $\hat{\beta}^{(2)}_{LS}$
n=100	[1.706118,2.300196]	0.0446
n=200	[1.705211,2.299007]	0.0220
n=300	[1.699598,2.295972]	0.0142

$(2.003157; 0.297039)$ with a sample MSE is 0.0446. Here the sample mean square error of β is defined by $\frac{1}{1000}\sum_{i=1}^{1000} d_2^2(\beta, \hat{\beta}_{LS})$.

Then we let the quantity of observations n be 200 and 300. Regarding X, we let

$$x_i = 0.5 + 0.01i, \ i = 1, 2, \cdots, 100,$$

$$x_i = x_{i-100}, \ i = 101, 102, \cdots, 200,$$

$$x_i = x_{i-200}, \ i = 201, 202, \cdots, 300.$$

Similarly, we obtained estimators of $\hat{\beta}^{(1)}_{LS}, \hat{\beta}^{(2)}_{LS}$ by the same method. The results are presented in Tables 1 and 2, which give the average value and the sample MSE of 1000 estimators of $\hat{\beta}^{(1)}_{LS}$ (real value is $[1, 2]$) and $\hat{\beta}^{(2)}_{LS}$ (real value is $[1.7, 2.3]$) respectively. We can see that the sample MSE decreases as the number of observations increases.

5.2 Comparison with Other Models

When handling the point-valued input and interval-valued output data, an easy and intuitive solution is to fit the left- and right-endpoints (or the center and the radius) of the interval-valued data to two point-valued linear model, respectively (e.g., [5],[14] and [18]). As a matter of fact, it is easy to see these two methods are equivalent. As already mentioned in the introduction, a drawback of using two separate point-valued linear model is that it is possible to obtain an inter-valued estimation or forecast result such that the left-endpoint is larger than the right-endpoint (or the radius is negative). In this section, we present the advantage of our model from another view via a simulation experiment: comparing the efficiency of the forecast.

We generated the data in the same way as in Section 5.1 with $\beta_1 = [1, 2] = (1.5; 0.5)$, $\beta_2 = [1.7, 2.1] = (1.9; 0.2)$ and

$$y_i = \beta_1 + x_i\beta_2 + \varepsilon_i, \tag{8}$$

in which $x_i = (-3 : 0.05 : 6)$ and c_{ε_i}, r_{ε_i} are $N(0, 0.1^2)$ independent random variables.

We then obtained the parameter estimation using the least square estimation for interval-valued linear model (3): $\hat{\beta}_{LS} = ([0.9979, 2.0062], [1.7017, 2.1000])^T$, and the regression function

$$y = [0.9979, 2.0062] + [1.7017, 2.1000]x. \tag{9}$$

In a second step, we fit (a_{y_i}, x_i) and (b_{y_i}, x_i), where a_{y_i} and b_{y_i} are the left- and right-endpoints of y_i, using two traditional point-valued linear models. Using the least square estimation for the traditional linear model, we obtain two fitted lines with equations:

$$\begin{cases} a_y = 0.6398 + 1.8061x \\ b_y = 2.3642 + 1.9956x. \end{cases} \tag{10}$$

Finally, we generated some new data from (8) and use (9) and (10) to forecast the output respectively. Letting $x_i = (-3 : 0.2 : 6)$, we put x_i back to (8), we obtain the (real) interval-valued output $y_i, i = 1, 2, \cdots, 46$. Then, we substitute $x_i = (-3 : 0.2 : 6)$ back to (9) and (10) and obtain the forecasts of $y_i, i = 1, 2, \cdots, 46$ using the interval-valued LS estimation (denoted by $\tilde{y}_i$) and two endpoints point-valued LS estimation (denoted by $\hat{y}_i$), respectively. The MSE of $\tilde{y}_i$ was $\frac{1}{46}\sum_{n=1}^{46} d_2^w(\tilde{y}_i, y_i) = 0.0352$ and the MSE of $\hat{y}_i$ was $\frac{1}{46}\sum_{n=1}^{46} d_2^w(\hat{y}_i, y_i) = 0.1290$. The box plots in Figure 2 show the median, the 25th and 75th percentiles and the extreme data points of the 46 forecasts using interval-valued linear model and using two separate linear models. Since the data are randomly generated, the above procedure (from data generation to forecast) is repeated 30 times, so that mean values of the MSEs of the forecasts may be computed, which are 0.0388 (using the interval-valued LS estimation) and 0.1321 (using two endpoints point-valued LS estimation). Obviously, we can see that the interval-valued linear model is better in the sense that it has smaller forecasting error.

6 Application to Real Data

In this section, we use the interval-valued linear model to investigate the relationship between temperature and latitude. The data we gather are the highest and

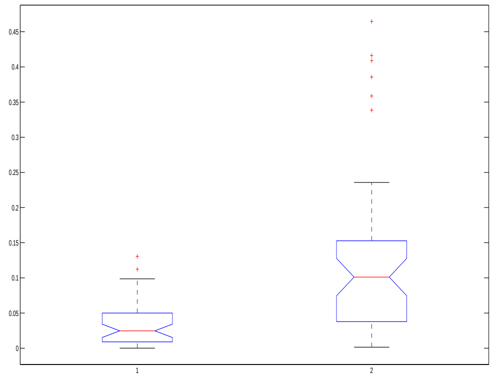


Figure 2: Box plots of forecasts results using interval-valued linear model (left) and left- and right-endpoints point-valued linear models (right).

Table 3: Temperatures and latitudes of 15 European cities on 14-th of August, 2012.

City	Latitude (°)	Highest Temp. (°C)	Lowest Temp. (°C)
Athens	38	24	34
Madrid	40.4	19	31
Istanbul	41	23	30
Roma	41.9	23	33
Marsaille	43.3	19	31
Geneve	46.25	13	28
Paris	48.8	19	26
Brussel	50.8	14	25
London	51.5	14	21
Berlin	52.5	13	23
Moscow	55.75	14	24
Stockholm	59.3	12	20
St. Petersburg	59.9	13	22
Bergen	60.4	14	20
Reykjavik	64	11	17

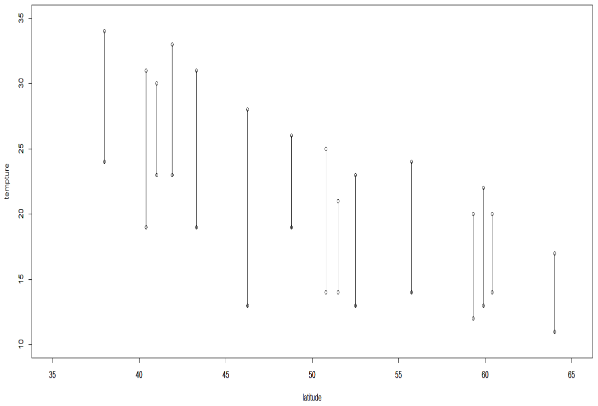


Figure 3: Temperatures (in the form of interval) of 15 European cities. Each line segment represents the temperature interval of a city.

the lowest temperatures of 15 cities in Europe on 14-th of August, 2012, as shown in Table 3 and Figure 3.

Suppose that temperature (interval-valued, y) and latitude (real-valued, x) follow the interval-valued linear model (1), that is

$$E(y_i) = \beta_1 + x_i\beta_2, i = 1, 2, \cdots, 15.$$

By least square estimation (3), which is also the best linear unbiased estimation by Theorem 4.1, we can get estimators of β_1, β_2. The linear relationship between temperature y and latitude x is

$$y = [39.03 - 0.45x, 56.01 - 0.60x],$$

which is also shown in Figure 4. From Figure 4, we can see that, as latitude increases the temperature decreases, and the daily difference in temperature also tends to decrease.

7 Conclusions

The linear model, which describes a random variable determined by a few variables and error in a linear way, plays an important role in statistics. However, in the real world, there are also a great deal of phenomena that are better described by an interval-valued random variable determined by a few real-valued random variables, e.g., temperature, stock price, service life of a kind of products. The relation between the interval-valued data and a few real-valued data can sometimes be expressed by a linear model. Therefore, we need a new type of statistical model to describe this kind of relation. In this paper, we introduced such a statistical model: the interval-valued linear

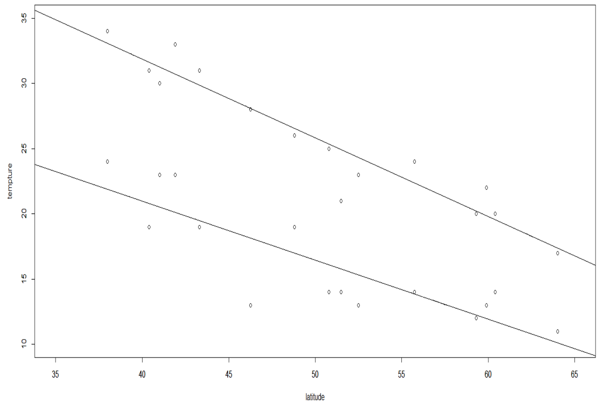


Figure 4: Data and linear relationship of temperature and latitude of 15 cities in Europe on 14-th of August, 2012. The two lines mean interval-valued linear regression function $y = [39.03196 - 0.451684x, 56.00954 - 0.6037982x]$.

model, which considers interval-valued observations determined by real-valued variables in a linear way.

Interval-valued random variables are a special kind of set-valued random variables, whose values are compact convex subsets of $\mathbb{R}^1$. In this paper, we investigated the theory in the general set-valued framework first, before focusing on the interval-valued random variables, in order to obtain some theoretical results in a wider range. In particular, we recalled the definition of variance and covariance of set-valued random variables based on the d_p metric of sets and the D_p metric of interval-valued random variables. We then introduced the interval-valued linear model and its least square estimation (LSE), proved the unbiasedness of the LSE and gave the covariance matrix of this estimator. We also showed that the best linear unbiased estimation does not exist in general, but the best binary linear unbiased estimation (BBLUE) exists and is unique, and the BBLUE is just the LSE. The performances of this estimator were illustrated using simulation experiments, and compared to those of the simple approach that consists in fitting two separate linear models using the endpoints of output intervals. The obtained results suggest that our approach yields better forecasting performance. Finally, we gave an example of the interval-valued linear model explaining how temperature is related by latitude. This short example shows how our model can be used and what type of practical problem can be solved using the interval-valued linear model.

References

[1] Aubin, J. P. and H. Franbowska, Set-Valued Analysis, Birkhauser, 1990.

[2] Aumann, R., Integrals of set valued functions, J. Math. Anal. Appl., vol: 12, pp. 1-12, 1965.

[3] Beresteanu, A. and F. Molinari, Asymptotic properties for a class of partially identified models, Econometrica, vol: 76, pp. 763-814, 2008.

[4] Blanco, A., N. Corral, G. Gonzalez-Redriguez and M. A. Lubiano, Some properties of the d_K-variance for interval-valued sets, D. Dubois et al. (Eds.): Soft Methods for Hand. Var. and Imprecision, ASC 48, pp. 331-337, 2008.

[5] Blanco-Fernandez, A., N. Corral and G. Gonzalez-Redriguez, Estimation of a flexible simple linear model for interval data based on set arithmetic, Computational Statistics and Data Analysis, vol: 55, pp. 2568-2578, 2011.

[6] Blanco-Fernandez, A., A. Colubi and G. Gonzalez-Redriguez, Confidence sets in a linear regression model for interval data, Journal of Statistical Planning and Inference, vol: 142, pp. 1320-1329, 2012.

[7] Clarke, B. R., Linear Model: the Theory and Application of Analysis of Variance, Wiley, 2008.

[8] Denoeux, T. and M.-H. Masson, Multidimensional scaling of interval-valued dissimilarity data, Pattern Recognition Letters, 21: 83-92, 2000.

[9] Denoeux, T. and M.-H. Masson, Principal component analysis of fuzzy data using autoassociative neural networks, IEEE Transactions on Fuzzy Systems, 12 (3): 336-349, 2004

[10] Diamond, P. and P. Kloeden, Metric Space of Fuzzy Sets, World Scientific, 1994.

[11] Hiai, F. and H. Umegaki, Integrals, conditional expectations and martingales of multivalued functions, J. Multivar. Anal., vol: 7, pp. 149-182, 1977.

[12] Maia, A., F. Carvalho and T. B. Ludermir, Forecasting models for interval-valued time series, Neurocomputing vol: 71 pp. 3344-3352, 2008.

[13] Masson, M.-H. and T. Denoeux, Multidimensional scaling of fuzzy dissimilarity data, Fuzzy Sets and Systems, 128 (3): 339-352, 2002.

[14] Hsu, H.L. and B. Wu, Evaluating forecasting performance for interval data, Computers and Mathematics with Applications, vol: 56, pp. 2155-2163, 2008.

[15] Lai, T. L. and H. Xing, Statistical Model and Methods for Financial Markets, Springer, 2007.

[16] Li, S., Y. Ogura and V. Kreinovich, Limit Theorems and Applications of Set-Valuded and Fuzzy

Set-Valued Random Variables, Kluwer Academic Publishers (Now Springer), Dordrecht, 2002.

[17] Molchanov, I., Theory of Random Sets, Springer, 2005.

[18] Sinova, B., A. Colubi, M. A. Gil and G. Gonzalez-Rodriguez, Interval arithmetic-based simple linear regression between interval data: Discussion and sensitivity analysis on the choice of the metric, Information Sciences, vol: 199, pp. 109-124, 2012.

[19] Tanaka, H. and H. Lee, Interval regression analysis by quadratic programming approach, IEEE Transactions on Fuzzy Systems, vol: 6, no. 4, 1998.

[20] Tseng, F., G. Tzeng, H. Wu and B. Yuan, Fuzzy ARIMA model for forecasting the foreign exchange market, Fuzzy Sets and Systems, vol: 118, pp. 9-19, 2001.

[21] Vital, R.A., L_p metrics for compact, convex sets, Journal of Approximation Theory, vol: 45, issue 3, pp. 280-287, 1985.

[22] Wang, X. and S. Li, The interval autoregressive time series model, in the proceeding of IEEE-FUZZ International Conference, pp. 2528-2533, 2011.

[23] Wang, X. and S. Li, Stationary set-valued and interval-valued time series, preprint, 2011.

[24] Yang, X. and S. Li, The D_p-metric space of set-valued random variables and its application to covariances, International Journal of Innovative Computing, Information and Control, vol: 1, pp. 73-82, 2005.

[25] Yang, X, The D_p-metric space of set-valued random variables and its applications, Dissertation for Sciences Master's Degree, in May, 2005.

[26] Zhang, W., S. Li, Z. Wang and Y. Gao, Set-Valued Stochastic Processes, Science Publisher (in Chinese), 2007.

Conference Poster Abstracts

Geometries of Inference

Miķelis Bickis
Department of Mathematics and Statistics, University of Saskatchewan
bickis@snoopy.usask.ca

Abstract

Inferential processes, having to do with the closeness of models to data, lend themselves to geometric ideas. There are several geometries that are relevant to probability models, and one must avoid the temptation to attribute features of one geometry to another, but to keep in mind the images appropriate to the task at hand.

A probability measure can be represented as an expectation, i.e., a linear functional on random variables. The set of all probability measures thus inherits a linear structure, and can be viewed as a convex subset of the linear space (a simplex in the finite-dimensional case). Walley's lower prevision [4] can be represented as the infimum of a convex subset of this larger set. To define a geometry, a linear structure also needs a distance. The appealing Euclidean norm does not adequately describe the distance concepts that are appropriate to inferential problems.

Kullback-Leibler divergence [2], while lacking the properties of a norm (or even a metric), is an inferentially meaningful measure of distance between probability measures since it is the expectation of a log-likelihood ratio. It is appealing to quantify the imprecision of a lower prevision by the information diamger—i.e., the supremum of Kullback-Leibler divergences—of the set of probability measures. This diameter, however, would be infinite if the measures in the set have different null events.

Walley's imprecise Dirichlet model [5] and the imprecise exponential family models of Quaeghebeur and de Cooman [3] are based on a convex set of hyperparameters for prior distributions of the model parameters, which are then modified by Bayesian updating. Upper and lower previsions of future observations can then be described geometrically in terms of tangent planes to the hyperparameter set. This interpretation is complicated for other predictands, or for models outside the class discussed by Diaconis and Ylvisaker [1].

The various issues are illustrated graphically by reference to 2 × 2 contingency tables.

Keywords. Exponential family, information geometry,

References

[1] Diaconis, P. and Ylvisaker, D. 1979. Conjugate priors for exponential families. *Ann. Statist.* **7** , 269–281.

[2] Kullback, S., 1959. *Information Theory and Statistics.* John Wiley and Sons, Inc., New York;

[3] Quaeghebeur, E., and G. de Cooman, 2005. Imprecise probability models for inference in exponential families. *Proceedings of the Fourth International Symposium on Imprecise Probabilities and their Applications*, G. Coman, R. Nau, & T. Seidenfeld, eds., SIPTA.

[4] Walley, Peter 1991 *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, London.

[5] Walley, Peter., 1996. Inferences from multinomial data: learning about a bag of marbles. *JRSS* **B 58**, 3–57.

A Step Towards a Conflicting Part of a Belief Function on Three-element Frame of Discernment*

Milan Daniel
Institute of Computer Science, Academy of Sciences of the Czech Republic
milan.daniel@cs.cas.cz

Abstract

Frequently, belief functions [12] usually contain an internal conflict. Based on Hájek-Valdés algebraic analysis of belief functions [10], a unique decomposition of a belief function into its conflicting and non-conflicting part was introduced at ISIPTA'11 symposium for belief functions defined on two-element frame of discernment [3]. This contribution studies the conditions under which such a decomposition exists for belief functions (BFs) defined on three-element frame. For all necessary basic notions, illuminative figures, and references see [3].

When combining belief functions, a more complicated conflict often appears. Commonly used interpretation of the sum of conflicting masses $m_{\textcircled{\cap}}(\emptyset)$ as a conflict between BFs is not correct. The problem of this interpretation was mentioned by Almond in 1995 [1], further by Liu [11], but the nature of the conflict has not been captured.

In [2, 7, 8], new ideas concerning interpretation, definition, and measurement of conflicts of BFs were introduced. An important difference between conflicts between BFs and internal conflicts of single BFs was pointed out; further, a conflict between BFs was distinguished from the difference/distance between BFs. When analyzing mathematical properties of the three approaches to conflicts of BFs from [2], there appears a possibility of expression of a BF Bel as Dempster's sum of non-conflicting BF Bel_0 with the same plausibility decisional support as the original BF Bel has and of indecisive BF Bel_S which does not prefer any of the elements of frame of discernment.

As only structures are described in the introduction to generalization of Hájek-Valdés analysis of BFs [5, 6], this study begins with an effort to make a generalization of Hájek-Valdés operation $-(a, b) = (b, a)$ and of the important homomorphism $f : (D_0, \oplus, -, 0, 0') \longrightarrow (S, \oplus, -, 0)$ given by $f(a, b) = (a, b) \oplus -(a, b)$, where $\oplus$ is Demspter's rule of combination.

Considering function '$-$' as transposition (permutation) of bbms of elements of the frame of discernment, we have $f(a, b) = (a, b) \oplus (b, a)$ as Dempster's sum of all permutations of bbms of $Bel = (a, b)$ on Ω_2. Analogously we can define

$$f(Bel) = \bigoplus_{\pi\in\Pi_3} \pi(Bel)$$

where Π_3 is the set of all permutations of bbms of elements of Ω_3: $\Pi_3 = \{\pi_{123}, \pi_{213}, \pi_{231}, \pi_{132}, \pi_{312}, \pi_{321}\}$, i.e., $f(a, b, c, d, e, f; g) = \bigoplus_{\pi\in\Pi_3} \pi(a, b, c, d, e, f; g) = (a, b, c, d, e, f; g) \oplus (b, a, c, d, f, e; g) \oplus (b, c, a, f, d, e; g) \oplus (a, c, b, e, d, f; g) \oplus (c, a, b, e, f, d; g) \oplus (c, b, a, f, e, d; g)$. It was proven that this is really homomorphism $f : D_3 \longrightarrow S$ of Dempster's semigroup $\mathbf{D}_3$ to its subsemigroup $S = (\{(a, a, a, b, b, b; 1 - 3a - 3b)\}, \oplus)$.

Having this, a series of open questions appears which are related to relation of this generalization of f to the partial generalization using $-Bel_0$ constructed via group G_3 of Bayesian BFs on Ω_3 from [3], see the updated schema of decomposition on Fig. 2. Further, the necessity of analysis of S_{Pl}, i.e., of subsemigroup of general indecisive belief functions, see Fig. 1, has appeared. Besides these new open questions, a partial positive result was reached: a unique decomposition for special classes of quasi Bayesian BFs.

Keywords. Belief function, Dempster-Shafer theory, Dempster's semigroup, conflict between belief functions, uncertainty, non-conflicting part of belief function, conflicting part of belief function.

* This research is supported by the grant P202/10/1826 of the Grant Agency of the Czech Republic. The partial institutional support RVO: 67985807 is also acknowledged.

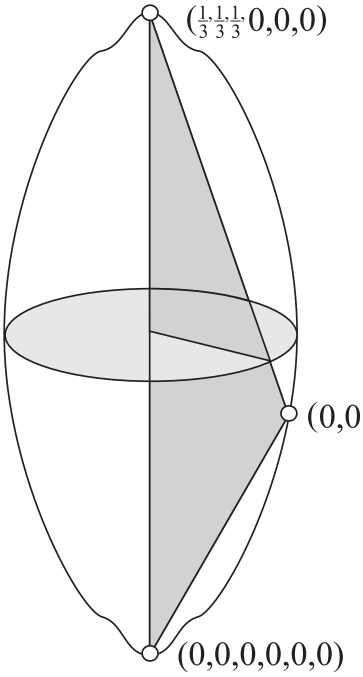

$(\frac{1}{3},\frac{1}{3},\frac{1}{3},0,0,0)$
$(0,0,0,\frac{1}{3},\frac{1}{3},\frac{1}{3})$
$(0,0,0,0,0,0)$

Figure 1: S_{Pl} — subsemigroup of general indecisive belief functions.

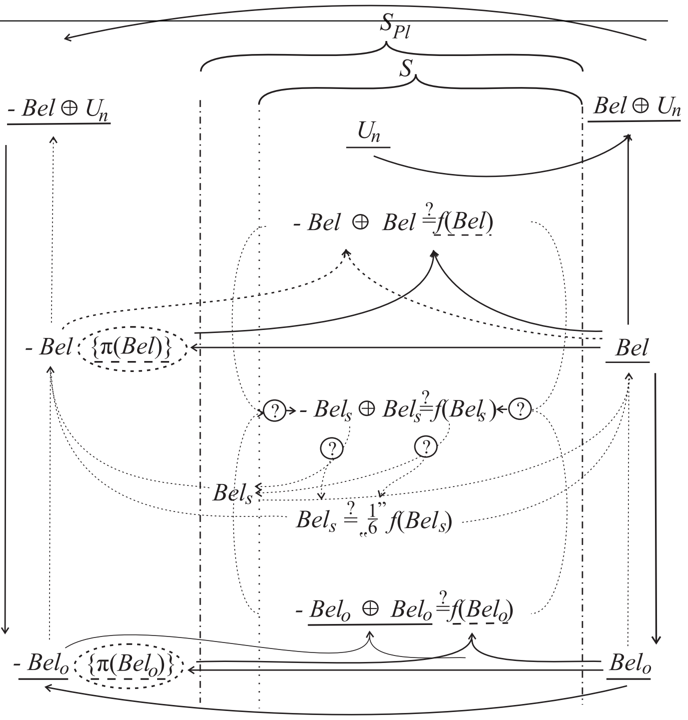

S_{Pl}
S
$-Bel \oplus U_n$
$Bel \oplus U_n$
U_n
$-Bel \oplus Bel \overset{?}{=} f(Bel)$
$-Bel$ $\{\pi(Bel)\}$
Bel
$-Bel_s \oplus Bel_s \overset{?}{=} f(Bel_s)$
?
Bel_s
$Bel_s \overset{?}{=} „\frac{1}{6}" f(Bel_s)$
$-Bel_o \oplus Bel_o \overset{?}{=} f(Bel_o)$
$-Bel_o$ $\{\pi(Bel_o)\}$
Bel_o

Figure 2: Updated detailed schema of a decomposition of BF *Bel*.

References

[1] R. G. Almond. Graphical Belief Modeling. Chapman & Hall, London, 1995.

[2] M. Daniel. Conflicts within and between Belief Functions. In: E. Hüllermeier, R. Kruse, E. Hoffmann (eds.) *IPMU 2010*. LNAI 6178: 696–705. Springer-Verlag, Berlin Heidelberg, 2010.

[3] M. Daniel. Non-conflicting and Conflicting Parts of Belief Functions. In: Coolen, F., de Cooman, G., Fetz, T., Oberguggenberger, M. (eds.) *ISIPTA'11; Proc. of the 7th ISIPTA*, pp 149–158. Innsbruck, 2011.

[4] M. Daniel. Morphisms of Dempster's Semigroup: A Revision and Interpretation. In: Barták, R. (ed.) *CSJ 2011; Proceedings of 14th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty CJS 2011*, 26–34. Matfyzpress, Prague, 2011.

[5] M. Daniel. Introduction to an Algebra of Belief Functions on Three-element Frame of Discernment — A Quasi Bayesian Case. In: S. Greco et al. (eds.) *IPMU 2012, Part III.* CCIS, vol. 299, pp 532–542. Springer, Berlin Heidelberg, 2012.

[6] M. Daniel. Introduction to Algebra of Belief Functions on Three-element Frame of Discernment — A General Case. In: T. Kroupa, J. Vejnarová (eds.) *WUPES 2012. Proc. of the 9th Workshop on Uncertainty Processing* University of Economics, Prague, 2012.

[7] M. Daniel. *Properties of Plausibility Conflict of Belief Functions.* In: L. Rutkowski et al. (eds.) *ICAISC 2013, Part I.* LNAI 7894: 235-246. Springer-Verlag, Berlin Heidelberg, 2013.

[8] M. Daniel. Belief Functions: a Revision of Plausibility Conflict and Pignistic Conflict. In: *Proceedings of 7th International conference SUM 2013.* LNCS, Springer, 2013 (in print).

[9] M. Daniel. *Steps Towards a Conflicting Part of a Belief Function.* Technical report V-1179, ICS AS CR, Prague, 2013.

[10] P. Hájek and J. J. Valdés. Generalized algebraic foundations of uncertainty processing in rule-based expert systems (dempsteroids). *Computers and Artificial Intelligence* **10** (1): 29–42, 1991.

[11] W. Liu. Analysing the degree of conflict among belief functions. Artificial Intelligence **170**: 909–924, 2006.

[12] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, New Jersey, 1976.

Parameter Dependent Uncertainty in Limit State Functions

Thomas Fetz
University of Innsbruck, Austria
Thomas.Fetz@uibk.ac.at

Abstract

In *reliability theory* an engineering system is given together with its *limit state function* $g : \mathcal{X} \subseteq \mathbb{R}^n \to \mathcal{Y} \subseteq \mathbb{R} : x \to y = g(x)$ where $x = (x_1, \dots, x_n) \in \mathcal{X}$ is a vector of basic variables (such as material properties, loads, etc.) and where $g(x) \leq 0$ means failure of the system. Then the *probability* p_f *of failure* of the system is obtained by

$$p_f = P(g(X) \leq 0) = \int_{\mathcal{X}} \chi(g(x) \leq 0) f^X(x)\, \mathrm{d}x \tag{1}$$

where f^X is the joint density function of the random variables $X = (X_1, \dots, X_n)$ and where χ is the indicator function. In the case of scarce information about the values of the basic variables x and the behavior of the system it is neither sufficient to model the uncertainty of x by a single probability density f^X nor to describe the system's reliability by a single deterministic limit state function g. A better way to model the uncertainty of the basic variables and the uncertainty in the limit state function is to use *sets of probability measures* (*credal sets*) which will result in *upper probabilities* $\overline{p}_f$ of failure. In our approach we parameterize the limit state function by additional parameters $z = (z_1, \dots, z_m) \in \mathcal{Z} \subseteq \mathbb{R}^m$ using a function $h : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y} : (x, z) \to h(x, z)$ where $h(x, z) \leq 0$ again means failure. Then a function $g_z : \mathcal{X} \to \mathcal{Y} : x \to g_z(x) = h(x, z)$ is one of the available limit state functions specified by a parameter value z. Both the basic variables x and these new additional parameters z are uncertain which means that we are not only uncertain in the choice of the values of the basic variables but also in the choice of an appropriate limit state function g_z. In [1] we assumed that the corresponding random variables X and Z are always independent and discussed the meaning of different notions of independence for sets of probability measures in the context of limit state functions. Such an assumption may be too restrictive, especially in cases where the preference we have for some limit state functions g_z may change with the values of the basic variables x.

As an extension of [1] the poster presentation is devoted to *parameter dependent uncertainty in limit state functions*. Our starting point is the formula $p_f = \int_{\mathcal{X}} \int_{\mathcal{Z}} \chi(h(x, z) \leq 0) f^{Z|x}(z)\, \mathrm{d}z\, f^X(x)\, \mathrm{d}x$ for the probability of failure with conditional density $f^{Z|x}$ of Z given x. We extend the probability of failure p_f to a mapping

$$p_f(a, b) = \int_{\mathcal{X}} \int_{\mathcal{Z}} \chi(h(x, z) \leq 0) f^{Z|x}_{b(x)}(z)\, \mathrm{d}z\, f^X_a(x)\, \mathrm{d}x \tag{2}$$

depending on parameters where $a = (a_1, \dots, a_{n_a}) \in \mathcal{A} \subseteq \mathbb{R}^{n_a}$ are the parameters of the density function f^X_a describing the uncertainty of the basic variables. The parameters b depend here on the basic variables x which means that b is a function $b : \mathcal{X} \to \mathcal{B} \subseteq \mathbb{R}^{n_b} : x \to b(x) = (b_1(x), \dots, b_{n_b}(x))$ which provides parameter values $b_1(x), \dots, b_{n_b}(x)$ to the densities of $Z|x$ for given x while in [1] b did not depend on x because of the independence of X and Z. In a next step a and b are assumed to be uncertain and sets or random sets are used to describe their uncertainty which leads to *sets of probability measures* for the random variables X and $Z|x$ and to *upper probabilities* $\overline{p}_f$ of failure. Further we will present an alternative approach using uncertain *random fields* defined on the set of basic variables to describe the uncertainty of the limit state function.

Keywords. Upper probability of failure, limit state functions, credal sets, parameterized probability measures.

References

[1] Th. Fetz. Modelling uncertainties in limit state functions. *Int. J. Approx. Reasoning*, 53(1):1–23, 2012.

First steps towards Little's Law with imprecise probabilities

Stavros Lopatatzidis, Jasper De Bock and Gert de Cooman
Ghent University, SYSTeMS Research Group
{stavros.lopatatzidis, jasper.debock, gert.decooman}@UGent.be

Abstract

In this research we take the first steps towards approaching the (distributional version of) Little's Law [1, 5, 6] from an imprecise-probabilistic point of view. We examine the law for a discrete-time, single-server queue where the arrivals and the servicing (departures) happen according to imprecise Bernoulli processes: forward irrelevant arrivals [3, 4] occur at each discrete time point with probability interval $[\underline{a}, \overline{a}]$ and, similarly, forward irrelevant departures occur at each discrete time point with probability interval $[\underline{d}, \overline{d}]$. Arrivals and departures are assumed to be epistemically independent [8].

We make two additional assumptions regarding the properties of the queue as well. The first one is that upon arriving, an item needs to remain in the queue till served. And secondly, departure is characterised by the FIFO (first in first out) property. These assumptions allow us to get closer to the distributional version of the Little's Law, as we can use them to relate (at any time point) the distribution of the size of the queue with the time spent in the queue.

We present our results using the framework of coherent lower and upper previsions [8]. Our main result is a relation between the lower (and upper) prevision of the waiting time D_t of the last item in the queue and the lower (and upper) prevision of the number L_t of items in the queue at any given time point t. More specifically, at any time t, we get $\underline{P}(L_t) = \overline{d}\underline{P}(D_t)$ and $\overline{P}(L_t) = \underline{d}\overline{P}(D_t)$. As a consequence, we find that this result also holds when, rather than forward irrelevance, we impose more stringent independence assumptions on the departure process, such as epistemic independence [7, 8], or strong independence [2].

We also address some questions related to the limit behaviour of the queuing system. What does the (imprecise) stationary distribution of the number of items look like? How can we use our main result above to derive the stationary distribution of the waiting time? And finally, how is this stationary behaviour influenced by the arrival process?

Keywords. Little's Law, Bernoulli processes, coherent lower (and upper) previsions, forward irrelevance, epistemic independence.

References

[1] Dimitris Bertsimas and Daisuke Nakazato. The distributional Little's law and its applications. *Operations Research*, 43(2):298–310, 1995.

[2] Fabio G. Cozman. Sets of probability distributions, independence, and convexity. *Synthese*, 186(2):577–600, 2012.

[3] Gert de Cooman and Enrique Miranda. Forward irrelevance. *Journal of Statistical Planning and Inference*, 139(2):256–276, 2009.

[4] Sebastien Destercke and Gert de Cooman. Relating epistemic irrelevance to event trees. In D. Dubois, M. Lubiano, H. Prade, M. Gil, P. Grzegiorzewski, and O. Hryniewicz, editors, *Soft Methods for Handling Variability and Imprecision*, pages 66–73. Springer, 2008.

[5] Rasoul Haji and Gordon F. Newell. A relation between stationary queue and waiting time distributions. *Journal of Applied Probability*, pages 617–620, 1971.

[6] John D.C. Little and Stephen C. Graves. Little's law. In *Building Intuition*, pages 81–100. Springer, 2008.

[7] Paolo Vicig. Epistemic independence for imprecise probabilities. *International Journal of Approximate Reasoning*, 24(2):235–250, 2000.

[8] Peter Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.

New Results on the Inferential Complexity of Credal Networks

Denis D. Mauá
IDSIA,
Switzerland,
denis@idsia.ch

Cassio P. de Campos
IDSIA,
Switzerland,
cassio@idsia.ch

Alessio Benavoli
IDSIA,
Switzerland,
alessio@idsia.ch

Alessandro Antonucci
IDSIA,
Switzerland,
alessandro@idsia.ch

Abstract

Credal networks are graph-based multivariate statistical models where irrelevance assessments between sets of variables are concisely described by means of an *acyclic directed graph* whose nodes are identified with variables. Here we focus on categorical variables [Cozman, 2000]. A credal network encodes a set of *Markov conditions*: the non-descendant non-parents of any variable are irrelevant to it once the value of the parents is known. The complete specification of a credal network requires the quantification of local conditional credal sets, closed and convex sets of conditional probability distributions [Levi, 1980]. Credal networks represent a joint credal set over all variables in the model, and thus allow for the distinction of randomness and ignorance, and facilitate the elicitation of the parameters from experts [Antonucci et al., 2007, Antonucci et al., 2009, Piatti et al., 2010].

To fully characterize the credal set induced by a credal network we need to settle on the concept of irrelevance adopted, and thus on the semantics of the arcs in the graph. The most commonly used concepts in the literature are *strong independence* and *epistemic irrelevance.* The former states that two variables X and Y are strongly independent if the joint credal set of X, Y can be regarded as originating from a number of precise models in each of which the two variables are stochastically independent. Strong independence is closely related to the sensitivity analysis interpretation of credal sets, which regards an imprecise model as arising out of partial ignorance of a precise one [Antonucci and Piatti, 2009, Zaffalon and Miranda, 2009]. A variable X is epistemically irrelevant to a variable Y if observing X does not affect our beliefs about Y. In other words, by making an epistemic irrelevance assessment, we are stating that our beliefs about Y do not change after receiving information about X [Walley, 1991].

Usually, credal networks are used to compute tight bounds on the expectation of some variable conditional on the value of some other variables, a task we call *predictive inference.* The complexity of this task varies greatly according to the topology of the underlying digraph and the irrelevance concept adopted. For instance, the 2U algorithm of [Fagiuoli and Zaffalon, 1998] can solve the problem in polynomial time if the digraph is a polytree, variables are binary and strong independence is assumed. When instead epistemic irrelevance is adopted, no analogous polynomial-time algorithm for the task is known. On the other hand, [de Cooman et al., 2010] developed a polynomial-time algorithm for predictive inferences in epistemic trees, that is, credal trees under epistemic irrelevance. No such algorithm is known to exist under strong independence.

The aim of this work is to present the following three new contributions that appeared in [Mauá et al., 2013], and to report on the current state of knowledge of the theoretical computational complexity of predictive inference in credal networks, as summarized in Table 1 (new results appear in boldface).

- Epistemic irrelevance and strong independence induce the same upper and lower predictive probabilities for the last (in topological order) hidden node in HMM-like credal trees. This implies that we can use the algorithm of [de Cooman et al., 2010] for credal trees under epistemic irrelevance to compute tight bounds on the posterior expectation of the last hidden node also under strong independence.

- Predictive inferences under strong independence in credal trees are NP-hard even if all variables are ternary, which shows that is unlikely that polynomial-time algorithms for these networks exist, in striking difference with the case of epistemic irrelevance.

Table 1: Summary of the computational complexity of predictive inference in credal networks.

Model	**Strong**	**Epistemic**
HMM	**P**	P
Tree	**NP-hard**	P
Polytree	NP-hard	**NP-hard**
General	NP$^{\text{PP}}$-hard	**NP$^{\text{PP}}$-hard**

- Predictive inference in networks where root nodes are vacuous and the remaining ones are precise is invariant to the irrelevance concept used. This in turn implies that the task in credal polytrees under epistemic irrelevance is NP-hard, even if all variables are at most ternary, as this is the case under strong independence unless all variables are binary [Fagiuoli and Zaffalon, 1998, de Campos and Cozman, 2005].

Keywords. Credal networks, graphical models, epistemic irrelevance, strong independence, coherent lower prevision, credal sets.

References

[Antonucci et al., 2009] Antonucci, A., Brühlmann, R., Piatti, A., and Zaffalon, M. (2009). Credal networks for military identification problems. *International Journal of Approximate Reasoning*, 50(4):666–679.

[Antonucci and Piatti, 2009] Antonucci, A. and Piatti, A. (2009). Modeling unreliable observations in Bayesian networks by credal networks. In Godo, L. and Pugliese, A., editors, *Scalable Uncertainty Management, Third International Conference, SUM 2009, Washington, DC, USA, September 28–30, 2009. Proceedings*, volume 5785 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

[Antonucci et al., 2007] Antonucci, A., Piatti, A., and Zaffalon, M. (2007). Credal networks for operational risk measurement and management. In *International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2007).*

[Cozman, 2000] Cozman, F. G. (2000). Credal networks. *Artificial Intelligence*, 120(2):199–233.

[de Campos and Cozman, 2005] de Campos, C. P. and Cozman, F. G. (2005). The inferential complexity of bayesian and credal networks. In *Proceedings of the 19th international joint conference on Artificial intelligence*, IJCAI'05, pages 1313–1318, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[de Cooman et al., 2010] de Cooman, G., Hermans, F., Antonucci, A., and Zaffalon, M. (2010). Epistemic irrelevance in credal nets: the case of imprecise markov trees. *International Journal of Approximate Reasoning*, 51(9):1029–1052.

[Fagiuoli and Zaffalon, 1998] Fagiuoli, E. and Zaffalon, M. (1998). 2U: An exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106(1):77–107.

[Levi, 1980] Levi, I. (1980). *The Enterprise of Knowledge.* MIT Press, London.

[Mauá et al., 2013] Mauá, D. D., de Campos, C. P., Benavoli, A., and Antonucci, A. (2013). On the complexity of strong and epistemic credal networks. Technical report, Dalle Molle Institute for Artificial Intelligence. IDSIA-02-13, http://www.idsia.ch/idsiareport/IDSIA-02-13.pdf.

[Piatti et al., 2010] Piatti, A., Antonucci, A., and Zaffalon, M. (2010). *Building Knowledge-Based Systems by Credal Networks: A Tutorial.* Nova Science.

[Walley, 1991] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London.

[Zaffalon and Miranda, 2009] Zaffalon, M. and Miranda, E. (2009). Conservative inference rule for uncertain reasoning under incompleteness. *Journal of Artificial Intelligence Research*, 34:757–821.

Bivariate P–Boxes

Enrique Miranda, Ignacio Montes
University of Oviedo, Spain
{mirandaenrique, imontes}@uniovi.es

Renato Pelessoni, Paolo Vicig
University of Trieste, Italy
{renato.pelessoni, paolo.vicig}@econ.units.it

Abstract

Given a random number X, a probability box or *p–box* $(\underline{F}_X, \overline{F}_X)$ is a couple of cumulative distribution functions (cdfs) s.t. $\underline{F}_X \leq \overline{F}_X$ [1, 4]. Here and in what follows, we impose no continuity property on any cdf, which is therefore a dF-coherent probability (a finitely, not necessarily σ-additive precise probability) on the monotone family of events $D_1 = \{A_x | x \in \mathbb{R}\} \cup \{\varnothing, \Omega\}, A_x = (X \leq x), \forall x \in \mathbb{R}$. A p–box therefore naturally extends to an imprecise probability framework the description of uncertainty about X by means of a cdf.

In this note we investigate properties of the generalisation of p–boxes, suited to describe couples (X, Y) of random numbers and to be called bivariate p–boxes. We focus on analogies between bivariate p–boxes and traditional joint distribution functions, and on how bivariate p-boxes may be obtained from marginal uncertainty judgements.

Definitions. Given (X, Y), let $A_{x,y} = (X \leq x \wedge Y \leq y)$. A map $F : D_2 = \{A_{x,y} : x, y \in \mathbb{R}\} \cup \{\varnothing, \Omega\} \to [0; 1]$ is *standardized* if F is non–negative, componentwise non–decreasing, $F(\varnothing) = 0$, $F(\Omega) = 1$. Later on, we shall also write $F(x, y)$ instead of $F(A_{x,y})$. $(\underline{F}, \overline{F})$ is a *bivariate p–box* if each of $\underline{F}$, $\overline{F}$ is standardized and $\underline{F} \leq \overline{F}$. $(\underline{F}, \overline{F})$ is a *coherent* p–box (a p–box that avoids sure loss (ASL)) iff, further, both $\underline{F}$ and $\overline{F}$ are *jointly* coherent (ASL) [5], lower and upper respectively, probabilities on D_2. We say that $\underline{F}$, $\overline{F}$ are jointly coherent (ASL) when the lower probability $\underline{P}$ defined as $\underline{P}(A_{x,y}) = \underline{F}(x, y)$ on $S = \{A_{x,y} | x, y \in \mathbb{R}\}$, $\underline{P}(A^c_{x,y}) = 1 - \overline{F}(x, y)$ on $S^- = \{A^c_{x,y} | x, y \in \mathbb{R}\}$ is coherent (ASL) on $S \cup S^-$.

A first major difference between coherent bivariate and univariate p–boxes is that $\underline{F}$, $\overline{F}$ need not be dF-coherent precise probabilities. This clearly depends on the structure of D_2, an only partially ordered set unlike D_1, but there are relationships with 2–monotonicity too:

Proposition 1 *Let $\underline{P}$ be a 2–monotone lower probability on some lattice $L \supset D_2$, and $\overline{P}$ its conjugate (hence, 2–alternating) upper probability.*

a) If $\underline{F}$ is the restriction of $\underline{P}$, $\underline{F}$ is dF-coherent [3].

b) If $\overline{F}$ is the restriction of $\overline{P}$, it is not necessarily dF-coherent, while its corresponding upper tail function is.

c) Conversely, if $(\underline{F}, \overline{F})$ is given and $\underline{F}$, $\overline{F}$ are jointly dF-coherent, the natural extension of $(\underline{F}, \overline{F})$ is not necessarily 2–monotone.

As well-known, a joint cdf F is characterised by some conditions, including a *rectangle inequality* $F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1) \geq 0, \forall x_1 \leq x_2, y_1 \leq y_2$. With a p-box $(\underline{F}, \overline{F})$, we have four rectangle inequalities:

[R1] $\underline{F}(x_2, y_2) - \underline{F}(x_1, y_2) - \underline{F}(x_2, y_1) + \overline{F}(x_1, y_1) \geq 0$

[R2] $\overline{F}(x_2, y_2) - \underline{F}(x_1, y_2) - \underline{F}(x_2, y_1) + \underline{F}(x_1, y_1) \geq 0$

[R3] $\overline{F}(x_2, y_2) - \underline{F}(x_1, y_2) - \overline{F}(x_2, y_1) + \overline{F}(x_1, y_1) \geq 0$

[R4] $\overline{F}(x_2, y_2) - \overline{F}(x_1, y_2) - \underline{F}(x_2, y_1) + \overline{F}(x_1, y_1) \geq 0.$

These inequalities interact variously with coherence or ASL of either a p–box $(\underline{F},\overline{F})$ or its components $\underline{F}$, $\overline{F}$, taken separately:

Proposition 2 *a) [R1]÷[R4] are necessary for coherence of* $(\underline{F},\overline{F})$.

b) Neither of them is, in general, necessary for ASL of $(\underline{F},\overline{F})$*;* $\underline{F}$ *(being standardized) always avoids sure loss, while* $\overline{F}$ *avoids sure loss if [R2] holds.*

c) In the case that X, Y *are both two–valued, [R1]÷[R4] are also sufficient for coherence of* $(\underline{F},\overline{F})$*, while [R1] is necessary and sufficient for* $(\underline{F},\overline{F})$ *to be ASL.*

An important situation originating bivariate p–boxes is when marginal cdfs for X and Y are given, and there is uncertainty about the kind of interaction between X and Y. More generally, we may think that marginal p-boxes $(\underline{F}_X,\overline{F}_X)$, $(\underline{F}_Y,\overline{F}_Y)$ are assessed for X and Y. Then, under these assumptions,

Proposition 3 *Let* $\mathcal{C}$ *be a set of copulas. Define the bivariate p–box* $(\underline{F},\overline{F})$ *as* $\underline{F}(x,y) = \inf_{C\in\mathcal{C}} C(\underline{F}_X(x),\underline{F}_Y(y)), \overline{F}(x,y) = \sup_{C\in\mathcal{C}} C(\overline{F}_X(x),\overline{F}_Y(y))$*. Then* $(\underline{F},\overline{F})$ *is coherent.*

While the above proposition may be viewed as a sort of imprecise counterpart of Sklar's Theorem [2], in the part ensuring that a certain function (copula) of two univariate cdfs returns a joint distribution having the given cdfs as marginals, it has to be stated that the correspondence breaks down on the reverse side, when wishing to view any bivariate p–box as depending on its arguments through a function (not necessarily a copula or subcopula) of its marginals. This is in general not possible, outside some special cases.

Fréchet upper and lower bounds also play a very important role in obtaining joint p–boxes from marginal ones, even in the n–variate case. In fact,

Proposition 4 *a) Given* $F_1, F_2, \ldots, F_n$ *(marginal cdfs, for* $X_1, X_2, \ldots, X_n$ *respectively), the lower Fréchet bound* $\underline{F}^L(x_1,x_2,\ldots,x_n) = \max(F_1(x_1)+F_1(x_2)+\ldots+F_n(x_n)) - n + 1, 0)$ *is a coherent lower probability (also dF-coherent, as well–known [2], when* $n = 2$*).*

b) Given the n *marginal p–boxes* $(\underline{F}_1,\overline{F}_1),\ldots,(\underline{F}_n,\overline{F}_n)$*, their natural extension on* $D_n = \{X_1 \leq x_1 \wedge \ldots \wedge X_n \leq x_n | x_1,\ldots,x_n \in \mathbb{R}\} \cup \{\varnothing,\Omega\}$ *is the n-dimensional p-box* $(\underline{F}^L,\overline{F}^U)$*, where* $\underline{F}^L(x_1,x_2,\ldots,x_n) = \max(\underline{F}_1(x_1)+\underline{F}_2(x_2)+\ldots+\underline{F}_n(x_n))-n+1,0)$*, while* $\overline{F}^U(x_1,x_2,\ldots,x_n) = \min(\overline{F}_1(x_1),\overline{F}_2(x_2),\ldots,\overline{F}_n(x_n))$ *is the Fréchet upper bound (which is dF–coherent,* $\forall n$*).*

Keywords. P–boxes, coherent lower/upper probabilities, rectangle inequalities, copulas, Fréchet bounds.

References

[1] S. Ferson, V. Kreinovich, L. Ginzburg, D.S. Myers and K. Sentz. *Constructing probability boxes and Dempster-Shafer structures.* Technical Report SAND2002-4015, Sandia National Laboratories, 2003.

[2] R. B. Nelsen. *An Introduction to Copulas.* Springer, 2006.

[3] M. Scarsini. Copulae of Capacities on Product Spaces. *Distributions with Fixed Marginals and Related Topics.* IMS Lecture Notes, Institute of Mathematical Statistics, 28:307–318, 1996.

[4] M. C. M. Troffaes and S.Destercke. Probability boxes on totally preordered spaces for multivariate modelling. *International Journal of Approximate Reasoning*, 52(6):767-791, 2011.

[5] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991

Stochastic PDEs with Random Set Coefficients

Jelena Nedeljković
University of Innsbruck, Austria
jelena.nedeljkovic@student.uibk.ac.at

Michael Oberguggenberger
University of Innsbruck, Austria
michael.oberguggenberger@uibk.ac.at

Abstract

This contribution addresses stochastic PDEs with random set coefficients. A typical example is the elliptic PDE

$$-\operatorname{div}\big(A(x)\operatorname{grad}u(x)\big) = f(x)$$

where the excitation and the coefficient matrix are given by any of the following: (a) a random field (a stochastic process with respect to the spatial variable); (b) a random set; (c) a random field whose parameters are random sets; (d) a combination thereof. As soon as random sets and stochastic processes are involved, the solution u is a set-valued process. The question arises in what sense it can be viewed as a random set.

For a stationary, Gaussian random field A it suffices to specify the expectation values $\mu \equiv \mathrm{E}(A(x))$ and the autocovariance function $C(\rho) = \mathrm{COV}(A(x), A(y))$ which then depends only on the distance $\rho = |x - y|$. As a starting point, we consider a parametrized autocovariance function of the form $C(\rho) = \sigma^2 \exp\big(-|\rho|/L\big)$ with the field variance σ^2 and the correlation length L as parameters. A useful feature of this type of random field is that it can be obtained as solution to the Langevin equation, W_t denoting Wiener process,

$$\mathrm{d}X_t = -\tfrac{1}{L}X_t + \sqrt{\tfrac{2}{L}}\,\sigma\,\mathrm{d}W_t, \quad X_0 \sim \mathcal{N}(0,\sigma^2). \tag{1}$$

A random set is a map X which assigns to every ω from a probability space (Ω, Σ, P) a subset $X(\omega)$ of a target space $\mathbb{E}$ such that the upper inverses $X^-(B) = \{\omega \in \Omega : X(\omega) \cap B \neq \emptyset\}$ are measurable for every Borel subset B of $\mathbb{E}$. An important tool is the *fundamental measurability theorem* that states (if $\mathbb{E}$ is a Polish space) the equivalence of the defining measurability property of $X^-(B)$ for Borel, open, and closed subsets B as well as the equivalence with the existence of a *Castaing representation.* A set-valued random variable such that $X^-(B)$ is measurable for every open set B is called *Effros-measurable.* Starting from a random field whose correlation length, e.g., is an interval, the assignment

$$\omega \to \{A_L(x,\omega) : L \in [\underline{L}, \overline{L}]\},$$

where x is a point in space and $A_L(x,\omega)$ is a realization at point x of the field with correlation length L, defines a random set. It is the purpose of this contribution to present a proof of this fact. Thanks to the representation (1), the continuity of the map $L \to A_L(x,\omega)$ can be derived from the results of [1, 2]. From there, a Castaing representation can be immediately obtained, which leads to the Effros measurability; the fundamental measurability theorem completes the argument. The methods will be demonstrated at the hand of a numerical example, employing polynomial chaos expansion as a computational device.

Keywords. Random fields, random sets, set-valued stochastic processes.

References

[1] B. Schmelzer. On solutions to stochastic differential equations with parameters modeled by random sets. *International Journal of Approximate Reasoning* 51:1159–1171, 2010.

[2] B. Schmelzer. Set-valued assessments of solutions to stochastic differential equations with random set parameters. *Journal of Mathematical Analysis and Applications* 400:425–438, 2013.

The Stochastic Wave Equation with an Interval Valued Parameter

Lukas Wurzer
University of Innsbruck, Austria
lukas.wurzer@uibk.ac.at

Michael Oberguggenberger
University of Innsbruck
michael.oberguggenberger@uibk.ac.at

Abstract

We consider the solution u_c of the stochastic wave equation in three space dimensions

$$\partial_t^2 u_c - c^2 \Delta u_c = \dot{W} \qquad\qquad u_c : \Omega \to \mathcal{S}'\left(\mathbb{R}^4\right)$$

denoting by $\dot{W}$ the white noise with support in $[0, \infty) \times \mathbb{R}^3$. It is a generalized stochastic process on a probability space (Ω, Σ, μ). A suitable choice for Ω is the space of tempered distributions $\mathcal{S}'(D)$.

Modelling the parameter c as an interval means to investigate the function

$$\begin{aligned} X : \Omega &\to P\left(\mathcal{S}'\left(\mathbb{R}^4\right)\right) \\ \omega &\mapsto \{u_c(\omega), c_1 \leq c \leq c_2\} \end{aligned}$$

In this contribution we show that X fulfils the Borel measurability condition

$$X^-(B) := \{\omega \in \Omega : X(\omega) \cap B \neq \emptyset\} \in \mathcal{B}(\Omega) \qquad\qquad \forall B \in \mathcal{B}\left(\mathcal{S}'\left(\mathbb{R}^4\right)\right)$$

and therefore is a random set in the general sense of Molchanov [1].

Keywords. Stochastic PDEs, Random Sets, Distributions.

References

[1] Ilya Molchanov. *Theory of random sets.* Probability and its Applications (New York). Springer-Verlag London Ltd., London, 2005.

Index