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 $i = 1,\dots,m$, 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:

$$\Delta_m(p) = \left\{ (p_1,\dots,p_m) : p_i \geq 0, \quad \sum_{j=1}^{m} p_j = 1 \right\}. \qquad (1)$$

A Bayesian approach consists in assuming a prior Dirichlet distribution for the vector of variables $(P_1,\dots,P_m)$, 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 $(t_1,\dots,t_m)$, 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 $(t_1,\dots,t_m)$ 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 $i = 1,\dots,m$. 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 $P_1,\dots,P_m$, for example

$$\underline{E}[P_iP_j] = 0, \ \overline{E}[P_iP_j] = \frac{1}{4}\frac{s}{s+1}, \qquad (4)$$

while a prior ignorance model for P_iP_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 $P_1,\dots,P_m$ 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 $i = 1, \ldots, m$ with $(t_1, \ldots, t_m)$ 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:

$$E[P_i] = t_i, \quad E[P_i | n_1, \ldots, n_m] = \frac{n_i + st_i}{N+s}, \tag{5}$$

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

$$\begin{array}{cccl} \frac{n_i+st_i}{N+s} & \overset{t_i \to 0}{=} & \frac{n_i}{N+s} & = \underline{E}[P_i | n_1, \ldots, n_m], \\ \frac{n_i+st_i}{N+s} & \overset{t_i \to 1}{=} & \frac{n_i+s}{N+s} & = \overline{E}[P_i | n_1, \ldots, n_m]. \end{array} \tag{6}$$

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 $t_1, \ldots, t_m$ 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 $X_1, \ldots, X_m$ which are assumed to be independent and distributed according to a Gamma distribution with parameters $(\alpha_1, 1), \ldots, (\alpha_m, 1)$, 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 $i = 1, \ldots, m$, then it can be shown that

$$(P_1, \ldots, P_m) \sim Dir(\alpha_1, \ldots, \alpha_m),$$

where $Dir(\alpha_1, \ldots, \alpha_m)$ denotes the Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_m$. 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 $X \overset{d}{=} Y_1 + \dots + Y_n$ 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:

$$\psi_i(u) := E[e^{-uX_i}] = \exp\left(-\int\limits_0^\infty (1-e^{-ux})\nu_i(dx)\right) \quad u \geq 0, \tag{7}$$

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 $W = X_1 + \dots + X_m$, 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 $W = X_1 + \dots + X_m$, $P_i = X_i/W$ for $i = 1, \dots, m$. Then, the PDF of the (NID) vector $(P_1, \dots, P_m)$ is:

$$\begin{gathered} g(p_1, \dots, p_{m-1}) = \\ \int\limits_0^\infty \prod_{i=1}^{m-1} f_i(p_i w) f_m\left(w - \sum_{i=1}^{m-1} p_i w\right) w^{m-1} dw. \end{gathered} \tag{9}$$

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:

$$\begin{aligned} &\int\limits_0^\infty \prod_{i=1}^{m-1} (p_i w)^{\alpha_i - 1} \exp(-p_i w) \\ &\cdot (w - w\textstyle\sum p_i)^{\alpha_m - 1} \exp(-(w - w\sum_{i=1}^{m-1} p_i)) w^{m-1} dw \\ &\propto p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} \cdots (1 - \textstyle\sum_{i=1}^{m-1} p_i)^{\alpha_m - 1}. \end{aligned} \tag{10}$$

■

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

$$\begin{gathered} g(p_1, \dots, p_{m-1}) = \\ \Gamma(s) \prod_{i=1}^m \tfrac{\beta_i}{\Gamma(a_i)} \prod_{i=1}^{m-1} p_i^{\alpha_i - 1} \left(1 - \sum_{j=1}^{m-1} p_j\right)^{\alpha_m - 1} \\ \cdot \left(\sum_{i=1}^{m-1} \beta_i p_i + \beta_m \left(1 - \sum_{j=1}^{m-1} p_j\right)\right)^{-s} \end{gathered} \tag{11}$$

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]:

$$E[P_j^r] = \frac{\Gamma(\alpha_j+r)\prod_{i=1}^m \beta_i^{\alpha_i}}{\Gamma(\alpha_j)\Gamma(r)} \int\limits_0^\infty \frac{u^{r-1}}{(\beta_j+u)^r \prod\limits_{i=1}^m (\beta_i+u)^{\alpha_i}} du. \tag{12}$$

To model prior near-ignorance, we consider the set of PDFs in (11) obtained by taking

$$\begin{array}{l} \alpha_i = st_i, \;\; \beta_i = t_i' \;\text{ for } i=1,\ldots,m \text{ with} \\ (t_1,\ldots,t_m) \in \Delta_m, \;\; (t_1',\ldots,t_m') \in \Delta_m; \end{array} \tag{13}$$

we call this model *Normalized 2dof Gamma* (N2dG).[4]

Proposition 1. *N2dG model satisfies:*

$$\begin{array}{lclclcl} \underline{E}[P_i^r] & = & 0, & \overline{E}[P_i^r] & = & 1 \\ \underline{E}[P_iP_j] & = & 0, & \overline{E}[P_iP_j] & \geq & \frac{1}{4}\frac{s}{s+1}. \end{array} \tag{14}$$

for any i,j and $r = 1,2,\ldots$. ■

The lower and upper expectations in (14) can be derived by noticing that for $t_i' = 1/m$ for $i = 1,\ldots,m$ 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 $(n_1,\ldots,n_m)$, i.e.,

$$L(n_1,\ldots,n_m|p_1,\ldots,p_{m-1}) = p_1^{n_1} p_2^{n_2} \cdots \left(1-\sum_{i=1}^{m-1} p_i\right)^{n_m}, \tag{15}$$

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 $i = 1,\ldots,m$.

Theorem 1. *The lower and upper posterior expectations of P_i are:*

$$\begin{array}{lcl} \underline{E}[P_i|n_1,\ldots,n_m] & = & \max\left(0, \frac{n_i-s}{N}\right), \\ \overline{E}[P_i|n_1,\ldots,n_m] & = & \min\left(1, \frac{n_i+s}{N}\right), \end{array} \tag{16}$$

for any $i = 1,\ldots,m$. ■

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 $\underline{E}[P_i|n_1,\ldots,n_m], \overline{E}[P_i|n_1,\ldots,n_m] \to \frac{n_i}{N}$ for $N \to \infty$, it also satisfies LCP.

Corollary 1. *The lower and upper posterior expectations of $\sum_{i\in J} P_i$, where J denotes a subset of $\{1,\ldots,m\}$, are:*

$$\begin{array}{lcl} \underline{E}[\sum_{i\in J} P_i|n_1,\ldots,n_m] & = & \max\left(0, \frac{\sum_{i\in J} n_i-s}{N}\right), \\ \overline{E}[\sum_{i\in J} P_i|n_1,\ldots,n_m] & = & \min\left(1, \frac{\sum_{i\in J} n_i+s}{N}\right). \end{array} \tag{17}$$

for any $i = 1,\ldots,m$. ■

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 $X_1,\ldots,X_m$ 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 $(P_1,\ldots,P_m)$ arising from the normalization of the m 1/2-stable distributed variables $X_1,\ldots,X_m$ has the *Normalized* $1/2-$*Stable distribution* (N1/2S) with PDF [15]:

$$g(p_1,\ldots,p_{m-1}) = \frac{\Gamma(\frac{m}{2})\prod\limits_{i=1}^m \alpha_i}{\pi^{\frac{m}{2}}} \frac{\prod\limits_{i=1}^{m-1} p_i^{-\frac{3}{2}}(1-\sum\limits_{i=1}^{m-1} p_i)^{-\frac{3}{2}}}{[\mathscr{A}(p_1,\ldots,p_{m-1})]^{\frac{m}{2}}}, \tag{19}$$

where $\mathscr{A}(p_1,\ldots,p_{m-1}) = \sum_{i=1}^{m-1} \frac{\alpha_j^2}{p_i} + \frac{\alpha_m^2}{1-\sum_{i=1}^{m-1} p_i}$.

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]),

$$E[P_i] = \frac{\alpha_i}{s}, \quad E[P_iP_j] = \frac{\alpha_i\alpha_j}{s^2}(1-\gamma), \tag{20}$$

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:*

$$\begin{array}{rclrcl} \underline{E}[P_i^r] & = & 0, & \overline{E}[P_i^r] & = & 1 \\ \underline{E}[P_iP_j] & = & 0, & \overline{E}[P_iP_j] & = & \frac{1}{4}(1-\gamma). \end{array} \tag{22}$$

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:

$$f(x_i|\alpha_i) = \frac{\alpha_i}{(2\pi)^{1/2}} \exp\left[-\frac{1}{2}\left(\frac{\alpha_i^2}{x_i} + x_i\right) + \alpha_i\right], \quad x_i \in \mathbb{R}^+. \tag{24}$$

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:

$$E[P_i] = \frac{\alpha_i}{s}, \quad E[P_iP_j] = \frac{\alpha_i\alpha_j}{s^2}(1 - s^2 e^s \Gamma(-2,s)), \tag{26}$$

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:*

$$\begin{array}{ll} \underline{E}[P_i] = 0, & \overline{E}[P_i] = 1, \\ \underline{E}[P_iP_j] = 0, & \overline{E}[P_iP_j] = \frac{1}{4}(1 - s^2 e^s \Gamma(-2,s)). \end{array} \tag{28}$$

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.

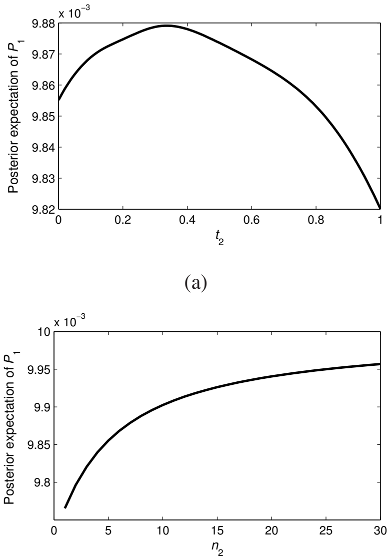
Conjecture 1. *Consider the set of priors defined by (25) and (27). Given the set of counts $(n_1,\ldots,n_m)$, the lower*

[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 $E[P_i|n_1,\ldots,n_m]$ 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 $(n_1,\ldots,n_m)$, with $n_j > 0$ for $j = 1,\ldots,m$, 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,*

$$\underline{E}[P_i|n_1,\ldots,n_m] = \frac{n_i - \frac{1}{2}}{N - n_k - \frac{1}{2}(m-1)} \times$$
$$\times \frac{\int_0^1 p_k^{n_k + \frac{m-6}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_k}}\right)(1-p_k)^{N-n_k-\frac{m-1}{2}}}{\int_0^1 p_k^{n_k + \frac{m-6}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_k}}\right)(1-p_k)^{N-n_k-\frac{m+1}{2}}}, \tag{29}$$

$$\overline{E}[P_i|n_1,\ldots,n_m] =$$
$$= \frac{\int_0^1 p_i^{n_i + \frac{m-2}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_i}}\right)(1-p_i)^{N-n_i-\frac{m+1}{2}}}{\int_0^1 p_i^{n_i + \frac{m-6}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_i}}\right)(1-p_i)^{N-n_i-\frac{m+1}{2}}}.$$

■

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 $j = 1,\ldots,m$ 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 $\frac{\hat{m}}{2N}$.

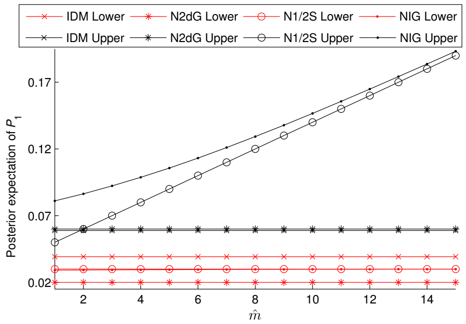


Figure 2: Posterior expectation of P_1 for $n_1 = 2$, $n_k = 3$ with $k = \arg_{j \neq 1}\min(n_j)$, $N = 50$, $s = 1$, and $\hat{m}$ ranging from 1 to 15.

	$\underline{P}(Z = red \vert n_1,\dots,n_m)$			$\overline{P}(Z = red \vert n_1,\dots,n_m)$		
	$\mathscr{Z}_1$	$\mathscr{Z}_2$	$\mathscr{Z}_3$	$\mathscr{Z}_1$	$\mathscr{Z}_2$	$\mathscr{Z}_3$
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 $\mathscr{Z}_1 = \{red, blue, green, yellow\}$, $\mathscr{Z}_2 = \{red, \text{ } all \text{ } other \text{ } colors\}$, $\mathscr{Z}_3 = \{light \text{ } red, \text{ } dark \text{ } red, \text{ } all \text{ } other \text{ } colors\}$. 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:

$$\int_0^1 dp_1 p_1^{n_1+\alpha_1-1} \int_0^{1-p_1} \cdots \int_0^{1-p_1-\cdots-p_{m-1}} \frac{p_2^{n_2+\alpha_2-1} p_3^{n_3+\alpha_3-1} (1-p_1-\cdots-p_{m-1})^{n_m+\alpha_m-1}}{\left(\sum_{i=1}^{m-1} \beta_i p_i + \beta_m (1-p_1-\cdots-p_{m-1})\right)^s} dp_2 \cdots dp_{m-1} \tag{30}$$

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:

$$\int_0^1 p_1^{n_1+\alpha_1-1} (1-p_1)^{N-n_1-\alpha_1-1} dp_1. \tag{31}$$

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:

$$\int_0^1 dp_1 p_1^{n_1-3/2+1} \int_0^{1-p_1} \cdots \int_0^{1-p_1-\ldots-p_{m-1}} \frac{\prod_{i=2}^{m-1} p_i^{n_i-3/2+m/2} (1-\sum_{i=1}^{m-1} p_i)^{n_m-3/2+m/2}}{\left[(1-\sum_{i=1}^{m-1} p_i) \sum_{i=2}^{m-1} \left(t_i^2 \prod_{i \neq j=2}^{m-1} p_j\right) + t_m^2 \prod_{i=2}^{m-1} p_i\right]^{m/2}} dp_2 \cdots dp_{m-1} \tag{33}$$

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:

$$E[P_1|n_1, \ldots, n_m] = \frac{\int_0^1 p_1^{n_1-\frac{3}{2}+1}(1-p_1)^{N-n_1-\frac{1}{2}}}{\int_0^1 p_1^{n_1-\frac{3}{2}}(1-p_1)^{N-n_1-\frac{1}{2}}} = \frac{n_1 - \frac{1}{2}}{N} \tag{34}$$

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:

$$\int_0^1 dp_2 p_2^{n_2-\frac{3}{2}+\frac{m}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_2}}\right) \int_0^{1-p_1} p_2^{n_2-\frac{3}{2}+1} dp_1 \int_0^{1-p_1-p_2} \cdots \int_0^{1-p_1-\cdots-p_{m-1}} \prod_{i=3}^{m-1} p_i^{n_i-\frac{3}{2}} (1 - \sum_{i=1}^{m-1} p_i)^{n_m-\frac{3}{2}} dp_3 \cdots dp_{m-1} \tag{35}$$

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:

$$\frac{\int_0^1 p_1^{n_1+\frac{m-6}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_1}}\right)(1-p_1)^{N-n_1-\frac{m-1}{2}}}{\int_0^1 p_1^{n_1+\frac{m-6}{4}} K_{-m/2}\left(\frac{s}{\sqrt{p_1}}\right)(1-p_1)^{N-n_1-\frac{m+1}{2}}} \times \frac{\int_0^1 {p'}_2^{n_2-\frac{3}{2}+1}(1-{p'}_2)^{N-n_1-n_2-\frac{m}{2}} dp'_2}{\int_0^1 {p'}_2^{n_2-\frac{3}{2}}(1-{p'}_2)^{N-n_1-n_2-\frac{m}{2}} dp'_2}, \tag{36}$$

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.