Learning from imperfect observations with prior ignorance

Alberto Piatti
Institute of Finance
University of Lugano
Via G.Buffi 19
CH-6900 Lugano
Switzerland
alberto.piatti@lu.unisi.ch

Marco Zaffalon
IDSIA
Stabile Galleria 2
CH-6928 Manno
Switzerland
zaffalon@idsia.ch

Fabio Trojani
Istitute of Banking and Finance
University of St.Gallen
Rosenbergstr. 52
CH-9000 St.Gallen
Switzerland
fabio.trojani@unisg.ch

Abstract

In this paper we analyze the behavior of the Imprecise Dirichlet Model when data are not perfectly observable. The results show that the existence of a perfect observation mechanism is a crucial assumption. In fact if the observation mechanism can never be assumed to be perfect, then the Imprecise Dirichlet Model produces vacuous predictive probabilities. At the other side, if we assume that there is a positive probability of having a perfect observation mechanism, then the IDM produces non vacuous predictive probabilities.

Keywords. Predictive Bayesian Inference, Imprecise Dirichlet Model, Vacuous Predictive Probabilities, Perfect and Imperfect Observation Mechanism.

1 Introduction

...

2 Setup

In this paper we consider an infinite population of individuals which can be classified in k *categories* (or *types*) from the set $\mathcal{X} = \{x_1, \ldots, x_k\}$. The proportion of units of type x_i is denoted by θ_i and called the chance of x_i. The population is therefore completely characterized by the chances $\vartheta = (\theta_1, \ldots, \theta_k)$, where the vector ϑ is a point in the closed k-dimensional unit simplex[1]

$$\Theta := \{\vartheta = (\theta_1, \ldots, \theta_k) \mid \sum_{i=1}^{k} \theta_i = 1,\ 0 \leq \theta_i \leq 1\}.$$

We define a random variable X with values in $\mathcal{X}$ which consists in drawing an individual at random from the population and observing its category. Clearly the chances of X are given by ϑ, i.e., θ_i is the chance that X is equal to x_i. Our aim is to predict the chance of drawing an individual of type x_i from a population of unknown chances ϑ after having observed N independent random draws. Having observed a dataset $\mathbf{x}$, we can summarize the observation with the counts $\mathbf{a} = (a_1, \ldots, a_k)$ where a_i is the number of individuals of type x_i observed in dataset $\mathbf{x}$ and $\sum_{i=1}^{k} a_i = N$. The chance of observing a dataset $\mathbf{x}$ with counts $\mathbf{a}$ given ϑ is equal to $P(\mathbf{x} \mid \vartheta) = \theta_1^{a_1} \cdots \theta_k^{a_k}$. In above setting each individual in the population is perfectly observable, i.e., the observer can determine the exact category of each individual without committing mistakes. In Section 4 we relax this assumption.

[1] For the rest of the paper we denote with := a definition.

3 The Imprecise Dirichlet Model

3.1 Bayesian Inference and Dirichlet Prior Density

In the Bayesian setting we learn from observed data using Bayes rule, which can be formulated as follows. Consider dataset $\mathbf{x}$ and the unknown chances ϑ. Then

$$p(\vartheta|\mathbf{x}) = \frac{P(\mathbf{x}|\vartheta) \cdot p(\vartheta)}{P(\mathbf{x})}, \qquad (1)$$

where $p(\vartheta)$ is some density measure on Θ and

$$P(\mathbf{x}) = \int_{\Theta} P(\mathbf{x}|\vartheta) p(\vartheta) d\vartheta.$$

This rule can be used only if $P(\mathbf{x}) \neq 0$. The probability measure $P(\mathbf{x}|\vartheta)$ is called the *likelihood*, the density measure $p(\vartheta)$ is called the *prior density* and the density measure $p(\vartheta|\mathbf{x})$ is called the *posterior density*. The aim of Bayesian inference in our setting is to learn the value of ϑ. We must therefore

specify as prior density a density measure p on Θ. A common choice of prior density in the multinomial setting is the *Dirichlet* density measure that is defined as follows.

Definition 1. *The* Dirichlet density $dir(s, \mathbf{t})$ *is defined on the closed k-dimensional simplex Θ and is given by the density measure*

$$p(\vartheta) := \frac{\Gamma(s)}{\prod_{i=1}^{k} \Gamma(st_i)} \prod_{i=1}^{k} \theta_i^{st_i - 1},$$

where s is a positive real number, Γ is the usual Gamma-function and $\mathbf{t} = (t_1, \ldots, t_k) \in \mathcal{T}$, where $\mathcal{T}$ is the open k-dimensional simplex

$$\mathcal{T} := \{\mathbf{t} = (t_1, \ldots, t_k) \,|\, \sum_{j=1}^{k} t_k = 1,\, 0 < t_j < 1\}.$$

We review first some important properties of Dirichlet densities.

Lemma 1 (First moment). *The first moments of a $dir(s, \mathbf{t})$ density are given by $E(\theta_i) = t_i$ for each $i \in \{1, \ldots, k\}$.*

Proof. See: (Kotz 2000), p.485 and following.

Lemma 2. *Consider an experiment with outcomes in $\mathcal{X} = \{x_1, \ldots, x_k\}$. Suppose further that the outcomes are distributed according to an unknown vector of chances ϑ and that our knowledge about the parameters ϑ can be summarized by a $dir(s, t)$-density. Then the probability of the outcome x_i according to our knowledge is given by $P(x_i) = E(\theta_i) = t_i$ for each $i \in \{1, \ldots, k\}$.*

Proof. See: (Walley 1996).

Proposition 1. *Consider a dataset $\mathbf{x}$ with corresponding counts $\mathbf{a} = (a_1, \ldots, a_k)$, where $\sum_{j=1}^{k} a_j = N$. Then the following equality holds*

$$\prod_{j=1}^{k} \theta_j^{a_j} \cdot dir(s, \mathbf{t}) = \\ = \frac{\prod_{j=1}^{k} \cdot \prod_{i=1}^{a_j} (st_j + i - 1)}{\prod_{i=1}^{N} (s + i - 1)} \cdot dir(s^{\mathbf{x}}, \mathbf{t}^{\mathbf{x}}),$$

where $s^{\mathbf{x}} = N + s$ and $t_j^{\mathbf{x}} = \frac{a_j + st_j}{N+s}$. Here and in the rest of the paper when $a_j = 0$ we set $\prod_{i=1}^{a_j=0} (st_j + i - 1) = 1$ by definition.

Proof. See Appendix A.

Remark 1. *Using a $dir(s, \mathbf{t})$ density measure as prior density in a problem involving Bayesian learning from a dataset $\mathbf{x}$ of multinomial data we have $p(\vartheta) = dir(s, \mathbf{t})$ and*

$$P(\mathbf{x}|\vartheta) = \prod_{j=1}^{k} \theta_j^{a_j}. \tag{2}$$

Proposition 1 states that the posterior density is then given by $P(\vartheta|\mathbf{x}) = dir(s^{\mathbf{x}}, \mathbf{t}^{\mathbf{x}})$ and therefore

$$P(x_i \,|\, \mathbf{x}) = E(\theta_i \,|\, \mathbf{x}) = \frac{a_i + st_i}{N + s}. \tag{3}$$

Furthermore, confronting Bayes rule (1) with the equality of the proposition we conclude that

$$P(\mathbf{x}) = \frac{\prod_{j=1}^{k} \prod_{i=1}^{a_j} (st_j + i - 1)}{\prod_{i=1}^{N} (s + i - 1)}. \tag{4}$$

3.2 The Imprecise Dirichlet Model

The Imprecise Dirichlet Model (IDM) (Walley 1996) is a model for Bayesian learning from multinomial data when there is prior near-ignorance about ϑ. Prior ignorance is modelled using the set of all Dirichlet densities $dir(s, \mathbf{t})$ for a fixed s and all $\mathbf{t}$ in $\mathcal{T}$ as set of prior densities instead of a single prior density. Because of Lemma 2 the probability of each category x_i a priori is vacuous, i.e. $P(x_i) \in [\inf_{\mathcal{T}} t_i, \sup_{\mathcal{T}} t_i] = [0, 1]$. The ignorance is therefore modeled assigning vacuous prior probabilities to each category of $\mathcal{X}$. For each prior density one calculates, using Bayes rule, a posterior density and obtains, taking into accounts the whole set of priors, a set of posterior densities. Let now s be a given positive constant number. Consider the set of prior densities $\mathcal{M}_s := \{dir(s, \mathbf{t}) \,|\, \mathbf{t} \in T, \text{s given}\}$. Suppose that we observe the dataset $\mathbf{x}$ with corresponding counts $\mathbf{a} = (a_1, \ldots, a_k)$. Then the set of posterior densities follows from Proposition 1 and is given by

$$\mathcal{M}_{N,s} := \left\{ dir(N + s, \mathbf{t}^*) \;\middle|\; t_j^* = \frac{a_j + st_j}{N + s},\, \mathbf{t} \in T \right\}.$$

Definition 2. *Given a set of probability measures $\mathcal{P}$, the* upper probability $\overline{P}$ *is given by $\overline{P}(\cdot) := \sup_{P \in \mathcal{P}} P(\cdot)$, the* lower probability $\underline{P}$ *by $\underline{P}(\cdot) := \inf_{P \in \mathcal{P}} P(\cdot)$.*

Remark 2. *The upper and lower posterior probabilities of an observation x_i in the IDM after N observations are found letting $t_i \to 1$, resp. $t_i \to 0$, and are given by $\overline{P}(x_i \,|\, \mathbf{x}) = \frac{a_i + s}{N + s}$ and $\underline{P}(x_i \,|\, \mathbf{x}) = \frac{a_i}{N + s}$ for each i.*

4 Imperfect observation mechanism

4.1 Introduction

In practice, there is always a even small possibility of doing classification mistakes during the observation process. Usually, if this probability is small, one assumes that the data are perfectly observable in order to use a simple model. It is an implicit assumption that there is a sort of continuity between models with perfectly observable data and models with small probability of mistakes in the observations. In this section, we study the behavior of the IDM when the observation process is not perfect and we show that there is absolutely no continuity between the results of the IDM with perfect observation mechanism and the results with imperfect observation mechanism. We consider a two-step model. In the first step a random variable X is generated with chances ϑ. In the second step, given the value of X, a second multinomial random variable O with values in $\mathcal{X}$ is generated with chances $\xi = (\xi_1, \ldots, \xi_k)$ dependent on the value of X. We define $\lambda_{ij} = P(\mathrm{O} = x_i \,|\, \mathrm{X} = x_j)$. All such chances can be collected in a $k \times k$ matrix, called the *emission matrix*,

$$\Lambda := \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1k} \\ \vdots & \ddots & \vdots \\ \lambda_{k1} & \cdots & \lambda_{kk} \end{pmatrix}. \tag{5}$$

Then the chances of O conditional on ϑ are given by $\xi_i = \sum_{j=1}^{k} \lambda_{ij} \cdot \theta_j$. Matrix Λ is a stochastic matrix. I.e., the sum of the elements of each column is equal to one. We assume that each row of the emission matrix has at least an element different from zero; in the opposite case we could define O on a strict subset of $\mathcal{X}$. We also assume that

$$p(\vartheta \,|\, \mathbf{x}, \mathbf{o}) = p(\vartheta \,|\, \mathbf{x}), \tag{6}$$

i.e., an observed dataset $\mathbf{o}$ gives no additional information about the value of ϑ given the true dataset $\mathbf{x}$. In the following calculations we make use of the well known equalities

$$P(\mathbf{o}) = \sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}). \tag{7}$$

and

$$p(\vartheta, \mathbf{x} \,|\, \mathbf{o}) = p(\vartheta \,|\, \mathbf{x}, \mathbf{o}) \cdot P(\mathbf{x} \,|\, \mathbf{o}). \tag{8}$$

4.2 Predictive inference

Suppose that we have observed a dataset $\mathbf{o}$ and we want to construct the posterior density $P(\vartheta \,|\, \mathbf{o})$ using Bayes rule and a prior density $dir(s, \mathbf{t})$. We have

$$\begin{aligned} P(\vartheta \,|\, \mathbf{o}) &= \sum_{\mathbf{x} \in \mathcal{X}^N} p(\vartheta, \mathbf{x} \,|\, \mathbf{o}) = \\ &\overset{(8)}{=} \sum_{\mathbf{x} \in \mathcal{X}^N} p(\vartheta \,|\, \mathbf{x}, \mathbf{o}) \cdot P(\mathbf{x} \,|\, \mathbf{o}) = \\ &\overset{(6)}{=} \sum_{\mathbf{x} \in \mathcal{X}^N} p(\vartheta \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \mathbf{o}) = \\ &\overset{(1)}{=} \sum_{\mathbf{x} \in \mathcal{X}^N} \frac{P(\mathbf{x} \,|\, \vartheta) \cdot p(\vartheta)}{P(\mathbf{x})} \cdot \frac{P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}{P(\mathbf{o})} = \\ &= \sum_{\mathbf{x} \in \mathcal{X}^N} \frac{P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \vartheta) \cdot p(\vartheta)}{P(\mathbf{o})} = \\ &\overset{(7)}{=} \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \vartheta) \cdot p(\vartheta)}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}. \end{aligned}$$

This is possible with $P(\mathbf{x}) > 0$ and $P(\mathbf{o}) > 0$, a condition satisfied in our setting[2]. From Remark 1 we have

$$P(\mathbf{x} \,|\, \vartheta) \cdot P(\vartheta) = P(\mathbf{x}) \cdot dir(s^{\mathbf{x}}, \mathbf{t}^{\mathbf{x}}). \tag{9}$$

Therefore

$$\begin{aligned} P(\vartheta \,|\, \mathbf{o}) &= \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \vartheta) \cdot P(\vartheta)}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})} = \\ &\overset{(9)}{=} \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot dir(s^x, \mathbf{t}^x)}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}, \end{aligned}$$

which is a convex combination of Dirichlet density measures. The IDM, when data are not perfectly observable, consists in performing the calculation above for each prior density in the set $\mathcal{M}_s$, the set of posterior densities consists of convex combinations of Dirichlet density measures. For each posterior density $p(\vartheta \,|\, \mathbf{o})$ we calculate the probability that the next individual drawn will be of type x_i. Because $p(\vartheta \,|\, \mathbf{o})$ is a convex combination of Dirichlet density measures, we can use (3) and we

[2] Since $t_j > 0$ for all j and $s > 0$ it follows from (4) that $P(\mathbf{x}) > 0$. Because all the rows of Λ are assumed to have at least one element different from zero, for each x_i there exists at least one j such that $\lambda_{ij} \neq 0$, therefore there exists at least one $\mathbf{x}$ with $P(\mathbf{o} \,|\, \mathbf{x}) \neq 0$ and, because $P(\mathbf{x}) > 0$ for each $\mathbf{x}$ it follows from (7) that $P(\mathbf{o}) > 0$.

obtain

$$
\begin{aligned}
P(x_i \,|\, \mathbf{o}) &\overset{(3)}{=} E(\theta_i \,|\, \mathbf{o}) \\
&= \frac{\int_\Theta \theta_i \cdot \sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \vartheta) \cdot p(\vartheta) d\vartheta}{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})} = \\
&= \frac{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot \int_\Theta \theta_i \cdot P(\mathbf{x} \,|\, \vartheta) \cdot p(\vartheta) d\vartheta}{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})} = \\
&\overset{(1)}{=} \frac{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot \int_\Theta \theta_i \cdot p(\vartheta \,|\, \mathbf{x}) d\vartheta}{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})} = \\
&= \frac{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot E(\theta_i \,|\, \mathbf{x})}{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})} = \\
&\overset{(3)}{=} \frac{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot \frac{a_i^{\mathbf{x}} + st_i}{N+s}}{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}. \qquad (10)
\end{aligned}
$$

4.3 Vacuous predictive probabilities

Theorem 1. *Assume that we are doing Bayesian inference with the IDM. We have observed a dataset* $\mathbf{o}$ *with counts* $\mathbf{n} = (n_1, \ldots, n_k)$ *and our observation mechanism is characterized by an emission matrix* Λ*. Then following results hold.*

1. *If the matrix* Λ *is full*[3]*, the the IDM produces vacuous upper and lower predictive probabilities for each category in* $\mathcal{X}$*, i.e.,*

$$\overline{P}(x_i \,|\, \mathbf{o}) = 1 \quad \textit{and} \quad \underline{P}(x_i \,|\, \mathbf{o}) = 0.$$

2. *The IDM produces non-vacuous upper predictive probabilities for the category* x_i*, only if it exists at least a* $\lambda_{ji} = 0$ *with* $n_j > 0$*.*

3. *The IDM produces non-vacuous lower predictive probabilities only for the categories* x_i*, such that* $\lambda_{ii} = 1$ *and* $\lambda_{ij} = 0$ *for each* $j \neq i$*.*

Corollary 1. *The IDM produces non-vacuous predictive probabilities for each category in* $\mathcal{X}$*, only if* $\Lambda = I$*, i.e., in the case described by (Walley 1996).*

Proof. For given $i \in \{1, \ldots, k\}$ define the dataset

$$\overline{\mathbf{x}}^i := \{x_i, \ldots, x_i\},$$

i.e., the dataset with counts $a_i = N$ and $a_j = 0$ for each $j \neq i$. We study the behavior of $P(x_i \,|\, \mathbf{o})$ when the prior $dir(s, \mathbf{t})$ is characterized by extreme values of the parameter $\mathbf{t}$. Set $\mathcal{T}$ is the open k-dimensional simplex, we can therefore define sequences of priors with extreme values of the parameters, e.g., $\mathbf{t}$ with $t_i \to 1$ and, because $\sum_{j=1}^k t_j = 1$, $t_j \to 0$ for each $j \neq i$. We show that $\lim_{t_i \to 1} P(\mathbf{x}) = 0$ for each $\mathbf{x} \in \mathcal{X}^N \setminus \{\overline{\mathbf{x}}^i\}$. The numerator of (4) is a product of terms

$$\prod_{r=1}^{a_j^{\mathbf{x}}} (st_j + r - 1). \qquad (11)$$

If $a_j^{\mathbf{x}} = 0$ (11) is equal to one by definition, else, if $a_j^{\mathbf{x}} > 0$ for a $j \neq i$, then (11) is equal to

$$st_j \cdot \ldots \cdot (st_j + a_j^{\mathbf{x}} - 1),$$

and tends to zero, because of the first term of the product, as $t_j \to 0$. Therefore for each $\mathbf{x} \neq \overline{\mathbf{x}}^i$ (4) tends to zero as $t_i \to 1$. For $\overline{\mathbf{x}}^i$ we have

$$
\begin{aligned}
\lim_{t_i \to 1} E(\theta_i \,|\, \overline{\mathbf{x}}^i) &\overset{(3)}{=} \lim_{t_i \to 1} \frac{a_i^{\overline{\mathbf{x}}^i} + st_i}{N+s} \\
&= \frac{N+s}{N+s} = 1,
\end{aligned}
$$

$$
\begin{aligned}
\lim_{t_i \to 1} E(\theta_j \,|\, \overline{\mathbf{x}}^i) &\overset{(3)}{=} \lim_{t_j \to 0} \frac{a_j^{\overline{\mathbf{x}}^i} + st_j}{N+s} \\
&= \frac{0 + s \cdot 0}{N+s} = 0,
\end{aligned}
$$

$$\lim_{t_i \to 1} P(\overline{\mathbf{x}}^i) \overset{(11)}{=} \lim_{t_j \to 0} \frac{\prod_{r=1}^N (st_i + r - 1)}{\prod_{j=1}^N (s + j - 1)} = 1.$$

For $P(\mathbf{o} \,|\, \overline{\mathbf{x}}^i) \neq 0$ it follows

$$
\begin{aligned}
&\lim_{t_i \to 1} P(x_i \,|\, \mathbf{o}) \overset{(10)}{=} \\
&\overset{(10)}{=} \lim_{t_i \to 1} \frac{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot \frac{a_i^{\mathbf{x}} + st_i}{N+s}}{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})} \\
&= \lim_{t_i \to 1} \frac{P(\mathbf{o} \,|\, \overline{\mathbf{x}}^i) \cdot P(\overline{\mathbf{x}}^i) \cdot \frac{a_i^{\overline{\mathbf{x}}^i} + st_i}{N+s}}{P(\mathbf{o} \,|\, \overline{\mathbf{x}}^i) \cdot P(\overline{\mathbf{x}}^i)} = 1,
\end{aligned}
$$

and therefore the IDM produces vacuous upper predictive probabilities for the category x_i. For $P(\mathbf{o} \,|\, \overline{\mathbf{x}}^i) = 0$, $\lim_{t_i \to 1} P(x_i \,|\, \mathbf{o})$ is non vacuous because all $\mathbf{x} \neq \overline{\mathbf{x}}^i$ have

$$\lim_{t_i \to 1} E(\theta_j \,|\, \mathbf{x}) = \frac{a_i^{\mathbf{x}} + s \cdot 1}{N+s} < \frac{N+s}{N+s} = 1.$$

Because

$$P(\mathbf{o} \,|\, \overline{\mathbf{x}}^i) = \prod_{j=1, n_j > 0}^k \lambda_{ji}^{n_j}, \qquad (12)$$

[3] Without zero elements.

the condition $(12) \neq 0$ is satisfied iff $\lambda_{ji} \neq 0$ for each j such that $n_j > 0$.

Consider now another extreme value $\mathbf{t}$ for the parameters of the prior density, this time with $t_i \to 0$ and $t_j \not\to 0$ for each $j \neq i$. In this case all the dataset $\mathbf{x} \in \mathcal{X}^N$ with $a_i^{\mathbf{x}} > 0$ have $lim_{t_i \to 0} P(\mathbf{x}) = 0$ and all dataset with $a_i^{\mathbf{x}} = 0$ have $lim_{t_i \to 0} P(\mathbf{x}) \neq 0$. For each $\mathbf{x} \in \mathcal{X}^N$ with $a_i^{\mathbf{x}} = 0$ we have

$$\lim_{t_i \to 0} \frac{a_i^{\mathbf{x}} + st_i}{N+s} = \frac{0 + s \cdot 0}{N+s} = 0,$$

and therefore, if $P(\mathbf{o} \,|\, \mathbf{x}) \neq 0$ for at least a dataset with $a_i^{\mathbf{x}} = 0$, it follows that

$$\begin{aligned}
&\lim_{t_i \to 0} P(x_i \,|\, \mathbf{o}) = \\
&= \frac{\sum_{\mathbf{x} \in \mathcal{X}^N, a_i^{\mathbf{x}}=0} \overbrace{P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}^{\text{At least one} \not\to 0} \cdot \overbrace{\frac{a_i^{\mathbf{x}} + st_i}{N+s}}^{\to 0}}{\sum_{\mathbf{x} \in \mathcal{X}^N, a_i^{\mathbf{x}}=0} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})} \\
&= 0.
\end{aligned}$$

Therefore, if there is at least one dataset $\mathbf{x}$ with $a_i^{\mathbf{x}} = 0$ and $P(\mathbf{o} \,|\, \mathbf{x}) \neq 0$, the model produces lower probability 0 for the category x_i. For each j with $n_j > 0$ there is at least a r with $\lambda_{jr} \neq 0$ because all the rows in Λ are non-zero. If $n_i = 0$ we can therefore construct easily a dataset $\mathbf{x}$ with $a_i^{\mathbf{x}} = 0$ and $P(\mathbf{o} \,|\, \mathbf{x}) \neq 0$. We cannot construct such a dataset only if $n_i > 0$ and the unique element different from zero on the i-th row of the matrix Λ is $\lambda_{ii} = 1$. This concludes the proof of the theorem.

4.4 Examples

We illustrate the results with two examples in the binary case.

Example 1. *Consider a situation with $k = 2$, $s = 2$, $N = 2$ and an emission matrix*

$$\Lambda_\varepsilon = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix},$$

where $\varepsilon > 0$. Suppose that we have observed the dataset $\mathbf{o} = (x_1, x_1)$ and therefore the count $\mathbf{n} = (2, 0)$. The probabilities of the observed dataset given the different dataset of $\mathcal{X}^2$ are given by

$$\begin{aligned}
&P(\mathbf{o}|(x_1, x_1)) = (1-\varepsilon) \cdot (1-\varepsilon) > 0, \\
&P(\mathbf{o}|(x_1, x_2)) = (1-\varepsilon) \cdot \varepsilon > 0, \\
&P(\mathbf{o}|(x_2, x_1)) = (1-\varepsilon) \cdot \varepsilon > 0, \\
&P(\mathbf{o}|(x_2, x_2)) = \varepsilon \cdot \varepsilon > 0.
\end{aligned}$$

Calculating the posterior probability $P(x_1|\mathbf{o})$ using (10) we obtain

$$\begin{aligned}
&P(x_1|\mathbf{o}) = \\
&= \Bigg((1-\varepsilon) \cdot (1-\varepsilon) \cdot st_1(1+st_1) \cdot \frac{2+st_1}{2+s} + \\
&+ 2 \cdot (1-\varepsilon) \cdot \varepsilon \cdot st_1 \cdot st_2 \cdot \frac{1+st_1}{2+s} + \\
&+ \varepsilon \cdot \varepsilon \cdot st_2 \cdot (1+st_2) \cdot \frac{0+st_1}{2+s} \Bigg) \cdot \\
&\cdot \Bigg((1-\varepsilon) \cdot (1-\varepsilon) \cdot st_1(1+st_1) + \\
&+ 2 \cdot (1-\varepsilon) \cdot \varepsilon \cdot st_1 \cdot st_2 + \\
&+ \varepsilon \cdot \varepsilon \cdot st_2 \cdot (1+st_2) \Bigg)^{-1}.
\end{aligned}$$

It follows that

$$\lim_{t_1 \to 1} P(x_1|\mathbf{o}) = \frac{(1-\varepsilon)^2 \cdot s(1+s)}{(1-\varepsilon)^2 \cdot s(1+s)} = 1,$$

and

$$\lim_{t_1 \to 0} P(x_1|\mathbf{o}) = \frac{\varepsilon^2 \cdot s(1+s) \cdot 0}{\varepsilon^2 \cdot s(1+s)} = 0,$$

implying

$$\underline{P}(x_1|\mathbf{o}) = 0,\ \overline{P}(x_1|\mathbf{o}) = 1.$$

The same result holds for $P(x_2|\mathbf{o})$.

Remark 3. *The result of Example 1 holds for each positive, even very small, value of ε. The IDM with $\Lambda = I$ and the same $\mathbf{o}$ produces*

$$\begin{aligned}
\overline{P}(x_1 \,|\, \mathbf{o}) &= \frac{2+2}{2+2} = 1, \\
\underline{P}(x_1 \,|\, \mathbf{o}) &= \frac{2}{2+2} = 0.5, \\
\overline{P}(x_2 \,|\, \mathbf{o}) &= \frac{0+2}{2+2} = 0.5, \\
\underline{P}(x_2 \,|\, \mathbf{o}) &= \frac{0}{2+2} = 0.
\end{aligned}$$

It is evident that there absolutely no continuity between the result for $\Lambda = I$ and the results for $\Lambda = \Lambda_\varepsilon$, even for very small ε.

Example 2. *Suppose that we have observed a dataset* $\mathbf{o}$ *with corresponding counts* $\mathbf{n} = (12, 23)$ *and assume that the observation mechanism is characterized by the emission matrix*

$$\Lambda = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}.$$

Figure 1 displays the results for $P(x_1|\mathbf{o})$ obtained with the IDM for $s = 2$. It is interesting to remark that the problem of vacuous probabilities arises very near the boundaries of $\mathcal{T}$. In the first plot,where the function is plotted in the interval $t_1 \in [0, 1]$, it seems that $\overline{P}(x_1|\mathbf{o})$ is around 0.34. But if we look at the second plot, where the function is plotted more precisely in the interval $t_1 \in [0.99999, 1]$ we see clearly that $\overline{P}(x_1|\mathbf{o}) = 1$ as confirmed by theoretical results.

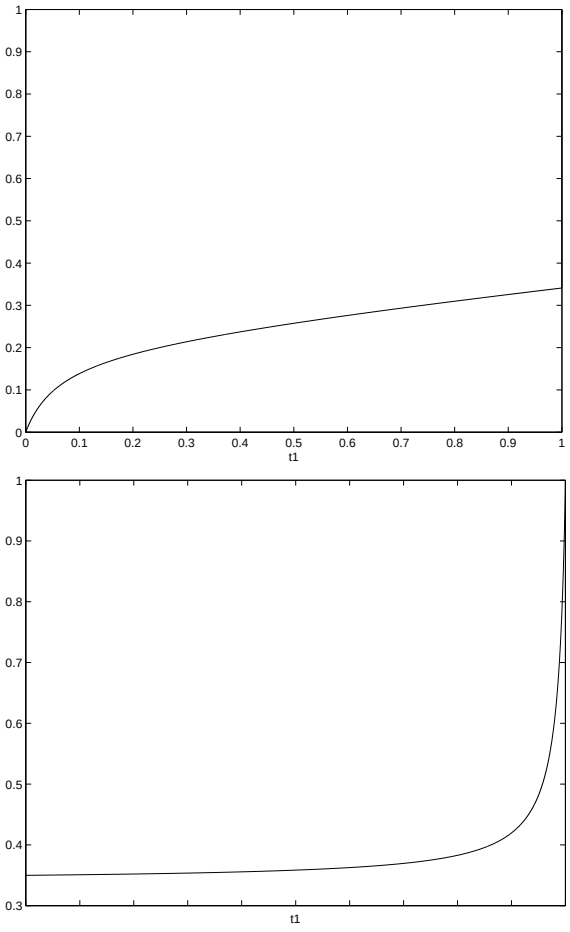

Figure 1: The function $P(x_1|\mathbf{o})$ for $t_1 \in [0, 1]$ and for $t_1 \in [0.99999, 1]$.

4.5 Discussion

Consider an observer with a unique extreme prior density measure $p(\vartheta) = dir(s, t)$ with $s > 0$ and $t_i \to 1$ for some $i \in \{1, \ldots, k\}$. The observer believes a-priori that the population is formed almost completely by individuals of category x_i. Now, if he observes an individual of category x_j and $\lambda_{ji} \neq 0$, then he will tend to believe that the individual observed is actually of category x_i and that there was a mistake in the observation mechanism. Only if $\lambda_{ji} = 0$ he has to rationally realize that observing something different from x_i can only be consistent with a strong modification of his prior beliefs.

Consider now an observer with $t_i \to 0$. Such an observer believes a-priori that there is almost no individuals of category x_i in the population. If he observes an individual of category x_i, he will believe that the actual category is another category x_j such that $t_j > 0$ and $\lambda_{ij} > 0$. The observer cannot believe that, only if $\lambda_{ij} = 0$ for all $i \neq j$.

When letting the prior density of an observer converge to a degenerate one, the model with imperfect observation mechanism produces trivial results because of the degeneration in the behavior of the observer. Such a feature arises only with prior densities that are extreme. To avoid vacuous inferences it would be sufficient to restrict the set of the

prior densities by excluding the problematic ones. However this is not compatible with the idea of complete prior ignorance: each restriction of the set of priors without a specific knowledge is arbitrary and hence impossible to motivate.

4.6 Inference on ξ

Consider the binary case ($k = 2$) with emission matrix

$$\Lambda_\varepsilon := \begin{pmatrix} 1-\varepsilon_i & \varepsilon_i \\ \varepsilon_i & 1-\varepsilon_i \end{pmatrix}, \qquad (13)$$

where $\varepsilon \neq 0.5$. The IDM with $k = 2$ is usually called *Imprecise Beta Model* (IBM) because the Dirichlet density measures with $k = 2$ are beta density measures (see (Walley 1991) and (Bernard 1996)). Consider the chances $\vartheta = (\theta_1, \theta_2)$ of the unobservable process X and the chances $\xi = (\xi_1, \xi_2)$ of the observable process O. Because the matrix (13) is non singular, we can reconstruct the values of ϑ starting from the values of ξ using the relation $\theta_i = \frac{\xi_i - \varepsilon}{1-2\varepsilon}$ and viceversa with $\xi_i = (1-2\varepsilon)\theta_i + \varepsilon$. Apparently, it should be possible to do inference on ξ using the observable values of O with the standard IBM and then reconstruct the values of ϑ. But this would contradict our previous result about vacuous probabilities. Actually, we show that our result is still valid, also from the point of view of the observable data. In particular we show that, at one side the application of the IBM on ξ ignoring the emission matrix produces sinless results, at the other side the application of the IBM on ξ taking correctly into account the emission matrix produces vacuous probabilities as the IBM on ϑ.

Proposition 2. *The inference on ξ with the standard IBM can produce sinless values for ϑ, i.e. negative values or values greater than 1.*

Proof. See appendix B. In fact the density measure of the parameters ξ is not a standard beta density. If we model our knowledge about the parameters ϑ with a $beta(s,t)$ density, then taking into consideration the emission matrix (13) we obtain for ξ a scaled $beta_{[\varepsilon,1-\varepsilon]}(s,\mathbf{t})$ density. Therefore the IBM on ξ should be performed using, as set of prior density measures, the set of all beta densities scaled on $[\varepsilon, 1-\varepsilon]$ with $\mathbf{t} \in \mathcal{T}$ and not the standard beta densities used usually in the IBM. But in this case following theorem holds:

Theorem 2. *The IBM on ξ, with, as set of prior densities, the set of all scaled beta densities described above, produces vacuous probabilities.*

Proof. See appendix C. We can conclude therefore that our result about vacuous probabilities is still valid also from the point of view of the observable data.

5 The binary case

In the previous section we have assumed an observation mechanism with known and constant emission matrix. In this section we study in detail the behavior of the IBM if the observation mechanism is not known, or when the observation mechanism changes with time, in order to generalize Theorem 1. We show that a crucial assumption is the possibility of a perfect observation mechanism. In particular we show that, if the observation mechanism varies over time but it it is surely never perfect, then the IBM produces vacuous predictive probabilities. At the other side we show that, if there is a probability, even small, that for an observation the observation mechanism is perfect, then the IBM produces non-vacuous probabilities. At the end of the section we illustrate this feature with a paradoxical example.

5.1 Vacuous probabilities in the binary case

We generalize theorem 1 to a situation where the observation mechanism is characterized by a non-constant, stochastically distributed emission matrix.

Corollary 2. *The IBM with observation mechanism defined by the emission matrix (13), where $\varepsilon \neq 0$, produces vacuous probabilities.*

Proof. This is a particular case of theorem 1. Now we allow the observation mechanism to vary over time, we obtain however the same result:

Theorem 3. *The IBM with observation mechanism for the i-th observation defined by the emission matrix*

$$\Lambda_{\varepsilon_i} := \begin{pmatrix} 1-\varepsilon_i & \varepsilon_i \\ \varepsilon_i & 1-\varepsilon_i \end{pmatrix}, \qquad (14)$$

where $\varepsilon_i \neq 0$ for each $i \in \{1, \ldots, N\}$ produces vacuous probabilities.

Proof. The proof is equal to the proof of theorem 1 except for the terms $P(\mathbf{o} \,|\, \mathbf{x})$ that contain $\varepsilon_1, \ldots, \varepsilon_N$ instead of a single ε.

Lemma 3 (Lebesgue Theorem). *Let $\{f_n\}$ be a series of functions on the domain A such that $f_n \to f$ pointwise. If for each n we have*

$$|f_n(x)| \leq \phi(x),$$

and

$$\int_A \phi(x)dx < \infty,$$

then

$$\lim_{n\to\infty} \int_A f_n(x)dx = \int_A f(x)dx.$$

In the following Theorem we allow the emission matrices to be unknown and we summarize our knowledge about ε_i with a continuous density measure. We obtain once more the same result.

Theorem 4. *Suppose that we want to perform predictive inference using the IBM and our observation mechanism for the i-th observation is defined by the emission matrix (14), where $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_N)$ is distributed according to a continuous density $f(\varepsilon)$ defined on $[0,1]^N$. Then, the IBM produces vacuous predictive probabilities.*

Proof. We know from Theorem 3 that given $\varepsilon_1, \ldots, \varepsilon_N \neq 0$ we have

$$\lim_{t_1\to 1} P(x_1 \,|\, \mathbf{o}, \varepsilon) = 1,$$

and

$$\lim_{t_1\to 1} P(x_2 \,|\, \mathbf{o}, \varepsilon) = 0,$$

for each $j \neq i$. We have

$$\lim_{t_1\to 1} P(x_1 \,|\, \mathbf{o}) =$$
$$= \lim_{t_1\to 1} \int_{[0,1]^N} P(x_1 \,|\, \mathbf{o}, \varepsilon) \cdot f(\varepsilon)d\varepsilon.$$

Furthermore

$$P(x_j \,|\, \mathbf{o}, \varepsilon) \cdot f(\varepsilon) \leq f(\varepsilon),$$

for any $j, \varepsilon, \mathbf{o}$ where

$$\int_{[0,1]^N} f(\varepsilon)d\varepsilon = 1.$$

Because of the continuity of f we know that $P(\varepsilon_i = 0) = 0$ for each i and each $\varepsilon \in [0,1]^2$. Applying Lemma 3 we conclude that

$$\lim_{t_1\to 1} P(x_1 \,|\, \mathbf{o}) =$$
$$= \lim_{t_1\to 1} \int_{[0,1]^N} P(x_1 \,|\, \mathbf{o}, \varepsilon) \cdot f(\varepsilon)d\varepsilon =$$
$$= \int_{[0,1]^N} \lim_{t_1\to 1} P(x_1 \,|\, \mathbf{o}, \varepsilon) \cdot f(\varepsilon)d\varepsilon =$$
$$= \int_{[0,1]^N} 1 \cdot f(\varepsilon)d\varepsilon = 1,$$

and, similarly,

$$\lim_{t_1\to 1} P(x_2 \,|\, \mathbf{o}) = 0.$$

Remark 4. *The result can be easily generalized to the k-dimensional case. It is sufficient to assume that $P(\lambda_{ij} = 0) = 0$ for each element in (5). This condition is satisfied by all continuous density measures defined on the components of the emission matrix.*

5.2 An alternative approach with relaxed assumptions

We have shown that we are unable to obtain non-vacuous predictive inferences with prior ignorance if we know that the observation process is not perfect. Vacuous probabilities arise because of the combination of two factors:

(i) Extreme, quasi-degenerate, prior densities.

(ii) Persistent doubt about the quality of observations.

therefore, in order to produce non-vacuous predictive probabilities, we can follow two approaches:

(i) We can restrict the set of prior densities according to some criteria. However, it is very difficult to do so, while maintaining a complete prior ignorance and without applying arbitrary criteria.

(ii) We can model the observation process allowing it to be perfect with some probability. In other words, without excluding a-priori the possibility of a perfect observation mechanism.

We follow the second approach. All results presented in the sequel for the binary case can be extended to the k-dimensional case. Given a N-dimensional vector of observations $\mathbf{o}$ we assume that some observations were made under a perfect observation mechanism, while the other ones have been obtained under an imperfect observation mechanism with emission matrix (13). Define a binary random variable E such that $P(\mathrm{E} = 1) = 1 - p$ and $P(\mathrm{E} = 0) = p$. For each observed dataset $\mathbf{o}$ there exists an unobservable vector $\mathbf{e}$ of length N consisting of independent realizations of random variable E. Set $\mathcal{E} := [0,1]^N$, $\mathcal{E}$ is equal to the set of all possible vectors $\mathbf{e}$ of length N. With

$e_i = 0$ ($e_i = 1$), we mean that the i-th observation is obtained under a perfect (imperfect) observation mechanism. Assume that

$$P(\vartheta, \mathbf{x}, \mathbf{o}, \mathbf{e}) = P(\mathbf{o} \,|\, \mathbf{e}, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \vartheta) \cdot P(\mathbf{e}) \cdot P(\vartheta). \tag{15}$$

From (15) it follows that

$$P(\mathbf{e} \,|\, \mathbf{o}) = P(\mathbf{e}). \tag{16}$$

In order to obtain predictive probabilities, we calculate

$$\begin{aligned} & P(x_i \,|\, \mathbf{o}) = E(\theta_i \,|\, \mathbf{o}) = \\ &= \int_\Theta \theta_i P(\vartheta \,|\, \mathbf{o}) d\vartheta = \\ &\overset{(7)}{=} \int_\Theta \theta_i \sum_{\mathbf{e} \in \mathcal{E}} P(\vartheta \,|\, \mathbf{o}, \mathbf{e}) P(\mathbf{e} \,|\, \mathbf{o}) d\vartheta = \\ &\overset{(16)}{=} \sum_{\mathbf{e} \in \mathcal{E}} \left(\int_\Theta \theta_i P(\vartheta \,|\, \mathbf{o}, \mathbf{e}) d\vartheta \right) \cdot P(\mathbf{e}) = \\ &= \sum_{\mathbf{e} \in \mathcal{E}} E(\theta_i \,|\, \mathbf{o}, \mathbf{e}) \cdot P(\mathbf{e}) = \\ &= \sum_{\mathbf{e} \in \mathcal{E}} P(x_i \,|\, \mathbf{o}, \mathbf{e}) \cdot P(\mathbf{e}), \end{aligned} \tag{17}$$

for $i = 1, 2$. Explicit calculations (see Appendix D) lead to the following result.

Theorem 5. *If $p > 0$, then the IBM produces non-vacuous inferences.*

Therefore the crucial assumption, in order to obtain non vacuous predictive inferences, is the possibility of a perfect observation mechanism with some positive probability.

5.3 A paradoxical example

We illustrate the above results with a paradoxical example. To this end we introduce two data generating processes.

Process 1: In a first step a random variable X with values in $\mathcal{X}^2$ and probability $\vartheta = \{\theta_1, \theta_2\}$ is generated. After X has been generated, a random variable O is generated, such that

$$\begin{aligned} &P(\mathrm{O} = x_1 \,|\, \mathrm{X} = x_1) = 1 - \varepsilon, \\ &P(\mathrm{O} = x_1 \,|\, \mathrm{X} = x_2) = \varepsilon, \\ &P(\mathrm{O} = x_2 \,|\, \mathrm{X} = x_1) = \varepsilon, \\ &P(\mathrm{O} = x_2 \,|\, \mathrm{X} = x_2) = 1 - \varepsilon. \end{aligned}$$

Process 2: In a first step a random variable X with values in $\mathcal{X}^2$ and probability $\vartheta = \{\theta_1, \theta_2\}$ is generated. Independently of X, a binary random variable E is then generated such that $P(\mathrm{E} = 0) = 1 - \varepsilon$ and $P(\mathrm{E} = 1) = \varepsilon$ where $\varepsilon > 0$. After X and E have been generated, a further random variable O is drawn, such that $\mathrm{O} = \mathrm{X}$ if $\mathrm{E} = 0$. Else, when $\mathrm{E} = 1$, if $\mathrm{X} = x_1$, then $\mathrm{O} = x_2$, if $\mathrm{X} = x_2$, then $\mathrm{O} = x_1$.

The paradoxon

The processes 1 and 2 produce exactly the same probabilities for the observed data $\mathbf{o}$. In fact for both processes

$$\begin{aligned} &P(\mathrm{O} = x_1) = (1 - \varepsilon)\theta_1 + \varepsilon\, \theta_2, \\ &P(\mathrm{O} = x_2) = (1 - \varepsilon)\theta_2 + \varepsilon\, \theta_1. \end{aligned}$$

It follows that the two processes are indistinguishable from the point of view of the observer. However, the first process is a process of the type described in section 4.1. Therefore the IBM produces vacuous predictive probabilities. The second process belongs to the category of processes described in section 5.2. In such a setting the IBM produce non-vacuous predictive probabilities. In particular, the possibility of producing non vacuous inferences depends not on the form of the distribution of observed data, but rather on some structural assumption on the process generating the observations. The difference between the two models lies in the assumptions. In the first model we assume the existence of perfect observation mechanisms, whereas the second one assumes only the existence of perfect observations. The assumptions of the second model are therefore much weaker than those of the first model.

6 Conclusions

Acknowledgements

Alberto Piatti and Fabio Trojani gratefully acknowledge the financial support of the Swiss National Science Foundation (NCCR FINRISK).

A Proof of Proposition 1

Proof. The Gamma function satisfies the property

$$\Gamma(x + 1) = x \cdot \Gamma(x).$$

Lemma.

$$\Gamma(s^*) \quad = \quad \prod_{i=1}^{N}(s+i-1)\cdot\Gamma(s).$$

Proof of the Lemma. If $N=0$ then $\Gamma(s^*)=\Gamma(s)$. Now assume that for $N-1$ the following equality holds

$$\Gamma(N-1+s)=\prod_{i=1}^{N-1}(s+i-1)\cdot\Gamma(s),$$

then

$$\begin{aligned}&\Gamma(N+s)=(N-1+s)\cdot\Gamma(N-1+s)=\\&=(N-1+s)\cdot\prod_{i=1}^{N-1}(s+i-1)\cdot\Gamma(s)=\\&=\prod_{i=1}^{N}(s+i-1)\cdot\Gamma(s).\end{aligned}$$

Lemma.

$$\Gamma(s^*\cdot t_j^*) \quad = \quad \prod_{i=1}^{a_j}(st_j+i-1)\cdot\Gamma(st_j).$$

Proof of the Lemma. If $a_j=0$ then $\Gamma(s^*\cdot t_j^*)=\Gamma(st_j)$. Now assume that for $a_j=n-1$ the following equality holds

$$\begin{aligned}&\Gamma(s^*\cdot t_j^*)=\Gamma(a_j+st_j)=\Gamma(n-1+st_j)=\\&=\prod_{i=1}^{n-1}(st_j+i-1)\cdot\Gamma(st_j).\end{aligned}$$

then with $a_j=n$

$$\begin{aligned}&\Gamma(s^*\cdot t_j^*)=\Gamma(a_j+st_j)=\\&=\Gamma(n+st_j)=\\&=(n-1+st_j)\cdot\Gamma(n-1+st_j)=\\&=(n-1+st_j)\cdot\prod_{i=1}^{n-1}(st_j+i-1)\cdot\Gamma(st_j)=\\&=\prod_{i=1}^{n}(st_j+i-1)\cdot\Gamma(st_j)=\\&=\prod_{i=1}^{a_j}(st_j+i-1)\cdot\Gamma(st_j).\end{aligned}$$

It follows that

$$\begin{aligned}&dir(s^*,\mathbf{t}^*)=\\&=\frac{\Gamma(s^*)}{\prod_{j=1}^{k}\Gamma(s^*t_j^*)}\cdot\prod_{j=1}^{k}\theta_j^{s^*t_j^*-1}=\\&=\frac{\Gamma(s^*)}{\prod_{j=1}^{k}\Gamma(s^*t_j^*)}\cdot\prod_{j=1}^{k}\theta_j^{a_j}\cdot\prod_{j=1}^{k}\theta_j^{st_j-1}=\\&=\frac{\prod_{i=1}^{N}(s+i-1)\cdot\Gamma(s)}{\prod_{j=1}^{k}\cdot(\prod_{i=1}^{a_j}(st_j+i-1))\cdot\Gamma(st_j)}\cdot\\&\quad\cdot\prod_{j=1}^{k}\theta_j^{a_j}\cdot\prod_{j=1}^{k}\theta_j^{st_j-1}=\\&=\frac{\prod_{i=1}^{N}(s+i-1)}{\prod_{j=1}^{k}\cdot\prod_{i=1}^{a_j}(st_j+i-1)}\cdot\\&\quad\cdot\left(\prod_{j=1}^{k}\theta_j^{a_j}\right)\cdot\frac{\Gamma(s)}{\prod_{j=1}^{k}\Gamma(st_j)}\cdot\prod_{j=1}^{k}\theta_j^{st_j-1}=\\&=\frac{\prod_{i=1}^{N}(s+i-1)}{\prod_{j=1}^{k}\cdot\prod_{i=1}^{a_j}(st_j+i-1)}\cdot\\&\quad\cdot\left(\prod_{j=1}^{k}\theta_j^{a_j}\right)\cdot dir(s,\mathbf{t}),\end{aligned}$$

and therefore

$$\begin{aligned}&\prod_{j=1}^{k}\theta_j^{a_j}\cdot dir(s,\mathbf{t})=\\&=\frac{\prod_{j=1}^{k}\cdot\prod_{i=1}^{a_j}(st_j+i-1)}{\prod_{i=1}^{N}(s+i-1)}\cdot dir(s^*,\mathbf{t}^*).\end{aligned}$$

B Proof of Proposition 2

Suppose that we have observed a dataset $\mathbf{o}$ with corresponding counts $\mathbf{n}=(n_1,n_2)$. If we perform the standard IBM on this dataset then we obtain

$$\xi_i\in\left[\frac{n_i}{N+s};\frac{n_i+s}{N+s}\right]$$

and therefore

$$\theta_i\in\left[\frac{n_i-\varepsilon(N+s)}{(N+s)(1-2\varepsilon)};\frac{n_i+s-\varepsilon(N+s)}{(N+s)(1-2\varepsilon)}\right].$$

But if $n_i<\varepsilon(N+s)$ then $\frac{n_i-\varepsilon(N+s)}{(N+s)(1-2\varepsilon)}<0$ and if $n_i<\varepsilon(N+s)-s$ then $\frac{n_i+s-\varepsilon(N+s)}{(N+s)(1-2\varepsilon)}<0$. At

the other side if $n_i > (1-\varepsilon)(N+s)$ then we find values for θ_i that are greater than 1. Therefore the standard IBM can produce sinless results.

Substituting $\theta_i = \frac{\xi_i - \varepsilon}{1-2\varepsilon}$ in the $beta(s,t)$ density $C \cdot \theta_1^{st_1-1} \cdot \theta_2^{st_2-1}$ of ϑ defined on $[0;1]$ we obtain for ξ the density

$$\frac{C}{1-2\varepsilon}\left(\frac{\xi_1-\varepsilon}{1-2\varepsilon}\right)^{st_1-1}\left(\frac{\xi_2-\varepsilon}{1-2\varepsilon}\right)^{st_2-1}, \quad (18)$$

defined on the interval $[\varepsilon, 1-\varepsilon]$. This density is called a *scaled beta density* and referred to with $beta_{[\varepsilon,1-\varepsilon]}(s,\mathbf{t})$. The first moment of a scaled beta density are given by $E(\theta_i) = (1-2\varepsilon)t_i + \varepsilon$. For scaled beta densities a result similar to proposition 1 holds, i.e. if $\mathbf{o}$ is a dataset with counts $\mathbf{n} = (n_1, n_2)$ then

$$\left(\frac{\xi_1-\varepsilon}{1-2\varepsilon}\right)^{n_1}\left(\frac{\xi_2-\varepsilon}{1-2\varepsilon}\right)^{n_2} \cdot beta(s,\mathbf{t}) =$$
$$= \frac{\prod_{i=1}^{n_1}(n_1+st_1-i)\cdot\prod_{j=1}^{n_2}(n_2+st_2-j)}{\prod_{i=1}^{N}(N+s-i)}\cdot$$
$$\cdot\, beta(s^{\mathbf{o}}, \mathbf{t}^{\mathbf{o}}).$$

C Proof of Theorem 2

D Proof of Theorem 5

) *Proof.* Suppose that we have observed a dataset $\mathbf{o}$. For given vector $\mathbf{e}$ and observed dataset $\mathbf{o}$ we introduce following notations:

- N_e is the number of components equal to one in the vector $\mathbf{e}$.
- n_{1e} is the number of ones in the vector $\mathbf{e}$ that correspond to an x_1 in dataset $\mathbf{o}$.
- n_{2e} is the number of ones in the vector $\mathbf{e}$ that correspond to an x_2 in dataset $\mathbf{o}$.
- $n_e := (n_{1e}, n_{2e})$.

From (15) it follows that

$$P(\mathbf{e} \,|\, \vartheta) = P(\mathbf{e}), \quad (19)$$
$$p(\vartheta \,|\, \mathbf{e}) = p(\vartheta), \quad (20)$$
$$P(\mathbf{e} \,|\, \mathbf{x}) = P(\mathbf{e}), \quad (21)$$

and therefore

$$P(\mathbf{o} \,|\, \mathbf{e}, \mathbf{x}) \overset{(8)}{=} \frac{P(\mathbf{o}, \mathbf{e} \,|\, \mathbf{x})}{P(\mathbf{e} \,|\, \mathbf{x})} \overset{(21)}{=} \frac{P(\mathbf{o}, \mathbf{e} \,|\, \mathbf{x})}{P(\mathbf{e})}. \quad (22)$$

We have that

$$P(\mathbf{o} \,|\, \mathbf{e}, \vartheta) \overset{(8)}{=} \frac{P(\mathbf{o}, \mathbf{e} \,|\, \vartheta)}{P(\mathbf{e} \,|\, \vartheta)} \overset{(19)}{=} \frac{P(\mathbf{o}, \mathbf{e} \,|\, \vartheta)}{P(\mathbf{e})} =$$
$$\overset{(7)}{=} \sum_{\mathbf{x}\in\mathcal{X}} \frac{P(\mathbf{o}, \mathbf{e} \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \vartheta)}{P(\mathbf{e})} =$$
$$\overset{(22)}{=} \sum_{\mathbf{x}\in\mathcal{X}} P(\mathbf{o} \,|\, \mathbf{e}, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \vartheta),$$

and therefore

$$p(\mathbf{o}, \vartheta, \mathbf{e}) = \sum_{\mathbf{x}\in\mathcal{X}} P(\vartheta, \mathbf{x}, \mathbf{o}, \mathbf{e}) =$$
$$= \sum_{\mathbf{x}\in\mathcal{X}} P(\mathbf{o} \,|\, \mathbf{e}, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \vartheta) \cdot P(\mathbf{e}) \cdot p(\vartheta) =$$
$$= P(\mathbf{o} \,|\, \mathbf{e}, \vartheta) \cdot P(\mathbf{e}) \cdot p(\vartheta).$$

thus

$$p(\vartheta \,|\, \mathbf{o}, \mathbf{e}) \overset{(8)}{=} \frac{p(\mathbf{o}, \vartheta, \mathbf{e})}{P(\mathbf{o}, \mathbf{e})} =$$
$$= \frac{P(\mathbf{o} \,|\, \mathbf{e}, \vartheta) \cdot P(\mathbf{e}) \cdot p(\vartheta)}{P(\mathbf{o} \,|\, \mathbf{e}) \cdot P(\mathbf{e})} =$$
$$= \frac{P(\mathbf{o} \,|\, \mathbf{e}, \vartheta) \cdot p(\vartheta)}{P(\mathbf{o} \,|\, \mathbf{e})} =$$
$$\overset{(20)+(1)}{=} \frac{P(\mathbf{o} \,|\, \mathbf{e}, \vartheta) \cdot p(\vartheta)}{\int_\Theta P(\mathbf{o} \,|\, \mathbf{e}, \vartheta) \cdot p(\vartheta) d\vartheta},$$

and therefore it follows from (17) that

$$P(x_i \,|\, \mathbf{o}, \mathbf{e}) = \frac{\int_\Theta \theta_i P(\mathbf{o} \,|\, \mathbf{e}, \vartheta) \cdot p(\vartheta) d\vartheta}{\int_\Theta P(\mathbf{o} \,|\, \mathbf{e}, \vartheta) \cdot p(\vartheta) d\vartheta}. \quad (23)$$

For each $\mathbf{e} \in \mathcal{E}$ the function $p(\vartheta \,|\, \mathbf{o}, \mathbf{e})$ is a likelihood function of the form

$$P(\mathbf{o} \,|\, \mathbf{e}, \vartheta) =$$
$$= \theta_1^{n_1 - n_{1e}} \cdot \theta_2^{n_2 - n_{2e}} \cdot ((1-\varepsilon)\cdot\theta_1 + \varepsilon\cdot\theta_2)^{n_{1e}} \cdot$$
$$\cdot ((1-\varepsilon)\cdot\theta_2 + \varepsilon\cdot\theta_1)^{n_{2e}} =$$
$$= \sum_{i=0}^{n_{1e}} \sum_{j=0}^{n_{2e}} \binom{n_{1e}}{i} \cdot \binom{n_{2e}}{j} \cdot$$
$$\cdot (1-\varepsilon)^{N_e - (i+j)} \cdot \varepsilon^{(i+j)} \cdot \theta_1^{n_1 - i + j} \cdot \theta_2^{n_2 - j + i} =$$
$$=: \sum_{i=0}^{n_{1e}} \sum_{j=0}^{n_{2e}} c(i, j, n_{1e}, n_{2e}) \cdot \theta_1^{n_1 - i + j} \cdot \theta_2^{n_2 - j + i}.$$

From Proposition 1 we know that

$$\theta_1^{n_1-i+j} \cdot \theta_2^{n_2-j+i} \cdot beta(s,t_1,t_2) =$$
$$= \frac{\prod_{a=1}^{n_1-i+j}(st_1+a-1)\cdot\prod_{b=1}^{n_2-i+j}(st_2+b-1)}{\prod_{c=1}^{N}(s+c-1)} \cdot$$
$$\cdot beta(s^{\mathbf{o}}, \mathbf{t}^{\mathbf{o}}_{1ij}, \mathbf{t}^{\mathbf{o}}_{2ij}) =$$
$$=: \frac{C(i,j,n_e)}{\prod_{c=1}^{N}(s+c-1)} \cdot beta(s^{\mathbf{o}}, \mathbf{t}^{\mathbf{o}}_{1ij}, \mathbf{t}^{\mathbf{o}}_{2ij}),$$

where

$$s^{\mathbf{o}} = N + s,$$
$$t^{\mathbf{o}}_{1ij} = \frac{n_1 - i + j + st_1}{N+s},$$

and

$$t^{\mathbf{o}}_{2ij} = \frac{n_2 - j + i + st_2}{N+s}.$$

Therefore using (23) we have

$$E(\theta_1 \,|\, \mathbf{o}, \mathbf{e}) =$$
$$= \frac{\sum_{i=0}^{n_{1e}}\sum_{j=0}^{n_{2e}} c(i,j,n_e)\cdot C(i,j,n_e)\cdot \frac{n_1-i+j+st_1}{N+s}}{\sum_{i=0}^{n_{1e}}\sum_{j=0}^{n_{2e}} c(i,j,n_e)\cdot C(i,j,n_e)},$$

and

$$E(\theta_2 \,|\, \mathbf{o}, \mathbf{e}) =$$
$$= \frac{\sum_{i=0}^{n_{1e}}\sum_{j=0}^{n_{2e}} c(i,j,n_e)\cdot C(i,j,n_e)\cdot \frac{n_2-j+i+st_2}{N+s}}{\sum_{i=0}^{n_{1e}}\sum_{j=0}^{n_{2e}} c(i,j,n_e)\cdot C(i,j,n_e)}.$$

Consider a vector $\mathbf{e}$ with $n_1 > n_{1e}$ and $n_2 > n_{2e}$ and the corresponding likelihood function $P(\mathbf{o} \,|\, \vartheta, \mathbf{e})$. We know that

$$E(\theta_1 \,|\, \mathbf{o}, \mathbf{e}) =$$
$$= \frac{\sum_{i=0}^{n_{1e}}\sum_{j=0}^{n_{2e}} c(i,j,n_e)\cdot C(i,j,n_e)\cdot \frac{n_1-i+j+st_1}{N+s}}{\sum_{i=0}^{n_{1e}}\sum_{j=0}^{n_{2e}} c(i,j,n_e)\cdot C(i,j,n_e)} \leq$$
$$\leq \frac{n_1 - 0 + n_{2e} + st_1}{N+s} <$$
$$< \frac{n_1 + n_2 + s\cdot 1}{N+s} = 1,$$

and in the same manner

$$E(\theta_1 \,|\, \mathbf{o}, \mathbf{e}) \geq$$
$$\geq \frac{n_1 - n_{1e} + st_1}{N+s} >$$
$$> \frac{n_1 - n_1 + s\cdot 0}{N+s} = 0.$$

Therefore each likelihood function of this type leads to non vacuous predictive probabilities for x_1. For x_2 we obtain the same result. Because the predictive probabilities that we obtain using the IDM are a weighted average of the predictions obtained with each single likelihood the model produce vacuous predictions only if the probabilities of all $\mathbf{e}$ with $n_1 > n_{1e}$ and $n_2 > n_{2e}$ are equal to zero. However, since

$$P(\mathbf{e}) = p^{N-N_e}(1-p)^{N_e},$$

and in this case $N - N_e > 0$, this holds only if $p = 0$.

References

[1] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *J. R. Statistic. Soc. B* , 58(1): 3-57, 1996.

[2] P. Walley. *Statistical Reasoning with Imprecise Probability.* Chapman and Hall, New York, 1991.

[3] S. Kotz, N. Balakrishnan, N. L. Johnson. *Continuous Multivariate Distributions, Volume 1: Models and Applications.* Wiley series in Probability and Statistics, New York, 2000.