Limits of Learning from Imperfect Observations under Prior Ignorance: the Case of the Imprecise Dirichlet Model

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

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

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

Abstract

Consider a relaxed multinomial setup, in which there may be mistakes in observing the outcomes of the process—this is often the case in real applications. What can we say about the next outcome if we start learning about the process in conditions of prior ignorance? To answer this question we extend the imprecise Dirichlet model to the case of imperfect observations and we focus on posterior predictive probabilities for the next outcome. The results are very surprising: the posterior predictive probabilities are vacuous, irrespectively of the amount of observations we do, and however small is the probability of doing mistakes. In other words, the imprecise Dirichlet model cannot help us to learn from data when the observational mechanism is imperfect. This result seems to rise a serious question about the use of the imprecise Dirichlet model for practical applications, and, more generally, about the possibility to learn from imperfect observations under prior ignorance.

Keywords. Predictive Bayesian Inference, imprecise Dirichlet model, Vacuous Predictive Probabilities, Imperfect Observational Mechanism.

1 Introduction

Consider the basic multinomial setup: an unknown process produces a sequence of symbols, from a finite alphabet, in an identically and independently distributed way. What is the probability of the next symbol produced? Walley's *imprecise Dirichlet model* (IDM) [7] offers an appealing solution to the predictive problem: it yields lower and upper probabilities of the next symbol that are initially vacuous and that converge to a precise probability as the sequence grows. The IDM can be regarded as a generalization of Bayesian inference to imprecise probability (Sect. 2), originated by the attempt to model prior ignorance about the process in an objective-minded way. The IDM is an important model as it yields credible inferences under prior ignorance, and because the multinomial setup is an abstraction of many important real problems. The IDM has indeed attracted considerable attention in the recent years; see, for example, the application of the IDM to classification [9, 11], nonparametric inference [2], robust estimation [4], analysis of contingency tables [3], discovery of dependency structures [10], and game theory [6].

But in real problems there is a, perhaps very small, probability of doing mistakes in the process of observing the sequence. It seems therefore worth relaxing the basic multinomial setup in order to consider the occurrence of *imperfect observations*, as in Section 3. We imagine a two-steps process to this extent: a multinomial process produces so-called *ideal symbols* from the alphabet, that we cannot observe; a subsequent *observational mechanism* takes the ideal symbols and produces the so-called *actual symbols*, which we do observe. The more accurate the observational mechanism, the more the ideal sequence will coincide with the actual sequence, and vice versa. But in any case, we assume that there exists a non-zero probability of mistake: the probability that the observational mechanism turns an ideal symbol into a different symbol of the alphabet.

We are interested in the following problem: can we compute the probability of the next ideal symbol, starting in a state of prior ignorance and observing only the actual sequence? To answer this question, we model prior ignorance at the ideal level with the IDM, and combine it with the imperfect observational mechanism at the actual level. The overall model generalizes the IDM, which is recovered in the case the probability of mistake is set to zero.

The outcome of the newly created model in Section 3.2 is very surprising: the predictive probabilities of the next ideal symbol are vacuous, irrespectively of the amount of symbols in the actual sequence, and of the accuracy of the observational mechanism! In other words, the model tells that it is not possible to learn with prior ignorance and an imperfect observational mechanism, no matter how small is the probability of error—provided that it is not zero, as in IDM. In the attempt to attack the vacuity problem we consider a weaker model for the observational mechanism: in Section 3.5 we assume that the probability of mistake, rather than being a constant, lies between 0 and 1 according to some distribution. The situation is unchanged: the probabilities are vacuous whatever precise distribution we choose.
This strong kind of discontinuity seems to rise a serious question about the IDM: what is the meaning of using the IDM for real problems? Indeed, the result seems to tell us that we cannot use the IDM as an approximation to more realistic models that admit the possibility of an imperfect observational mechanisms, just because the transition between these and the IDM is not at all continuous. One might say that this does not need to be a serious problem, as in the real world we are only concerned with actual symbols, rather than ideal ones. But in Section 4 it turns out that even the probabilities of the next actual symbol are vacuous for any length of the observed sequence and any accuracy of the observational mechanism.

2 The Imprecise Dirichlet Model

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

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

We define a random variable X with values in $\mathcal{X}$ which consists in drawing an individual at random from the population. Clearly the chance that $\mathrm{X} = x_i$ is θ_i. Our problem is to predict the probability of drawing an individual of type x_i from a population of unknown chances θ after having observed N independent random draws and starting from prior ignorance. Having observed a dataset $\mathbf{x}$, we can summarize the observation with the counts $\mathbf{a} = (a_1, \dots, a_k)$ where a_i is the number of individuals of type x_i observed in the dataset $\mathbf{x}$ and with $\sum_{i=1}^{k} a_i = N$. For given θ, the probability of observing a dataset $\mathbf{x}$ with counts $\mathbf{a}$ given θ is equal to $P(\mathbf{x} \mid \theta) = \theta_1^{a_1} \cdots \theta_k^{a_k}$. In this section we assume that each individual in the population is perfectly observable, i.e., the observer can determine the exact category of each individual without committing mistakes, and we solve our problem using the standard imprecise Dirichlet model.

2.1 Bayesian Inference and Dirichlet Prior Density

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

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

provided that

$$P(\mathbf{x}) = \int_{\Theta} P(\mathbf{x}|\theta) p(\theta) d\theta \neq 0,$$

where $p(\theta)$ is some density measure on Θ. The probability measure $P(\mathbf{x}|\theta)$ is called the *likelihood*, $p(\theta)$ is called the *prior density* and $p(\theta|\mathbf{x})$ is called the *posterior density*. Bayesian inference enables us to update our confidence on θ given the data by representing it as $P(\theta \mid \mathbf{x})$. Bayesian inference relies on the specification of a prior density on Θ. A common choice of prior 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 expression*

$$dir(s, \mathbf{t})(\theta) := \frac{\Gamma(s)}{\prod_{i=1}^{k} \Gamma(st_i)} \prod_{i=1}^{k} \theta_i^{st_i - 1},$$

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

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

[1] The symbol ':=' denotes a definition.

We recall first some important properties of Dirichlet densities.

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

Proof. See [5]. □

Remark 1. *In a multinomial setting we have*

$$P(x_i) = \int_\Theta P(x_i \,|\, \theta) \cdot p(\theta) d\theta = \\ = \int_\Theta \theta_i \cdot p(\theta) d\theta = E(\theta_i).$$

In particular, if $p(\theta)$ is a $dir(s, \mathbf{t})$ density, $P(x_i) = E(\theta_i) = t_i$.

Proposition 1. *Consider a dataset $\mathbf{x}$ with counts $\mathbf{a} = (a_1, \ldots, a_k)$. 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 $t_j^{\mathbf{x}} := \frac{a_j + st_j}{N+s}$. When $a_j = 0$ we set $\prod_{i=1}^{0} (st_j + i - 1) := 1$, for each $0 < t_j < 1$, by definition.

Remark 2. *Using a $dir(s, \mathbf{t})$ density measure as prior in a Bayesian learning problem with multinomial data we have $p(\theta) = dir(s, \mathbf{t})$ and*

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

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

$$P(x_i \,|\, \mathbf{x}) = t_i^{\mathbf{x}} = \frac{a_i + st_i}{N + s}. \qquad (3)$$

Moreover, comparing (1) with the equality of Proposition 1, we conclude that

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

2.2 The Imprecise Dirichlet Model

The Imprecise Dirichlet Model (IDM) (see [1] and [7]) is a model that generalizes Bayesian learning from multinomial data to the case when there is prior near-ignorance about θ. Prior ignorance about θ is modeled using the set of all the Dirichlet densities $dir(s, \mathbf{t})$ for a fixed s and all $\mathbf{t}$ in $\mathcal{T}$; that is, the IDM uses a set of prior densities instead of a single prior. The probability of each category x_i a priori is vacuous, i.e., $P(x_i) \in [\inf_{\mathcal{T}} t_i, \sup_{\mathcal{T}} t_i] = [0, 1]$. Prior ignorance is therefore modeled by 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 posteriors. Let now $s > 0$ be given and consider the set of prior densities $\mathcal{M}_s := \{dir(s, \mathbf{t}) \,|\, \mathbf{t} \in \mathcal{T}\}$. Suppose that we observe the dataset $\mathbf{x}$ with corresponding counts $\mathbf{a} = (a_1, \ldots, a_k)$. Then, the set of resulting posterior densities follows from Proposition 1 and is given by

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

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 3. *The upper and lower posterior predictive probabilities of a category x_i in the IDM 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.*

Remark 4. *The IDM with $k = 2$ is usually called* Imprecise Beta Model *(IBM), because the Dirichlet densities with $k = 2$ are beta densities (see [1] and [8]).*

3 The Imprecise Dirichlet Model with Imperfect Observational Mechanism

The standard IDM was originally defined for perfect observational mechanisms. But, in practice, there is always a (perhaps small) probability of making mistakes during the observational process. Often, if this probability is small, one assumes that the data are perfectly observable in order to use a simple model; doing so, one implicitly assumes that there is a sort of continuity between models with perfectly observable data and models with small probability of errors in the observations. In this section, our aim is to generalize the IDM to the case of imperfect observational mechanisms, and construct posterior predictive probabilities in order to verify if the implicit assumption described above is acceptable in practice. We model our imperfect observational mechanism with 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 from X. We define the chances $\lambda_{ij} := P(\mathrm{O} = x_i \,|\, \mathrm{X} = x_j)$. 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 $\xi = (\xi_1, \ldots, \xi_k)$ of the random variable O are given by

$$\xi_i = \sum_{j=1}^{k} \lambda_{ij} \cdot \theta_j. \tag{6}$$

Matrix Λ is stochastic, that is in each column the elements sum 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 $\mathcal{X}$. Consider a dataset $\mathbf{o}$ generated by the above two-step model. For each dataset $\mathbf{o}$ generated at the actual level and composed by realizations of the random variable O, there exists at the ideal level an unobservable dataset $\mathbf{x}$, of realizations of X, such that $\mathbf{o}$ was generated from $\mathbf{x}$ by the observational mechanism. Knowing $\mathbf{x}$, makes $\mathbf{o}$ not to depend on the chances θ of X. We can therefore summarize the two step model with the independence assumption

$$p(\mathbf{o}, \mathbf{x}, \theta) = P(\mathbf{o} \,|\, \mathbf{x}) P(\mathbf{x} \,|\, \theta) p(\theta). \tag{7}$$

3.1 The IDM with Imperfect Observational Mechanism

We use now the above two-step model to generalize the IDM to the case of imperfect observational mechanism. We begin calculating the posterior predictive probabilities for a given prior.

Lemma 2. *Suppose that we have observed a dataset* $\mathbf{o}$ *and we construct the posterior predictive probabilities* $p(\mathrm{X} = x_i \,|\, \mathbf{o})$ *using Bayes rule and a prior* $dir(s, \mathbf{t})$. *Then*

$$P(\mathrm{X} = x_i \,|\, \mathbf{o}) = \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot \frac{a_i^{\mathbf{x}} + st_i}{N+s}}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}. \tag{8}$$

Proof.

$$\begin{aligned} p(\theta \,|\, \mathbf{o}) &= \sum_{\mathbf{x} \in \mathcal{X}^N} p(\theta, \mathbf{x} \,|\, \mathbf{o}) = \\ &= \sum_{\mathbf{x} \in \mathcal{X}^N} p(\theta \,|\, \mathbf{x}, \mathbf{o}) \cdot P(\mathbf{x} \,|\, \mathbf{o}) = \\ &\overset{(7)}{=} \sum_{\mathbf{x} \in \mathcal{X}^N} p(\theta \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \mathbf{o}) = \\ &= \sum_{\mathbf{x} \in \mathcal{X}^N} \frac{P(\mathbf{x} \,|\, \theta) \cdot dir(s, \mathbf{t})}{P(\mathbf{x})} \cdot \frac{P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}{P(\mathbf{o})} = \\ &= \sum_{\mathbf{x} \in \mathcal{X}^N} \frac{P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \theta) \cdot dir(s, \mathbf{t})}{P(\mathbf{o})} = \\ &= \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \theta) \cdot dir(s, \mathbf{t})}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}. \end{aligned}$$

This is possible if $P(\mathbf{x}) > 0$ and $P(\mathbf{o}) > 0$. 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 that $P(\mathbf{o}) > 0$. From Remark 2 we have $P(\mathbf{x} \,|\, \theta) \cdot dir(s, \mathbf{t}) = P(\mathbf{x}) \cdot dir(s^{\mathbf{x}}, \mathbf{t}^{\mathbf{x}})$. Therefore,

$$P(\theta \,|\, \mathbf{o}) = \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot dir(s^{\mathbf{x}}, \mathbf{t}^{\mathbf{x}})}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})},$$

which is a convex combination of Dirichlet density measures, and, using (3), we obtain

$$\begin{aligned} &P(\mathrm{X} = x_i \,|\, \mathbf{o}) = \\ &= \frac{\int_\Theta \theta_i \cdot \sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot dir(s^{\mathbf{x}}, \mathbf{t}^{\mathbf{x}}) d\theta}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})} = \\ &= \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot \int_\Theta \theta_i \cdot dir(s^{\mathbf{x}}, \mathbf{t}^{\mathbf{x}}) d\theta}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})} = \\ &\overset{(3)}{=} \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot \frac{a_i^{\mathbf{x}} + st_i}{N+s}}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}. \end{aligned}$$

□

If we consider now each prior density in the set $\mathcal{M}_s$ and we calculate the posterior predictive probabilities $P(\mathrm{X} = x_i \,|\, \mathbf{o})$ using (8) we obtain a generalization of the IDM to the case of imperfect observational mechanism. It is interesting to remark that, in this case, the set of posterior densities consists of convex combinations of Dirichlet density measures and not of Dirichlet densities as in the IDM with perfect observational mechanism.

3.2 Vacuous Predictive Probabilities

In this section we study the behavior of the above generalization of the IDM in order to compare it with the standard IDM. The results are surprising: we show that there is a drastic discontinuity between the results obtained with the IDM with perfect observational mechanism and those obtained assuming an imperfect observational mechanism. In particular, the IDM with an emission matrix without zero elements produces vacuous predictive probabilities for each category in $\mathcal{X}$. This effect is observed also if the elements not on the diagonal of Λ are very small. It follows that, using a model with perfect observational mechanism in order to approximate a model with imperfect observational mechanism but very small probability of errors, does not seem to be justifiable from a theoretical point of view. Our results are summarized by the following theorem.

Theorem 1. *Assume that we have observed a dataset* $\mathbf{o}$ *with counts* $\mathbf{n} = (n_1, \ldots, n_k)$ *and that the observational mechanism is characterized by an emission matrix* Λ. *Then, for each* $i \in \{1, \ldots, k\}$*, the following results hold.*

1. *If all the elements of* Λ *are nonzero, then the IDM produces vacuous predictive probabilities, i.e.,* $\overline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) = 1$ *and* $\underline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) = 0$.

2. *The IDM produces* $\overline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) < 1$*, iff* $\exists j \in \{1, \ldots, k\}$*, such that* $n_j > 0$ *and* $\lambda_{ji} = 0$.

3. *The IDM produces* $\underline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) > 0$*, iff* $\exists j \in \{1, \ldots, k\}$*, such that* $n_j > 0$*,* $\lambda_{ji} \neq 0$ *and* $\lambda_{jr} = 0$ *for each* $r \neq i$.

Proof. 1. Assume that all the elements of Λ are nonzero. We show that in this case $\lim_{t_i \to 1} P(\mathrm{X} = x_i \,|\, \mathbf{o}) = 1$ and $\lim_{t_i \to 0} P(\mathrm{X} = x_i \,|\, \mathbf{o}) = 0$, in other words $\overline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) = 1$ and $\underline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) = 0$. From (8) we know that

$$\lim_{t_i \to 1} P(\mathrm{X} = x_i \,|\, \mathbf{o}) = $$
$$= \lim_{t_i \to 1} \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot \frac{a_i^{\mathbf{x}} + st_i}{N+s}}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}.$$

Because all the elements of Λ are nonzero, it follows immediately that $P(\mathbf{o} \,|\, \mathbf{x}) \neq 0$ for each $\mathbf{o}$ and each $\mathbf{x}$ in $\mathcal{X}^N$. Define $\overline{\mathbf{x}}^i$ as the dataset with $a_i^{\overline{\mathbf{x}}^i} = N$ and $a_j^{\overline{\mathbf{x}}^i} = 0$ for each $j \neq i$. We show that $\lim_{t_i \to 1} P(\mathbf{x}) = 0$ for each $\mathbf{x} \in \mathcal{X}^N \setminus \{\overline{\mathbf{x}}^i\}$. Actually, the numerator of (4) is a product of terms

$$\prod_{r=1}^{a_j^{\mathbf{x}}} (st_j + r - 1). \tag{9}$$

If $a_j^{\mathbf{x}} = 0$, then (9) is equal to one by definition. Otherwise, if $a_j^{\mathbf{x}} > 0$ for a $j \neq i$, then (9) is equal to

$$st_j \cdot \ldots \cdot (st_j + a_j^{\mathbf{x}} - 1). \tag{10}$$

If $t_i \to 1$, since $\mathbf{t} \in \mathcal{T}$, we have $t_j \to 0$ for each $j \neq i$. Because of the first term of the product (10), it follows that (10) tends to zero as $t_i \to 1$ and thus $P(\mathbf{x}) \to 0$. At the other side we have

$$\lim_{t_i \to 1} P(\overline{\mathbf{x}}^i) \overset{(9)}{=} \lim_{t_i \to 1} \frac{\prod_{r=1}^{N} (st_i + r - 1)}{\prod_{j=1}^{N} (s + j - 1)} = 1.$$

It follows that

$$\lim_{t_i \to 1} P(\mathrm{X} = x_i \,|\, \mathbf{o}) \overset{(8)}{=}$$
$$\overset{(8)}{=} \lim_{t_i \to 1} \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot \frac{a_i^{\mathbf{x}} + st_i}{N+s}}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}$$
$$= \lim_{t_i \to 1} \frac{P(\mathbf{o} \,|\, \overline{\mathbf{x}}^i) \cdot 1 \cdot \frac{a_i^{\overline{\mathbf{x}}^i} + st_i}{N+s}}{P(\mathbf{o} \,|\, \overline{\mathbf{x}}^i) \cdot 1} =$$
$$= \lim_{t_i \to 1} \frac{a_i^{\overline{\mathbf{x}}^i} + st_i}{N+s} = \frac{N+s}{N+s} = 1.$$

We calculate now $\lim_{t_i \to 0} P(\mathrm{X} = x_i \,|\, \mathbf{o})$. In this case all the datasets in $\mathcal{X}^N$ with $a_i^{\mathbf{x}} > 0$ have $\lim_{t_i \to 0} P(\mathbf{x}) = 0$, because $\lim_{t_i \to 0} st_i \cdot \ldots \cdot (a_i^{\mathbf{x}} + st_i - 1) = 0$. Assume for simplicity that $t_j \not\to 0$ for each $j \neq i$, then

$\lim_{t_i \to 0} P(\mathbf{x}) \neq 0$ for each $\mathbf{x} \in \mathcal{X}^N$ with $a_i^{\mathbf{x}} = 0$. It follows that

$$\lim_{t_i \to 0} P(\mathrm{X} = x_i \,|\, \mathbf{o}) = \\ = \frac{\sum_{\mathbf{x} \in \mathcal{X}^N : a_i^{\mathbf{x}} = 0} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot \frac{a_i^{\mathbf{x}} + st_i}{N+s}}{\sum_{\mathbf{x} \in \mathcal{X}^N : a_i^{\mathbf{x}} = 0} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}, \tag{11}$$

and because, with $a_i^{\mathbf{x}} = 0$,

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

we obtain $\lim_{t_i \to 0} P(\mathrm{X} = x_i \,|\, \mathbf{o}) = 0$.

2. If there exists j, such that $\lambda_{ji} = 0$, then it is impossible to observe $\mathrm{O} = x_j$ if $\mathrm{X} = x_i$. It follows, because $n_j > 0$, that $P(\mathbf{o} \,|\, \overline{\mathbf{x}}^i) = 0$. With $P(\mathbf{o} \,|\, \overline{\mathbf{x}}^i) = 0$, we show that $P(\mathrm{X} = x_i \,|\, \mathbf{o}) < 1$ for each $\mathbf{t} \in \mathcal{T}$, in particular $\lim_{t_i \to 1} P(x_i \,|\, \mathbf{o}) < 1$. Actually, for each $\mathbf{x} \neq \overline{\mathbf{x}}^i$ and each $\mathbf{t} \in \mathcal{T}$, we have $\frac{a_i^{\mathbf{x}} + st_i}{N+s} \leq \frac{a_i^{\mathbf{x}} + s}{N+s} < \frac{N+s}{N+s} = 1$, and (8) becomes thus a convex sum of fractions smaller than 1, and is therefore smaller than 1.

3. If there exists j such that $n_j > 0$, $\lambda_{ji} \neq 0$ and $\lambda_{jr} = 0$ for each $r \neq i$, then $P(\mathbf{o} \,|\, \mathbf{x}) \neq 0 \Leftrightarrow a_i^{\mathbf{x}} > 0$. Actually, in this case, we have $P(\mathrm{X} = x_i \,|\, \mathrm{O} = x_j) = 1$ and it is therefore impossible that $n_j > 0$ if $a_i = 0$. From (8) it follows that

$$P(\mathrm{X} = x_i \,|\, \mathbf{o}) = \\ = \frac{\sum_{\mathbf{x} \in \mathcal{X}^N : a_i^{\mathbf{x}} > 0} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot \frac{a_i^{\mathbf{x}} + st_i}{N+s}}{\sum_{\mathbf{x} \in \mathcal{X}^N : a_i^{\mathbf{x}} > 0} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})},$$

which is a convex combination of terms $\frac{a_i^{\mathbf{x}} + st_i}{N+s} \geq \frac{a_i^{\mathbf{x}}}{N+s} > \frac{0}{N+s} = 0$, and is therefore greater than 0 for each $\mathbf{t} \in \mathcal{T}$, in particular for $t_i \to 0$. If the condition above about the emission matrix is not satisfied, then for each j with $n_j > 0$ there exists an r, such that $\lambda_{jr} \neq 0$ and $r \neq i$. Therefore it is possible to construct a dataset $\mathbf{x}$ substituting x_j with x_r in $\mathbf{o}$ for each j with $n_j > 0$, such that $P(\mathbf{o} \,|\, \mathbf{x}) \neq 0$ and $a_i^{\mathbf{x}} = 0$. It follows from (11) that $\underline{P}(x_i \,|\, \mathbf{o}) = 0$.

□

Corollary 1. *Assume that Λ has non-zero elements on the diagonal. Then the IDM produces non-vacuous predictive probabilities for each category, iff $\Lambda = I$, i.e., in the case described by Walley in [7].*

Proof. Since the elements on the diagonal of Λ are non-zero, the condition of the third part of Theorem 1 is satisfied only if $\lambda_{ir} = 0$ for each i and each $r \neq i$. It follows that the elements on the diagonal are the unique elements different from 0, and because Λ is stochastic, $\lambda_{11} = \ldots = \lambda_{kk} = 1$. □

3.3 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*

$$\Lambda_\varepsilon := \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}, \tag{12}$$

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 possible datasets of $\mathcal{X}^2$ are given by

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

Using (8), the posterior probability $P(x_1|\mathbf{o})$ is given by

$$P(\mathrm{X} = x_1|\mathbf{o}) = \\ = \Bigg((1-\varepsilon) \cdot (1-\varepsilon) \cdot st_1(1+st_1) \cdot \frac{2+st_1}{2+s} + \\ +2 \cdot (1-\varepsilon) \cdot \varepsilon \cdot st_1 \cdot st_2 \cdot \frac{1+st_1}{2+s} + \\ + \varepsilon \cdot \varepsilon \cdot st_2 \cdot (1+st_2) \cdot \frac{0+st_1}{2+s} \Bigg) \cdot \\ \cdot \Bigg((1-\varepsilon) \cdot (1-\varepsilon) \cdot st_1(1+st_1) + \\ +2 \cdot (1-\varepsilon) \cdot \varepsilon \cdot st_1 \cdot st_2 + \\ + \varepsilon \cdot \varepsilon \cdot st_2 \cdot (1+st_2) \quad \Bigg)^{-1}.$$

It follows that

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

and

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

implying

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

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

Remark 5. *The result of Example 1 holds for each positive, even very small, value of* ε. *With* $\varepsilon = 0$ *we obtain* $\Lambda = I$, *therefore*

$$P(\mathrm{X} = x_1|\mathbf{o}) = \frac{2 + st_1}{2 + s},$$

$$P(\mathrm{X} = x_2|\mathbf{o}) = \frac{0 + st_2}{2 + s},$$

and the same $\mathbf{o}$ *yields*

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

This makes it clear that there is a strong kind of discontinuity 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 counts* $\mathbf{n} = (12, 23)$ *and assume that the emission matrix is*

$$\Lambda = \left(\begin{array}{cc} 0.8 & 0.2 \\ 0.2 & 0.8 \end{array} \right).$$

Figure 1 displays the results for $P(\mathrm{X} = x_1|\mathbf{o})$ *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* $\overline{P}(\mathrm{X} = x_1|\mathbf{o})$ *is about 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* $\overline{P}(\mathrm{X} = x_1|\mathbf{o}) = 1$ *as confirmed by theoretical results.*

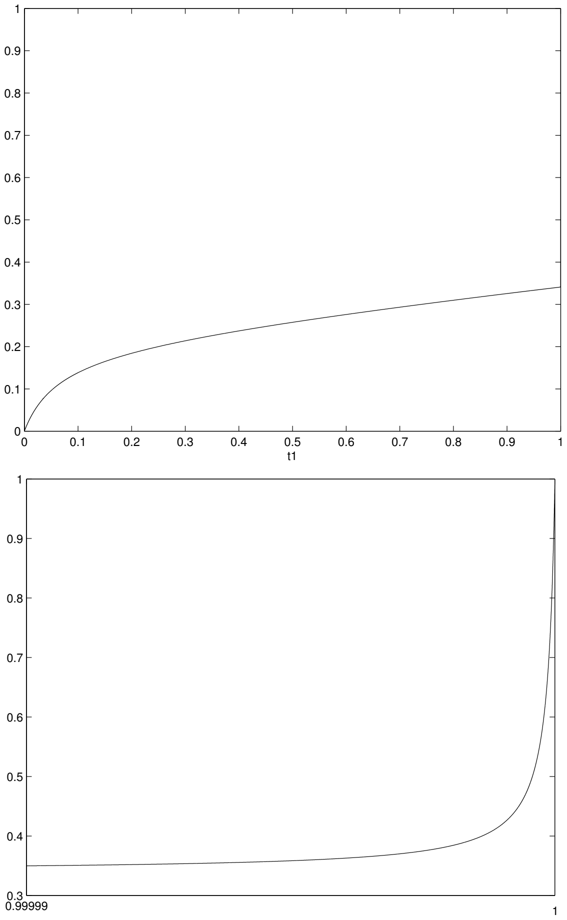

Figure 1: The function $P(\mathrm{X} = x_1|\mathbf{o})$ for $t_1 \in [0, 1]$ and for $t_1 \in [0.99999, 1]$.

3.4 Discussion

The results stated in Theorem 1 can be explained in an intuitive way. To understand the meaning of Statement 2 of Theorem 1, consider an observer with a unique extreme prior density $p(\theta) = dir(s, \mathbf{t})$, such that $s > 0$ and $t_i \to 1$ for an $i \in \{1, \ldots, k\}$. The observer believes a priori that the population is formed (almost) completely by individuals of category x_i. If he observes an individual of category x_j and $\lambda_{ji} \neq 0$, he will tend to believe that the individual observed is actually of category x_i and that there was a mistake in the observational mechanism. Only if $\lambda_{ji} = 0$ he has to rationally realize that observing something different from x_i can only be consistent with a modification of his strong prior beliefs.

To understand the meaning of Statement 3 of Theorem 1, consider now an observer with $t_i \to 0$. Such an observer believes a priori that there are 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 rationally believe that $\mathrm{X} = x_j$, only if $\lambda_{ij} = 0$

for all $j \neq i$. Similarly, if there exists a j, such that $n_j > 0$, $\lambda_{ji} \neq 0$ and $\lambda_{jr} = 0$ for all $r \neq i$, then observing $\mathrm{O} = x_j$ we know for sure that $\mathrm{X} = x_i$.

When letting the prior density of an observer converge to a degenerate one, the model with imperfect observational mechanism produces trivial results because of the degeneration in the behavior of the observer. Such a feature arises only with extreme prior densities. To avoid vacuous inferences it would be sufficient to restrict the set of prior densities closing the simplex $\mathcal{T}$ in a way to exclude these degenerate priors. However, this is not compatible with the idea of prior ignorance, which a priori should lead to

$$\underline{P}(\mathrm{X} = x_i) = 0, \quad \overline{P}(\mathrm{X} = x_i) = 1,$$

for each $i = 1, \ldots, k$.

3.5 The Case of Non-Deterministic Emission Matrix

Up to this point we have assumed an observational mechanism with known and constant emission matrix. In this section, in order to generalize Theorem 1, we study in detail the behavior of the IDM when the emission matrix is not deterministic and changes over time. We show that the IDM produces also in this case vacuous predictive probabilities. We prove firstly some results about the imprecise Beta model, and then extend the results to the IDM.

Corollary 2. *The IBM with observational mechanism defined by the emission matrix (12), where $\varepsilon \neq 0$, produces vacuous probabilities.*

Proof. This is a particular case of Theorem 1. □

Now we allow the observational mechanism to vary over time, we obtain however the same result:

Theorem 2. *The IBM with observational mechanism for the i-th observation defined by the emission matrix*

$$\Lambda_{\varepsilon_i} := \begin{pmatrix} 1-\varepsilon_i & \varepsilon_i \\ \varepsilon_i & 1-\varepsilon_i \end{pmatrix}, \tag{13}$$

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. With $\varepsilon_1, \ldots, \varepsilon_N \neq 0$, $P(\mathbf{o} \,|\, \mathbf{x}) \neq 0$ for each $\mathbf{o}$ and $\mathbf{x}$ in $\mathcal{X}^N$ and therefore we obtain the same results. □

Lemma 3 (Lebesgue Theorem). *Let $\{f_n\}$ be a series of functions on the domain A such that $f_n \rightarrow f$ 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 non-deterministic and we summarize our knowledge about ε_i with a continuous density measure. We obtain once more the same result.

Theorem 3. *The IBM with observational mechanism for the i-th observation defined by the emission matrix (13) , where $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_N)$ is distributed according to a continuous density $f(\varepsilon)$ defined on $[0,1]^N$, produces vacuous predictive probabilities.*

Proof. We know from Theorem 2 that, given $\varepsilon_1, \ldots, \varepsilon_N \neq 0$, we have $\lim_{t_1\to 1} P(\mathrm{X} = x_1 \,|\, \mathbf{o}, \varepsilon) = 1$, and $\lim_{t_1\to 1} P(\mathrm{X} = x_2 \,|\, \mathbf{o}, \varepsilon) = 0$.We have

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

Furthermore $P(\mathrm{X} = x_j \,|\, \mathbf{o}, \varepsilon) \cdot f(\varepsilon) \leq f(\varepsilon)$, 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 $P(\varepsilon_i \neq 0) = 1$ for each i. Applying Lemma 3 we conclude that

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

and, similarly,

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

□

Theorem 3 can be easily generalized to the k-dimensional case. Define the set $S^{k\times k}$ of $k \times$

k stochastic matrices. Assume that N observations are characterized by N emission matrices $\Lambda_1, \ldots, \Lambda_N \in S^{k \times k}$. Define $\Delta := (\Lambda_1, \ldots, \Lambda_N)$. The following theorem holds.

Theorem 4. *If Δ is distributed according to a continuous distribution function $f(\Delta)$ defined on $(S^{k \times k})^N$, then the IDM produces vacuous predictive probabilities.*

The proof is very similar to the proof of Theorem 3 and is omitted.

4 The Actual Level

One might say that the problem of vacuous predictive probabilities for the ideal symbols could be avoided considering only the actual symbols and applying therefore the standard IDM at the actual level. In fact the random variable O, defined in Section 3, is perfectly observable by definition. Therefore, having observed a dataset $\mathbf{o}$, apparently it should be possible to produce useful inferences on the chances $\xi = (\xi_1, \ldots, \xi_k)$ of O using the standard IDM. Assuming the emission matrix Λ to be given, it would then be possible to reconstruct the chances $\theta = (\theta_1, \ldots, \theta_k)$ using ξ and Λ. In particular, from (6), it follows that $\xi = \Lambda \cdot \theta$. If Λ is a non-singular matrix, we have $\theta = \Lambda^{-1} \cdot \xi$. In this section we show why the approach described above does not work. We restrict the discussion for simplicity to the binary case ($k = 2$) with emission matrix (12) and $\varepsilon \neq 0.5$. Consider the chances $\theta = (\theta_1, \theta_2)$ of the unobservable random variable X and the chances $\xi = (\xi_1, \xi_2)$ of the observable random variable O. Since the matrix (12) with $\varepsilon \neq 0.5$ is non-singular, we can reconstruct the values of θ starting from the values of ξ. We have $\xi_1 = (1 - \varepsilon)\theta_1 + \varepsilon\theta_2$ and $\xi_2 = (1 - \varepsilon)\theta_2 + \varepsilon\theta_1$. For simplicity we assume in the calculations that $\varepsilon < 0.5$, such that $1 - 2\varepsilon > 0$. All results are valid also for $0.5 < \varepsilon < 1$. Because $\theta_1 + \theta_2 = 1$, we have $\xi_i = (1 - 2\varepsilon)\theta_i + \varepsilon$, $i = 1, 2$, and

$$\theta_i = \frac{\xi_i - \varepsilon}{1 - 2\varepsilon}. \qquad (14)$$

It follows that

$$\mathrm{E}(\theta_i) = \frac{\mathrm{E}(\xi_i) - \varepsilon}{1 - 2\varepsilon}. \qquad (15)$$

4.1 Inference on O ignoring the Emission Matrix

We follow the approach described above in order to show that meaningless results are produced. In particular we apply the standard IBM at the actual level disregarding the fact that O is produced from X by the observational mechanism. Consider an observed dataset $\mathbf{o}$ with counts $\mathbf{n} = (n_1, n_2)$ and length $N = n_1 + n_2$. Applying the standard IBM we obtain

$$\underline{P}(\mathrm{O} = x_i \,|\, \mathbf{o}) = \frac{n_i}{N + s},$$

$$\overline{P}(\mathrm{O} = x_i \,|\, \mathbf{o}) = \frac{n_i + s}{N + s}.$$

Now we use (15) to construct $\underline{P}(\mathrm{X} = x_i \,|\, \mathbf{o})$ and $\overline{P}(\mathrm{X} = x_i \,|\, \mathbf{o})$, we obtain

$$\underline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) = \frac{n_i - \varepsilon(N + s)}{(N + s)(1 - 2\varepsilon)},$$

$$\overline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) = \frac{n_i + s - \varepsilon(N + s)}{(N + s)(1 - 2\varepsilon)}.$$

It is easy to see that, if $n_i < \varepsilon(N + s)$, then $\underline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) < 0$ and, if $n_i + s < \varepsilon(N + s)$, then $\overline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) < 0$. Therefore this approach produces meaningless results in general.

Example 3. *Suppose that we have observed the dataset $\mathbf{o}$ with counts $n_1 = 0$ and $n_2 = 10$ and that our observational mechanism is characterized by (12) with $\varepsilon = 0.2$. Applying the standard IBM with $s = 2$ at the actual level we obtain at the ideal level,*

$$\overline{P}(\mathrm{X} = x_1 \,|\, \mathbf{o}) = -0.0\overline{5},$$

$$\underline{P}(\mathrm{X} = x_2 \,|\, \mathbf{o}) = 1.0\overline{5}.$$

4.2 Inference on O considering the Emission Matrix

What is the problem of the approach described in Section 4.1? The problem is the following: we know that $\mathrm{E}(\theta_i) \in [0, 1]$ and $\mathrm{E}(\xi_i) = (1 - 2\varepsilon)\mathrm{E}(\theta_i) + \varepsilon$, it follows immediately that $\mathrm{E}(\xi_i) \in [\varepsilon, 1 - \varepsilon]$. But if we use the standard IBM to make inference on ξ we are implicitly assuming that, a priori, $\mathrm{E}(\xi_i \in [0, 1]$ and therefore we are doing a wrong assumption. If we model our knowledge about θ using a $beta(s, \mathbf{t})$ density, then our knowledge about ξ is modeled by a scaled beta density on the interval $[\varepsilon, 1 - \varepsilon]$. In fact, substituting (14) in the $beta(s, \mathbf{t})$ density for θ, since $d\theta = \frac{d\xi}{1 - 2\varepsilon}$, we obtain for ξ the density

$$\frac{C}{1 - 2\varepsilon} \left(\frac{\xi_1 - \varepsilon}{1 - 2\varepsilon} \right)^{st_1 - 1} \left(\frac{\xi_2 - \varepsilon}{1 - 2\varepsilon} \right)^{st_2 - 1}, \qquad (16)$$

where $C := \frac{\Gamma(s)}{\Gamma(st_1)\Gamma(st_2)}$. We call this density *scaled beta density*. The first moments of a scaled beta density are given by

$$E(\theta_i) = (1 - 2\varepsilon)t_i + \varepsilon. \tag{17}$$

To be consistent with the given data-generating process, the IBM on ξ should be performed using, as set of prior densities, the set of all beta densities scaled on $[\varepsilon, 1 - \varepsilon]$ with $\mathbf{t} \in \mathcal{T}$ and not the standard beta densities used in the IBM. In this way we assume a priori that $\xi_1, \xi_2 \in [\varepsilon, 1 - \varepsilon]$. But in this case the following theorem holds.

Theorem 5. *The IBM on ξ, with, as set of prior densities, the set of all scaled beta densities described above, produces vacuous*[2] *predictive probabilities.*

The complete proof is rather technical and is omitted. We sketch briefly the main idea of the proof. The effect observed in this case is very similar to the effect observed in the proof of Theorem 1. The likelihood function in this case is given by

$$P(\mathbf{o} \,|\, \xi) = \prod_{i=1}^{2} \xi_i^{n_i},$$

but, because $\xi_i \in [\varepsilon, 1 - \varepsilon]$, the likelihood function is strictly positive for each $\mathbf{o}$. Choosing extreme values for the parameters of the prior, the likelihood is unable to reduce this value because it cannot tend to zero, and therefore we obtain also extreme posterior predictive probabilities.

5 Conclusions

In this paper we have described the behavior of the imprecise Dirichlet model when the observations are nor perfect. We have modeled a situation characterized by an imperfect observational mechanism and prior near ignorance, using a two step process. We have shown, in Sections 3 and 4, that the IDM produces in general, both at the ideal and actual levels, vacuous predictive probabilities, also for very small probability of errors. Vacuous predictive probabilities are not produced only for very particular emission matrices Λ. There are some interesting questions arising from the results, in particular about the application of the IDM in practice, the assumptions on the observational mechanism and more generally about the possibility of learning with prior ignorance and imperfect observations.

1. In the light of our results, a person that uses the IDM in real applications can produce non-vacuous predictive probabilities only if he assumes a perfect observational mechanism. But in practice this assumption seems not to be tenable: we can never exclude the possibility of an error in the observational mechanism. How can we justify using the IDM for practical problems?

2. The behavior observed in the case of imperfect observations for the imprecise Dirichlet model seems not to be strictly related to its particular structure. The suspicion emerges, that the behavior observed by the IDM is only a particular case of a more general phenomenon concerning the inference models with prior ignorance and imperfect observations. Is it really possible to learn something, starting from prior ignorance and with imperfect observations?

Acknowledgements

This work was partially supported by the Swiss NSF programme NCCR FINRISK (Alberto Piatti and Fabio Trojani), by the Swiss NSF grant 100012-105745/1 (Fabio Trojani) and by the Swiss NSF grant 2100-067961 (Marco Zaffalon). The authors are grateful to Marcus Hutter for having spotted a mistake in an early version of the paper.

References

[1] J.-M. Bernard. Bayesian interpretation of frequentist procedures for a Bernoulli process. *Amer. Statist.*, 50: 7-13, 1996.

[2] J.-M. Bernard. Non-parametric inference about an unknown mean using the imprecise Dirichlet model, in: G. de Cooman, T. Fine, T. Seidenfeld (Eds.), Proc. 2nd Int. Symp. on Imprecise Probabilities and their Applications (ISIPTA '01), Shaker, Ithaca, New York, USA, 40-50, 2001.

[3] J.-M. Bernard. Analysis of local and asymmetric dependencies in contingency tables using the imprecise Dirichlet model, in: J.-M. Bernard, T. Seidenfeld, M. Zaffalon (Eds.), Proc. 3rd Int. Symp. on Imprecise Probabilities

[2] Note that we are abusing terminology here, as the predictive upper and lower prior and posterior probabilities are identical, but not equal to 1 and 0.

and their Applications (ISIPTA '03), Proceedings in Informatics, Vol. 18, Carleton Scientific, Waterloo, Ontario, Canada, 46-61, 2003.

[4] M. Hutter. Robust estimators under the imprecise Dirichlet model, in: J.-M. Bernard, T. Seidenfeld, M. Zaffalon (Eds.), Proc. 3rd Int. Symp. on Imprecise Probabilities and their Applications (ISIPTA '03), Proceedings in Informatics, Vol. 18, Carleton Scientific, Waterloo, Ontario, Canada, 274-289, 2003.

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

[6] E. Quaeghebeur, G. de Cooman. Game-theoretic learning using the imprecise Dirichlet model, in: J.-M. Bernard, T. Seidenfeld, M. Zaffalon (Eds.), Proc. 3rd Int. Symp. on Imprecise Probabilities and their Applications (ISIPTA '03), Proceedings in Informatics, Vol. 18, Carleton Scientific, Waterloo, Ontario, Canada, 450-464, 2003.

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

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

[9] M. Zaffalon. Statistical inference of the naive credal classifier, in: G. de Cooman, T. Fine, T. Seidenfeld (Eds.), Proc. 2nd Int. Symp. on Imprecise Probabilities and their Applications (ISIPTA '01), Shaker, Ithaca, New York, USA , 384-393, 2001.

[10] M. Zaffalon. Robust discovery of tree-dependency structures, in: G. de Cooman, T. Fine, T. Seidenfeld (Eds.), Proc. 2nd Int. Symp. on Imprecise Probabilities and their Applications (ISIPTA '01), Shaker, Ithaca, New York, USA , 394-403, 2001.

[11] M. Zaffalon. The naive credal classifier. *J. Statist. Plann. Inference*, 105(1): 5-21, 2002.