University of Lugano
Faculty of economics

LEARNING UNDER PRIOR IGNORANCE

A thesis submitted by

Alberto Piatti

for the degree of
PhD in Economics

October 2006

Accepted on the recommendation of:
Dr J. M. Bernard, Université de Paris V, External Member,
Prof. E. Ronchetti, Universities of Lugano and Geneva, President,
Prof. F. Trojani, Universities of Lugano and St. Gallen, Advisor,
Dr M. Zaffalon, IDSIA (University of Lugano), Advisor.

A Mamma, Barba, Fra, Ila, Andrea e Anna.

Ringraziamenti

Non ho parole sufficienti per esprimere la mia gratitudine nei confronti di Fabio Trojani e Marco Zaffalon per la loro guida e i loro insegnamenti: sono stati per me dei veri maestri, nel senso più profondo del termine.

Ringrazio il Prof. Elvezio Ronchetti e il Dr Jean-Marc Bernard per aver accettato di far parte della giuria.

Non sarei riuscito a portare a termine il mio lavoro di dottorato senza l'appoggio di numerose persone. Voglio ringraziare in primo luogo il direttore dell'Istituto di Finanza dell'USI, Prof. Giovanni Barone Adesi, e la direzione del Dipartimento Tecnologie Innovative della SUPSI, in particolare il direttore Prof. Giambattista Ravano, per la flessibilità e l'appoggio che mi hanno dimostrato nel conciliare la mia attività professionale con la mia attività di ricerca. Desidero ringraziare anche i colleghi e la direzione dell'Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA) per l'appoggio e la simpatia.

Ringrazio i colleghi e amici che mi hanno sostenuto e hanno reso piacevoli questi anni. Tutti i colleghi della SUPSI, in particolare i miei colleghi matematici, Martina Solder, Janos Barta, Paola Rezzonico, Andrea Graf, Daniela Pellizzari e Paolo Stoppa. Tutti i miei colleghi e amici dell'istituto di finanza, in particolare Francesca Bellini, Ilaria Finzi e Claudio Ortelli. I miei colleghi della Facoltà di Scienze della Comunicazione, in particolare Marco Colombetti e Eddo Rigotti. I miei amici della Società Matematica della Svizzera Italiana: Paolo Hägler, Edo Montella, Claudio Beretta, Giorgio Mainini, Gianfranco Arrigo e Alberto Vancheri.

Numerose persone hanno lasciato una traccia indelebile nella mia formazione scientifica; voglio ringraziare in particolare, Gianfranco Arrigo, René Sperb, Marco Martucci, Marco Moretti, Angela Macciocchi e Roberto Eggmann.

Desidero ricordare con particolare gratitudine il Prof Pietro Balestra, prematuramente scomparso lo scorso anno.

Ringrazio la mia famiglia, per avermi sempre appoggiato e motivato nello studio. A loro dedico questo lavoro.

Ringrazio infine Barbara, per essermi sempre vicina con il suo amore e la sua intelligenza. Un ringraziamento va anche alla sua famiglia, per la simpatia e l'ospitalità che mi hanno sempre riservato, in particolare negli ultimi mesi di dottorato.

Bellinzona, 19 ottobre 2006

Contents

1 Introduction **1**

2 Limits of learning from imperfect observations under prior near-ignorance: the case of the imprecise Dirichlet model **11**
2.1 Summary 11
2.2 Introduction 12
2.3 The imprecise Dirichlet model 14
2.3.1 Bayesian inference and Dirichlet prior density 15
2.3.2 The imprecise Dirichlet model 17
2.4 The imprecise Dirichlet model with imperfect observational mechanism 17
2.4.1 The IDM with imperfect observational mechanism . . . 18
2.4.2 Vacuous predictive probabilities 19
2.4.3 Examples 20
2.4.4 Discussion 22
2.4.5 The case of non-deterministic emission matrix 23
2.5 The actual level 24
2.5.1 Inference on O ignoring the emission matrix 25
2.5.2 Inference on O considering the emission matrix 26
2.6 Conclusions 27
2.7 Proofs 27

3 Limits of learning about a categorical latent variable under prior near-ignorance **35**
3.1 Summary 35
3.2 Introduction 36
3.3 Categorical Latent Variables 38
3.4 Near-Ignorance Priors 40

3.5 Limits of Learning under Prior Near-Ignorance 42
3.5.1 General parametric inference 43
3.5.2 An important special case: predictive probabilities . . . 47
3.5.3 Predicting the next outcome with categorical manifest variables 48
3.6 Conclusions 52
3.7 Technical preliminaries 52
3.8 Proofs of the main results 57
3.8.1 Proof of Theorem 22 and Corollary 23 57
3.8.2 Proof of Theorem 27 and Corollary 28 58

4 Learning from quasi perfect observations under prior near-ignorance: the binary case 62
4.1 Summary 62
4.2 Introduction 63
4.3 The imprecise Beta model for imperfect observations 65
4.4 Quasi perfect observations 67
4.5 Comparison between the IBM and the MIBM 70
4.6 Conclusions 72
4.7 Technical results 75
4.8 Proofs 78

5 Conclusions and outlook 93

Chapter 1

Introduction

Consider a generic sample space Ω. Following Knight (1933), we can distinguish between two types of uncertainty about Ω: a determinate uncertainty, called *determinacy*, and an indeterminate uncertainty, called *indeterminacy*. We say that our beliefs about Ω are *determinate* if, given two gambles (i.e., bounded random variables) X and Y on Ω, either we prefer one gamble to the other, or the two gambles are equivalent. We say that our beliefs about Ω are *indeterminate* if, given two gambles X and Y on Ω, it is possible that the two gambles are not equivalent given our beliefs but we cannot specify any preference between them. In other words: our beliefs are indeterminate if our knowledge about Ω is not sufficient to express a preference for any arbitrary pair of gambles X and Y on Ω. In practical decision problems, indeterminacy can lead to a situation of indecision that cannot occur in a situation of determinacy. Because indecision is a very common situation in practical decision problems, indeterminacy deserves to be modelled mathematically in an adequate way.

Determinacy and indeterminacy can both be modelled using probabilities. The difference between the two types of uncertainty lies in the way probabilities are specified. As explained by Walley (1991), determinacy is usually modelled using precise probabilities, while to model indeterminacy we need imprecise probabilities. Consider for example two event indicators $\mathbb{I}_A$ and $\mathbb{I}_B$ defined on the same sample space Ω. We interpret the two event indicators as 0-1 gambles: with the gamble $\mathbb{I}_A$ ($\mathbb{I}_B$) we get 1 dollar if an event A (B) occurs and nothing elsewhere. If we are able to specify precise probabilities for the events A and B, denoted $P(A)$ and $P(B)$, then whether we can ex-

press a preference or the two gambles are equivalent. In fact we prefer $\mathbb{I}_A$ to $\mathbb{I}_B$ if $P(A) > P(B)$, we prefer $\mathbb{I}_B$ to $\mathbb{I}_A$ if $P(B) > P(A)$, and we consider the two gambles equivalent if $P(B) = P(A)$. We follow De Finetti (1970) in the interpretation of probabilities as betting rates. Following this interpretation, the probability $P(A)$ corresponds to the price that we would be willing to pay to buy the gamble $\mathbb{I}_A$ and to the price that we would be willing to accept to sell the same gamble. But this situation is not realistic in practice, because usually the Infimum selling price and the Supremum buying price for a gamble of the type described above are not identical. In particular the Infimum selling price is usually higher than the Supremum buying price. In this case we cannot specify a single precise probability $P(A)$ for the event A, but we have to specify a lower probability $\underline{P}(A)$, corresponding to the Supremum buying price for $\mathbb{I}_A$, and an upper probability $\overline{P}(A)$, corresponding to the Infimum selling price. The probabilities assigned to the event A in this case are said to be *imprecise*, because we assign to the event A a probability interval $[\underline{P}(A), \overline{P}(A)]$ instead of a single value $P(A)$. Imprecise probabilities are used to model indeterminacy. Suppose that we specify lower and upper probabilities for the events A and B of the lotteries above. If $\underline{P}(A) > \overline{P}(B)$ we prefer $\mathbb{I}_A$ to $\mathbb{I}_B$, if $\underline{P}(B) > \overline{P}(A)$ we prefer $\mathbb{I}_B$ to $\mathbb{I}_A$, but if $\underline{P}(A) \leq \overline{P}(B)$ and $\underline{P}(B) \leq \overline{P}(A)$, then we are unable to express a preference between the two lotteries, although the two lotteries are not necessarily equivalent. We are therefore in a situation of indecision.

According to Walley (1991), there are many possible sources of indeterminacy. For example: lack of information concerning Ω, conflicting information and beliefs, information of limited relevance, physical indeterminacy. But, again according to Walley, the most important source of indeterminacy is the lack of information concerning Ω. The lack of information about a gamble $\mathbb{I}_A$ can be measured by quantifying the difference between the upper and the lower probability assigned to the event A, the so called *degree of imprecision*. The degree of imprecision, in this case, ranges from 0 to 1, where 0 reflects complete knowledge (and therefore precise probabilities) about the gamble $\mathbb{I}_A$ and 1 reflects complete ignorance. The degree of imprecision is equal to 1 only with *vacuous probabilities*, i.e., only if the lower probability is set to 0 and the upper probability is set to 1. According to Walley (1991), vacuous probabilities about a 0-1 gamble mean that we would pay up to 0 dollars to buy the gamble $\mathbb{I}_A$ and we would be payed at least 1 dollar to sell the same

gamble. In fact, this approach is suited to describe a situation of complete ignorance about A. This is the way complete ignorance about the event A is modelled, using probabilities, in the field of imprecise probabilities.

We focus now on a categorical random variable X with outcomes in $\mathcal{X}$, where $\mathcal{X} = \{x_1, \ldots, x_k\}$, and unknown chances $\vartheta \in \Theta$, where

$$\Theta := \{\vartheta = (\vartheta_1, \ldots, \vartheta_k) \mid \sum_{i=1}^{k} \vartheta_i = 1,\ 0 \leq \vartheta_i \leq 1\}.$$

Suppose that we have no relevant prior information about ϑ and we are therefore in a situation of prior ignorance about X. How should we model our prior beliefs in order to reflect the initial lack of knowledge?

Let us give a brief overview of this topic in the case of coherent models of uncertainty, such as Bayesian probability and Walley's theory of *coherent lower previsions*.

In the traditional Bayesian setting, prior beliefs are modelled using a single prior probability distribution. The problem of defining a standard prior probability distribution modeling a situation of prior ignorance, a so-called *noninformative prior*, has been an important research topic in the last two centuries[1] and, despite the numerous contributions, it remains an open research issue, as illustrated by Kass and Wassermann (1996). See also Hutter (2006) for recent developments and complementary considerations. There are many principles and properties that are desirable to model a situation of prior ignorance and that have been used in past research to define noninformative priors. For example Laplace's *symmetry or indifference* principle has suggested, in case of finite possibility spaces, the use of the uniform distribution. Other principles, like for example the principle of *invariance under group transformations*, the *maximum entropy* principle, the *conjugate priors* principle, etc., have suggested the use of other noninformative priors, in particular for continuous possibility spaces, satisfying one or more of these principles. But, in general, it has proven to be difficult to define a standard noninformative prior satisfying, at the same time, all the desirable principles.

[1]Starting from the work of Laplace at the beginning of the 19^{th} century (Laplace (1820)).

We follow De Cooman and Miranda (2006) when they say that there are at least two principles that should be satisfied to model a situation of prior ignorance: the *symmetry* and the *embedding principles*. The *symmetry principle* states that, if we are ignorant a priori about ϑ, then we have no reason to favour one possible outcome of X to another, and therefore our probability model on X should be symmetric. This principle recalls Laplace's *symmetry or indifference* principle that, in the past decades, has suggested the use of the *uniform prior* as standard noninformative prior. The *embedding principle* states that, for each possible event A, the probability assigned to A should not depend on the possibility space $\mathcal{X}$ in which A is embedded. In particular, the probability assigned a priori to the event A should be invariant with respect to refinements and coarsenings of $\mathcal{X}$. It is easy to show that the embedding principle is not satisfied by the uniform distribution. How should we model our prior ignorance in order to satisfy these two principles? Walley (1991) gives a compelling answer to this question: he proves[2] that the only coherent probability model on X consistent with the two principles is the *vacuous probability model*, i.e., the model that assigns, for each non-trivial event A, lower probability $\underline{\mathrm{P}}(A) = 0$ and upper probability $\overline{\mathrm{P}}(A) = 1$. Clearly, the vacuous probability model cannot be expressed using a single probability distribution. It follows that, if we agree that the symmetry and the embedding principles are characteristics of prior ignorance, then we need *imprecise probabilities* to model such a state of beliefs.[3] Unfortunately, it is easy to show that updating the vacuous probability model on X produces only vacuous posterior probabilities. Therefore, the vacuous probability model alone is not a viable way to address our initial problem. Walley (1991) suggests, as an alternative, the use of *near-ignorance priors*.

A near-ignorance prior is a probability model on the chances θ of X, modelling a very weak state of knowledge about θ. In practice, a near-ignorance prior is a large closed convex set $\mathcal{M}_0$ of prior probability densities[4] on θ which produces *vacuous expectation*s for various functions f on Θ, i.e., such that $\underline{\mathbf{E}}(f) = \inf_{\vartheta\in\Theta} f(\vartheta)$ and $\overline{\mathbf{E}}(f) = \sup_{\vartheta\in\Theta} f(\vartheta)$. The key point is that near-ignorance priors can be designed so as to satisfy both the symmetry and the

[2] In Note 7, p. 526. See also Section 5.5 of the same book.

[3] For a complementary point of view, see Hutter (2006).

[4] A set of probability masses or densities is often called *credal set*, according to Levi (1980).

embedding principles. In fact, if a near-ignorance prior produces vacuous expectations for all the functions $f(\theta) = \theta_i$ for each $i \in \{1, \ldots, k\}$, then, because a priori $P(\mathrm{X} = x_i) = E(\theta_i)$, the near-ignorance prior implies the vacuous probability model on X and satisfies therefore both the symmetry and the embedding principle, thus delivering a satisfactory model of prior (near-)ignorance.[5] Updating a near-ignorance prior consists in updating all the probability densities in $\mathcal{M}_0$ using the Bayes rule. Because the beliefs on θ are not vacuous, we obtain thus a non-vacuous set of posterior probability densities on θ that can be used to calculate posterior probabilities for X.

Walley (1996) has proposed an important model for learning under prior near-ignorance: the *imprecise Dirichlet model* (IDM). Suppose that the categorical variable X takes values in the set $\mathcal{X} = \{x_1, \ldots, x_k\}$, and denote with $\vartheta = (\vartheta_1, \ldots, \vartheta_k) \in \Theta$ the unknown chances of X. Let us focus on calculating the predictive posterior probabilities $P(\mathrm{X} = x_i \,|\, \mathbf{x})$ for each $x_i \in \mathcal{X}$, given an observed sample $\mathbf{x}$ of length N of realizations of X, and following the approach outlined above. The IDM models prior near-ignorance with the set of all *Dirichlet densities* $dir(s, \mathbf{t})$, defined on Θ for a fixed $s \geq 0$ and all $\mathbf{t} \in \mathcal{T}$, where the density $dir(s, \mathbf{t})$ is defined by the expression

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

and

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

Denote this prior set of densities with $\mathcal{M}_0$. The first moments of a $dir(s, \mathbf{t})$ density are $E(\theta_i) = t_i$, for each $i = 1, \ldots, n$. Because X is a categorical variable, we have

$$\underline{P}(\mathrm{X} = x_i) = \inf_{\mathcal{M}_0} E(\theta_i) = \inf_{\mathcal{T}} t_i = 0,$$

and

$$\overline{P}(\mathrm{X} = x_i) = \sup_{\mathcal{M}_0} E(\theta_i) = \sup_{\mathcal{T}} t_i = 1,$$

[5] We call this state near-ignorance because, although we are completely ignorant a priori about X, we are not completely ignorant about θ (Walley, 1991, Section 5.3, Note 4).

for each $x_i \in \mathcal{X}$. Therefore, a priori the IDM produces a vacuous probability model for X and satisfies therefore both the symmetry and the embedding principles. Posterior inference is obtained by updating each Dirichlet prior using the observed sample. Because Dirichlet densities and the multinomial likelihood are conjugate, this approach leads to a set of posterior Dirichlet densities, defined by

$$\mathcal{M}_N := \left\{ dir(N+s, \mathbf{t}^{\mathbf{x}}) \;\middle|\; t_i^{\mathbf{x}} = \frac{a_i^{\mathbf{x}} + st_i}{N+s}, \mathbf{t} \in \mathcal{T} \right\},$$

where $a_i^{\mathbf{x}}$ denotes for each $x_i \in \mathcal{X}$ the number of times that we have observed the outcome x_i. The upper and lower posterior predictive probabilities are then given by

$$\underline{P}(\mathrm{X} = x_i \,|\, \mathbf{x}) = \inf_{\mathcal{M}_N} E(\theta_i \,|\, \mathbf{x}) = \inf_{\mathcal{T}} \frac{a_i^{\mathbf{x}} + st_i}{N+s} = \frac{a_i^{\mathbf{x}}}{N+s}$$

and

$$\overline{P}(\mathrm{X} = x_i \,|\, \mathbf{x}) = \sup_{\mathcal{M}_N} E(\theta_i \,|\, \mathbf{x}) = \sup_{\mathcal{T}} \frac{a_i^{\mathbf{x}} + st_i}{N+s} = \frac{a_i^{\mathbf{x}} + s}{N+s}.$$

It follows that the posterior degree of imprecision is

$$\overline{P}(\mathrm{X} = x_i \,|\, \mathbf{x}) - \underline{P}(\mathrm{X} = x_i \,|\, \mathbf{x}) = \frac{a_i^{\mathbf{x}} + s}{N+s} - \frac{a_i^{\mathbf{x}}}{N+s} = \frac{s}{N+s}$$

and depends therefore solely on N. The parameter s is a fixed parameter that can be interpreted as a degree of caution in inference. With large values of s the degree of imprecision decreases more slowly than with smaller values of s. For a discussion on the choice of possible values of s see Walley (1996) or Bernard (2005). The degree of imprecision in the IDM can be interpreted as follows. Initially, with $N = 0$, the degree of imprecision is equal to 1 and reflects the initial state of ignorance. As the number of observations increases, the probability intervals narrow. Therefore, the degree of imprecision decreases, reflecting our increased knowledge about X. Note that the IDM leads to precise probabilities in the limit, after observing infinitely many data. The IDM has gained in the recent years considerable attention, and is recognized as an important milestone of the field of imprecise probability, with applications such as classification (Zaffalon (2001a), Zaffalon et al. (2003)), nonparametric inference (Bernard (2001)), robust estimation (Hutter (2003)), analysis of contingency tables (Bernard (2003)), discovery

of dependency structures (Zaffalon et al. (2005)), and game theory (Quaghebeur et al. (2003)). For a detailed overview of theory and applications of the IDM see Bernard (2005).

The IDM is the starting point of this thesis. Our research starts from the consideration that, in real world applications, any observed sample may contain observation errors. The possibility of errors in the observations is not taken into consideration by the IDM, which assumes *perfect observations*. In Chapter 2, we relax and extend the IDM to *imperfect observations*. We assume that the categorical variable X is unobservable. For each realization of X, we can observe a realization of another random variable O, which takes values in the same possibility space of X and represents the observed value. We model the imperfect observational mechanism using a stochastic matrix Λ, called *emission matrix*, containing all the probabilities $\lambda_{ij} := P(\mathrm{O} = x_i \,|\, \mathrm{X} = x_j)$, where λ_{ij} is known for each $i, j = 1, \ldots, k$. With this generalization, the case of perfect observational mechanism is recovered when $\Lambda = I$. We focus on constructing posterior predictive probabilities $P(\mathrm{X} = x_i \,|\, \mathbf{o})$ for each $x_i \in \mathcal{X}$, having observed a sample $\mathbf{o}$ of realizations of O and knowing Λ. This extension of the IDM to imperfect observations yields the following, at first sight surprising, results:

1. If all elements of Λ are nonzero then, for every $x_i \in \mathcal{X}$, $\overline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) = 1$ and $\underline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) = 0$.

2. $\overline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) < 1$ for some $x_i \in \mathcal{X}$, iff we observed at least once $\mathrm{O} = x_j$ such that $\lambda_{ji} = 0$.

3. $\underline{P}(\mathrm{X} = x_i \,|\, \mathbf{o}) > 0$ for some $x_i \in \mathcal{X}$, iff we observed at least once $\mathrm{O} = x_j$ such that $\lambda_{ji} \neq 0$ and $\lambda_{jr} = 0$ for each $r \neq i$ in $\{1, \ldots, k\}$.

In other words, if the observational process has an arbitrary small non-zero probability to generate any sort of mistake, then the posterior probabilities of the extended IDM are vacuous, irrespective of the observed sample. Partial learning is possible in this setting only for very special observational processes having some zero probabilities in the emission matrix. In other words, a component of perfection is needed to generate useful inferences in the IDM. These results have two different implications. The first is concrete: the IDM can yield useful conclusions only by assuming (at least partial) perfection of the observational process. The second implication is more philosophical

and points to a fundamental statistical problem: prior near-ignorance seems to be incompatible with learning, at least in settings relevant for real-world problems.

To understand the generality of the incompatibility of prior near-ignorance and imperfect observations, we consider, in Chapter 3, a more general statistical setup than the one assumed by the IDM. We consider a sequence of independent and identically distributed (IID) categorical latent variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ with outcomes in $\mathcal{X}$ and unknown chances ϑ, and a sequence of independent manifest variables $(\mathrm{S}_i)_{i\in\mathbf{N}}$. We assume that a realization of the manifest variable S_i can be observed only after a (hidden) realization of the latent variable X_i. Furthermore, we assume S_i to be independent of the chances ϑ of X_i conditional on X_i, i.e.,

$$P(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j, \vartheta) = P(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j), \tag{1.1}$$

for each $x_j \in \mathcal{X}$ and $\vartheta \in \Theta$. These assumptions model what we call an *observational process*, i.e., a two-step process where the variable S_i is used to acquire information about the realized value of X_i for each i, independently on the chances of X_i. We focus on a very general problem of parametric inference. Suppose that we observe a dataset $\mathbf{s}$ of realizations of manifest variables $\mathrm{S}_1, \ldots, \mathrm{S}_N$ related to the (unobservable) dataset $\mathbf{x} \in \mathcal{X}^N$ of realizations of the variables $\mathrm{X}_1, \ldots, \mathrm{X}_N$. Defining the random variables $\mathbf{X} := (\mathrm{X}_1, \ldots, \mathrm{X}_N)$ and $\mathbf{S} := (\mathrm{S}_1, \ldots, \mathrm{S}_N)$ we have $\mathbf{S} = \mathbf{s}$ and $\mathbf{X} = \mathbf{x}$. Given a bounded function $f(\theta)$, our aim is to calculate $\underline{\mathbf{E}}(f \,|\, \mathbf{s})$ and $\overline{\mathbf{E}}(f \,|\, \mathbf{s})$ starting from a condition of near-ignorance about f, i.e., using a generic near-ignorance prior $\mathcal{M}_0$ such that $\underline{\mathbf{E}}(f) = f_{\min} := \inf_{\vartheta\in\Theta} f(\vartheta)$ and $\overline{\mathbf{E}}(f) = f_{\max} : \sup_{\vartheta\in\Theta} f(\vartheta)$. In such a setting, we introduce a condition, related to the likelihood of the observed data, that is shown to be sufficient to prevent learning about X under prior near-ignorance. The condition is very general as it is developed for any prior that models near-ignorance (not only the one used in the IDM), and for very general kinds of relation between X and S. We show then, by simple examples, that such a condition is easily satisfied, even in the most elementary and common statistical problems.

The results of Chapter 3 raise, in general, serious criticisms about the use of near-ignorance priors in practical applications. To produce non-vacuous posterior predictive probabilities for a categorical latent variable starting

from a condition of very weak knowledge, it is necessary to develop new models with a very weak specification of prior knowledge that are stronger than prior near-ignorance.

In Chapter 4, we propose such a model. We obtain a new model modifying the IDM with an additional assumption that restricts slightly the set of prior densities by eliminating the problematic quasi-deterministic ones. We focus on the two-dimensional version of the IDM, the imprecise Beta model (IBM) (see Bernard (1996), Walley (1991) and Walley (1996)). We consider a setting similar to the one of Chapter 2, but assume that the probabilities of errors are small. We assume a symmetric emission matrix (in this case, a 2×2 matrix), so that the probability of error is the probability of confounding one outcome with the other. This setting is relevant for applications because in the practice one is often tempted to use a model for perfect observations like the IBM when the probability of an observation error can be assumed to be small. This approach, however, is inconsistent with our previous theoretical findings. In Chapter 4 we propose an additional assumption to the IBM, called *quasi perfection*, that is natural when the probability of error is small. The assumption is based on the intuition that, having observed a small sample with very small probability of error, one does not usually expect to have errors in the sample. We show that this assumption yields to a restriction of the set of priors in the IBM, where the restriction depends on the probability of error and the strength of the assumption. We show that the restricted IBM is actually able to learn from imperfect observations under a very weak specification of prior knowledge. Furthermore, the results produced by the restricted IBM can be arbitrarily close to those produced by the IBM for perfect observations considering the observed sample as perfect, depending on the probability of error and the strength of the additional assumption. This last finding has some potential for the IBM in applications characterized by a small probability of observation errors. The model in Chapter 4 is only a first step towards a theory of learning from imperfect observations under weak prior knowledge. There are still many research issues and questions that remain to be investigated. These topics, as well as the relevance and the limits of the present work, are discussed in the concluding Chapter 5.

Outline. This work consists of three main chapters organized as independent papers. Differences across chapters are therefore possible in the nota-

tion and terminology. However, the three chapters can also be read as a unique discussion about the problem of the incompatibility between prior near-ignorance and imperfect observations. From this perspective, Chapter 2 represents the discovery of the problem, Chapter 3 its generalization and conceptualization and Chapter 4 a first attempt to solve it in a simple, particular but important case. Chapter 2 has been published as Piatti et al. (2005). Chapters 3 and 4 are the papers Piatti et al. (2006a) and Piatti et al. (2006b). Both papers are actually submitted.

Chapter 2

Limits of learning from imperfect observations under prior near-ignorance: the case of the imprecise Dirichlet model

2.1 Summary

History

The problem of vacuous probabilities in the IDM with imperfect observations was first discovered in the month of July of 2004 thanks to computational simulations. The effect was then proved theoretically in the following months. A first version of the paper was submitted to the Society for Imprecise Probability Theory and Applications (SIPTA) in February 2005 for presentation at the International Symposium on Imprecise Probabilities Theory and their Applications (ISIPTA '05). The paper was accepted in April 2005 and the definitive version, reproduced in this chapter, was presented at ISIPTA '05 in a plenary session in Pittsburgh in the month of July 2005. The paper was then published in the Proceedings of the Symposium as Piatti et al. (2005).

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 near-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 near-ignorance.

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.

2.2 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) Walley (1996) 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. 2.3), 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 near-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 (Zaffalon (2001a), Zaffalon (2002)), nonparametric inference (Bernard (2001)), robust estimation (Hutter (2003)), analysis of contingency tables (Bernard (2003)), discovery of dependency structures (Zaffalon (2001b)), and game theory (Quaghebeur et al. (2003)).

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 2.4. 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 near-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 2.4.2 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 near-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 2.4.5 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 2.5 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.

2.3 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 $\mathcal{X} = \{x_1, \ldots, x_k\}$. 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 $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$ is a point in the closed k-dimensional unit simplex[1]

$$\Theta := \{\vartheta = (\vartheta_1, \ldots, \vartheta_k) \,|\, \sum_{i=1}^{k} \vartheta_i = 1,\ 0 \leq \vartheta_i \leq 1\}.$$

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 near-ignorance. Having observed a dataset $\mathbf{x}$, we can summarize the observation with the counts $\mathbf{a} = (a_1^{\mathbf{x}}, \ldots, a_k^{\mathbf{x}})$ 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 $P(\mathbf{x} \,|\, \vartheta) = \vartheta_1^{a_1} \cdots \vartheta_k^{a_k}$. 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.

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

2.3.1 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

$$p(\theta|\mathbf{x}) = \frac{P(\mathbf{x}|\theta) \cdot p(\theta)}{P(\mathbf{x})}, \tag{2.1}$$

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 $P(\theta \,|\, \mathbf{x})$. 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* $\mathbf{t} = (t_1, \ldots, t_k) \in \mathcal{T}$*, where* $\mathcal{T}$ *is the open* k*-dimensional simplex*

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

We recall first some important properties of Dirichlet densities.

Lemma 2 (First moment) *The first moments of a* $dir(s, \mathbf{t})$ *density are given by* $E(\theta_i) = t_i$*,* $i \in \{1, \ldots, k\}$*.*

For a proof of Lemma 2 see Kotz et al. (2000).

Remark 3 *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 4 *Consider a dataset $\mathbf{x}$ with counts $\mathbf{a} = (a_1^{\mathbf{x}}, \ldots, a_k^{\mathbf{x}})$. 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.

The proof of Proposition 4 and all the other proofs of this chapter are in Section 2.7.

Remark 5 *Using a $dir(s, \mathbf{t})$ density measure as prior in a Bayesian learning problem with categorical data we have $p(\theta) = dir(s, \mathbf{t})$ and*

$$P(\mathbf{x}|\theta) = \prod_{j=1}^{k} \theta_j^{a_j}. \tag{2.2}$$

According to Proposition 4, the posterior density is then given by $P(\theta|\mathbf{x}) = dir(s^{\mathbf{x}}, \mathbf{t}^{\mathbf{x}})$ and therefore

$$P(x_i \,|\, \mathbf{x}) = t_i^{\mathbf{x}} = \frac{a_i + st_i}{N + s}. \tag{2.3}$$

Moreover, comparing (2.1) with the equality of Proposition 4, we conclude that

$$P(\mathbf{x}) = \frac{\prod_{j=1}^{k} \prod_{i=1}^{a_j} (st_j + i - 1)}{\prod_{i=1}^{N} (s + i - 1)}. \tag{2.4}$$

2.3.2 The imprecise Dirichlet model

The Imprecise Dirichlet Model (IDM) (see Bernard (1996) and Walley (1996)) is a model that generalizes Bayesian learning from categorical data to the case when there is prior near-ignorance about θ. Prior near-ignorance about θ is modelled 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 modelled 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 $\mathcal{M}_0 := \{dir(s, \mathbf{t}) \,|\, \mathbf{t} \in \mathcal{T}\}$. Suppose that we observe the dataset $\mathbf{x}$ with corresponding counts $\mathbf{a} = (a_1^{\mathbf{x}}, \ldots, a_k^{\mathbf{x}})$. Then, the set of resulting posterior densities follows from Proposition 4 and is given by

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

Definition 6 *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 7 *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 8 *The IDM with $k = 2$ is usually called* Imprecise Beta Model *(IBM), because the Dirichlet densities with $k = 2$ are beta densities (see Bernard (1996) and Walley (1991)).*

2.4 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 categorical 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*,

$$\Lambda := \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1k} \\ \vdots & \ddots & \vdots \\ \lambda_{k1} & \cdots & \lambda_{kk} \end{pmatrix}. \tag{2.5}$$

Then, the chances $\xi = (\xi_1, \ldots, \xi_k)$ of the random variable O are given by

$$\xi_i = \sum_{j=1}^{k} \lambda_{ij} \cdot \theta_j. \tag{2.6}$$

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

$$p(\mathbf{o}, \mathbf{x}, \theta) = P(\mathbf{o} \,|\, \mathbf{x}) P(\mathbf{x} \,|\, \theta) p(\theta). \tag{2.7}$$

2.4.1 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 9 *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*

$$P(\mathrm{X} = x_i \,|\, \mathbf{o}) = \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot \frac{a_i^{\mathbf{x}} + st_i}{N+s}}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}. \tag{2.8}$$

If we consider now each prior density in the set $\mathcal{M}_0$ and we calculate the posterior predictive probabilities $P(\mathrm{X} = x_i \,|\, \mathbf{o})$ using (2.8) 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.

2.4.2 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 10 *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* $r \neq i$*.*

2.4.3 Examples

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

Example 11 *Consider a situation with $k = 2$, $s = 2$, $N = 2$ and an emission matrix*

$$\Lambda_\varepsilon := \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}, \tag{2.9}$$

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

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

Using (2.8), the posterior probability $P(x_1|\mathbf{o})$ is given by

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

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 12 *The result of Example 11 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 13 *Suppose that we have observed a dataset* $\mathbf{o}$ *with counts* $\mathbf{n} = (12, 23)$ *and assume that the emission matrix is*

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

Figure 2.1 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*

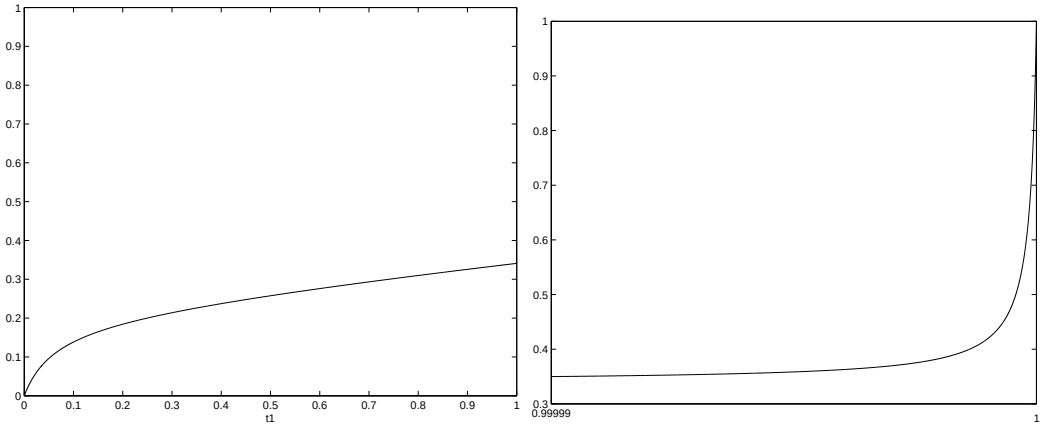

Figure 2.1: The function $P(\mathrm{X} = x_1|\mathbf{o})$ for $t_1 \in [0, 1]$ and for $t_1 \in [0.99999, 1]$.

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.*

2.4.4 Discussion

The results stated in Theorem 10 can be explained in an intuitive way. To understand the meaning of Statement 2 of Theorem 10, 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 10, 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 $i = 1, \ldots, k$.

2.4.5 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 10, 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 14 *The IBM with observational mechanism defined by the emission matrix (2.9), where $\varepsilon \neq 0$, produces vacuous probabilities.*

Allowing now the observational mechanism to vary over time, we obtain however the same result:

Theorem 15 *The IBM with observational mechanism for the i-th observation defined by the emission matrix*

$$\Lambda_{\varepsilon_i} := \begin{pmatrix} 1-\varepsilon_i & \varepsilon_i \\ \varepsilon_i & 1-\varepsilon_i \end{pmatrix}, \tag{2.10}$$

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

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 16 *The IBM with observational mechanism for the i-th observation defined by the emission matrix (2.10) , 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.*

Theorem 16 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 17 *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.*

2.5 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 2.4, 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 $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$ using ξ and Λ. In particular, from (4.25), 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 (2.9) 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 (2.9) 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

$$\theta_i = \frac{\xi_i - \varepsilon}{1 - 2\varepsilon}. \tag{2.11}$$

It follows that

$$\mathrm{E}(\theta_i) = \frac{\mathrm{E}(\xi_i) - \varepsilon}{1 - 2\varepsilon}. \tag{2.12}$$

2.5.1 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 (2.12) 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 18 *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 (2.9) with* $\varepsilon = 0.2$*. Applying the standard IBM with* $s = 2$ *at the actual level we obtain at the ideal level,*

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

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

2.5.2 Inference on O considering the emission matrix

What is the problem of the approach described in Section 2.5.1? 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 modelled by a scaled beta density on the interval $[\varepsilon, 1-\varepsilon]$. In fact, substituting (2.11) in the $beta(s,\mathbf{t})$ density for θ, since $d\theta = \frac{d\xi}{1-2\varepsilon}$, we obtain for ξ the density

$$\frac{C}{1-2\varepsilon}\left(\frac{\xi_1-\varepsilon}{1-2\varepsilon}\right)^{st_1-1}\left(\frac{\xi_2-\varepsilon}{1-2\varepsilon}\right)^{st_2-1}, \tag{2.13}$$

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

$$E(\theta_i) = (1-2\varepsilon)t_i + \varepsilon. \tag{2.14}$$

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 19 *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 10. The likelihood function in this case is given by

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

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.

[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.

2.6 Conclusions

In this paper we have described the behavior of the imprecise Dirichlet model when the observations are not perfect. We have modelled a situation characterized by an imperfect observational mechanism and prior near-ignorance, using a two step process. We have shown, in Sections 2.4 and 2.5, 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 near-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 near-ignorance and imperfect observations. Is it really possible to learn something, starting from prior near-ignorance and with imperfect observations?

2.7 Proofs

Proof of Proposition 4

The Gamma function satisfies the property $\Gamma(x+1) = x \cdot \Gamma(x)$. We begin proving, by induction, following equation.

$$\Gamma(s^{\mathbf{x}}) = \prod_{i=1}^{N} (s+i-1) \cdot \Gamma(s). \tag{2.15}$$

If $N = 0$ then $\Gamma(s^{\mathbf{x}}) = \Gamma(s)$. Now assume that for $N-1$ the equality $\Gamma(N-1+s) = \prod_{i=1}^{N-1}(s+i-1)\cdot\Gamma(s)$ holds, then

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

We prove now, always by induction, following equation.

$$\Gamma(s^{\mathbf{x}}\cdot t_j^{\mathbf{x}}) = \prod_{i=1}^{a_j}(st_j+i-1)\cdot\Gamma(st_j). \tag{2.16}$$

If $a_j = 0$ then $\Gamma(s^{\mathbf{x}}\cdot t_j^{\mathbf{x}}) = \Gamma(st_j)$. Now, assume that for $a_j = n-1$ the equality

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

holds. Then, for $a_j = n$ we have

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

We are now ready to prove Proposition 34. We have that

$$
\begin{aligned}
&dir(s^{\mathbf{x}}, \mathbf{t}^{\mathbf{x}}) = \\
&= \frac{\Gamma(s^{\mathbf{x}})}{\prod_{j=1}^{k} \Gamma(s^{\mathbf{x}} t_j^{\mathbf{x}})} \cdot \prod_{j=1}^{k} \theta_j^{s^{\mathbf{x}} t_j^{\mathbf{x}} - 1} = \\
&= \frac{\Gamma(s^{\mathbf{x}})}{\prod_{j=1}^{k} \Gamma(s^{\mathbf{x}} t_j^{\mathbf{x}})} \cdot \prod_{j=1}^{k} \theta_j^{a_j} \cdot \prod_{j=1}^{k} \theta_j^{st_j - 1} = \\
&\overset{(2.15)+(2.16)}{=} \frac{\prod_{i=1}^{N} (s + i - 1) \cdot \Gamma(s)}{\prod_{j=1}^{k} \cdot (\prod_{i=1}^{a_j} (st_j + i - 1)) \cdot \Gamma(st_j)} \cdot \\
&\cdot \prod_{j=1}^{k} \theta_j^{a_j} \cdot \prod_{j=1}^{k} \theta_j^{st_j - 1} = \\
&= \frac{\prod_{i=1}^{N} (s + i - 1)}{\prod_{j=1}^{k} \cdot \prod_{i=1}^{a_j} (st_j + i - 1)} \cdot \\
&\cdot \prod_{j=1}^{k} \theta_j^{a_j} \cdot \frac{\Gamma(s)}{\prod_{j=1}^{k} \Gamma(st_j)} \cdot \prod_{j=1}^{k} \theta_j^{st_j - 1} = \\
&= \frac{\prod_{i=1}^{N} (s + i - 1)}{\prod_{j=1}^{k} \cdot \prod_{i=1}^{a_j} (st_j + i - 1)} \cdot \\
&\cdot \prod_{j=1}^{k} \theta_j^{a_j} \cdot dir(s, \mathbf{t}),
\end{aligned}
$$

Therefore

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

Proof of Lemma 9

$$
\begin{aligned}
p(\theta \,|\, \mathbf{o}) &= \sum_{\mathbf{x} \in \mathcal{X}^N} p(\theta, \mathbf{x} \,|\, \mathbf{o}) = \\
&= \sum_{\mathbf{x} \in \mathcal{X}^N} p(\theta \,|\, \mathbf{x}, \mathbf{o}) \cdot P(\mathbf{x} \,|\, \mathbf{o}) = \\
&\overset{(2.7)}{=} \sum_{\mathbf{x} \in \mathcal{X}^N} p(\theta \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \mathbf{o}) = \\
&= \sum_{\mathbf{x} \in \mathcal{X}^N} \frac{P(\mathbf{x} \,|\, \theta) \cdot dir(s, \mathbf{t})}{P(\mathbf{x})} \cdot \frac{P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}{P(\mathbf{o})} = \\
&= \sum_{\mathbf{x} \in \mathcal{X}^N} \frac{P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \theta) \cdot dir(s, \mathbf{t})}{P(\mathbf{o})} = \\
&= \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \theta) \cdot dir(s, \mathbf{t})}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})}.
\end{aligned}
$$

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 (2.4) 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 5 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 (2.3), we obtain

$$
\begin{aligned}
&P(\mathrm{X} = x_i \,|\, \mathbf{o}) = \\
&= \frac{\int_\Theta \theta_i \cdot \sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o}\,|\,\mathbf{x}) \cdot P(\mathbf{x}) \cdot dir(s^{\mathbf{x}}, \mathbf{t}^{\mathbf{x}}) d\theta}{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o}\,|\,\mathbf{x}) \cdot P(\mathbf{x})} = \\
&= \frac{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o}\,|\,\mathbf{x}) \cdot P(\mathbf{x}) \cdot \int_\Theta \theta_i \cdot dir(s^{\mathbf{x}}, \mathbf{t}^{\mathbf{x}}) d\theta}{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o}\,|\,\mathbf{x}) \cdot P(\mathbf{x})} = \\
&\overset{(2.3)}{=} \frac{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o}\,|\,\mathbf{x}) \cdot P(\mathbf{x}) \cdot \frac{a_i^{\mathbf{x}} + st_i}{N+s}}{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{o}\,|\,\mathbf{x}) \cdot P(\mathbf{x})}.
\end{aligned}
$$

Proof of Theorem 10

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 (2.8) we know that

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

   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 (2.4) is a product of terms

$$
\prod_{r=1}^{a_j^{\mathbf{x}}} (st_j + r - 1). \tag{2.17}
$$

   If $a_j^{\mathbf{x}} = 0$, then (2.17) is equal to one by definition. Otherwise, if $a_j^{\mathbf{x}} > 0$ for a $j \neq i$, then (2.17) is equal to

$$
st_j \cdot \ldots \cdot (st_j + a_j^{\mathbf{x}} - 1). \tag{2.18}
$$

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 (2.18), it follows that (2.18) tends to zero as $t_i \to 1$ and thus $P(\mathbf{x}) \to 0$. At the other side we have

$$\lim_{t_i \to 1} P(\overline{\mathbf{x}}^i) \overset{(2.17)}{=} \lim_{t_i \to 1} \frac{\prod_{r=1}^{N}(st_i + r - 1)}{\prod_{j=1}^{N}(s + j - 1)} = 1.$$

It follows that

$$\begin{aligned}
&\lim_{t_i \to 1} P(\mathrm{X} = x_i \,|\, \mathbf{o}) \overset{(2.8)}{=} \\
&\overset{(2.8)}{=} \lim_{t_i \to 1} \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot \frac{a_i^{\mathbf{x}} + st_i}{N+s}}{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})} \\
&= \lim_{t_i \to 1} \frac{P(\mathbf{o} \,|\, \overline{\mathbf{x}}^i) \cdot 1 \cdot \frac{a_i^{\overline{\mathbf{x}}^i} + st_i}{N+s}}{P(\mathbf{o} \,|\, \overline{\mathbf{x}}^i) \cdot 1} = \\
&= \lim_{t_i \to 1} \frac{a_i^{\overline{\mathbf{x}}^i} + st_i}{N+s} = \frac{N+s}{N+s} = 1.
\end{aligned}$$

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

$$\begin{aligned}
&\lim_{t_i \to 0} P(\mathrm{X} = x_i \,|\, \mathbf{o}) = \\
&= \frac{\sum_{\mathbf{x} \in \mathcal{X}^N : a_i^{\mathbf{x}} = 0} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x}) \cdot \frac{a_i^{\mathbf{x}} + st_i}{N+s}}{\sum_{\mathbf{x} \in \mathcal{X}^N : a_i^{\mathbf{x}} = 0} P(\mathbf{o} \,|\, \mathbf{x}) \cdot P(\mathbf{x})},
\end{aligned} \tag{2.19}$$

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 (2.8) 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 (2.8) 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 (2.19) that $\underline{P}(x_i\,|\,\mathbf{o}) = 0$.

Proofs in Section 2.4.5

Proof of Corollary 14: Corollary 14 is a particular case of Theorem 10.

Proof of Theorem 15: the proof of Theorem 15 is equal to the proof of Theorem 10, 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.

To prove the other Theorems of Section 2.4.5 we need the following well-known Lemma:

Lemma 20 (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*

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

Proof of Theorem 16: we know from Theorem 15 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

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

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 20 we conclude that

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

and, similarly,

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

The proof of Theorem 17 is very similar to the proof of Theorem 16 and is omitted.

Chapter 3

Limits of learning about a categorical latent variable under prior near-ignorance

3.1 Summary

History

The necessary condition for learning under prior ignorance, stated in this chapter, was discovered in April 2005. This condition was the starting point for the generalization of the incompatibility between imperfect signals and prior near-ignorance.

Abstract

It is well known that a state of complete prior ignorance is not compatible with learning, at least in a coherent theory of (epistemic) uncertainty. What is less widely known, is that there is another state of beliefs, called *near-ignorance*, that is very similar to complete ignorance and that allows learning to take place. What this paper does is to provide new and substantial evidence that also near-ignorance cannot be really regarded as a way out of the problem of starting statistical inference in conditions of very weak beliefs. The key to this result is focusing on a setting characterized by a variable of interest that is *latent*. We argue that such a setting is by far the most common case in practice, and we provide, for the case of categorical

latent variables (and general *manifest* variables) a condition that, if satisfied, prevents learning to take place under prior near-ignorance. This condition is shown to be easily satisfied even in the most common statistical problems. We regard these results as a strong form of evidence against the possibility to adopt a condition of prior near-ignorance in real statistical problems.

Acknowledgements

This work was partially supported by the Swiss NSF programme NCCR FINRISK, by Swiss NSF grants 200021-113820/1 (Alberto Piatti), 200020-109295/1 (Marco Zaffalon) and 100012-105745/1 (Fabio Trojani).

3.2 Introduction

Epistemic theories of statistics are often confronted with the question of *prior ignorance*. Prior ignorance means that a subject, who is about to perform a statistical analysis, has not any substantial belief about the underlying data-generating process. Yet, the subject would like to exploit the available sample to draw some statistical inference, i.e., the subject would like to use the data to learn, moving away from the initial condition of ignorance. This situation is very important as it is often desirable to start a statistical analysis with weak assumptions about the problem of interest, thus trying to implement an objective-minded approach to statistics.

A fundamental question is if prior ignorance is compatible with learning. Walley gives a negative answer for the case of his self-consistent (or *coherent*) theory of statistics: he shows, in a very general sense, that *vacuous* prior beliefs lead to vacuous posterior beliefs, irrespective of the type and amount of observed data (Walley, 1991, Section 7.3.7).[1] At the same time, he proposes focusing on a slightly different state of beliefs, called *near-ignorance*, that does enable learning to take place (Walley, 1991, Section 4.6.9). Loosely speaking, a near-ignorance prior is a probability model on the chances of a categorical random variable, modelling a state of very weak knowledge on the chances (see Section 3.4). The fact that learning is possible under prior near-ignorance is shown, for instance, in the special case of the *imprecise Dirichlet*

[1] We recall that Walley's theory can be regarded, to a large extent, as providing the foundations of *robust Bayesian statistics*, for which the same consideration then applies.

model (IDM). This is a popular model, based on a near-ignorance prior, used in the case of inference from categorical data generated by a multinomial process (Walley (1996); Bernard (2005)).

What is important to realize, at this point, is that near-ignorance plays a crucially unique role in the question of modeling prior ignorance; the key point is that near-ignorance priors can be made to satisfy two principles: the *symmetry* and the *embedding principles*. The first is well known and is related to Laplace's *indifference principle*; the second states, loosely speaking, that if we are ignorant a priori, our beliefs on an event of interest should not depend on the space of possibilities in which the event is embedded (see Section 3.4 for a discussion about these two principles). Walley (1991) and later De Cooman and Miranda (2006) have argued extensively on the necessity of both the symmetry and the embedding principles in order to characterize a condition of ignorance.

In this paper, we investigate whether near-ignorance can be really regarded as a possible way to model ignorance in real statistical problems. To this extent, we focus, as in the case of the IDM, on a categorical random variable X expressing the outcomes of a multinomial process, but, as opposed to the IDM, we assume that such a variable is *latent*. This means that we cannot observe the realizations of X, so we can learn about it only by means of another, not necessarily categorical, variable S, related to X in some known way. Variable S is assumed to be *manifest*, in the sense that its realizations can be observed (see Section 3.3).

In such a setting, we introduce a condition in Section 3.5, related to the likelihood of the observed data, that is shown to be sufficient to prevent learning about X under prior near-ignorance. The condition is very general as it is developed for any prior that models near-ignorance (not only the one used in the IDM), and for very general kinds of relation between X and S. We show then, by simple examples, that such a condition is easily satisfied, even in the most elementary and common statistical problems.

In order to appreciate this result, it is important to realize that latent variables are ubiquitous in problems of uncertainty. It can be argued, indeed, that there is a persistent distinction between (latent) facts (e.g., health, state

of economy, color of a ball) and (manifest) observations of facts: one can regard them as being related by a so-called *observational process*; and the point is that these kinds of processes are imperfect in practice. Observational processes are often neglected in statistics, when their imperfection is deemed to be tiny. But a striking outcome of the present research is that, no matter how tiny the imperfection, provided it exists, learning is not possible under prior near-ignorance. This is shown in a definite sense in Example 29 of Section 3.5.3, where we analyze the relevance of our results for the special case of the IDM.

On our view, the present results raise serious doubts about the possibility to adopt a condition of prior near-ignorance in real, as opposed to idealized, applications of statistics. As a consequence, it may make sense to consider re-focusing the research about this subject on developing models of very weak states of belief that are, however, stronger than near-ignorance. This would involve dropping the idea that both the symmetry and the embedding principles can be realistically met in practice.

3.3 Categorical Latent Variables

In this paper, we follow the general definition of *latent* and *manifest variables* given by Skrondal and Rabe-Hesketh (2004): a *latent variable* is a random variable whose realizations are unobservable (hidden), while a *manifest variable* is a random variable whose realizations can be directly observed. The concept of latent variable is central in many sciences, like for example psychology and medicine. Skrondal and Rabe-Hesketh (2004) list several fields of application and several phenomena that can be modelled using latent variables, and conclude that latent variable modeling "*pervades modern mainstream statistics*," although "*this omni-presence of latent variables is commonly not recognized, perhaps because latent variables are given different names in different literatures, such as random effects, common factors and latent classes*," or hidden variables.

But what are latent variables in practice? According to Boorsbom et al. (2002), there may be different interpretations of latent variables. A latent variable can be regarded, for example, as an unobservable random variable

that exists independent of the observation. An example is the unobservable health status of a patient that is subject to a medical test. Another possibility is to regard a latent variable as a product of the human mind, a construct that does not exist independent of the observation. For example the *unobservable state of the economy*, often used in economic models. In this paper, we assume the existence of a latent categorical random variable X, with outcomes in $\mathcal{X} = \{x_1, \ldots, x_k\}$ and unknown chances[2] $\vartheta \in \Theta := \{\vartheta = (\vartheta_1, \ldots, \vartheta_k) \mid \sum_{i=1}^{k} \vartheta_i = 1,\ 0 \leq \vartheta_i \leq 1\}$, without stressing any particular interpretation.

Now, let us focus on a bounded real-valued function f defined on Θ, where $\vartheta \in \Theta$ are the unknown chances of X. We aim at learning the value $f(\vartheta)$ using N realizations of the variable X. Because the variable X is latent and therefore unobservable by definition, the only way to learn $f(\vartheta)$ is to observe the realizations of some manifest variable S related, in a known way, to the (unobservable) realizations of X. An example of known relationship between latent and manifest variables is the following.

Example 21 Consider a binary medical diagnostic test used to assess the health status of a patient with respect to a given disease. The accuracy of a diagnostic test[3] is determined by two probabilities: the *sensitivity* of a test is the probability of obtaining a positive result if the patient is diseased; the *specificity* is the probability of obtaining a negative result if the patient is healthy. Medical tests are assumed to be imperfect indicators of the unobservable true disease status of the patient. Therefore, we assume that the probability of obtaining a positive result when the patient is healthy, respectively of obtaining a negative result if the patient is diseased, are non-zero. Suppose, to make things simpler, that the sensitivity and the specificity of the test are known. In this example, the unobservable health status of the patient can be considered as a binary latent variable X with values in the set $\{\text{Healthy}, \text{Ill}\}$, while the result of the test can be considered as a binary manifest variable S with values in the set $\{\text{Negative result}, \text{Positive result}\}$. Because the sensitivity and the specificity of the test are known, we know how X and S are related. $\diamondsuit$

[2]Throughout the paper, we denote with ϑ a particular vector of chances in Θ and with θ a (random) variable on Θ.

[3]For further details about the modeling of diagnostic accuracy with latent variables see Yang and Becker (1997).

We continue discussion about this example later on, in the light of our results, in Example 24 of Section 3.5.

3.4 Near-Ignorance Priors

Consider a categorical random variable X with outcomes in $\mathcal{X} = \{x_1, \ldots, x_k\}$ and unknown chances $\vartheta \in \Theta$. Suppose that we have no relevant prior information about ϑ and we are therefore in a situation of prior ignorance about X. How should we model our prior beliefs in order to reflect the initial lack of knowledge?

Let us give a brief overview of this topic in the case of coherent models of uncertainty, such as Bayesian probability and Walley's theory of *coherent lower previsions*.

In the traditional Bayesian setting, prior beliefs are modelled using a single prior probability distribution. The problem of defining a standard prior probability distribution modeling a situation of prior ignorance, a so-called *noninformative prior*, has been an important research topic in the last two centuries[4] and, despite the numerous contributions, it remains an open research issue, as illustrated by Kass and Wassermann (1996). See also Hutter (2006) for recent developments and complementary considerations. There are many principles and properties that are desirable to model a situation of prior ignorance and that have been used in past research to define noninformative priors. For example Laplace's *symmetry or indifference* principle has suggested, in case of finite possibility spaces, the use of the uniform distribution. Other principles, like for example the principle of *invariance under group transformations*, the *maximum entropy* principle, the *conjugate priors* principle, etc., have suggested the use of other noninformative priors, in particular for continuous possibility spaces, satisfying one or more of these principles. But, in general, it has proven to be difficult to define a standard noninformative prior satisfying, at the same time, all the desirable principles.

[4]Starting from the work of Laplace at the beginning of the 19^{th} century (Laplace (1820)).

We follow De Cooman and Miranda (2006) when they say that there are at least two principles that should be satisfied to model a situation of prior ignorance: the *symmetry* and the *embedding principles*. The *symmetry principle* states that, if we are ignorant a priori about ϑ, then we have no reason to favour one possible outcome of X to another, and therefore our probability model on X should be symmetric. This principle recalls Laplace's *symmetry or indifference* principle that, in the past decades, has suggested the use of the *uniform prior* as standard noninformative prior. The *embedding principle* states that, for each possible event A, the probability assigned to A should not depend on the possibility space $\mathcal{X}$ in which A is embedded. In particular, the probability assigned a priori to the event A should be invariant with respect to refinements and coarsenings of $\mathcal{X}$. It is easy to show that the embedding principle is not satisfied by the uniform distribution. How should we model our prior ignorance in order to satisfy these two principles? Walley (1991) gives a compelling answer to this question: he proves[5] that the only coherent probability model on X consistent with the two principles is the *vacuous probability model*, i.e., the model that assigns, for each non-trivial event A, lower probability $\underline{\mathrm{P}}(A) = 0$ and upper probability $\overline{\mathrm{P}}(A) = 1$. Clearly, the vacuous probability model cannot be expressed using a single probability distribution. It follows that, if we agree that the symmetry and the embedding principles are characteristics of prior ignorance, then we need *imprecise probabilities* to model such a state of beliefs.[6] Unfortunately, it is easy to show that updating the vacuous probability model on X produces only vacuous posterior probabilities. Therefore, the vacuous probability model alone is not a viable way to address our initial problem. Walley (1991) suggests, as an alternative, the use of *near-ignorance priors*.

A near-ignorance prior is a probability model on the chances θ of X, modelling a very weak state of knowledge about θ. In practice, a near-ignorance prior is a large closed convex set $\mathcal{M}_0$ of prior probability densities on θ which produces *vacuous expectation*s for various functions f on Θ, i.e., such that $\underline{\mathbf{E}}(f) = \inf_{\vartheta\in\Theta} f(\vartheta)$ and $\overline{\mathbf{E}}(f) = \sup_{\vartheta\in\Theta} f(\vartheta)$. The key point is that near-ignorance priors can be designed so as to satisfy both the symmetry and the embedding principles. In fact, if a near-ignorance prior produces vacuous expectations for all the functions $f(\theta) = \theta_i$ for each $i \in \{1, \ldots, k\}$, then,

[5] In Note 7, p. 526. See also Section 5.5 of the same book.

[6] For a complementary point of view, see Hutter (2006).

because a priori $P(\mathrm{X} = x_i) = E(\theta_i)$, the near-ignorance prior implies the vacuous probability model on X and satisfies therefore both the symmetry and the embedding principle, thus delivering a satisfactory model of prior (near-)ignorance.[7] Updating a near-ignorance prior consists in updating all the probability densities in $\mathcal{M}_0$ using the Bayes rule. Because the beliefs on θ are not vacuous, we obtain thus a non-vacuous set of posterior probability densities on θ that can be used to calculate posterior probabilities for X.

A good example of near-ignorance prior is the set $\mathcal{M}_0$ used in the *imprecise Dirichlet model* (IDM). The IDM models a situation of prior ignorance about the chances θ of a categorical random variable X. The near-ignorance prior $\mathcal{M}_0$ used in the IDM consists of the set of all Dirichlet densities $p(\theta) = dir_{s,\mathbf{t}}(\theta)$ for a fixed $s > 0$ and all $\mathbf{t} \in \mathcal{T}$, where

$$dir_{s,\mathbf{t}}(\theta) := \frac{\Gamma(s)}{\prod_{i=1}^{k} \Gamma(st_i)} \prod_{i=1}^{k} \theta_i^{st_i - 1}, \tag{3.1}$$

and

$$\mathcal{T} := \{\mathbf{t} = (t_1, \ldots, t_k) \mid \sum_{j=1}^{k} t_j = 1,\ 0 < t_j < 1\}. \tag{3.2}$$

The particular choice of $\mathcal{M}_0$ in the IDM implies vacuous prior expectations for all functions $f(\theta) = \theta_i^{N'}$, for all $N' \geq 1$ and all $i \in \{1, \ldots, k\}$, i.e., $\underline{\mathbf{E}}(\theta_i^{N'}) = 0$ and $\overline{\mathbf{E}}(\theta_i^{N'}) = 1$. Choosing $N' = 1$, we have, a priori,

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

It follows that the particular near-ignorance prior $\mathcal{M}_0$ used in the IDM implies a priori the vacuous probability model on X and, therefore, satisfies both the symmetry and embedding principles. In Walley (1996), it is shown that the IDM produces, for each observed dataset, non-vacuous posterior probabilities for X.

3.5 Limits of Learning under Prior Near-Ignorance

Consider a sequence of independent and identically distributed (IID) categorical latent variables $(\mathrm{X}_i)_{i \in \mathbf{N}}$ with outcomes in $\mathcal{X}$ and unknown chances ϑ,

[7] We call this state near-ignorance because, although we are completely ignorant a priori about X, we are not completely ignorant about θ (Walley, 1991, Section 5.3, Note 4).

and a sequence of independent manifest variables $(\mathrm{S}_i)_{i\in\mathbf{N}}$. We assume that a realization of the manifest variable S_i can be observed only after a (hidden) realization of the latent variable X_i. Furthermore, we assume S_i to be independent of the chances ϑ of X_i conditional on X_i, i.e.,

$$P(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j, \vartheta) = P(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j), \tag{3.3}$$

for each $x_j \in \mathcal{X}$ and $\vartheta \in \Theta$. These assumptions model what we call an *observational process*, i.e., a two-step process where the variable S_i is used to acquire information about the realized value of X_i for each i, independently on the chances of X_i. For simplicity, we assume the probability mass function $P(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j)$ to be precise and known for each $x_j \in \mathcal{X}$ and each $i \in \mathbf{N}$.

We divide the discussion about the limits of learning under prior near-ignorance in three subsections. In Section 3.5.1 we discuss our general parametric problem and we obtain a condition that, if satisfied, prevents learning to take place. In Section 3.5.2 we study the consequences of our theoretical results in the particular case of predictive probabilities. Finally, in Section 3.5.3, we focus on the particular near-ignorance prior used in the IDM and we obtain necessary and sufficient conditions for learning with categorical manifest variables.

3.5.1 General parametric inference

We focus on a very general problem of parametric inference. Suppose that we observe a dataset $\mathbf{s}$ of realizations of manifest variables $\mathrm{S}_1, \ldots, \mathrm{S}_N$ related to the (unobservable) dataset $\mathbf{x} \in \mathcal{X}^N$ of realizations of the variables $\mathrm{X}_1, \ldots, \mathrm{X}_N$. Defining the random variables $\mathbf{X} := (\mathrm{X}_1, \ldots, \mathrm{X}_N)$ and $\mathbf{S} := (\mathrm{S}_1, \ldots, \mathrm{S}_N)$ we have $\mathbf{S} = \mathbf{s}$ and $\mathbf{X} = \mathbf{x}$. To simplify notation, when no confusion can arise, we denote in the rest of the paper $\mathbf{S} = \mathbf{s}$ with $\mathbf{s}$. Given a bounded function $f(\theta)$, our aim is to calculate $\underline{\mathbf{E}}(f \,|\, \mathbf{s})$ and $\overline{\mathbf{E}}(f \,|\, \mathbf{s})$ starting from a condition of ignorance about f, i.e., using a near ignorance prior $\mathcal{M}_0$, such that $\underline{\mathbf{E}}(f) = f_{\min} := \inf_{\vartheta\in\Theta} f(\vartheta)$ and $\overline{\mathbf{E}}(f) = f_{\max} : \sup_{\vartheta\in\Theta} f(\vartheta)$.

Is it really possible to learn something about the function f, starting from a condition of prior near-ignorance and having observed a dataset of manifest variables $\mathbf{S} = \mathbf{s}$? The following theorem shows that, very often, this is not the case. In particular, Corollary 23 shows that there is a condition that, if satisfied, prevents learning to take place.

Theorem 22 *Let* $\mathbf{s}$ *be given. Consider a bounded continuous function* f *defined on* Θ*. Then following statements hold.*[8]

1. *If the likelihood function* $\mathrm{P}(\mathbf{s} \,|\, \theta)$ *is strictly positive*[9] *in each point in which* f *reaches its maximum value* $f_{\max}$*, is continuous in an arbitrary small neighborhood of those points, and* $\mathcal{M}_0$ *is such that a priori* $\overline{\mathbf{E}}(f) = f_{\max}$*, then*
$$\overline{\mathbf{E}}(f \,|\, \mathbf{s}) = \overline{\mathbf{E}}(f) = f_{\max}.$$

2. *If the likelihood function* $\mathrm{P}(\mathbf{s} \,|\, \theta)$ *is strictly positive in each point in which* f *reaches its minimum value* $f_{\min}$*, is continuous in an arbitrary small neighborhood of those points, and* $\mathcal{M}_0$ *is such that a priori* $\underline{\mathbf{E}}(f) = f_{\min}$*, then*
$$\underline{\mathbf{E}}(f \,|\, \mathbf{s}) = \underline{\mathbf{E}}(f) = f_{\min}.$$

Corollary 23 *Let* $\mathbf{s}$ *be given and let* $\mathrm{P}(\mathbf{s} \,|\, \theta)$ *be a continuous strictly positive function on* Θ*. If* $\mathcal{M}_0$ *is such that* $\underline{\mathbf{E}}(f) = f_{\min}$ *and* $\overline{\mathbf{E}}(f) = f_{\max}$*, then*

$$\underline{\mathbf{E}}(f \,|\, \mathbf{s}) = \underline{\mathbf{E}}(f) = f_{\min},$$

$$\overline{\mathbf{E}}(f \,|\, \mathbf{s}) = \overline{\mathbf{E}}(f) = f_{\max}.$$

In other words, given $\mathbf{s}$, if the likelihood function is strictly positive, then the functions f that, according to $\mathcal{M}_0$, have vacuous expectations a priori, have vacuous expectations also a posteriori, after having observed $\mathbf{s}$. It follows that, if this sufficient condition is satisfied, we cannot use near-ignorance priors to model a state of prior ignorance because only vacuous posterior expectations are produced. The sufficient condition described above is satisfied very often in practice, as illustrated by the following striking examples.

Example 24 Consider the medical test introduced in Example 21 and an (ideally) infinite population of individuals. Denote with the binary variable $\mathrm{X}_i \in \{H, I\}$ the health status of the i-th individual of the population and

[8] The proof of this theorem is given in the appendix, together with all the other proofs of the paper.

[9] In the appendix it is shown that the assumptions of positivity of $\mathrm{P}(\mathbf{s} \,|\, \theta)$ in Theorem 22 can be substituted by the following weaker assumptions. For a given arbitrary small $\delta > 0$, denote with Θ_δ the measurable set, $\Theta_\delta := \{\vartheta \in \Theta \,|\, f(\vartheta) \geq f_{\max} - \delta\}$. If $\mathrm{P}(\mathbf{s} \,|\, \theta)$ is such that, $\lim_{\delta \to 0} \inf_{\vartheta \in \Theta_\delta} \mathrm{P}(\mathbf{s} \,|\, \vartheta) = c > 0$, then Statement 1 of Theorem 22 holds. The same holds for the second statement, substituting Θ_δ with $\widetilde{\Theta}_\delta := \{\vartheta \in \Theta \,|\, f(\vartheta) \leq f_{\min} + \delta\}$.

with $\mathrm{S}_i \in \{+,-\}$ the results of the diagnostic test applied to the same individual. We assume that the variables in the sequence $(\mathrm{X}_i)_{i\in\mathbf{N}}$ are IID with unknown chances $(\vartheta, 1-\vartheta)$, where ϑ corresponds to the (unknown) proportion of diseased individuals in the population. Denote with $1-\varepsilon_1$ the specificity and with $1-\varepsilon_2$ the sensitivity of the test. Then it holds that

$$\mathrm{P}(\mathrm{S}_i = + \,|\, \mathrm{X}_i = H) = \varepsilon_1 > 0, \quad \mathrm{P}(\mathrm{S}_i = - \,|\, \mathrm{X}_i = I) = \varepsilon_2 > 0,$$

where $(I, H, +, -)$ denote (patient ill, patient healthy, test positive, test negative).

Suppose that we observe the results of the test applied to N different individuals of the population; using our previous notation we have $\mathbf{S} = \mathbf{s}$. For each individual we have,

$$\begin{aligned}
&\mathrm{P}(\mathrm{S}_i = + \,|\, \vartheta) = \\
=&\mathrm{P}(\mathrm{S}_i = + \,|\, \mathrm{X}_i = I)\mathrm{P}(\mathrm{X}_i = I \,|\, \vartheta) + \mathrm{P}(\mathrm{S}_i = + \,|\, \mathrm{X}_i = H)\mathrm{P}(\mathrm{X}_i = H \,|\, \vartheta) = \\
=&\underbrace{(1-\varepsilon_2)}_{>0}\cdot\vartheta + \underbrace{\varepsilon_1}_{>0}\cdot(1-\vartheta) > 0.
\end{aligned}$$

Analogously,

$$\begin{aligned}
&\mathrm{P}(\mathrm{S}_i = - \,|\, \vartheta) = \\
=&\mathrm{P}(\mathrm{S}_i = - \,|\, \mathrm{X}_i = I)\mathrm{P}(\mathrm{X}_i = I \,|\, \vartheta) + \mathrm{P}(\mathrm{S}_i = - \,|\, \mathrm{X}_i = H)\mathrm{P}(\mathrm{X}_i = H \,|\, \vartheta) = \\
=&\underbrace{\varepsilon_2}_{>0}\cdot\vartheta + \underbrace{(1-\varepsilon_1)}_{>0}\cdot(1-\vartheta) > 0.
\end{aligned}$$

Denote with $n^{\mathbf{s}}$ the number of positive tests in the observed sample $\mathbf{s}$. Then, because the variables S_i are independent, we have

$$\mathrm{P}(\mathbf{S} = \mathbf{s} \,|\, \vartheta) = ((1-\varepsilon_2)\cdot\vartheta + \varepsilon_1\cdot(1-\vartheta))^{n^{\mathbf{s}}} \cdot (\varepsilon_2\cdot\vartheta + (1-\varepsilon_1)\cdot(1-\vartheta))^{N-n^{\mathbf{s}}} > 0$$

for each $\vartheta \in [0,1]$ and each $\mathbf{s} \in \mathcal{X}^N$. Therefore, according to Corollary 23, all the functions f that, according to $\mathcal{M}_0$, have vacuous expectations a priori have vacuous expectations also a posteriori. It follows that, if we want to

avoid vacuous posterior expectations, then we cannot model our prior knowledge (ignorance) using a near-ignorance prior. This simple example shows that our previous theoretical results raise serious questions about the use of near-ignorance priors also in very simple, common, and important situations.

The situation presented in this example can be extended, in a straightforward way, to a more general categorical case and has been studied, in the special case of the near-ignorance prior used in the imprecise Dirichlet model, in Piatti et al. (2005). ♢

Example 24 focuses on categorical latent and manifest variables. In the next example, we show that our theoretical results have important implications also in models with categorical latent variables and continuous manifest variables.

Example 25 Consider a sequence of IID categorical variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ with outcomes in $\mathcal{X}^N$ and unknown chances $\vartheta \in \Theta$. Suppose that, for each $i \geq 1$, after a realization of the latent variable X_i, we can observe a realization of a continuous manifest variable S_i. Assume that $p(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j)$ is a continuous positive probability density, e.g., a normal $N(\mu_j, \sigma_j^2)$ density, for each $x_j \in \mathcal{X}$. We have

$$p(\mathrm{S}_i \,|\, \vartheta) = \sum_{x_j \in \mathcal{X}^N} p(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j) \cdot \mathrm{P}(\mathrm{X}_i = x_j \,|\, \vartheta) = \sum_{x_j \in \mathcal{X}^N} \underbrace{p(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j)}_{>0} \cdot \vartheta_j > 0,$$

because ϑ_j is positive for at least one $j \in \{1, \ldots, N\}$ and we have assumed S_i to be independent of θ given X_i. Because we have assumed $(\mathrm{S}_i)_{i\in\mathbf{N}}$ to be a sequence of independent variables, we have

$$p(\mathbf{S} = \mathbf{s} \,|\, \vartheta) = \prod_{i=1}^{N} \underbrace{p(\mathrm{S}_i = \mathbf{s}_i \,|\, \vartheta)}_{>0} > 0.$$

Therefore, according to Corollary 23, if we model our prior knowledge using a near-ignorance prior $\mathcal{M}_0$, the vacuous prior expectations implied by $\mathcal{M}_0$ remain vacuous a posteriori. It follows that, if we want to avoid vacuous posterior expectations, we cannot model our prior knowledge using a near-ignorance prior. ♢

Examples 24 and 25 raise, in general, serious criticisms about the use of near-ignorance priors in real applications.

3.5.2 An important special case: predictive probabilities

We focus now on a particular and very important particular case: the case of predictive inference.[10] Suppose that our aim is to predict the outcomes of the next N' variables $\mathrm{X}_{N+1}, \ldots, \mathrm{X}_{N+N'}$. Let $\mathbf{X}' := (\mathrm{X}_{N+1}, \ldots, \mathrm{X}_{N+N'})$. If no confusion is possible, we denote $\mathbf{X}' = \mathbf{x}'$ by $\mathbf{x}'$. Given $\mathbf{x}' \in \mathcal{X}^{N'}$, our aim is to calculate $\underline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s})$ and $\overline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s})$.

Modelling our prior ignorance about the parameters θ with a near-ignorance prior $\mathcal{M}_0$ and denoting by $\mathbf{n}' := (n'_1, \ldots, n'_k)$ the frequencies of the dataset $\mathbf{x}'$, we have

$$\begin{aligned} \underline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s}) &= \inf_{p \in \mathcal{M}_0} \mathrm{P}_p(\mathbf{x}' \,|\, \mathbf{s}) = \inf_{p \in \mathcal{M}_0} \int_\Theta \prod_{i=1}^k \theta_i^{n'_i} p(\theta \,|\, \mathbf{s}) d\theta = \\ &= \inf_{p \in \mathcal{M}_0} \mathbf{E}_p \left(\prod_{i=1}^k \theta_i^{n'_i} \,|\, \mathbf{s} \right) = \underline{\mathbf{E}} \left(\prod_{i=1}^k \theta_i^{n'_i} \,|\, \mathbf{s} \right), \end{aligned} \tag{3.4}$$

where, according to Bayes rule,

$$p(\theta \,|\, \mathbf{s}) = \frac{\mathrm{P}(\mathbf{s} \,|\, \theta) p(\theta)}{\int_\Theta \mathrm{P}(\mathbf{s} \,|\, \theta) p(\theta) d\theta},$$

provided that $\int_\Theta \mathrm{P}(\mathbf{s} \,|\, \theta) p(\theta) d\theta \neq 0$. Analogously, substituting sup to inf in (3.4), we obtain

$$\overline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s}) = \overline{\mathbf{E}} \left(\prod_{i=1}^k \theta_i^{n'_i} \,|\, \mathbf{s} \right). \tag{3.5}$$

Therefore, the lower and upper probabilities assigned to the dataset $\mathbf{x}'$ a priori (a posteriori) correspond to the prior (posterior) lower and upper expectations of the continuous bounded function $f(\theta) = \prod_{i=1}^k \theta_i^{n'_i}$.

[10] For a general presentation of predictive inference see Geisser (1993); for a discussion of the imprecise probability approach to predictive inference see Walley and Bernard (1999)

It is easy to show that, in this case, the minimum of f is 0 and is reached in all the points $\vartheta \in \Theta$ with $\vartheta_i = 0$ for some i such that $n'_i > 0$, while the maximum of f is reached in a single point of Θ corresponding to the relative frequencies $\mathbf{f}'$ of the sample $\mathbf{x}'$, i.e., at $\mathbf{f}' = \left(\frac{n'_1}{N'}, \ldots, \frac{n'_k}{N'}\right) \in \Theta$, and the maximum of f is given by $\prod_{i=1}^{k} \left(\frac{n'_i}{N'}\right)^{n'_i}$. It follows that vacuous probabilities regarding the dataset $\mathbf{x}'$ are given by

$$\underline{\mathrm{P}}(\mathbf{x}') = \underline{\mathbf{E}}\left(\prod_{i=1}^{k} \theta_i^{n'_i}\right) = 0, \quad \overline{\mathrm{P}}(\mathbf{x}') = \overline{\mathbf{E}}\left(\prod_{i=1}^{k} \theta_i^{n'_i}\right) = \prod_{i=1}^{k} \left(\frac{n'_i}{N'}\right)^{n'_i}.$$

The general results stated in Section 3.5.1 hold also in the particular case of predictive probabilities. In particular, Corollary 23 can be rewritten as follows.

Corollary 26 *Let* $\mathbf{s}$ *be given and let* $\mathrm{P}(\mathbf{s} \,|\, \theta)$ *be a continuous strictly positive function on* Θ*. Then, if* $\mathcal{M}_0$ *implies vacuous prior probabilities for a dataset* $\mathbf{x}' \in \mathcal{X}^{N'}$*, the predictive probabilities of* $\mathbf{x}'$ *are vacuous also a posteriori, after having observed* $\mathbf{s}$*, i.e.,*

$$\underline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s}) = \underline{\mathrm{P}}(\mathbf{x}') = 0, \quad \overline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s}) = \overline{\mathrm{P}}(\mathbf{x}') = \prod_{i=1}^{k} \left(\frac{n'_i}{N'}\right)^{n'_i}.$$

3.5.3 Predicting the next outcome with categorical manifest variables

In this section we consider a special case for which we give necessary and sufficient conditions to learn under prior near-ignorance. These conditions are then used to analyze the IDM.

We assume that all the manifest variables in $\mathbf{S}$ are categorical. Given an arbitrary categorical manifest variable S_i, denote with $\mathcal{S}^i = \{s_1, \ldots, s_{n^i}\}$ the finite set of possible outcomes of S_i. The probabilities of S_i are defined conditional on the realized value of X_i and are given by

$$\lambda_{hj}^{\mathrm{S}_i} := P(\mathrm{S}_i = s_h \,|\, \mathrm{X}_i = x_j),$$

where $h \in \{1, \ldots, n^i\}$ and $j \in \{1, \ldots, k\}$. The probabilities of S_i can be collected in a $n^i \times k$ stochastic matrix Λ^{S_i} defined by

$$\Lambda^{\mathrm{S}_i} := \begin{pmatrix} \lambda_{11}^{\mathrm{S}_i} & \ldots & \lambda_{1k}^{\mathrm{S}_i} \\ \vdots & \ddots & \vdots \\ \lambda_{n^i 1}^{\mathrm{S}_i} & \ldots & \lambda_{n^i k}^{\mathrm{S}_i} \end{pmatrix},$$

which is called *emission matrix* of S_i.

Our aim, given $\mathbf{s}$, is to predict the next (latent) outcome starting from prior near-ignorance. In other words, our aim is to calculate $\underline{\mathrm{P}}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s})$ and $\overline{\mathrm{P}}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s})$ for each $x_j \in \mathcal{X}$, using a set of priors $\mathcal{M}_0$ such that $\underline{\mathrm{P}}(\mathrm{X}_{N+1} = x_j) = 0$ and $\overline{\mathrm{P}}(\mathrm{X}_{N+1} = x_j) = 1$ for each $x_i \in \mathcal{X}$.

A possible near-ignorance prior for this problem is the set $\mathcal{M}_0$ used in the IDM. We have seen, in Section 3.4, that this particular near-ignorance prior is such that $\underline{\mathrm{P}}(\mathrm{X}_{N+1} = x_j) = 0$ and $\overline{\mathrm{P}}(\mathrm{X}_{N+1} = x_j) = 1$ for each $x_i \in \mathcal{X}$. For this particular choice, the following theorem[11] states necessary and sufficient conditions for learning.

Theorem 27 *Let Λ^{S_i} be the emission matrix of S_i for $i = 1, \ldots, N$. Let $\mathcal{M}_0$ be the near-ignorance prior used in the IDM. Given an arbitrary observed dataset $\mathbf{s}$, we obtain a posteriori the following inferences.*

1. *If all the elements of matrices Λ^{S_i} are nonzero, then, $\overline{P}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) = 1$, $\underline{P}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) = 0$, for every $x_j \in \mathcal{X}$.*
2. *$\overline{P}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) < 1$ for some $x_j \in \mathcal{X}$, iff we observed at least one manifest variable $\mathrm{S}_i = s_h$ such that $\lambda_{hj}^{\mathrm{S}_i} = 0$.*
3. *$\underline{P}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) > 0$ for some $x_j \in \mathcal{X}$, iff we observed at least one manifest variable $\mathrm{S}_i = s_h$ such that $\lambda_{hj}^{\mathrm{S}_i} \neq 0$ and $\lambda_{hr}^{\mathrm{S}_i} = 0$ for each $r \neq j$ in $\{1, \ldots, k\}$.*

In other words, to avoid vacuous posterior predictive probabilities for the next outcome, we need at least a partial perfection of the observational process. Some simple criteria to recognize settings producing vacuous inferences are the following.

[11] Theorem 27 is a slightly extended version of Theorem 1 in Piatti et al. (2005).

Corollary 28 *Under the assumptions of Theorem 27, the following criteria hold:*

1. *If the j-th columns of matrices Λ^{S_i} have all nonzero elements, then, for each $\mathbf{s}$, $\overline{P}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) = 1$.*

2. *If the j-th rows of matrices Λ^{S_i} have more than one nonzero element, then, for each $\mathbf{s}$, $\underline{P}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) = 0$.*

Example 29 Consider again the medical test of Example 24. The manifest variable S_i (the result of the medical test applied to the i-th individual) is a binary variable with outcomes *positive* $(+)$ or *negative* $(-)$. The underlying latent variable X_i (the health status of the i-th individual) is also a binary variable, with outcomes *ill* (I) or *healthy* (H). The emission matrix in this case is the same for each $i \in \mathbf{N}$ and is the 2×2 matrix,

$$\Lambda = \begin{pmatrix} 1-\varepsilon_2 & \varepsilon_1 \\ \varepsilon_2 & 1-\varepsilon_1 \end{pmatrix}.$$

All the elements of Λ are different from zero. Therefore, using as set of priors the near-ignorance prior $\mathcal{M}_0$ of the IDM, according to Theorem 27, we are unable to move away from the initial state of ignorance. This result confirms, in the case of the near-ignorance prior of the IDM, the general result of Example 24.

It is interesting to remark that it is impossible to learn for arbitrarily small values of ε_1 and ε_2, provided that they are positive. It follows that there are situations where the observational process cannot be neglected, even when we deem it to be imperfect with tiny probability.

$\diamondsuit$

The previous example has been concerned with the case in which the IDM is applied to a latent categorical variable. Now we focus on the original setup for which the IDM was conceived, where there are no latent variables. In this case, it is well known that the IDM leads to to non-vacuous posterior predictive probabilities for the next outcome. In the next example, we show how such a setup makes the IDM avoid the theoretical limitations stated in Section 3.5.1.

Example 30 In the IDM, we assume that the IID categorical variables $(\mathrm{X}_i)_{i \in \mathbf{N}}$ are observable. In other words, we have $\mathrm{S}_i = \mathrm{X}_i$ for each $i \geq 1$

and therefore the IDM is not a latent variable model. The IDM is equivalent to a model with categorical manifest variables and emission matrices equal to the identity matrix I. Therefore, according to the second and third statements of Theorem 27, if $\mathbf{x}$ contains only observations of the type x_j, then

$$\underline{\mathrm{P}}(\mathrm{X}_{N+1} = x_j) > 0 \,, \overline{\mathrm{P}}(\mathrm{X}_{N+1} = x_j) = 1 \,, \underline{\mathrm{P}}(\mathrm{X}_{N+1} = x_h) = 0 \,, \overline{\mathrm{P}}(\mathrm{X}_{N+1} = x_h) < 1,$$

for each $h \neq j$. Otherwise, for all the other possible observed dataset $\mathbf{x}$,

$$\underline{\mathrm{P}}(\mathrm{X}_{N+1} = x_j) > 0 \,, \overline{\mathrm{P}}(\mathrm{X}_{N+1} = x_j) < 1 \,, \underline{\mathrm{P}}(\mathrm{X}_{N+1} = x_h) > 0 \,, \overline{\mathrm{P}}(\mathrm{X}_{N+1} = x_h) < 1.$$

It follows that, in general, the IDM produces, for each observed dataset $\mathbf{x}$, non-vacuous posterior predictive probabilities for the next outcome.

The IDM avoids the theoretical limitations highlighted in Section 3.5.1 thanks to its particular likelihood function. Having observed $\mathbf{S} = \mathbf{X} = \mathbf{x}$, we have

$$\mathrm{P}(\mathbf{S} = \mathbf{x} \,|\, \theta) = \mathrm{P}(\mathbf{X} = \mathbf{x} \,|\, \theta) = \prod_{i=1}^{k} \theta_i^{n_i},$$

where n_i denotes the number of times that $x_i \in \mathcal{X}$ has been observed in $\mathbf{x}$. We have $\mathrm{P}(\mathbf{X} = \mathbf{x} \,|\, \vartheta) = 0$ for all ϑ such that $\vartheta_j = 0$ for at least one j such that $n_j > 0$ and $\mathrm{P}(\mathbf{X} = \mathbf{x} \,|\, \vartheta) > 0$ for all the other $\vartheta \in \Theta$, in particular for all ϑ in the interior of Θ.

Consider, to make things simpler, that in $\mathbf{x}$ at least two different outcomes have been observed. The posterior predictive probabilities for the next outcome are obtained calculating the lower and upper expectations of the function $f(\theta) = \theta_j$ for all $j \in \{1, \ldots, k\}$. This function reaches its minimum ($f_{\min} = 0$) if $\theta_j = 0$ and its maximum ($f_{\min} = 1$) if $\theta_j = 1$. Therefore, the points where the function $f(\theta) = \theta_j$ reaches its minimum, resp. its maximum, are on the boundary of Θ and it is easy to show that the likelihood function equals zero at least in one of these points. It follows that the positivity assumptions of Theorem 22 are not met. $\diamondsuit$

Example 30 shows that we are able to learn, using a near-ignorance prior, only if the likelihood function $\mathrm{P}(\mathbf{s} \,|\, \theta)$ is equal to zero in some critical points. The likelihood function of the IDM is very peculiar, being in general equal

to zero on some parts of the boundary of Θ, and allows therefore to use a near-ignorance prior $\mathcal{M}_0$ that models in a satisfactory way a condition of prior (near-)ignorance.[12]

Yet, since the variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ are assumed to be observable, the successful application of a near-ignorance prior in the IDM is not helpful in addressing the doubts raised by our theoretical results about the applicability of near-ignorance priors in situations, where the variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ are latent, as shown in Example 29.

3.6 Conclusions

In this paper we have proved a sufficient condition that prevents learning about a latent categorical variable to take place under prior near-ignorance regarding the data-generating process.

The condition holds as soon as the likelihood is strictly positive (and continuous), and so is satisfied frequently, even in the simplest settings. Taking into account that the considered framework is very general and pervasive of statistical practice, we regard this result as a form of strong evidence against the possibility to use prior near-ignorance in real statistical problems.

As a consequence, we suggest that future research efforts should be directed to study and develop new forms of knowledge that are close to near-ignorance but that do not coincide with it. This might involve modifying the symmetry and the embedding principles so as to capture a notion of quasi-ignorance that can be effectively implemented in real statistical problems.

3.7 Technical preliminaries

In this appendix we prove some technical results that are used to prove the theorems in the paper. First of all, we introduce some notation used in

[12] See Walley (1996) and Bernard (2005) for a more detailed discussion on the theoretical properties of the IDM.

this appendix. Consider a sequence of probability densities $(p_n)_{n\in\mathbf{N}}$ and a function f defined on a set Θ. Then, we use the notation,

$$\mathbf{E}_n(f) := \int_\Theta f(\theta)p_n(\theta)d\theta, \quad \mathrm{P}_n(\widetilde{\Theta}) := \int_{\widetilde{\Theta}} p_n(\theta)d\theta, \quad \widetilde{\Theta} \subseteq \Theta,$$

and with $\to$ we denote $\lim_{n\to\infty}$.

Theorem 31 *Let $\Theta \subset \mathbf{R}^k$ be the closed k-dimensional simplex and let $(p_n)_{n\in\mathbf{N}}$ be a sequence of probability densities defined on Θ w.r.t. the Lebesgue measure. Let $f \geq 0$ be a bounded continuous function on Θ and denote with $f_{\max}$ the Supremum of f on Θ and with $f_{\min}$ the Infimum of f on Θ. For this function define the measurable sets*

$$\Theta_\delta = \{\vartheta \in \Theta \,|\, f(\vartheta) \geq f_{\max} - \delta\}, \tag{3.6}$$

$$\widetilde{\Theta}_\delta = \{\vartheta \in \Theta \,|\, f(\vartheta) \leq f_{\min} + \delta\}. \tag{3.7}$$

1. *Assume that $(p_n)_{n\in\mathbf{N}}$ concentrates on a maximum of f for $n \to \infty$, in the sense that*

$$\mathbf{E}_n(f) \to f_{\max}, \tag{3.8}$$

then, for all $\delta > 0$, it holds

$$\mathrm{P}_n(\Theta_\delta) \to 1.$$

2. *Assume that $(p_n)_{n\in\mathbf{N}}$ concentrates on a minimum of f for $n \to \infty$, in the sense that*

$$\mathbf{E}_n(f) \to f_{\min}, \tag{3.9}$$

then, for all $\delta > 0$, it holds

$$\mathrm{P}_n(\widetilde{\Theta}_\delta) \to 1.$$

Proof. We begin by proving the first statement. Let $\delta > 0$ be arbitrary and $\bar{\Theta}_\delta := \Theta \setminus \Theta_\delta$. From (3.6) we know that on Θ_δ it holds $f(\theta) \geq f_{\max} - \delta$, and therefore on $\bar{\Theta}_\delta$ we have $f(\theta) \leq f_{\max} - \delta$, and thus

$$\frac{f_{\max} - f(\theta)}{\delta} \geq 1. \tag{3.10}$$

It follows that

$$\begin{aligned} 1 - \mathrm{P}_n(\Theta_\delta) &= \mathrm{P}_n(\bar{\Theta}_\delta) = \int_{\bar{\Theta}_\delta} p_n(\theta)d\theta \overset{(3.10)}{\leq} \int_{\bar{\Theta}_\delta} \frac{f_{\max} - f(\theta)}{\delta} p_n(\theta)d\theta \\ &\leq \int_{\Theta} \frac{f_{\max} - f(\theta)}{\delta} p_n(\theta)d\theta = \frac{1}{\delta}(f_{\max} - \mathbf{E}_n(f)) \overset{(3.9)}{\longrightarrow} 0, \end{aligned}$$

and therefore $\mathrm{P}_n(\Theta_\delta) \to 1$ and thus the first statement is proved. To prove the second statement, let $\delta > 0$ be arbitrary and $\widehat{\Theta}_\delta := \Theta \setminus \widetilde{\Theta}_\delta$. From (3.7) we know that on $\widetilde{\Theta}_\delta$ it holds $f(\theta) \leq f_{\min} + \delta$, and therefore on $\widehat{\Theta}_\delta$ we have $f(\theta) \geq f_{\min} + \delta$, and thus

$$\frac{f(\theta) - f_{\min}}{\delta} \geq 1. \tag{3.11}$$

It follows that

$$\begin{aligned} 1 - \mathrm{P}_n(\widetilde{\Theta}_\delta) &= \mathrm{P}_n(\widehat{\Theta}_\delta) = \int_{\widehat{\Theta}_\delta} p_n(\theta)d\theta \overset{(3.11)}{\leq} \int_{\widehat{\Theta}_\delta} \frac{f(\theta) - f_{\min}}{\delta} p_n(\theta)d\theta \\ &\leq \int_{\Theta} \frac{f(\theta) - f_{\min}}{\delta} p_n(\theta)d\theta = \frac{1}{\delta}(\mathbf{E}_n(f) - f_{\min}) \overset{(3.9)}{\longrightarrow} 0, \end{aligned}$$

and therefore $\mathrm{P}_n(\widetilde{\Theta}_\delta) \to 1$. ■

Theorem 32 *Let $L(\theta) \geq 0$ be a bounded measurable function and suppose that the Assumptions of Theorem 31 hold. Then the following two statements hold.*

1. *If the function $L(\theta)$ is such that*

$$\lim_{\delta \to 0} \inf_{\theta \in \Theta_\delta} L(\theta) =: c > 0, \tag{3.12}$$

and $(p_n)_{n \in \mathbf{N}}$ concentrates on a maximum of f for $n \to \infty$, then

$$\frac{\mathbf{E}_n(Lf)}{\mathbf{E}_n(L)} = \frac{\int_\Theta f(\theta)L(\theta)p_n(\theta)d\theta}{\int_\Theta L(\theta)p_n(\theta)d\theta} \to f_{\max}. \tag{3.13}$$

2. If the function $L(\theta)$ is such that

$$\lim_{\delta \to 0} \inf_{\theta \in \widetilde{\Theta}_\delta} L(\theta) =: c > 0, \tag{3.14}$$

and $(p_n)_{n \in \mathbf{N}}$ concentrates on a minimum of f for $n \to \infty$, then

$$\frac{\mathbf{E}_n(Lf)}{\mathbf{E}_n(L)} \longrightarrow f_{\min}. \tag{3.15}$$

Remark 33 *If L is strictly positive in each point in Θ where the function f reaches its maximum, resp. minimum, and is continuous in an arbitrary small neighborhood of those points, then (3.12), resp. (3.14), are satisfied.*

Proof. We begin by proving the first statement of the theorem. Fix ε and δ arbitrarily small, but δ small enough such that $\inf_{\vartheta \in \Theta_\delta} L(\vartheta) \geq \frac{c}{2}$. Denote with $L_{\max}$ the supremum of the function $L(\theta)$ in Θ. From Theorem 31, we know that $\mathrm{P}_n(\Theta_\delta) \geq 1 - \varepsilon$, for n sufficiently large. This implies, for n sufficiently large,

$$\mathbf{E}_n(L) = \int_\Theta L(\theta) p_n(\theta) d\theta \geq \int_{\Theta_\delta} L(\theta) p_n(\theta) d\theta \geq \frac{c}{2}(1 - \varepsilon), \tag{3.16}$$

$$\mathbf{E}_n(Lf) \leq \mathbf{E}_n(L f_{\max}) = f_{\max} \mathbf{E}_n(L), \tag{3.17}$$

$$\begin{aligned} \mathbf{E}_n(L) &= \int_{\bar{\Theta}_\delta} L(\theta) p_n(\theta) d\theta + \int_{\Theta_\delta} L(\theta) p_n(\theta) d\theta \\ &\leq L_{\max} \int_{\bar{\Theta}_\delta} p_n(\theta) d\theta + \int_{\Theta_\delta} \underbrace{\frac{f(\theta)}{f_{\max} - \delta}}_{\geq 1 \text{ on } \Theta_\delta} L(\theta) p_n(\theta) d\theta \\ &\leq L_{\max} \cdot \varepsilon + \frac{1}{f_{\max} - \delta} \mathbf{E}_n(Lf). \end{aligned} \tag{3.18}$$

Combining (3.16), (3.17) and (3.18), we have

$$f_{\max} \geq \frac{\mathbf{E}_n(Lf)}{\mathbf{E}_n(L)} \geq (f_{\max} - \delta) \frac{\mathbf{E}_n(L) - L_{\max} \cdot \varepsilon}{\mathbf{E}_n(L)} \geq (f_{\max} - \delta) \left(1 - \frac{L_{\max} \cdot \varepsilon}{\frac{c}{2}(1 - \varepsilon)} \right).$$

Since the right hand side of the last inequality tends to $f_{\max}$ for $\delta, \varepsilon \to 0$, and both δ, ε can be chosen arbitrarily small, we have

$$\frac{\mathbf{E}_n(Lf)}{\mathbf{E}_n(L)} \to f_{\max}.$$

To prove the second statement of the theorem, fix ε and δ arbitrarily small, but δ small enough such that $\inf_{\vartheta \in \widetilde{\Theta}_\delta} L(\vartheta) \geq \frac{c}{2}$. From Theorem 31, we know that $\mathrm{P}_n(\widetilde{\Theta}_\delta) \geq 1 - \varepsilon$, for n sufficiently large and therefore $\mathrm{P}_n(\widehat{\Theta}_\delta) \leq \varepsilon$. This implies, for n sufficiently large,

$$\mathbf{E}_n(L) = \int_\Theta L(\theta) p_n(\theta) d\theta \geq \int_{\widetilde{\Theta}_\delta} L(\theta) p_n(\theta) d\theta \geq \frac{c}{2}(1 - \varepsilon), \tag{3.19}$$

$$\mathbf{E}_n(Lf) \geq \mathbf{E}_n(Lf_{\min}) = f_{\min} \mathbf{E}_n(L) \Rightarrow f_{\min} \leq \frac{\mathbf{E}_n(Lf)}{\mathbf{E}_n(L)}. \tag{3.20}$$

Define the function

$$K(\theta) := \left(1 - \frac{f(\theta)}{f_{\min} + \delta}\right) L(\theta).$$

By definition, the function K is negative on $\widehat{\Theta}_\delta$ and is bounded. Denote with $K_{\min}$ the (negative) minimum of K. We have

$$\begin{aligned}
\mathbf{E}_n(L) &= \int_{\widehat{\Theta}_\delta} L(\theta) p_n(\theta) d\theta + \int_{\widetilde{\Theta}_\delta} L(\theta) p_n(\theta) d\theta \\
&\geq \int_{\widehat{\Theta}_\delta} L(\theta) p_n(\theta) d\theta + \int_{\widetilde{\Theta}_\delta} \underbrace{\frac{f(\theta)}{f_{\min} + \delta}}_{\leq 1 \text{ on } \widetilde{\Theta}_\delta} L(\theta) p_n(\theta) d\theta \\
&= \int_{\widehat{\Theta}_\delta} \underbrace{\left(L(\theta) - \frac{f(\theta)}{f_{\min} + \delta} L(\theta)\right)}_{=K(\theta)} p_n(\theta) d\theta + \frac{1}{f_{\min} + \delta} \underbrace{\int_\Theta f(\theta) L(\theta) p_n(\theta) d\theta}_{=\mathbf{E}_n(Lf)} \\
&\geq K_{\min} \cdot \mathrm{P}_n(\widehat{\Theta}_\delta) + \frac{1}{f_{\min} + \delta} \cdot \mathbf{E}_n(Lf).
\end{aligned}$$

It follows that

$$\left(\mathbf{E}_n(L) - K_{\min} \cdot \mathrm{P}_n(\widehat{\Theta}_\delta)\right)(f_{\min} + \delta) \geq \mathbf{E}_n(Lf),$$

and thus, combining the last inequality with (3.19) and (3.20), we obtain

$$\begin{aligned} f_{\min} \leq \frac{\mathbf{E}_n(Lf)}{\mathbf{E}_n(L)} &\leq (f_{\min}+\delta)\left(1+\frac{|K_{\min}|\cdot P_n(\widehat{\Theta}_\delta)}{\mathbf{E}_n(L)}\right) \\ &\leq (f_{\min}+\delta)\left(1+\frac{|K_{\min}|\cdot\varepsilon}{\frac{c}{2}(1-\varepsilon)}\right). \end{aligned}$$

Since the right hand side of the last inequality tends to $f_{\min}$ for $\delta, \varepsilon \to 0$, and both δ, ε can be chosen arbitrarily small, we have

$$\frac{\mathbf{E}_n(Lf)}{\mathbf{E}_n(L)} \to f_{\min}.$$

■

3.8 Proofs of the main results

3.8.1 Proof of Theorem 22 and Corollary 23

Define, $f_{\min} := \inf_{\vartheta\in\Theta} f(\vartheta)$, $f_{\max} := \sup_{\vartheta\in\Theta} f(\vartheta)$, and define the bounded non-negative function $\tilde{f}(\theta) := f(\theta) - f_{\min} \geq 0$. We have, $\tilde{f}_{\max} = f_{\max} - f_{\min}$. If $\mathcal{M}_0$ is such that a priori, $\overline{\mathbf{E}}(f) = f_{\max}$, then we have also that $\overline{\mathbf{E}}(\tilde{f}) = \tilde{f}_{\max}$, because,

$$\overline{\mathbf{E}}(\tilde{f}) = \sup_{p\in\mathcal{M}_0} E_p(f - f_{\min}) = \sup_{p\in\mathcal{M}_0} E_p(f) - f_{\min} = \overline{\mathbf{E}}(f) - f_{\min} = f_{\max} - f_{\min} = \tilde{f}_{\max}.$$

Then, it is possible to define a sequence $(p_n)_{n\in\mathbf{N}} \subset \mathcal{M}_0$ such that $\mathbf{E}_n(\tilde{f}) \to \tilde{f}_{\max}$. According to Theorem 32, substituting $L(\theta)$ with $\mathrm{P}(\mathbf{s}\,|\,\theta)$ in (3.13), we see that $\mathbf{E}_n(\tilde{f}\,|\,\mathbf{s}) \to \tilde{f}_{\max} = \overline{\mathbf{E}}(\tilde{f})$ and therefore $\overline{\mathbf{E}}(\tilde{f}\,|\,\mathbf{s}) = \overline{\mathbf{E}}(\tilde{f})$, from which follows that,

$$\overline{\mathbf{E}}(f\,|\,\mathbf{s}) - f_{\min} = \overline{\mathbf{E}}(f) - f_{\min} = f_{\max} - f_{\min}.$$

We can conclude that, $\overline{\mathbf{E}}(f\,|\,\mathbf{s}) = \overline{\mathbf{E}}(f) = f_{\max}$. In the same way, substituting $\underline{\mathbf{E}}$ to $\overline{\mathbf{E}}$, we can prove that $\underline{\mathbf{E}}(f\,|\,\mathbf{s}) = \underline{\mathbf{E}}(f) = f_{\min}$.

Corollary 23 is a direct consequence of Theorem 22. ■

3.8.2 Proof of Theorem 27 and Corollary 28

To prove Theorem 27 we need following lemma.

Lemma 34 *Consider a dataset* $\mathbf{x}$ *with frequencies* $\mathbf{a} = (a_1^{\mathbf{x}}, \ldots, a_k^{\mathbf{x}})$. *Then, the following equality holds,*

$$\prod_{h=1}^{k} \theta_h^{a_h^{\mathbf{x}}} \cdot dir_{s,\mathbf{t}}(\theta) = \frac{\prod_{h=1}^{k} \cdot \prod_{j=1}^{a_h^{\mathbf{x}}} (st_h + j - 1)}{\prod_{j=1}^{N} (s + j - 1)} \cdot dir_{s^{\mathbf{x}},\mathbf{t}^{\mathbf{x}}}(\theta),$$

where $s^{\mathbf{x}} := N + s$ *and* $t_h^{\mathbf{x}} := \frac{a_h^{\mathbf{x}} + st_h}{N+s}$. *When* $a_h^{\mathbf{x}} = 0$, *we set* $\prod_{j=1}^{0}(st_h + j - 1) := 1$ *by definition.*

A proof of Lemma 34 is in Piatti et al. (2005). Because $\mathrm{P}(\mathbf{x} \,|\, \theta) = \prod_{h=1}^{k} \theta_h^{a_h^{\mathbf{x}}}$, according to Bayes rule, we have $p(\theta \,|\, \mathbf{x}) = dir_{s^{\mathbf{x}},\mathbf{t}^{\mathbf{x}}}(\theta)$ and

$$P(\mathbf{x}) = \frac{\prod_{h=1}^{k} \prod_{l=1}^{a_h^{\mathbf{x}}} (st_h + l - 1)}{\prod_{l=1}^{N} (s + l - 1)}. \tag{3.21}$$

Given a Dirichlet distribution $dir_{s,\mathbf{t}}(\theta)$, the expected value $\mathbf{E}(\theta_j)$ is given by $\mathbf{E}(\theta_j) = t_j$ (see Kotz et al. (2000)). It follows that

$$\mathbf{E}(\theta_j \,|\, \mathbf{x}) = \mathbf{t}_j^{\mathbf{x}} = \frac{a_j^{\mathbf{x}} + st_j}{N + s}.$$

We are now ready to prove Theorem 27.

1. The first statement of Theorem 27 is a consequence of Corollary 26. Because S_i is independent of θ given X_i for each $i \in \mathbf{N}$, we have

   $$\mathrm{P}(\mathbf{s} \,|\, \mathbf{x}, \theta) = \mathrm{P}(\mathbf{s} \,|\, \mathbf{x}), \tag{3.22}$$

   and therefore, using (3.22) and Bayes rule, we obtain the likelihood function,

   $$L(\theta) = P(\mathbf{s} \,|\, \theta) = \sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{s} \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \theta) = \sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{s} \,|\, \mathbf{x}) \cdot \prod_{h=1}^{k} \theta_h^{a_h^{\mathbf{x}}}. \tag{3.23}$$

Because all the elements of the matrices Λ^{S_i} are nonzero, we have $P(\mathbf{s}\,|\,\mathbf{x}) > 0$, for each $\mathbf{s}$ and each $\mathbf{x} \in \mathcal{X}^N$. For each $\vartheta \in \Theta$, there is at least one $\mathbf{x} \in \mathcal{X}^N$ such that $\prod_{h=1}^{k} \vartheta_h^{a_h^{\mathbf{x}}} > 0$. It follows that,

$$L(\vartheta) = \sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{s}\,|\,\mathbf{x}) \cdot \prod_{j=1}^{k} \vartheta_j^{a_j^{\mathbf{x}}} > 0,$$

for each $\vartheta \in \Theta$ and therefore, according to Corollary 26 with $N' = 1$, the predictive probabilities that are vacuous a priori remain vacuous also a posteriori.

2. We have $\mathrm{P}(\mathrm{X}_{N+1} = x_j\,|\,\mathbf{s}) = \mathbf{E}(\theta_j\,|\,\mathbf{s})$, and therefore, according to Lemma 34 and Bayes rule,

$$\begin{aligned}
P(\mathrm{X}_{N+1} = x_j\,|\,\mathbf{s}) &= \frac{\int_\Theta \theta_j P(\mathbf{s}\,|\,\theta)p(\theta)d\theta}{\int_\Theta P(\mathbf{s}\,|\,\theta)p(\theta)d\theta} = \\
&\overset{(3.22)}{=} \frac{\sum_{\mathbf{x}\in\mathcal{X}^N} \int_\Theta \theta_j P(\mathbf{s}\,|\,\mathbf{x})P(\mathbf{x}\,|\,\theta)p(\theta)d\theta}{\sum_{\mathbf{x}\in\mathcal{X}^N} \int_\Theta P(\mathbf{s}\,|\,\mathbf{x})P(\mathbf{x}\,|\,\theta)p(\theta)d\theta} = \\
&= \frac{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{s}\,|\,\mathbf{x})P(\mathbf{x}) \int_\Theta \theta_j p(\theta\,|\,\mathbf{x})d\theta}{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{s}\,|\,\mathbf{x})P(\mathbf{x})} = \\
&= \sum_{\mathbf{x}\in\mathcal{X}^N} \left(\frac{P(\mathbf{s}\,|\,\mathbf{x})P(\mathbf{x})}{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{s}\,|\,\mathbf{x})P(\mathbf{x})}\right) \cdot \mathbf{E}(\theta_j\,|\,\mathbf{x}), \\
&= \sum_{\mathbf{x}\in\mathcal{X}^N} \left(\frac{P(\mathbf{s}\,|\,\mathbf{x})P(\mathbf{x})}{\sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{s}\,|\,\mathbf{x})P(\mathbf{x})}\right) \cdot \frac{a_j^{\mathbf{x}} + st_j}{N+s} \qquad (3.24)
\end{aligned}$$

(3.24) is a convex sum of fractions and is therefore a continuous function of t on $\mathcal{T}$. Denote with $\overline{\mathbf{x}}^j$ the dataset of length N composed only by outcomes x_j, i.e., the dataset with $a_j^{\overline{\mathbf{x}}^j} = N$ and $a_h^{\overline{\mathbf{x}}^j} = 0$ for each $h \neq j$. For all $\mathbf{x} \neq \overline{\mathbf{x}}^j$ we have

$$\frac{a_j^{\mathbf{x}} + st_j}{N+s} \leq \frac{N-1+st_j}{N+s} \leq \frac{N-1+s}{N+s} < 1,$$

on $\overline{\mathcal{T}}$ (the closure of $\mathcal{T}$), only $\overline{\mathbf{x}}^j$ has

$$\sup_{t\in\mathcal{T}} \frac{a_j^{\overline{\mathbf{x}}^j} + st_j}{N+s} = \sup_{t\in\mathcal{T}} \frac{N+st_j}{N+s} = 1.$$

A convex sum of fractions smaller than or equal to one is equal to one, only if the weights associated to fractions smaller than one are all equal to zero and there are some positive weights associated to fractions equal to one. If $P(\mathbf{s} \,|\, \overline{\mathbf{x}}^j) = 0$, then (3.24) is a convex combination of fractions strictly smaller than 1 on $\overline{\mathcal{T}}$ and therefore $\overline{P}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) < 1$. If $P(\mathbf{s} \,|\, \overline{\mathbf{x}}^j) \neq 0$, then letting $t_j \to 1$, and consequently $t_h \to 0$ for all $h \neq j$, according to (3.21), we have $P(\overline{\mathbf{x}}^j) \to 1$ and $P(\mathbf{x}) \to 0$ for all $\mathbf{x} \neq \overline{\mathbf{x}}^j$, and thus, using (3.24),

$$1 \geq \overline{P}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) \geq \lim_{t_j \to 1} P(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) = \frac{P(\mathbf{s} \,|\, \overline{\mathbf{x}}^j) P(\overline{\mathbf{x}}^j) \frac{N+s}{N+s}}{P(\mathbf{s} \,|\, \overline{\mathbf{x}}^j) P(\overline{\mathbf{x}}^j)} = 1.$$

If we have observed a manifest variable $\mathrm{S}_i = s_h$ with $\lambda_{hj}^{\mathrm{S}_t} = 0$, it means that the observation excludes the possibility that the underlying value of X_i is x_j, therefore $P(\mathbf{s} \,|\, \overline{\mathbf{x}}^j) = 0$ and thus

$$\overline{P}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) < 1.$$

On the other hand, if $\overline{P}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) < 1$, it must hold that $P(\mathbf{s} \,|\, \overline{\mathbf{x}}^j) = 0$, i.e., that we have observed a realization of a manifest that is incompatible with the underlying (latent) outcome x_j. But a realization of a manifest that is incompatible with the underlying (latent) outcome only if the observed manifest variable was $\mathrm{S}_i = s_h$ with $\lambda_{hj}^{\mathrm{S}_i} = 0$.

3. Having observed a manifest variable $\mathrm{S}_i = s_h$, such that $\lambda_{hj}^{\mathrm{S}_i} \neq 0$ and $\lambda_{hr}^{\mathrm{S}_i} = 0$ for each $r \neq j$ in $\{1, \ldots, k\}$, we are sure that the underlying value of X_i is x_j. Therefore, $P(\mathbf{s} \,|\, \mathbf{x}) = 0$ for all $\mathbf{x}$ with $a_j^{\mathbf{x}} = 0$. It follows from (3.24) that

$$P(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) = \frac{\sum_{\mathbf{x} \in \mathcal{X}^N,\, a_j^{\mathbf{x}} > 0} P(\mathbf{s} \,|\, \mathbf{x}) P(\mathbf{x}) \cdot \frac{a_j^{\mathbf{x}} + s t_j}{N+s}}{\sum_{\mathbf{x} \in \mathcal{X}^N,\, a_j^{\mathbf{x}} > 0} P(\mathbf{s} \,|\, \mathbf{x}) P(\mathbf{x})},$$

which is a convex combination of terms

$$\frac{a_j^{\mathbf{x}} + s t_j}{N+s} \geq \frac{a_j^{\mathbf{x}}}{N+s} \geq \frac{1}{N+s},$$

and is therefore greater than zero for each $t \in \overline{\mathcal{T}}$. It follows that

$$\underline{P}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) \geq \frac{1}{N+s} > 0.$$

On the other hand, if we do not observe a signal as described above, it exists surely at least one $\mathbf{x}$ with $a_j^{\mathbf{x}} = 0$ and $P(\mathbf{s} \,|\, \mathbf{x}) > 0$. In this case, using (3.24) and letting $t_j \to 0$, we have, because of (3.21), that $P(\mathbf{x}) \to 0$ for all $\mathbf{x}$ with $a_j^{\mathbf{x}} > 0$. It follows that

$$\lim_{t_j \to 0} P(\mathrm{X} = x_j \,|\, \mathbf{s}) = \lim_{t_j \to 0} \frac{\sum_{\mathbf{x} \in \mathcal{X}^N,\, a_j^{\mathbf{x}} = 0} P(\mathbf{s} \,|\, \mathbf{x}) P(\mathbf{x}) \cdot \frac{a_j^{\mathbf{x}} + s t_j}{N+s}}{\sum_{\mathbf{x} \in \mathcal{X}^N,\, a_j^{\mathbf{x}} = 0} P(\mathbf{s} \,|\, \mathbf{x}) P(\mathbf{x})}.$$

Assume for simplicity that, for all $h \neq j$, $t_h \not\to 0$, then $P(\mathbf{x}) > 0$ for all $\mathbf{x}$ with $a_j^{\mathbf{x}} = 0$ and $P(\mathbf{x}) \not\to 0$. Because, with $a_j^{\mathbf{x}} = 0$, we have

$$\lim_{t_j \to 0} \frac{a_j^{\mathbf{x}} + s t_j}{N+s} = \lim_{t_j \to 0} \frac{0 + s t_i}{N+s} = 0,$$

we obtain directly,

$$0 \leq \underline{P}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) = \inf_{t \in \mathcal{T}} P(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) \leq \lim_{t_j \to 0} P(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) = 0.$$

Corollary 28 is a direct consequence of Theorem 27. ■

Chapter 4

Learning from quasi perfect observations under prior near-ignorance: the binary case

4.1 Summary

History

The condition of Quasi Perfection, that is the starting point of the present chapter, was defined in March 2006. A first version of the paper was submitted in June 2006, as Piatti et al. (2006b). The paper is currently under revision.

Abstract

The imprecise Beta model (IBM) of Bernard (1996) and Walley (1996) is the most popular model for learning about a binary random variable under prior near-ignorance. Piatti et al. (2005) show that there is a fundamental issue with the interpretation of results produced by the IBM in applications. When the possibility that data may contain errors can be excluded, the IBM is able to learn from each sequence of observations of the variable of interest. However, in the more realistic case in which observations may be affected by errors, the IBM is unable to learn. In this paper, we propose a modified approach that allows learning from imperfect observations under a weak specification of prior knowledge if the probability of error is small.

The approach is based on an additional assumption that seems natural and acceptable in applications with moderate probabilities of observation errors. We show that the results produced by the modified model are arbitrarily close to those produced by the IBM, when the probability of observation errors is smaller than a pre-specified threshold that depends on the desired accuracy level. This last finding yields a possible explanation for the usefulness of the IBM in applications characterized by a small probability of observation errors.

Acknowledgements

This work was partially supported by the Swiss NSF programme NCCR FINRISK, by the Swiss NSF grant 100012-105745/1 (Fabio Trojani) and by the Swiss NSF grant 200020-109295/1 (Marco Zaffalon).

4.2 Introduction

Modelling prior ignorance is a fundamental issue in Bayesian statistics and its generalizations to robust statistics or imprecise probabilities.[1] Walley (1996) has proposed an appealing model of prior ignorance, which has been implemented in one of the most popular imprecise-probability models: the imprecise Dirichlet model (IDM).[2] In the IDM, prior ignorance about the chances[3] of a categorical random variable is modelled by a set of prior probability densities. These prior densities imply *vacuous* predictive probabilities, i.e., the probabilities of the underlying categorical variable are only known to be between 0 and 1. After having observed a sequence of realizations of the variable of interest, the prior densities are updated using Bayes rule to obtain a set of posterior probability densities that imply non-vacuous posterior probabilities for the underlying random variable.

In this paper, we focus on the two-dimensional version of the IDM, the imprecise Beta model (IBM) (Bernard (1996) and Walley (1996)). As the IDM, the IBM assumes that data are perfectly observed. This assumption is unrealistic for many applications. Piatti et al. (2005) have relaxed the

[1] See Walley (1991) for a comprehensive introduction to this topic.

[2] Bernard (2005) provides a detailed overview of the IDM.

[3] We call chances the physical probabilities of a random variable and we call probabilities the epistemic probabilities.

assumption of perfect observations and considered the IBM in a setting of *imperfect observations*. They show that, in such a relaxed framework, vacuous posterior probabilities arise irrespective of the amount of available data. This feature prevents the IBM to learn from the data. Piatti et al. (2006a) have extended that result by showing that learning is impossible as soon as prior near-ignorance and imperfect observations arise jointly. This incompatibility is mainly due to the presence of extreme priors, arbitrarily close to the degenerate ones, in any set of prior densities specified according to Walley (1996). To learn from imperfect observations under a weak specification of prior knowledge, it is therefore necessary to modify the IBM by some additional assumption which restricts the set of priors, to drop those arbitrarily near to the degenerate ones.

This paper studies a modification of the IBM, the modified IBM (MIBM), for the case in which the probability ε of observation errors is small. In Section 4.4 we propose an additional assumption for the IBM, called *quasi perfection* (QP), which seems natural and acceptable when ε is small. We show that Assumption (QP) is incompatible with near-degenerate priors and leads to a restriction of the set of prior densities. The severity of the restriction depends on the strength of Assumption (QP) and the probability of error ε. The approach resulting from the addition of Assumption (QP) to the IBM enables to learn from imperfect observations under a specification of prior knowledge which can be, depending on the size of ε, arbitrarily weak, although never compatible with the definition of prior ignorance.

To study the relation between the MIBM and the IBM, we compare in Section 4.5 the results produced by these two settings. We first show that the difference between the posterior probabilities of the two models is bounded in a way that depends only on the strength of Assumption (QP) and the size of the data set. Moreover, given an arbitrary small tolerance $\delta > 0$, there exists a maximum probability of error ε_{max} such that the distance between the prior, the posterior and the lower and upper probabilities produced by the two models is bounded by δ. This last finding offers a possible explanation for the usefulness of the IBM in applications when the probability of error is small.

4.3 The imprecise Beta model for imperfect observations

Consider a sequence of independent and identically distributed (IID) binary random variables $\mathrm{X}_1, \ldots, \mathrm{X}_{N+1}$, for $N \geq 1$, taking values in the set $\mathcal{X} = \{x_1, \ldots, x_k\}$. We assume that X_i is unobservable for each $i = 1, \ldots, N$. If X_i is realized, then a random variable O_i with values in $\mathcal{X}$ is observed, such that

$$P(\mathrm{O}_i = x \,|\, \mathrm{X}_i = \overline{x}) = P(\mathrm{O}_i = \overline{x} \,|\, \mathrm{X}_i = x) = \varepsilon,$$

for each $i = 1, \ldots, N$. In this setting, X_i represents the ideal *realized* value of a binary variable and O_i its *observed* value. If $\varepsilon = 0$, then $\mathrm{O}_i = \mathrm{X}_i$ almost surely and the observations are called *perfect*. If $\varepsilon > 0$, O_i can be different from X_i, and the observations are called *imperfect*.

Define the random variables $\mathbf{X} := (\mathrm{X}_1, \ldots, \mathrm{X}_N)$ and $\mathbf{O} := (\mathrm{O}_1, \ldots, \mathrm{O}_N)$. For simplicity, we re-label by X the random variable of interest X_{N+1}. Suppose that $\mathbf{X}$ is realized with a corresponding observation $\mathbf{O}$. Our aim is to calculate the posterior predictive probability

$$P(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}),$$

starting from a condition of near-ignorance a priori.

To model prior near-ignorance, we extend the *Imprecise Beta Model* (IBM) (Bernard (1996) and Walley (1996)) to our partial information setting. In this context, the relevant prior densities are beta densities parameterized by a parameter $t \in]0, 1[$. For any given $t \in]0, 1[$ it is possible to calculate the desired predictive probabilities using the formula:[4]

$$P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \sum_{\mathbf{x} \in \mathcal{X}^N} \left(\frac{P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{X} = \mathbf{x}) P_t(\mathbf{X} = \mathbf{x})}{\sum_{\mathbf{y} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{X} = \mathbf{y}) P_t(\mathbf{X} = \mathbf{y})} \right) \cdot \frac{a^{\mathbf{x}} + st}{N + s}, \tag{4.1}$$

with[5]

$$P_t(\mathbf{X} = \mathbf{x}) = \frac{\prod_{i=1}^{a^{\mathbf{x}}} (st + i - 1) \prod_{j=1}^{N - a^{\mathbf{x}}} (s(1-t) + j - 1)}{\prod_{k=1}^{N} (s + k - 1)}, \tag{4.2}$$

[4] See Piatti et al. (2005) for a formal derivation of equation (4.1).

[5] For $a^{\mathbf{x}} = 0$, we set $\prod_{i=1}^{0} (st + i - 1) = 1$, by definition. For $a^{\mathbf{x}} = N$, we set $\prod_{j=1}^{0} (s(1-t) + j - 1) = 1$, by definition.

and

$$P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{X} = \mathbf{x}) = \varepsilon^{\mathbf{n}} \cdot (1 - \varepsilon)^{N - \mathbf{n}}, \tag{4.3}$$

where $a^{\mathbf{x}}$ is the number of outcomes equal to x in the data set $\mathbf{x}$, $\mathbf{n}$ is the number of wrong observations in $\mathbf{o}$ with respect to $\mathbf{x}$ and $s > 0$ is the hyperparameter of the IBM. In this setting, the parameter t is the probability assigned a priori to the outcome x.

The function $P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})$ has the following important property, which allows us to compute easily lower and upper probabilities with respect to t.

Theorem 35 *For each $\mathbf{o} \in \mathcal{X}^N$ and $\varepsilon > 0$, the function $P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})$ is strictly increasing in t .*

The proof of Theorem 35 and all the other proofs of this paper are in Section 4.7 and Section 4.8. From Theorem 35, it follows that

$$\overline{P}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) := \sup_{t \in]0,1[} P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \lim_{t \to 1} P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}),$$

$$\underline{P}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) := \inf_{t \in]0,1[} P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \lim_{t \to 0} P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}).$$

But, in Piatti et al. (2005), we have shown that for each $\varepsilon > 0$ and $\mathbf{o} \in \mathcal{X}^N$ it follows

$$\underline{P}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \lim_{t \to 0} P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = 0,$$

$$\overline{P}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \lim_{t \to 1} P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = 1.$$

Therefore, the direct extension of the IBM to a setting with imperfect observations does not lead to useful results: if there is a positive probability ε of an observation error the extension produces only vacuous probabilities.

As shown in Piatti et al. (2006a), the inability to learn from imperfect observations is not due to the particular structure of the set of prior densities in the IBM, but is a more general incompatibility between the definition of prior near-ignorance and the existence of imperfect observations. This incompatibility is mainly due to the presence of prior densities that are arbitrarily close to the deterministic degenerate ones in any set of prior densities specified according to Walley (1996). In the IBM, these degenerate priors correspond to values of t very close to 0 or 1. In order to learn from imperfect observations, it would be therefore enough to restrict slightly the admissible values of t. This task can be achieved by introducing some weak assumption which restricts the prior near-ignorance in a way that avoids quasi degenerate prior densities.

4.4 Quasi perfect observations

Given an observed data set $\mathbf{o}$, we say that $\mathbf{o}$ was observed *quasi perfectly* if,

$$P(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \geq K \cdot P(\mathbf{X} \neq \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}), \tag{QP}$$

for some $K > 1$, where K represents the strength of the assumption (QP). A person that assumes (QP) puts a higher conditional likelihood on the hypothesis that the observed data set $\mathbf{o}$ is free from observation errors, with respect to the alternative of the existence of observation errors. We refer to Assumption (QP) as the assumption of *quasi perfection*. This assumption seems appropriate for applications in which the probability of observation errors ε is very small. However, one has to be careful in interpreting (QP) as an assumption only about ε. Indeed, by Bayes rule we have

$$P(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o}) = \frac{P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{X} = \mathbf{x}) P(\mathbf{X} = \mathbf{x})}{\sum_{\mathbf{y} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{X} = \mathbf{y}) P(\mathbf{X} = \mathbf{y})}, \tag{4.4}$$

where

$$P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{X} = \mathbf{x}) = \varepsilon^{\mathbf{n}} \cdot (1 - \varepsilon)^{N - \mathbf{n}}.$$

In addition,

$$P(\mathbf{X} = \mathbf{x}) = \int_0^1 P(\mathbf{X} = \mathbf{x} \,|\, \theta) p(\theta) d\theta,$$

where

$$P(\mathbf{X} = \mathbf{x} \,|\, \theta) = \theta^{\,a^{\mathbf{x}}} \cdot (1 - \theta)^{N - a^{\mathbf{x}}},$$

and p is the prior density on the unknown chance of X. Therefore, Assumption (QP) depends on both ε and the prior belief p about the chances of X.

In settings of prior near-ignorance the prior ignorance is modelled using a set $\mathcal{P}$ of prior densities. In this case, (QP) holds only if it holds for each $p \in \mathcal{P}$. In the IBM, the set $\mathcal{P}$ is the set of all $\text{beta}_{s,t}(\theta)$ densities indexed by a fixed $s > 0$ and all $t \in]0, 1[$, defined by:

$$\text{beta}_{s,t}(\theta) := \frac{\Gamma(s)}{\Gamma(st)\Gamma(s(1-t))} \theta^{st-1} (1 - \theta)^{s(1-t)-1}.$$

In this case, we know from (4.2) that

$$P_t(\mathbf{X} = \mathbf{x}) = \frac{\prod_{i=1}^{a^{\mathbf{x}}} (st + i - 1) \prod_{j=1}^{N - a^{\mathbf{x}}} (s(1-t) + j - 1)}{\prod_{k=1}^{N} (s + k - 1)}.$$

It follows that the IBM satisfies the assumption (QP) if and only if

$$P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \geq K \cdot P_t(\mathbf{X} \neq \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}), \tag{4.5}$$

or, equivalently, if and only if

$$P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \geq \frac{K}{K+1}, \tag{4.6}$$

for each $t \in]0,1[$. As shown in Theorem 36 and Corollary 37, the IBM does not satisfy the condition (4.6) and consequently does not satisfy Assumption (QP) if $\varepsilon > 0$. This result is surprising at first sight because ε can be arbitrarily small. However, it is quite intuitive if we consider that t can assume extreme values, arbitrarily close to 0 or 1.

Theorem 36 *Consider the function $P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o})$, with $t \in]0,1[$, where $\mathbf{o} \in \mathcal{X}^N$ is a given data set and $\varepsilon > 0$.*

1. *If $0 < a^{\mathbf{o}} < N$, $P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o})$ has a unique global maximum in $]0,1[$ and no local extremes. Furthermore,*

$$\lim_{t \to 0} P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) = \lim_{t \to 1} P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) = 0.$$

2. *If $a^{\mathbf{o}} = N$, $P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o})$ is strictly monotone increasing in $]0,1[$. Furthermore,*

$$\lim_{t \to 0} P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) = 0, \quad \lim_{t \to 1} P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) = 1.$$

3. *If $a^{\mathbf{o}} = 0$, $P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o})$ is strictly monotone decreasing in $]0,1[$. Furthermore,*

$$\lim_{t \to 0} P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) = 1, \quad \lim_{t \to 1} P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) = 0.$$

In particular, Theorem 36 implies that Assumption (QP) is satisfied in the IBM only by prior densities such that t belongs to a proper subset of $]0,1[$.

Corollary 37 *Let (QP) be satisfied by some $t \in]0,1[$. Given the observed data set $\mathbf{o}$, the following results hold.*

1. *If $0 < a^{\mathbf{o}} < N$, there exist unique thresholds $t_{\min} > 0$ and $t_{\max} < 1$ such that (QP) holds if and only if $t \in [t_{\min}, t_{\max}]$.*

2. *if* $a^{\mathbf{o}} = N$*, there exists a unique threshold* $t_{\min} > 0$ *such that (QP) holds if and only if* $t \in [t_{\min}, 1[$.

3. *if* $a^{\mathbf{o}} = 0$*, there exists a unique threshold* $t_{\max} < 1$ *such that (QP) holds if and only if* $t \in]0, t_{\max}]$.

It follows that Assumption (QP) is equivalent to a restriction on the set of admissible prior densities. We thus obtain a new approach to learning under weak prior knowledge. For the rest of the paper we call the IBM with the additional assumption (QP) the *modified IBM* (MIBM).

Let I_ε be the restricted interval of admissible values for t implied by (QP). I_ε depends on the observed sample $\mathbf{o}$, the probability of errors ε and the value K in the assumption. The dependence of I_ε on the parameters ε and K is characterized by Theorem 38.

Theorem 38 *For each observed data set* $\mathbf{o}$*, if the interval* I_ε *is not empty, then it is monotone decreasing in* ε *and* K *with respect to the partial order* $\subset$*. Furthermore*

$$\lim_{\varepsilon \to 0} I_\varepsilon =]0, 1[.$$

From Theorem 38, larger values of ε imply more severe restrictions on the admissible values of t. The same argument can be expressed in the reverse way. For instance, we can ask which is the largest value of ε in the MIBM consistent with an arbitrarily weak restriction $\delta > 0$ on the admissible values of t, i.e., such that

$$[\delta, 1 - \delta] \subset I_\varepsilon.$$

From Theorem 38 it follows immediately that such a value of ε exists for each observed data set $\mathbf{o}$, as summarized by the following corollary.

Corollary 39 *For each observed data set* $\mathbf{o}$*, each* $K > 1$ *and each* $\delta < \frac{1}{2}$*, there exists* $\varepsilon_{\max} > 0$ *such that*

$$[\delta, 1 - \delta] \subset I_\varepsilon \Leftrightarrow \varepsilon < \varepsilon_{\max}.$$

According to Corollary 37, I_ε is a proper subset of $]0, 1[$ for each $\varepsilon > 0$. Therefore, the prior probabilities of the MIBM are not vacuous. It follows that the MIBM is incompatible with the specification of prior near-ignorance. However, Corollary 39 ensures that the restriction of the set of priors densities can be arbitrarily weak. Consequently, the implied specification of prior

knowledge can be arbitrarily near to a condition of prior near-ignorance, provided that ε is sufficiently small. The restriction of the interval of admissible values of t is a pragmatic solution to the problem of vacuous posterior probabilities under imperfect observations. Assumption (QP) restricts the set of admissible values of t in a transparent and interpretable way, which produces non-vacuous posterior predictive probabilities in the MIBM.

4.5 Comparison between the IBM and the MIBM

From a fundamental point of view the IBM and the MIBM are very different models. The IBM assumes that observation errors have zero probability; this strong assumption allows to specify prior knowledge using the definition of prior near-ignorance. The result is a model that produces non-vacuous posterior probabilities. In particular, setting $\varepsilon = 0$ in equation (4.1), it follows for each $t \in]0, 1[$:

$$P_t^{\text{IBM}}(X = x \,|\, \mathbf{O} = \mathbf{o}) := \frac{a^{\mathbf{o}} + st}{N + s}. \tag{4.7}$$

Therefore,

$$\underline{P}^{\text{IBM}}(X = x \,|\, \mathbf{O} = \mathbf{o}) := \frac{a^{\mathbf{o}}}{N + s},$$
$$\overline{P}^{\text{IBM}}(X = x \,|\, \mathbf{O} = \mathbf{o}) := \frac{a^{\mathbf{o}} + s}{N + s}.$$

In the MIBM it is assumed that observation errors are possible with probability $\varepsilon > 0$. In addition, Assumption (QP) is imposed. This leads to a restriction of the interval of admissible values of t, which is incompatible with the definition of prior near-ignorance used in the IBM.

Despite their conceptual difference, the MIBM and the IBM produce, under suitable conditions on ε and K, similar results.

Theorem 40 *Given an observed data set* $\mathbf{o}$ *and the corresponding restricted interval* I_ε *implied by Assumption (QP), it follows for each* $t \in I_\varepsilon$*:*

$$|P_t(X = x \,|\, \mathbf{O} = \mathbf{o}) - P_t^{\text{IBM}}(X = x \,|\, \mathbf{O} = \mathbf{o})| < \frac{1}{K + 1} \frac{N}{N + s}.$$

Theorem 40 shows that, for large values of K and each $t \in I_\varepsilon$, the distance between the MIBM and the IBM is small. A stronger reverse argument is developed in Theorem 41. There, we specify an arbitrary maximal tolerance $\delta > 0$ and show that the difference of the prior, the posterior and the lower and upper predictive probabilities of the MIBM and the IBM can be made smaller than δ when ε is sufficiently small. At the same time, the specification of prior knowledge can be made arbitrarily weak for $\delta \to 0$.

Theorem 41 *For each observed data set* $\mathbf{o}$ *and* $0 < \delta < \frac{1}{2}$*, there exists* $\varepsilon_{\max} > 0$ *such that* $[\delta, 1-\delta] \subset I_\varepsilon$ *for each* $\varepsilon < \varepsilon_{\max}$ *and the following inequalities hold:*

$$|\underline{P}(X = x \,|\, \mathbf{O} = \mathbf{o}) - \underline{P}^{\mathrm{IBM}}(X = x \,|\, \mathbf{O} = \mathbf{o})| < \delta, \tag{4.8}$$

$$|\overline{P}(X = x \,|\, \mathbf{O} = \mathbf{o}) - \overline{P}^{\mathrm{IBM}}(X = x \,|\, \mathbf{O} = \mathbf{o})| < \delta. \tag{4.9}$$

Moreover,

$$|P_t(X = x \,|\, \mathbf{O} = \mathbf{o}) - P_t^{\mathrm{IBM}}(X = x \,|\, \mathbf{O} = \mathbf{o})| < \delta, \tag{4.10}$$

for all $t \in I_\varepsilon$*.*

The parameter δ in Theorem 41 can be interpreted as a measure of the size of the differences in the sets of prior and posterior predictive probabilities implied by the two approaches. If the distance between the prior, the posterior and the lower and upper probabilities of the two settings is smaller than δ, and δ is chosen sufficiently small, then we might consider the two approaches as equivalent for practical purposes. Theorem 41 ensures that, given an observed data set $\mathbf{o}$, there exists a threshold $\varepsilon_{\max}$ such that the MIBM and the IBM are equivalent if $\varepsilon < \varepsilon_{\max}$. This result yields a possible explanation for the usefulness of the IBM in applications characterized by imperfect observations with small probability of observation errors.

Example 42 *Suppose that we have observed a data set* $\mathbf{o}$ *with* $N = 10$ *and* $a^{\mathbf{o}} = 5$*. We specify a maximal tolerance* $\delta = 0.01$ *and verify if the MIBM and the IBM are equivalent at this level of tolerance, given a probability of observation errors* ε*.*

For $\varepsilon = 0.05$*, the interval* I_ε *is not empty for each* $K < 1.174$*. Setting* $K = 1$*, we obtain*

$$I_\varepsilon = [0.012, 0.988] \not\supseteq [\delta, 1-\delta] = [0.01, 0.99].$$

The construction of the interval I_ε in this and the next case is illustrated in Figure 4.1. Because the interval I_ε is monotone decreasing in K, for each $K > 1$ we have $I_\varepsilon = [0.012, 0.988] \not\supseteq [\delta, 1-\delta]$. Therefore the MIBM and the IBM are not equivalent in this case. According to our tolerance $\delta = 0.01$, the probability of observation errors $\varepsilon = 0.05$ is too large for the two models to be equivalent.

If $\varepsilon = 0.01$, the interval I_ε is not empty for each $K < 7.8$. Setting $K = 6.5$, we obtain

$$I_\varepsilon = [0.0000083, 0.9999917] \supset [0.01, 0.99].$$

Moreover,

$$|\underline{P}(X = x \,|\, \mathbf{O} = \mathbf{o}) - \underline{P}^{\mathrm{IBM}}(X = x \,|\, \mathbf{O} = \mathbf{o})| \cong 0.0092 < \delta = 0.01,$$

$$|\underline{P}(X = x \,|\, \mathbf{O} = \mathbf{o}) - \underline{P}^{\mathrm{IBM}}(X = x \,|\, \mathbf{O} = \mathbf{o})| \cong 0.0092 < \delta = 0.01,$$

and

$$\max_{t \in I_\varepsilon} |P_t(X = x \,|\, \mathbf{O} = \mathbf{o}) - P_t^{\mathrm{IBM}}(X = x \,|\, \mathbf{O} = \mathbf{o})| \cong 0.0092 < \delta = 0.01.$$

In this case, the two approaches are equivalent given the tolerance $\delta = 0.01$. This is illustrated in Figure 4.2. From these results, we deduce that in this example $\varepsilon_{\max}$ is larger than 0.01 and smaller than 0.05.

4.6 Conclusions

The specification of prior near-ignorance in Walley (1996) is incompatible with imperfect observations: a direct extension of the IBM to this case leads to vacuous posterior probabilities. This result is due to the presence of quasi-deterministic priors in the set of priors densities of the IBM.

In this paper we propose a modification of the IBM that enables to learn from imperfect observations for small values of the observation error probability ε. The proposed approach, called modified IBM (MIBM), is based on the additional assumption of quasi perfection. This assumption restricts in a natural way the set of admissible prior densities in the IBM and eliminates the problematic quasi-degenerate prior densities. The specification of prior knowledge in the MIBM can be arbitrarily weak, although never compatible

with the exact definition of prior near-ignorance in Walley (1996). Furthermore, the probabilities produced by the MIBM and the IBM can be made arbitrarily close when ε is sufficiently small. This result yields a possible explanation for the usefulness of the IBM in applications where the probability of observation errors is small.

The results of this paper are a first step towards a theory of learning from imperfect observations under weak prior knowledge. Further research is needed to understand the implications of the assumption of quasi perfection in the general multinomial case.

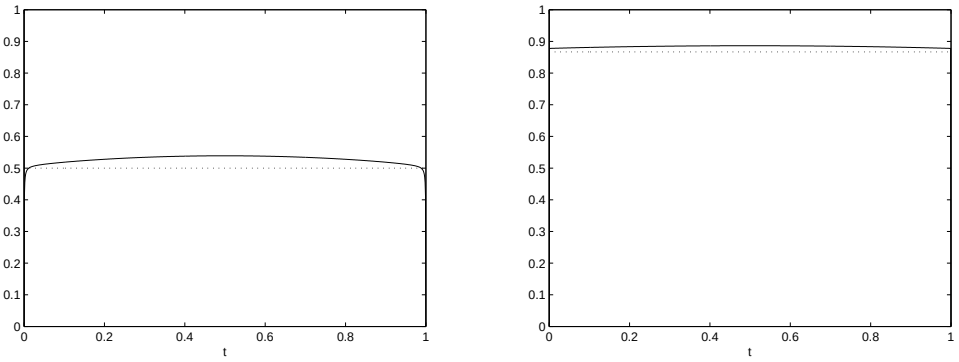

Figure 4.1: The function $P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o})$ for an observed dataset $\mathbf{o}$ of length $N = 10$ with $a^{\mathbf{o}} = 5$ and $s = 2$. In the left panel the function is plotted for $\varepsilon = 0.05$. Setting $K := 1$ in the Assumption (QP), we find I_ε searching the values of t for which the function is larger than $\frac{1}{1+1} = 0.5$. With a bisection algorithm with precision 10^{-14}, we find that $I_\varepsilon = [0.012, 0.988]$. In the right panel the same function is plotted for $\varepsilon = 0.01$. Setting $K := 6.5$ we search the values of t for which the function is larger than $\frac{6.5}{6.5+1} \sim 0.8667$. Using the same bisection algorithm as above, we find $I_\varepsilon = [0.0000083, 0.9999917]$.

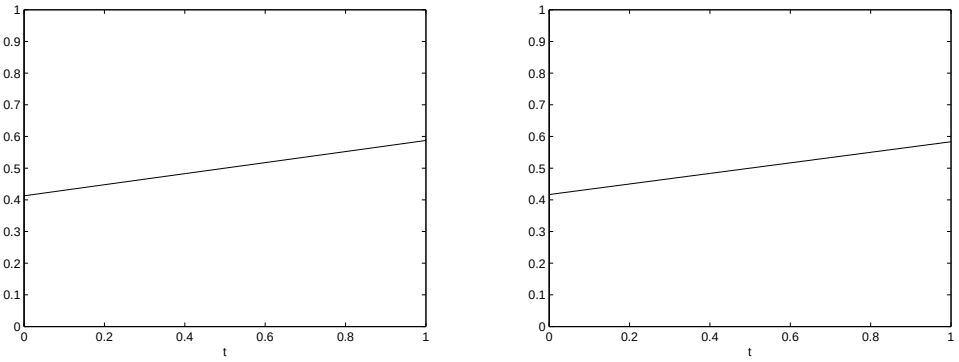

Figure 4.2: The values of $P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})$ (left panel) and of $P_t^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})$ (right panel) for $K = 6.5$. The results produced are so similar, and the restriction on the possible values of t implied by (QP) in this case so small, that the two plots are nearly indistinguishable.

4.7 Technical results

Lemma 43 *Consider the functions $f_i(t) \in C^2(]0,1[)$ for $i = 1, \ldots, m$, where $m \geq 1$.*

1. *If $f_i(t) > 0$, $f_i'(t) > 0$ and $f_i''(t) > 0$ for each $i = 1, \ldots, m$ and each $t \in]0,1[$, then*

$$\left(\prod_{i=1}^{m} f_i(t)\right) > 0, \qquad \left(\prod_{i=1}^{m} f_i(t)\right)' > 0, \qquad \left(\prod_{i=1}^{m} f_i(t)\right)'' > 0,$$

   for each $t \in]0,1[$.

2. *If $f_i(t) > 0$, $f_i'(t) < 0$ and $f_i''(t) > 0$ for each $i = 1, \ldots, m$ and each $t \in]0,1[$, then*

$$\left(\prod_{i=1}^{m} f_i(t)\right) > 0, \qquad \left(\prod_{i=1}^{m} f_i(t)\right)' < 0, \qquad \left(\prod_{i=1}^{m} f_i(t)\right)'' > 0,$$

   for each $t \in]0,1[$.

Proof. We prove the lemma by induction on m. We begin with $m = 2$. Clearly $f_1(t) \cdot f_2(t) > 0$ because both functions are positive. For the first derivative we have

$$(f_1(t) \cdot f_2(t))' = f_1'(t) \cdot f_2(t) + f_1(t) \cdot f_2'(t).$$

If $f_i'(t) > 0$ for $i = 1, 2$, $(f_1(t) \cdot f_2(t))' > 0$. If $f_i'(t) < 0$ for $i = 1, 2$, $(f_1(t) \cdot f_2(t))' < 0$. For the second derivative we have

$$(f_1(t) \cdot f_2(t))'' = f_1''(t) \cdot f_2(t) + 2f_1'(t) \cdot f_2'(t) + f_1(t) \cdot f_2''(t).$$

If $f_i'(t) > 0$ and $f_i''(t) > 0$ for $i = 1, 2$, $(f_1(t) \cdot f_2(t))'' > 0$. If $f_i'(t) < 0$ and $f_i''(t) > 0$ for $i = 1, 2$, $f_1'(t) \cdot f_2'(t) > 0$ and therefore $(f_1(t) \cdot f_2(t))'' > 0$.

Denote now with $f(t)$ the function

$$f(t) = \prod_{i=1}^{m-1} f_i(t).$$

Assume that $f(t) > 0$, $f'(t) > 0$ and $f''(t) > 0$. Substituting $f_1(t)$ with $f(t)$ and $f_2(t)$ with $f_m(t)$ in the calculations above we obtain

$$\left(\prod_{i=1}^{m} f_i(t)\right) > 0, \qquad \left(\prod_{i=1}^{m} f_i(t)\right)' > 0, \qquad \left(\prod_{i=1}^{m} f_i(t)\right)'' > 0.$$

In the same way, assuming $f(t) > 0$, $f'(t) < 0$ and $f''(t) > 0$ we obtain that

$$\left(\prod_{i=1}^{m} f_i(t)\right) > 0, \qquad \left(\prod_{i=1}^{m} f_i(t)\right)' < 0, \qquad \left(\prod_{i=1}^{m} f_i(t)\right)'' > 0.$$

■

Lemma 44 *Let $M, H \geq 0$ be two real constants, $s > 0$ and $t \in]0, 1[$. Then the function*

$$f(t) := \frac{st + H}{s(1-t) + M}$$

is such that $f(t) > 0$, $f'(t) > 0$ and $f''(t) > 0$.

Proof. We have $st+H > 0$, $s(1-t)+M > 0$ and hence $f(t) > 0$. Calculating the first derivative we obtain

$$\dot{f}(t) = \frac{s(s(1-t)+M) - (st+H)(-s)}{(s(1-t)+M)^2} = \frac{s(M+H)+s^2}{(s(1-t)+M)^2} > 0.$$

Finally, calculating the second derivative, we obtain

$$f''(t) = \underbrace{(s(M+H)+s^2)}_{>0} \cdot \underbrace{(-2(s(1-t)+M)^{-3})}_{<0} \cdot \underbrace{(-s)}_{<0} > 0.$$

■

Lemma 45 *Let $M, H \geq 0$ be two real constants, $s > 0$ and $t \in]0, 1[$. Then the function*

$$f(t) := \frac{s(1-t) + H}{st + M}$$

is such that $f(t) > 0$, $f'(t) < 0$ and $f''(t) > 0$.

Proof. We have $s(1-t)+H > 0$, $st+M > 0$ and therefore $f(t) > 0$. Calculating the first derivative we obtain

$$\dot{f}(t) = \frac{-s(st+M) - s(s(1-t)+H)}{(st+M)^2} = -\frac{(s(M+H)+s^2)}{(st+M)^2} < 0.$$

Finally, calculating the second derivative we obtain

$$f''(t) = \underbrace{-(s(M+H)+s^2)}_{<0} \cdot \underbrace{(-2(st+M)^{-3})}_{<0} \cdot \underbrace{s}_{>0} > 0.$$

■

Lemma 46 *Let $\mathbf{o}, \mathbf{x} \in \mathcal{X}^N$ be fixed but arbitrary.*

1. *If $a^{\mathbf{x}} = a^{\mathbf{o}}$, it follows*

$$\left(\frac{P_t(\mathbf{X}=\mathbf{x})}{P_t(\mathbf{X}=\mathbf{o})}\right)'' = 0.$$

2. *If $a^{\mathbf{x}} \neq a^{\mathbf{o}}$, it follows*

$$\left(\frac{P_t(\mathbf{X}=\mathbf{x})}{P_t(\mathbf{X}=\mathbf{o})}\right)'' > 0.$$

Proof. From (4.2) we know that

$$P_t(\mathbf{X}=\mathbf{x}) = \frac{\prod_{i=1}^{a^{\mathbf{x}}}(st+i-1)\prod_{j=1}^{N-a^{\mathbf{x}}}(s(1-t)+j-1)}{\prod_{k=1}^{N}(s+k-1)},$$

therefore

$$\frac{P_t(\mathbf{X}=\mathbf{x})}{P_t(\mathbf{X}=\mathbf{o})} = \frac{\prod_{i=1}^{a^{\mathbf{x}}}(st+i-1)\prod_{j=1}^{N-a^{\mathbf{x}}}(s(1-t)+j-1)}{\prod_{h=1}^{a^{\mathbf{o}}}(st+h-1)\prod_{k=1}^{N-a^{\mathbf{o}}}(s(1-t)+k-1)}.$$

If $a^{\mathbf{x}} = a^{\mathbf{o}}$

$$\frac{P_t(\mathbf{X}=\mathbf{x})}{P_t(\mathbf{X}=\mathbf{o})} = 1 \Rightarrow \left(\frac{P_t(\mathbf{X}=\mathbf{x})}{P_t(\mathbf{X}=\mathbf{o})}\right)'' = 0.$$

If $a^{\mathbf{x}} > a^{\mathbf{o}}$, setting $K := a^{\mathbf{x}} - a^{\mathbf{o}}$, we have

$$\begin{aligned}\frac{P_t(\mathbf{X}=\mathbf{x})}{P_t(\mathbf{X}=\mathbf{o})} &= \frac{\prod_{h=a^{\mathbf{o}}+1}^{a^{\mathbf{x}}}(st+h-1)}{\prod_{k=N-a^{\mathbf{o}}}^{N-a^{\mathbf{o}}}(s(1-t)+k-1)} \\ &= \frac{\prod_{i=1}^{K}(st+a^{\mathbf{o}}+i-1)}{\prod_{j=1}^{K}(s(1-t)+N-a^{\mathbf{x}}+j-1)} \\ &= \prod_{i=1}^{K}\left(\frac{st+(a^{\mathbf{o}}+i-1)}{s(1-t)+(N-a^{\mathbf{x}}+i-1)}\right). \qquad (4.11)\end{aligned}$$

Because of Lemma 44, each factor in (4.11) is a function that satisfies the assumptions of the first statement of Lemma 43. Therefore we conclude that

$$\left(\frac{P_t(\mathbf{X}=\mathbf{x})}{P_t(\mathbf{X}=\mathbf{o})}\right)'' > 0.$$

If $a^{\mathbf{x}} < a^{\mathbf{o}}$, setting $K := a^{\mathbf{o}} - a^{\mathbf{x}}$, we have

$$\begin{aligned}\frac{P_t(\mathbf{X}=\mathbf{x})}{P_t(\mathbf{X}=\mathbf{o})} &= \frac{\prod_{i=1}^{K}(s(1-t)+N-a^{\mathbf{o}}+i-1)}{\prod_{j=1}^{K}(st+a^{\mathbf{x}}+j-1)} \\ &= \prod_{i=1}^{K}\left(\frac{s(1-t)+(N-a^{\mathbf{o}}+i-1)}{st+(a^{\mathbf{x}}+i-1)}\right). \qquad (4.12)\end{aligned}$$

Because of Lemma 45, each factor in (4.12) is a function that satisfies the assumptions of the second statement of Lemma 43. Therefore we conclude that

$$\left(\frac{P_t(\mathbf{X}=\mathbf{x})}{P_t(\mathbf{X}=\mathbf{o})}\right)'' > 0.$$

■

4.8 Proofs

Proof of Theorem 35

For simplicity, we use following notation:

$$P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{x}) := P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{x}),$$

$$P_t(\mathbf{x}) := P_t(\mathbf{X} = \mathbf{x}),$$
$$P'_t(\mathbf{x}) := \frac{\delta}{\delta t} P_t(\mathbf{X} = \mathbf{x}).$$

We show that $\frac{\delta}{\delta t} P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) > 0$ for each $t \in]0,1[$, each $\mathbf{o} \in \mathcal{X}^N$ and each $\varepsilon > 0$.

We first note that Equation (4.1) can be rewritten as

$$P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \frac{1}{N+s} \frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{x}) P_t(\mathbf{x}) a^{\mathbf{x}}}{\sum_{\mathbf{y} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{y}) P_t(\mathbf{y})} + \frac{st}{N+s}.$$

Therefore,

$$\frac{\delta}{\delta t} P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \underbrace{\frac{1}{N+s}}_{>0} \frac{\delta}{\delta t} \left(\frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{x}) P_t(\mathbf{x}) a^{\mathbf{x}}}{\sum_{\mathbf{y} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{y}) P_t(\mathbf{y})} \right) + \underbrace{\frac{s}{N+s}}_{>0}. \tag{4.13}$$

In the rest of the proof, we show:

$$\frac{\delta}{\delta t} \left(\frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{x}) P_t(\mathbf{x}) a^{\mathbf{x}}}{\sum_{\mathbf{y} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{y}) P_t(\mathbf{y})} \right) \geq 0.$$

Calculating the derivative on the left hand side, we obtain

$$\frac{\delta}{\delta t} \left(\frac{\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{x}) P_t(\mathbf{x}) a^{\mathbf{x}}}{\sum_{\mathbf{y} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{y}) P_t(\mathbf{y})} \right) = \frac{1}{\left(\sum_{\mathbf{y} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{y}) P_t(\mathbf{y}) \right)^2} \cdot$$

$$\cdot \left(\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{x}) \dot{P}_t(\mathbf{x}) a^{\mathbf{x}} \cdot \sum_{\mathbf{y} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{y}) P_t(\mathbf{y}) - \right.$$

$$\left. - \sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{x}) P_t(\mathbf{x}) a^{\mathbf{x}} \cdot \sum_{\mathbf{y} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{y}) \dot{P}_t(\mathbf{y}) \right) =$$

$$= \frac{1}{\left(\sum_{\mathbf{y} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{y}) P_t(\mathbf{y}) \right)^2} \cdot$$

$$\cdot \left(\sum_{\mathbf{x} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{x}) \dot{P}_t(\mathbf{x}) a^{\mathbf{x}} \cdot \sum_{\mathbf{y} \in \mathcal{X}^N} P(\mathbf{O} = \mathbf{o} \,|\, \mathbf{y}) P_t(\mathbf{y})) - \right.$$

$$- \sum_{\mathbf{y}\in\mathcal{X}^N} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{y})P_t(\mathbf{y})a^{\mathbf{y}} \cdot \sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{x})\dot{P}_t(\mathbf{x})\Bigg) =$$

$$= \underbrace{\frac{1}{\left(\sum_{\mathbf{y}\in\mathcal{X}^N} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{y})P_t(\mathbf{y})\right)^2}}_{>0} \cdot$$

$$\cdot \sum_{(\mathbf{x},\mathbf{y})\in(\mathcal{X}^N)^2} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{x})P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{y})\dot{P}_t(\mathbf{x})P_t(\mathbf{y})(a^{\mathbf{x}}-a^{\mathbf{y}}).$$

It follows that, if

$$\sum_{(\mathbf{x},\mathbf{y})\in(\mathcal{X}^N)^2} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{x})P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{y})\dot{P}_t(\mathbf{x})P_t(\mathbf{y})(a^{\mathbf{x}}-a^{\mathbf{y}}) \geq 0,$$

then

$$\frac{\delta}{\delta t}P_t(\mathrm{X}=x\,|\,\mathbf{O}=\mathbf{o}) > 0.$$

The former expression can be written as

$$\sum_{(\mathbf{x},\mathbf{y})\in(\mathcal{X}^N)^2} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{x})P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{y})\dot{P}_t(\mathbf{x})P_t(\mathbf{y})(a^{\mathbf{x}}-a^{\mathbf{y}}) =$$

$$= \sum_{(\mathbf{x},\mathbf{y})\in(\mathcal{X}^N)^2,\, a^{\mathbf{x}}>a^{\mathbf{y}}} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{x})P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{y})\dot{P}_t(\mathbf{x})P_t(\mathbf{y})(a^{\mathbf{x}}-a^{\mathbf{y}})+$$

$$+ \sum_{(\mathbf{x},\mathbf{y})\in(\mathcal{X}^N)^2,\, a^{\mathbf{x}}=a^{\mathbf{y}}} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{x})P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{y})\dot{P}_t(\mathbf{x})P_t(\mathbf{y})\underbrace{(a^{\mathbf{x}}-a^{\mathbf{y}})}_{=0}+$$

$$+ \sum_{(\mathbf{x},\mathbf{y})\in(\mathcal{X}^N)^2,\, a^{\mathbf{x}}<a^{\mathbf{y}}} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{x})P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{y})\dot{P}_t(\mathbf{x})P_t(\mathbf{y})(a^{\mathbf{x}}-a^{\mathbf{y}}) =$$

$$= \sum_{(\mathbf{x},\mathbf{y})\in(\mathcal{X}^N)^2,\, a^{\mathbf{x}}>a^{\mathbf{y}}} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{x})P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{y})\dot{P}_t(\mathbf{x})P_t(\mathbf{y})(a^{\mathbf{x}}-a^{\mathbf{y}})+$$

$$+ \sum_{(\mathbf{y},\mathbf{x})\in(\mathcal{X}^N)^2,\, a^{\mathbf{y}}<a^{\mathbf{x}}} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{y})P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{x})\dot{P}_t(\mathbf{y})P_t(\mathbf{x})(a^{\mathbf{y}}-a^{\mathbf{x}}) =$$

$$= \sum_{(\mathbf{x},\mathbf{y})\in(\mathcal{X}^N)^2,\, a^{\mathbf{x}}>a^{\mathbf{y}}} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{x})P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{y})(a^{\mathbf{x}}-a^{\mathbf{y}})(\dot{P}_t(\mathbf{x})P_t(\mathbf{y})-\dot{P}_t(\mathbf{y})P_t(\mathbf{x})) =$$

$$= \sum_{(\mathbf{x},\mathbf{y})\in(\mathcal{X}^N)^2,\, a^{\mathbf{x}}>a^{\mathbf{y}}} \underbrace{P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{x})}_{>0}\underbrace{P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{y})}_{>0}\underbrace{(a^{\mathbf{x}}-a^{\mathbf{y}})}_{>0}\underbrace{P_t(\mathbf{y})^2}_{>0}\frac{\delta}{\delta t}\left(\frac{P_t(\mathbf{x})}{P_t(\mathbf{y})}\right).$$

Therefore, if for each couple $(\mathbf{x},\mathbf{y})\in(\mathcal{X}^N)^2$ such that $a^{\mathbf{x}}>a^{\mathbf{y}}$ we have

$$\frac{\delta}{\delta t}\left(\frac{P_t(\mathbf{x})}{P_t(\mathbf{y})}\right)>0,$$

then:

$$\frac{\delta}{\delta t}P_t(\mathrm{X}=x\,|\,\mathbf{O}=\mathbf{o})>0.$$

Using (4.2) for $a^{\mathbf{x}}>a^{\mathbf{y}}$ we have

$$\frac{P_t(\mathbf{x})}{P_t(\mathbf{y})}=\frac{\prod_{h=a^{\mathbf{y}}+1}^{a^{\mathbf{x}}}(st+h-1)}{\prod_{h=N-a^{\mathbf{x}}+1}^{N-a^{\mathbf{y}}}(s(1-t)+h-1)}. \tag{4.14}$$

It is evident that the numerator of (4.14) is strictly increasing in t and that the denominator is strictly decreasing in t. Therefore, (4.14) is strictly increasing in t,

$$\frac{\delta}{\delta t}\left(\frac{P_t(\mathbf{x})}{P_t(\mathbf{y})}\right)>0$$

and

$$\frac{\delta}{\delta t}P_t(\mathrm{X}=x\,|\,\mathbf{O}=\mathbf{o})>0.$$

This concludes the proof of Theorem 35.

Proof of Theorem 36 and Corollary 37

Let $\mathbf{o}\in\mathcal{X}^N$ be fixed. If $\varepsilon>0$, we know from (4.3) that for every $\mathbf{x}\in\mathcal{X}^N$, $P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{x})>0$. It then follows

$$P_t(\mathbf{X}=\mathbf{o}\,|\,\mathbf{O}=\mathbf{o})=\frac{P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{o})\cdot P_t(\mathbf{X}=\mathbf{o})}{P_t(\mathbf{O}=\mathbf{o})}=$$

$$=\frac{P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{o})\cdot P_t(\mathbf{X}=\mathbf{o})}{\sum_{\mathbf{x}\in\mathcal{X}^N}P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{x})\cdot P_t(\mathbf{X}=\mathbf{x})}= \tag{4.15}$$

$$=\frac{1}{1+\sum_{\mathbf{x}\in\mathcal{X}^N,\,\mathbf{x}\neq\mathbf{o}}\underbrace{\left(\frac{P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{x})}{P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{o})}\right)}_{>0}\underbrace{\left(\frac{P_t(\mathbf{X}=\mathbf{x})}{P_t(\mathbf{X}=\mathbf{o})}\right)}_{\text{Convex}}}. \tag{4.16}$$

According to Lemma 46, the denominator of (4.16) is a positive linear combination of convex and strictly convex functions, and is therefore strictly convex. It follows that either (4.16) has a unique maximum in $]0,1[$ and no local extremes, or it is a strictly monotone function. In order to distinguish these cases, we study the numerator and the denominator of (4.15). According to (4.2), we have

$$P_t(\mathbf{X}=\mathbf{o}) = \frac{\prod_{i=1}^{a^{\mathbf{o}}}(st+i-1)\prod_{j=1}^{N-a^{\mathbf{o}}}(s(1-t)+j-1)}{\prod_{k=1}^{N}(s+k-1)}. \tag{4.17}$$

If $0 < a^{\mathbf{o}} < N$, the numerator of (4.17) is a product containing as factors t, $(1-t)$ and other strictly positive and bounded factors. Therefore, $\lim_{t\to 0} P_t(\mathbf{X}=\mathbf{o}) = \lim_{t\to 1} P_t(\mathbf{X}=\mathbf{o}) = 0$. If $a^{\mathbf{o}} = N$,

$$P_t(\mathbf{X}=\mathbf{o}) = \frac{\prod_{k=1}^{N}(st+k-1)}{\prod_{k=1}^{N}(s+k-1)}, \tag{4.18}$$

$\lim_{t\to 0} P(\mathbf{X}=\mathbf{o}) = 0$ and $\lim_{t\to 1} P_t(\mathbf{X}=\mathbf{o}) = 1$. If $a^{\mathbf{o}} = 0$,

$$P_t(\mathbf{X}=\mathbf{o}) = \frac{\prod_{k=1}^{N}(s(1-t)+k-1)}{\prod_{k=1}^{N}(s+k-1)}. \tag{4.19}$$

Therefore $\lim_{t\to 0} P_t(\mathbf{X}=\mathbf{o}) = 1$ and $\lim_{t\to 1} P_t(\mathbf{X}=\mathbf{o}) = 0$. Denote with $\mathbf{o}_N$ the data set with $a^{\mathbf{o}_N} = N$ and with $\mathbf{o}_0$ the data set with $a^{\mathbf{o}_0} = 0$. When calculating the same type of limits for the denominator of (4.15) we obtain

following results.

$$\begin{aligned}
\lim_{t\to 0} \quad & \sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{x})\cdot P_t(\mathbf{X}=\mathbf{x}) = \\
= \quad & \sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{x})\cdot \underbrace{\lim_{t\to 0} P_t(\mathbf{X}=\mathbf{x})}_{=0,\,\forall \mathbf{x}\neq\mathbf{o}_0} = \\
= \quad & P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{o}_0)\cdot \underbrace{\lim_{t\to 0} P_t(\mathbf{X}=\mathbf{o}_0)}_{=1} = P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{o}_0) > 0, \\
\lim_{t\to 1} \quad & \sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{x})\cdot P_t(\mathbf{X}=\mathbf{x}) = \\
= \quad & \sum_{\mathbf{x}\in\mathcal{X}^N} P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{x})\cdot \underbrace{\lim_{t\to 1} P_t(\mathbf{X}=\mathbf{x})}_{=0,\,\forall \mathbf{x}\neq\mathbf{o}_N} = \\
= \quad & P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{o}_N)\cdot \underbrace{\lim_{t\to 1} P_t(\mathbf{X}=\mathbf{o}_N)}_{=1} = P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{o}_N) > 0.
\end{aligned}$$

We can now prove the three statements in the theorem. If $0 < a^{\mathbf{o}} < N$ it follows immediately from the above limits that

$$\lim_{t\to 0} P_t(\mathbf{X}=\mathbf{o}\,|\,\mathbf{O}=\mathbf{o}) = \frac{0}{P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{o}_0)} = 0,$$

$$\lim_{t\to 1} P_t(\mathbf{X}=\mathbf{o}\,|\,\mathbf{O}=\mathbf{o}) = \frac{0}{P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{o}_N)} = 0.$$

In addition, because the denominator of (4.16) is strictly convex, $P_t(\mathbf{X}=\mathbf{o}\,|\,\mathbf{O}=\mathbf{o})$ has in this case a unique maximum in the interior of $]0,1[$. If $a^{\mathbf{o}} = N$ and therefore $\mathbf{o} = \mathbf{o}_N$, we obtain with the same arguments as above:

$$\lim_{t\to 0} P_t(\mathbf{X}=\mathbf{o}_N\,|\,\mathbf{O}=\mathbf{o}_N) = \frac{0}{P(\mathbf{O}=\mathbf{o}_N\,|\,\mathbf{X}=\mathbf{o}_0)} = 0,$$

$$\lim_{t\to 1} P_t(\mathbf{X}=\mathbf{o}_N\,|\,\mathbf{O}=\mathbf{o}_N) = \frac{P(\mathbf{O}=\mathbf{o}_N\,|\,\mathbf{X}=\mathbf{o}_N)}{P(\mathbf{O}=\mathbf{o}_N\,|\,\mathbf{X}=\mathbf{o}_N)} = 1.$$

Because the denominator of (4.16) is strictly convex and $0 \leq P_t(\mathbf{X}=\mathbf{o}\,|\,\mathbf{O}=\mathbf{o}) \leq 1$, the latter function is strictly monotone increasing in $]0,1[$. If $a^{\mathbf{o}} = 0$ and therefore $\mathbf{o} = \mathbf{o}_0$, we have

$$\lim_{t\to 0} P_t(\mathbf{X}=\mathbf{o}_0\,|\,\mathbf{O}=\mathbf{o}_0) = \frac{P(\mathbf{O}=\mathbf{o}_0\,|\,\mathbf{X}=\mathbf{o}_0)}{P(\mathbf{O}=\mathbf{o}_0\,|\,\mathbf{X}=\mathbf{o}_0)} = 1,$$

$$\lim_{t\to 1} P_t(\mathbf{X}=\mathbf{o}_0 \,|\, \mathbf{O}=\mathbf{o}_0) = \frac{0}{P(\mathbf{O}=\mathbf{o}_0 \,|\, \mathbf{X}=\mathbf{o}_N)} = 0.$$

Because the denominator of (4.16) is strictly convex and $0 \leq P_t(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o}) \leq 1$, the latter function is strictly monotone decreasing in $]0,1[$. This concludes the proof of Theorem 36.

To prove Corollary 37 let

$$\max_{t\in]0,1[} P_t(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o}) > \frac{K}{K+1}.$$

According to Theorem 36, if $0 < a^{\mathbf{o}} < N$ there exist two values $t_{\min} < t_{\max}$ such that $t_{\min}, t_{\max} \in]0,1[$ and

$$P_{t_{\min}}(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o}) = P_{t_{\max}}(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o}) = \frac{K}{K+1}.$$

Because the function $P_t(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o})$ has no local maxima, it follows that

$$P_t(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o}) \geq \frac{K}{K+1} \Leftrightarrow t \in [t_{\min}, t_{\max}].$$

If $a^{\mathbf{o}} = N$, $P_t(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o})$ is a monotone increasing function of t, such that

$$\lim_{t\to 0} P_t(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o}) = 0,$$
$$\lim_{t\to 1} P_t(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o}) = 1.$$

Therefore, there exists $t_{\min} < 1$ such that

$$P_t(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o}) \geq \frac{K}{K+1} \Leftrightarrow t \in [t_{\min}, 1[.$$

Finally, if $a^{\mathbf{o}} = 0$, $P_t(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o})$ is a monotone decreasing function of t such that

$$\lim_{t\to 0} P_t(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o}) = 1,$$
$$\lim_{t\to 1} P_t(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o}) = 0.$$

Therefore, there exists $t_{\max} < 1$ such that

$$P_t(\mathbf{X}=\mathbf{o} \,|\, \mathbf{O}=\mathbf{o}) \geq \frac{K}{K+1} \Leftrightarrow t \in]0, t_{\max}].$$

Proof of Theorem 38 and Corollary 39

Consider a data set $\mathbf{x} \in \mathcal{X}^N$ such that $\mathbf{x} \neq \mathbf{o}$. Denote with $\mathbf{n^x}$ the number of outcomes in $\mathbf{x}$ different from outcomes in $\mathbf{o}$, where $\mathbf{n^x} \geq 1$. Then, using (4.3), we have

$$\frac{P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{x})}{P(\mathbf{O}=\mathbf{o}\,|\,\mathbf{X}=\mathbf{o})} = \frac{\varepsilon^{\mathbf{n^x}}(1-\varepsilon)^{N-\mathbf{n^x}}}{(1-\varepsilon)^N} = \left(\frac{\varepsilon}{1-\varepsilon}\right)^{\mathbf{n^x}}. \tag{4.20}$$

(4.20) is strictly monotone increasing in ε. Moreover, we can rewrite (4.16) as

$$P_t(\mathbf{X}=\mathbf{o}\,|\,\mathbf{O}=\mathbf{o}) = \frac{1}{1+\sum_{\mathbf{x}\in\mathcal{X}^N,\,\mathbf{x}\neq\mathbf{o}}\left(\frac{\varepsilon}{1-\varepsilon}\right)^{\mathbf{n^x}}\left(\frac{P_t(\mathbf{X}=\mathbf{x})}{P_t(\mathbf{X}=\mathbf{o})}\right)}. \tag{4.21}$$

Because the term $\frac{P_t(\mathbf{X}=\mathbf{x})}{P_t(\mathbf{X}=\mathbf{o})}$ does not depend on ε, the denominator of (4.21) is strictly monotone increasing in ε and therefore $P_t(\mathbf{X}=\mathbf{o}\,|\,\mathbf{O}=\mathbf{o})$ is strictly monotone decreasing in ε. Now, calculating $\lim_{\varepsilon\to 0}$ in (4.21), we obtain

$$\lim_{\varepsilon\to 0} P_t(\mathbf{X}=\mathbf{o}\,|\,\mathbf{O}=\mathbf{o}) = \lim_{\varepsilon\to 0} \frac{1}{1+\sum_{\mathbf{x}\in\mathcal{X}^N,\,\mathbf{x}\neq\mathbf{o}} \underbrace{\left(\frac{\varepsilon}{1-\varepsilon}\right)^{\mathbf{n^x}}}_{\to 0}\left(\frac{P_t(\mathbf{X}=\mathbf{x})}{P_t(\mathbf{X}=\mathbf{o})}\right)} = 1. \tag{4.22}$$

The interval I_ε is such that

$$P_t(\mathbf{X}=\mathbf{o}\,|\,\mathbf{O}=\mathbf{o}) > \frac{K}{K+1} \Leftrightarrow t \in I_\varepsilon.$$

It follows immediately that for a fixed ε, if the interval I_ε is non-empty, then it is monotone decreasing in K with respect to the partial order $\subset$. Analogously, because $P_t(\mathbf{X}=\mathbf{o}\,|\,\mathbf{O}=\mathbf{o})$ is decreasing in ε, for any fixed K, if the interval I_ε is non-empty, then it is monotone decreasing in ε. Finally, for each $K > 1$ we have $\frac{K}{K+1} < 1$. Therefore, because of (4.22), we have $\lim_{\varepsilon\to 0} I_\varepsilon =]0,1[$. This concludes the proof of Theorem 38

We prove now Corollary 39. The interval I_ε is such that

$$P_t(\mathbf{X}=\mathbf{o}\,|\,\mathbf{O}=\mathbf{o}) > \frac{K}{K+1} \Leftrightarrow t \in I_\varepsilon.$$

Therefore, according to Theorem 36, $[\delta, 1-\delta] \subset I_\varepsilon$ if and only if $P_{t=\delta}(\mathbf{X}=\mathbf{o}\,|\,\mathbf{O}=\mathbf{o}) > \frac{K}{K+1}$ and $P_{t=1-\delta}(\mathbf{X}=\mathbf{o}\,|\,\mathbf{O}=\mathbf{o}) > \frac{K}{K+1}$.

We know from Theorem 38 that I_ε increases monotonically with decreasing ε and that $\lim_{\varepsilon\to 0} I_\varepsilon =]0,1[$ for each $K > 1$. It follows immediately that there exists a unique $\varepsilon_{\max}$ such that:

$$[\delta, 1-\delta] \subseteq I_\varepsilon \Leftrightarrow \varepsilon \leq \varepsilon_{\max}.$$

Therefore

$$[\delta, 1-\delta] \subset I_\varepsilon \Leftrightarrow \varepsilon < \varepsilon_{\max}.$$

Proof of Theorem 40

The formula (4.1) can be rewritten as,

$$P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \sum_{\mathbf{x}\in\mathcal{X}^N} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o}) \cdot \frac{a^{\mathbf{x}} + st}{N+s}. \tag{4.23}$$

Assumption (QP) is equivalent to

$$P_t(\mathbf{X} \neq \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \leq \frac{1}{K+1}, \tag{4.24}$$

for each $t \in I$. We prove the statement of Theorem 40 in the three cases $0 < a^{\mathbf{o}} < N$, $a^{\mathbf{o}} = N$ and $a^{\mathbf{o}} = 0$. We begin with the case $0 < a^{\mathbf{o}} < N$. Without loss of generality, let ε be small enough to obtain for the given K a nonempty interval $I := [t_{\min}, t_{\max}]$. Using (4.23), we find, for each $t \in I$, the following bounds for $P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})$:

$$\begin{aligned}
& P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \sum_{\mathbf{x}\in\mathcal{X}^N} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o}) \frac{a^{\mathbf{x}} + st}{N+s} = \\
= \;& P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \frac{a^{\mathbf{o}} + st}{N+s} + \sum_{\mathbf{x}\in\mathcal{X}^N, \mathbf{x}\neq\mathbf{o}} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o}) \frac{a^{\mathbf{x}} + st}{N+s} < \\
< \;& P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \frac{a^{\mathbf{o}} + st}{N+s} + \sum_{\mathbf{x}\in\mathcal{X}^N, \mathbf{x}\neq\mathbf{o}} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o}) \frac{N + st}{N+s} = \\
= \;& P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \frac{a^{\mathbf{o}} + st}{N+s} + P_t(\mathbf{X} \neq \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \frac{N + st}{N+s}.
\end{aligned} \tag{4.25}$$

The maximum of (4.25) is attained when the weight assigned to $\frac{N+st}{N+s}$ is maximal and the weight assigned to $\frac{a^{\mathbf{o}}+st}{N+s}$ is minimal, because $a^{\mathbf{o}} < N$. According to Assumption 4.24, we obtain

$$P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) < \frac{K}{K+1}\frac{a^{\mathbf{o}}+st}{N+s} + \frac{1}{K+1}\frac{N+st}{N+s}. \tag{4.26}$$

In the same way, we obtain

$$\begin{aligned}
& P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \sum_{\mathbf{x}\in\mathcal{X}^N} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o})\frac{a^{\mathbf{x}}+st}{N+s} > \\
> \;& P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o})\frac{a^{\mathbf{o}}+st}{N+s} + \sum_{\mathbf{x}\in\mathcal{X}^N, \mathbf{x}\neq\mathbf{o}} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o})\frac{st}{N+s} = \\
= \;& P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o})\frac{a^{\mathbf{o}}+st}{N+s} + P_t(\mathbf{X} \neq \mathbf{o} \,|\, \mathbf{O} = \mathbf{o})\frac{st}{N+s} \geq \\
\geq \;& \frac{K}{K+1}\frac{a^{\mathbf{o}}+st}{N+s} + \frac{1}{K+1}\frac{st}{N+s}.
\end{aligned} \tag{4.27}$$

Remember that the value

$$P_t^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) := \frac{a^{\mathbf{o}}+st}{N+s},$$

is the posterior probability in the IBM for the same value of t as above, but when $\mathbf{o}$ is observed perfectly. It is easily verified that

$$P_t^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) < \frac{K}{K+1}\frac{a^{\mathbf{o}}+st}{N+s} + \frac{1}{K+1}\frac{N+st}{N+s},$$

$$P_t^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) > \frac{K}{K+1}\frac{a^{\mathbf{o}}+st}{N+s} + \frac{1}{K+1}\frac{st}{N+s}.$$

Therefore, the interval

$$\left[\frac{K}{K+1}\frac{a^{\mathbf{o}}+st}{N+s} + \frac{1}{K+1}\frac{st}{N+s}, \frac{K}{K+1}\frac{a^{\mathbf{o}}+st}{N+s} + \frac{1}{K+1}\frac{N+st}{N+s}\right],$$

contains both $P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})$ and $P_t^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})$, and it follows

$$\left|P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - P_t^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})\right| < \frac{1}{K+1}\frac{N}{N+s}.$$

Assume $a^{\mathbf{o}} = N$. We now have a non empty interval $I_\varepsilon := [t_{\min}, 1[\subset]0, 1[$ such that (4.24) is satisfied for each $t \in I_\varepsilon$. Therefore, for each $t \in I_\varepsilon$ we have

$$\begin{aligned} & P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \sum_{\mathbf{x} \in \mathcal{X}^N} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o}) \frac{a^{\mathbf{x}} + st}{N + s} = \\ = \; & P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \frac{a^{\mathbf{o}} + st}{N + s} + \sum_{\mathbf{x} \in \mathcal{X}^N, \mathbf{x} \neq \mathbf{o}} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o}) \frac{a^{\mathbf{x}} + st}{N + s} < \\ < \; & P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \frac{N + st}{N + s} + \sum_{\mathbf{x} \in \mathcal{X}^N, \mathbf{x} \neq \mathbf{o}} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o}) \frac{N + st}{N + s} = \\ = \; & P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \frac{N + st}{N + s} + P_t(\mathbf{X} \neq \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \frac{N + st}{N + s} = \frac{N + st}{N + s}, \end{aligned} \tag{4.28}$$

$$\begin{aligned} & P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \sum_{\mathbf{x} \in \mathcal{X}^N} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o}) \frac{a^{\mathbf{x}} + st}{N + s} > \\ > \; & P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \frac{N + st}{N + s} + \sum_{\mathbf{x} \in \mathcal{X}^N, \mathbf{x} \neq \mathbf{o}} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o}) \frac{st}{N + s} \geq \\ \geq \; & \frac{K}{K+1} \frac{N + st}{N + s} + \frac{1}{K+1} \frac{st}{N + s}. \end{aligned} \tag{4.29}$$

In this case

$$P_t^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) := \frac{N + st}{N + s}.$$

Therefore,

$$\begin{aligned} & \left| P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - P_t^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) \right| < \\ < \; & \frac{N + st}{N + s} - \left(\frac{K}{K+1} \frac{N + st}{N + s} + \frac{1}{K+1} \frac{st}{N + s} \right) = \\ = \; & \frac{1}{K+1} \frac{N}{N + s}. \end{aligned}$$

Finally, assume $a^{\mathbf{o}} = 0$. We now have a non empty interval $I_\varepsilon :=]0, t_{\max}]$ such that (4.24) is satisfied for each $t \in I_\varepsilon$. Therefore, for each $t \in I_\varepsilon$ it follows

$$\begin{aligned} & P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \sum_{\mathbf{x} \in \mathcal{X}^N} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o}) \frac{a^{\mathbf{x}} + st}{N + s} < \\ < \; & P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \frac{st}{N + s} + \sum_{\mathbf{x} \in \mathcal{X}^N, \mathbf{x} \neq \mathbf{o}} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o}) \frac{N + st}{N + s} \leq \\ \leq \; & \frac{K}{K+1} \frac{st}{N + s} + \frac{1}{K+1} \frac{N + st}{N + s}, \end{aligned} \tag{4.30}$$

$$\begin{aligned} & P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = \sum_{\mathbf{x} \in \mathcal{X}^N} P_t(\mathbf{X} = \mathbf{x} \,|\, \mathbf{O} = \mathbf{o}) \frac{a^{\mathbf{x}} + st}{N + s} > \\ > \; & P_t(\mathbf{X} = \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \frac{st}{N + s} + P_t(\mathbf{X} \neq \mathbf{o} \,|\, \mathbf{O} = \mathbf{o}) \frac{st}{N + s} = \\ = \; & \frac{st}{N + s}. \end{aligned} \tag{4.31}$$

In this case

$$P_t^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) := \frac{st}{N + s},$$

and therefore

$$\begin{aligned} & \left| P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - P_t^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) \right| < \\ < \; & \frac{K}{K+1} \frac{st}{N + s} + \frac{1}{K+1} \frac{N + st}{N + s} - \frac{st}{N + s} = \\ = \; & \frac{1}{K+1} \frac{N}{N + s}. \end{aligned}$$

Proof of Theorem 41

Given δ, denote by $\widetilde{\delta}$ the value such that

$$\delta = \frac{N + 2s}{N + s} \cdot \widetilde{\delta}.$$

Clearly, $\widetilde{\delta} < \delta$. Given $\mathbf{O} = \mathbf{o}$, denote by $\widetilde{K}$ the minimal $K > 1$ such that

$$\frac{1}{K+1}\frac{N}{N+s} \leq \widetilde{\delta}.$$

Given $\widetilde{K}$, according to Corollary 39, there exists an $\varepsilon_{\max} > 0$ such that, for each $\varepsilon < \varepsilon_{\max}$,

$$[\widetilde{\delta}, 1-\widetilde{\delta}] \subset I_\varepsilon,$$

and, because $\delta > \widetilde{\delta}$,

$$[\delta, 1-\delta] \subset I_\varepsilon.$$

According to Theorem 40, we have

$$|P_t(X = x \,|\, \mathbf{O} = \mathbf{o}) - P_t^{\mathrm{IBM}}(X = x \,|\, \mathbf{O} = \mathbf{o})| < \frac{1}{\widetilde{K}+1}\frac{N}{N+s} \leq \widetilde{\delta} < \delta, \quad (4.32)$$

for each $t \in I_\varepsilon$. We show now that for each $\varepsilon < \varepsilon_{\max}$, (4.8) and (4.9) hold. Define $\underline{t} := \inf_{t \in I_\varepsilon} t$ and $\overline{t} := \sup_{t \in I_\varepsilon} t$. We have

$$\begin{aligned} & \left|\overline{P}^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - P_{\overline{t}}^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})\right| = \\ = & \left|\frac{a+s}{N+s} - \frac{a+s\overline{t}}{N+s}\right| = \frac{s(1-\overline{t})}{N+s} < \frac{s\widetilde{\delta}}{N+s}, \end{aligned} \quad (4.33)$$

and

$$\begin{aligned} & \left|\underline{P}^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - P_{\underline{t}}^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})\right| = \\ = & \left|\frac{a}{N+s} - \frac{a+s\underline{t}}{N+s}\right| = \frac{s\underline{t}}{N+s} < \frac{s\widetilde{\delta}}{N+s}. \end{aligned} \quad (4.34)$$

$P_t(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})$ is continuous and increasing in t. For $\overline{t} < 1$ and $\underline{t} > 0$ it follows:

$$\overline{P}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = P_{\overline{t}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}), \quad (4.35)$$

$$\underline{P}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = P_{\underline{t}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}). \quad (4.36)$$

It follows that, if $\overline{t} < 1$,

$$\begin{aligned}
& \left|\overline{P}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - \overline{P}^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})\right| = \\
\leq\ & \left|\overline{P}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - P^{\mathrm{IBM}}_{\overline{t}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})\right| + \\
+\ & \left|P^{\mathrm{IBM}}_{\overline{t}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - \overline{P}^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})\right| = \\
\overset{(4.35)}{=}\ & \left|P_{\overline{t}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - P^{\mathrm{IBM}}_{\overline{t}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})\right| + \\
+\ & \left|P^{\mathrm{IBM}}_{\overline{t}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - \overline{P}^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})\right| < \\
\overset{(4.32)+(4.33)}{<}\ & \widetilde{\delta} + \frac{s\widetilde{\delta}}{N+s} = \left(\frac{N+2s}{N+s}\right)\widetilde{\delta} = \delta. \qquad (4.37)
\end{aligned}$$

Analogously, if $\underline{t} > 0$,

$$\begin{aligned}
& \left|\underline{P}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - \underline{P}^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})\right| = \\
\leq\ & \left|\underline{P}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - P^{\mathrm{IBM}}_{\underline{t}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})\right| + \\
+\ & \left|P^{\mathrm{IBM}}_{\underline{t}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - \underline{P}^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})\right| = \\
\overset{(4.36)}{=}\ & \left|P_{\underline{t}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - P^{\mathrm{IBM}}_{\underline{t}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})\right| + \\
+\ & \left|P^{\mathrm{IBM}}_{\underline{t}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) - \underline{P}^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o})\right| < \\
\overset{(4.32)+(4.34)}{<}\ & \widetilde{\delta} + \frac{s\widetilde{\delta}}{N+s} = \left(\frac{N+2s}{N+s}\right)\widetilde{\delta} = \delta. \qquad (4.38)
\end{aligned}$$

We distinguish between the three cases: $a^{\mathbf{o}} = 0$, $a^{\mathbf{o}} = N$, and $0 < a^{\mathbf{o}} < N$. According to Corollary 37 we have

- $a^{\mathbf{o}} = 0 \Rightarrow \underline{t} = 0$ and $\overline{t} < 1$,
- $0 < a^{\mathbf{o}} < N \Rightarrow \underline{t} > 0$ and $\overline{t} < 1$,
- $a^{\mathbf{o}} = N \Rightarrow \underline{t} > 0$ and $\overline{t} = 1$.

Furthermore, for $\underline{t} = 0$,

$$\underline{P}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = 0 = \underline{P}^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}). \qquad (4.39)$$

Analogously, for $\overline{t} = 1$,

$$\overline{P}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}) = 1 = \overline{P}^{\mathrm{IBM}}(\mathrm{X} = x \,|\, \mathbf{O} = \mathbf{o}). \qquad (4.40)$$

Combining the above results, we see that (4.8) and (4.9) are satisfied in each case: if $a^{\mathbf{o}} = 0$ because of (4.37) and (4.39), if $a^{\mathbf{o}} = N$ because of (4.38) and (4.40). Finally, if $0 < a^{\mathbf{o}} < N$, (4.8) and (4.9) are satisfied because of (4.37) and (4.38).

Chapter 5

Conclusions and outlook

The question of learning under prior near-ignorance is fundamental in statistics. Our results highlight a serious problem related to prior near-ignorance, which had, to the best of our knowledge, not been previously recognized: learning under prior near-ignorance with imperfect observational processes is generally impossible. This problem concerns several statistical models, because (i) statistics is often concerned with prior ignorance and (ii) imperfect observations and signals belong to most real applications.

Statistical models need a sound theoretical foundation to yield useful conclusions. For models concerned with prior near-ignorance this foundation seems to be missing to some extent, as pointed out by our research. It appears therefore important to develop a theory able to isolate situations of prior near-ignorance where learning from imperfect observations is possible, from others. In our research we have identified a general sufficient condition that, if satisfied, prevents learning under prior near-ignorance to take place. In the particular case of categorical signals with Dirichlet priors we have already derived necessary and sufficient conditions for learning under prior near-ignorance.

In the case of manifest variables not allowing for learning under prior near-ignorance, it will be necessary to identify additional assumptions and alternative models that permit learning to take place. In the present work, we have proposed such an assumption in the simple but important case of the imprecise Beta model. The proposed assumption, called quasi-perfection, is particularly suited for imperfect observation processes characterized by a very

small probability of error. We have shown that the IBM modified with our additional assumption is actually able to learn from imperfect observations with a specification of prior knowledge that can be arbitrary weak, depending on the given probability of error. Furthermore, we have shown that the results produced by the modified model are arbitrary close to those produced by the standard IBM considering the observations as perfect.

Although our results suggest a possible way to obtain models able to learn from imperfect signals, many questions and issues remain to be investigated. Firstly, it would be necessary to extend the results obtained in the binary case to the general multinomial case. Furthermore, it would be important to investigate the theoretical properties of the IBM (or IDM) modified with the assumption of quasi perfection. Actually, the IDM satisfies many important principles that are desirable for learning under prior ignorance, like the symmetry principle, the representation invariance principle, the likelihood principle, and others; it is not clear ex-ante which properties remain valid after the addition of the assumption of quasi perfection. To study this issue, it would be necessary to adapt the definitions of these principles to the case of imperfect observation, and verify the theoretical properties of the modified model.

In general, it would be important to identify additional assumptions and alternative models that permit learning under prior ignorance to take place in settings relevant for practical applications.

Bibliography

Berger J. O. et al. (1996) Bayesian Robustness, in: Berger et al. (Eds.), *Bayesian robustness*. Lecture Notes Vol. 29. Hayward: Institute of Mathematical Statistic.

Bernard J. M. (1996) Bayesian interpretation of frequentist procedures for a Bernoulli process. *Amer. Statist.*, 50: 7–13.

Bernard J. M. (2001) 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.

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

Bernard J. M. (2005) An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39 (2–3), 123–150.

Boorsbom D., Mellenbergh G. J., van Heerden J. (2002) The Theoretical Status of Latent Variables. *Psychological Review*, 110 (2), 203–219.

De Cooman G., Miranda E. (2006) Symmetry of models versus models of symmetry. In *Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr.*. Eds. Hofer W. and Wheeler G.. 82 pages. King's College Publications, London.

De Finetti B. (1970) *Teoria delle probabilità*. Einaudi, Torino.

De Finetti B. (1974-1975) *Theory of probability.* John Wiley & Sons, Chichester. English translations of De Finetti (1970), two volumes.

Geisser S. (1993) *Predictive Inference: An Introduction.* Monographs on Statistics and Applied Probability 55. Chapman and Hall, New York.

Hutter M. (2003) 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.

Hutter M. (2006) On The Foundations Of Universal Sequence Prediction. In *Proc. 3rd Annual Conference on Theory and Applications of Models of Computation (TAMC'06)*, 408–420, Beijing.

Kass R., Wassermann L. (1996) The selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, 91: 1343–1370.

Knight F. H. (1933) *Risk, Uncertainty and Profit.* Houghton Mifflin, Boston. (First published 1921.)

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

Laplace P. S. (1820), *Essai Philosophique sur les probabilités.* English translation: *Philosophical Essays on Probabilities* (1951), New York: Dover.

Levi I. (1980) *The Enterprise of Knowledge.* MIT Press, London.

Piatti A., Zaffalon M., Trojani F. (2005) Limits of learning from imperfect observations under prior ignorance: the case of the imprecise Dirichlet model, in: Cozman, F. G., Nau, B., Seidenfeld, T. (Eds), *ISIPTA '05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications.*, Manno (Switzerland), pp. 276-286.

Piatti A., Zaffalon M., Trojani F., Hutter M. (2006a) Learning under Prior Ignorance. Submitted.

Piatti A., Trojani F., Zaffalon M. (2006b) Learning from Quasi Perfect Observations under Prior Ignorance. Submitted.

Quaeghebeur E., de Cooman G. (2003) 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.

Skrondal A., Rabe-Hasketh S. (2004) *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models.* Chapman and Hall/CRC, Boca Raton.

Yang I., Becker M. P. (1997) Latent Variable Modeling of Diagnostic Accuracy. *Biometrics*, 53: 948–958.

Walley P. (1991) *Statistical Reasoning with Imprecise Probability.* Chapman and Hall, New York.

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

Walley P., Bernard J-M. (1999) Imprecise probabilistic prediction for categorical data. Tech. Rep. CAF-9901, Laboratoire Cognition et Activités Finalisées. Université Paris 8, Saint-Denis, France.

Walley P. (2002) Reconciling frequentist properties with the likelihood principle. *Journal of Statistical Planning and Inference*, 105: 35–65.

Zaffalon M. (2001a) 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.

Zaffalon M. (2001b) 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.

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

Zaffalon M., Fagiuoli E. (2003) Tree-based credal networks for classification. *Reliable Computing*, 9(6): 487–509.

Zaffalon M., Hutter M. (2005) Robust Inference of Trees. *Annals of Mathematics and Artificial Intelligence*, Accepted for publication.