Limits of Learning about a Categorical Latent Variable under Prior Near-Ignorance

Alberto Piatti[a], **Marco Zaffalon**[a], **Fabio Trojani**[b], **Marcus Hutter**[c]

alberto.piatti@idsia.ch, zaffalon@idsia.ch, fabio.trojani@unisg.ch, marcus@hutter1.net

[a]Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA),
Galleria 2, CH-6928 Manno-Lugano, Switzerland.

[b]Institute of Banking and Finance, University of St. Gallen,
Rosenbergstr. 52, CH-9000 St.Gallen, Switzerland.

[c]RSISE, Australian National University and NICTA,
[d] Canberra, ACT, 0200, Australia.

April 2009

Abstract

In this paper, we consider the coherent theory of (epistemic) uncertainty of Walley, in which beliefs are represented through sets of probability distributions, and we focus on the problem of modeling prior ignorance about a categorical random variable. In this setting, it is a known result that a state of prior ignorance is not compatible with learning. To overcome this problem, another state of beliefs, called *near-ignorance*, has been proposed. Near-ignorance resembles ignorance very closely, by satisfying some principles that can arguably be regarded as necessary in a state of ignorance, and 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.

Keywords

Near-ignorance set of priors; Latent variables; Imprecise Dirichlet model.

1 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, is missing substantial beliefs about the underlying data-generating process. Yet, the subject would like to exploit the available sample to draw some statistical conclusion, 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 whether prior ignorance is compatible with learning or not. Walley gives a negative answer for the case of his self-consistent (or *coherent*) theory of statistics based on the modeling of beliefs through sets of probability distributions. He shows, in a very general sense, that *vacuous* prior beliefs, i.e., beliefs that a priori are maximally imprecise, lead to vacuous posterior beliefs, irrespective of the type and amount of observed data [Wal91, Section 7.3.7]. At the same time, he proposes focusing on a slightly different state of beliefs, called *near-ignorance*, that does enable learning to take place [Wal91, Section 4.6.9]. Loosely speaking, near-ignorant beliefs are beliefs that are vacuous for a proper subset of the functions of the random variables under consideration (see Section 3). In this way, a near-ignorance prior still gives one the possibility to express vacuous beliefs for some functions of interest, and at the same time it maintains the possibility to learn from data. The fact that learning is possible under prior near-ignorance is shown, for instance, in the special case of the *imprecise Dirichlet model* (IDM) [Wal96, Ber05]. This is a popular model, based on a near-ignorance set of priors, used in the case of inference from categorical data generated by a multinomial process.

Our aim in this paper is to investigate whether near-ignorance can be really regarded as a possible way out of the problem of starting statistical inference in conditions of very weak beliefs. We carry out this investigation in a setting made of categorical data generated by a multinomial process, like in the IDM, but we consider near-ignorance sets of priors in general, not only that used in the IDM.

The interest in this investigation is motivated by the fact that near-ignorance sets of priors appear to play a crucially important role in the question of modeling prior ignorance about a categorical random variable. The key point is that near-ignorance sets of priors can be made to satisfy two principles: the *symmetry* and the *embedding principles*. The first is well known and is equivalent to Laplace's *indifference principle*; the second states, loosely speaking, that if we are ignorant a priori, our prior beliefs on an event of interest should not depend on the space of possibilities in which the event is embedded (see Section 3 for a discussion about these two principles). Walley [Wal91], and later de Cooman and Miranda [DCM06], have argued extensively on the necessity of both the symmetry and the embedding principles in order to characterize a condition of ignorance about a categorical random variable. This implies, if we agree that the symmetry and the embedding principles are necessary for ignorance, that near-ignorance sets of priors should be regarded as an especially important avenue for a subject who wishes to learn starting in a condition of ignorance.

Our investigation starts by focusing on a setting where the categorical variable X under consideration is *latent*. This means that we cannot observe the realizations of X, so that we can learn about it only by means of another, not necessarily categorical, variable S, related to X through a known conditional probability distribution $P(\mathrm{S}\,|\,\mathrm{X})$. Variable S is assumed to be *manifest*, in the sense that its realizations can be observed (see Section 2). The intuition behind the setup considered, made of X and S, is that in many real cases it is not possible to directly observe the value of a random variable in which we are interested, for instance when this variable represents a patient's health and we are observing the result of a diagnostic test. In these cases, we need to use a manifest variable (the medical test) in order to obtain information about the original latent variable (the patient's health). In this paper, we regard the passage from the latent to the manifest variable as made by a process that we call the *observational process*.[1]

Using the introduced setup, we give a condition in Section 4, related to the likelihood function, 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 set of priors that models near-ignorance (thus including the case of the IDM), and for very general kinds of probabilistic relations 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 fully appreciate this result, it is important to realize that latent variables are ubiquitous in problems of uncertainty. The key point here is that the scope of observational processes greatly extends if we consider that even when we directly obtain the value of a variable of interest, what we actually obtain is the observation of the value rather than the value itself. Doing this distinction makes sense because in practice an observational process is usually imperfect, i.e., there is very often (it could be argued that there is always) a positive probability of confounding the realized value of X with another possible value committing thus an observation error.

Of course, if the probability of an observation error is very small and we consider one of the common Bayesian model proposed to learn under prior ignorance, then there is little difference between the results provided by a latent variable model modeling correctly the observational process, and the results provided by a model where the observations are assumed to be perfect. For this reason, the observational process is often neglected in practice and the distinction between the latent variable and the manifest one is not enforced.

But, on the other hand, if we consider sets of probability distributions to model our prior beliefs, instead of a single probability distribution, and in particular if

[1] Elsewhere, this is also called the *measurement process*.

we consider near-ignorance sets of priors, then there can be an extreme difference between a latent variable model and a model where the observations are considered to be perfect, so that learning may be impossible in the first model and possible in the second. As a consequence, when dealing with sets of probability distributions, neglecting the observational process may be no longer justified even if the probability of observation error is tiny. This is shown in a definite sense in Example 9 of Section 4.3, where we analyze the relevance of our results for the special case of the IDM. From the proofs in this paper, it follows that this kind of behavior is mainly determined by the presence, in the near-ignorance set of priors, of extreme, almost-deterministic, distributions. And the question is that these problematic distributions, which are usually not considered when dealing with Bayesian models with a single prior, cannot be ruled out without dropping near-ignorance.

These considerations highlight the quite general applicability of the present results and raise hence 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 might also involve dropping the idea that both the symmetry and the embedding principles can be realistically met in practice.

2 Categorical Latent Variables

In this paper, we follow the general definition of *latent* and *manifest variables* given by Skrondal and Rabe-Hasketh [SRH04]: 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-Hasketh 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. [BMvH02], there may be different interpretations of latent variables. A latent variable can be regarded, for example, as an unobservable random variable that exists independently 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 independently 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

$\boldsymbol{\vartheta} \in \Theta := \{\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_k) \mid \sum_{i=1}^{k} \vartheta_i = 1,\ 0 \leq \vartheta_i \leq 1\}$, without stressing any particular interpretation. Throughout the paper, we denote by $\boldsymbol{\vartheta}$ a particular vector of chances in Θ and by $\boldsymbol{\theta}$ a random variable on Θ.

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

Example 1 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[2] 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 $P(\mathrm{S} \mid \mathrm{X})$. $\diamondsuit$

We continue discussion about this example later on, in the light of our results, in Example 4 of Section 4.

3 Near-ignorance sets of priors

Consider a categorical random variable X with outcomes in $\mathcal{X} = \{x_1, \ldots, x_k\}$ and unknown chances $\boldsymbol{\vartheta} \in \Theta$. Suppose that we have no relevant prior information about $\boldsymbol{\vartheta}$ 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 theory 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 *non-informative prior*,

[2]For further details about the modeling of diagnostic accuracy with latent variables see Yang and Becker [YB97].

has been an important research topic in the last two centuries[3] and, despite the numerous contributions, it remains an open research issue, as illustrated by Kass and Wassermann [KW96]. See also Hutter [Hut06] for recent developments and complementary considerations. There are many principles and properties that are desirable when the focus is on modeling a situation of prior ignorance, and that have indeed been used in past research to define non-informative 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 non-informative 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 non-informative prior satisfying, at the same time, all the desirable principles.

We follow Walley [Wal96] and de Cooman and Miranda [DCM06] 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 $\boldsymbol{\vartheta}$, then we have no reason to favour one possible outcome of X over another, and therefore our probability model on X should be symmetric. This principle is equivalent to Laplace's *symmetry or indifference* principle. 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[4] gives what we believe to be a compelling answer to this question: he proves 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 then, if we agree that the symmetry and the embedding principles are characteristics of prior ignorance, that we need *imprecise probabilities* to model such a state of beliefs.[5] 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 suggests, as an alternative, the use of *near-ignorance sets of priors*.[6]

[3]Starting from the work of Laplace at the beginning of the 19^{th} century [Lap51].

[4]In Walley [Wal91], Note 7 at p. 526. See also Section 5.5 of the same book.

[5]For a complementary point of view, see Hutter [Hut06].

[6]Walley calls a set of probability distributions modeling near-ignorance a *near-ignorance prior*. In this paper we use the term *near-ignorance set of priors* in order to avoid confusion with the

A near-ignorance set of priors is a probability model on the chances $\boldsymbol{\theta}$ of X, modeling a very weak state of knowledge about $\boldsymbol{\theta}$. In practice, a near-ignorance set of priors is a large closed convex set $\mathcal{M}_0$ of prior probability densities on $\boldsymbol{\theta}$ which produces *vacuous expectations* for various but not all functions f on Θ, i.e., such that $\underline{\mathbf{E}}(f) = \inf_{\boldsymbol{\vartheta} \in \Theta} f(\boldsymbol{\vartheta})$ and $\overline{\mathbf{E}}(f) = \sup_{\boldsymbol{\vartheta} \in \Theta} f(\boldsymbol{\vartheta})$.

The key point here is that near-ignorance sets of priors can be designed so as to satisfy both the symmetry and the embedding principles. In fact, if a near-ignorance set of priors produces vacuous expectations for all the functions $f(\boldsymbol{\vartheta}) = \vartheta_i$ for each $i \in \{1, \ldots, k\}$, then, because a priori $P(\mathrm{X} = x_i) = E(\theta_i)$, the near-ignorance set of priors 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 Bayes' rule. Since the beliefs on $\boldsymbol{\theta}$ are not vacuous, this makes it possible to calculate non-vacuous posterior probabilities for X.

A good example of near-ignorance set of priors is the set $\mathcal{M}_0$ used in the *imprecise Dirichlet model*. The IDM models a situation of prior near-ignorance about a categorical random variable X. The near-ignorance set of priors $\mathcal{M}_0$ used in the IDM consists of the set of all Dirichlet densities[8] $p(\boldsymbol{\vartheta}) = dir_{s,\mathbf{t}}(\boldsymbol{\vartheta})$ for a fixed $s > 0$ and all $\mathbf{t} \in \mathcal{T}$, where

$$dir_{s,\mathbf{t}}(\boldsymbol{\vartheta}) := \frac{\Gamma(s)}{\prod_{i=1}^{k} \Gamma(st_i)} \prod_{i=1}^{k} \vartheta_i^{st_i - 1}, \tag{1}$$

and

$$\mathcal{T} := \{\mathbf{t} = (t_1, \ldots, t_k) \mid \sum_{j=1}^{k} t_k = 1,\ 0 < t_j < 1\}. \tag{2}$$

The particular choice of $\mathcal{M}_0$ in the IDM implies vacuous prior expectations for all the functions $f(\boldsymbol{\vartheta}) = \vartheta_i^R$, for all integers $R \geq 1$ and all $i \in \{1, \ldots, k\}$, i.e., $\underline{\mathbf{E}}(\theta_i^R) = 0$ and $\overline{\mathbf{E}}(\theta_i^R) = 1$. Choosing $R = 1$, 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 set of priors $\mathcal{M}_0$ used in the IDM implies a priori the vacuous probability model on X and, therefore, satisfies both the symmetry and embedding principles. On the other hand, the particular set of priors used in the IDM does not imply vacuous prior expectations for all the functions $f(\boldsymbol{\vartheta})$. For example, vacuous expectations for the functions $f(\boldsymbol{\vartheta}) = \vartheta_i \cdot \vartheta_j$ for $i \neq j$ would be $\underline{\mathbf{E}}(\vartheta_i \cdot \vartheta_j) = 0$ and $\overline{\mathbf{E}}(\vartheta_i \cdot \vartheta_j) = 0.25$, but in the IDM we have

precise Bayesian case.

[7] We call this state near-ignorance because, although we are completely ignorant a priori about X, we are not completely ignorant about $\boldsymbol{\theta}$ [Wal91, Section 5.3, Note 4].

[8] Throughout the paper, if no confusion is possible, we denote the outcome $\boldsymbol{\theta} = \boldsymbol{\vartheta}$ by $\boldsymbol{\vartheta}$. For example, we denote $p(\boldsymbol{\theta} = \boldsymbol{\vartheta})$ by $p(\boldsymbol{\vartheta})$.

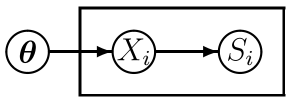
a priori $\overline{\mathbf{E}}(\vartheta_i \cdot \vartheta_j) < 0.25$ and the prior expectations are therefore not vacuous. In Walley [Wal96], it is shown that the IDM produces, for each observed dataset, non-vacuous posterior probabilities for X.

4 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 $\boldsymbol{\theta}$, and a sequence of independent manifest variables $(\mathrm{S}_i)_{i\in\mathbf{N}}$, which we allow to be defined either on finite or infinite spaces. 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 $\boldsymbol{\theta}$ of X_i conditional on X_i, i.e.,

$$P(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j, \boldsymbol{\theta} = \boldsymbol{\vartheta}) = P(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j), \tag{3}$$

for each $x_j \in \mathcal{X}$ and $\boldsymbol{\vartheta} \in \Theta$.[9] These assumptions model a two-step process where the variable S_i is used to convey information about the realized value of X_i for each i, independently of the chances of X_i. The (in)dependence structure can be depicted graphically as follows,



where the framed part of this structure is what we call an *observational process*.

To make things simpler, we assume the probability distribution $P(\mathrm{S}_i \,|\, \mathrm{X}_i = x_j)$ 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 4.1 we discuss our general parametric problem and we obtain a condition that, if satisfied, prevents learning to take place. In Section 4.2 we study the consequences of our theoretical results in the particular case of predictive probabilities. Finally, in Section 4.3, we focus on the particular near-ignorance set of priors used in the IDM and we obtain necessary and sufficient conditions for learning with categorical manifest variables.

4.1 General parametric inference

We focus on a very general problem of parametric inference. Suppose that we observe a dataset $\mathbf{s}$ of realizations of the 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

[9] We denote usually by P a probability (discrete case) and with p a probability density (continuous case). If an expression holds in both the discrete and the continuous case, like for example Equation (3), then we use P to indicate both cases.

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 $f(\boldsymbol{\vartheta})$, our aim is to calculate $\underline{\mathbf{E}}(f \mid \mathbf{s})$ and $\overline{\mathbf{E}}(f \mid \mathbf{s})$ starting from a condition of ignorance about f, i.e., using a near ignorance prior $\mathcal{M}_0$, such that $\underline{\mathbf{E}}(f) = f_{\min} := \inf_{\boldsymbol{\vartheta}\in\Theta} f(\boldsymbol{\vartheta})$ and $\overline{\mathbf{E}}(f) = f_{\max} := \sup_{\boldsymbol{\vartheta}\in\Theta} f(\boldsymbol{\vartheta})$.

Is it really possible to learn something about the function f, starting from a condition of prior near-ignorance and having observed a dataset $\mathbf{s}$? The following theorem shows that, very often, this is not the case. In particular, Corollary 3 shows that there is a condition that, if satisfied, prevents learning to take place.

Theorem 2 *Let* $\mathbf{s}$ *be given. Consider a bounded continuous function* f *defined on* Θ *and a near-ignorance set of priors* $\mathcal{M}_0$*. Then the following statements hold.*[10]

1. *If the likelihood function* $\mathrm{P}(\mathbf{s} \mid \boldsymbol{\vartheta})$ *is strictly positive*[11] *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*
$$\overline{\mathbf{E}}(f \mid \mathbf{s}) = \overline{\mathbf{E}}(f) = f_{\max}.$$

2. *If the likelihood function* $\mathrm{P}(\mathbf{s} \mid \boldsymbol{\vartheta})$ *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*
$$\underline{\mathbf{E}}(f \mid \mathbf{s}) = \underline{\mathbf{E}}(f) = f_{\min}.$$

Corollary 3 *Consider a near-ignorance set of priors* $\mathcal{M}_0$*. Let* $\mathbf{s}$ *be given and let* $\mathrm{P}(\mathbf{s} \mid \boldsymbol{\vartheta})$ *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*
$$\begin{aligned}\underline{\mathbf{E}}(f \mid \mathbf{s}) &= \underline{\mathbf{E}}(f) = f_{\min},\\ \overline{\mathbf{E}}(f \mid \mathbf{s}) &= \overline{\mathbf{E}}(f) = f_{\max}.\end{aligned}$$

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

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

[11] In the appendix it is shown that the assumptions of positivity of $\mathrm{P}(\mathbf{s} \mid \boldsymbol{\vartheta})$ in Theorem 2 can be substituted by the following weaker assumptions. For a given arbitrary small $\delta > 0$, denote by Θ_δ the measurable set, $\Theta_\delta := \{\boldsymbol{\vartheta} \in \Theta \mid f(\boldsymbol{\vartheta}) \geq f_{\max} - \delta\}$. If $\mathrm{P}(\mathbf{s} \mid \boldsymbol{\vartheta})$ is such that, $\lim_{\delta\to 0} \inf_{\boldsymbol{\vartheta}\in\Theta_\delta} \mathrm{P}(\mathbf{s} \mid \boldsymbol{\vartheta}) = c > 0$, then Statement 1 of Theorem 2 holds. The same holds for the second statement, substituting Θ_δ with $\widetilde{\Theta}_\delta := \{\boldsymbol{\vartheta} \in \Theta \mid f(\boldsymbol{\vartheta}) \leq f_{\min} + \delta\}$.

of prior ignorance because only vacuous posterior expectations are produced. The sufficient condition described above is met very easily in practice, as shown in the following two examples. In the first example, we consider a very simple setting where the manifest variables are categorical. In the second example, we consider a simple setting with continuous manifest variables. We show that, in both cases, the sufficient condition is satisfied and therefore we are unable to learn under prior near-ignorance.

Example 4 Consider the medical test introduced in Example 1 and an (ideally) infinite population of individuals. Denote by the binary variable $\mathrm{X}_i \in \{H, I\}$ the health status of the i-th individual of the population and 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 by $1-\varepsilon_1$ 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 by $n^{\mathbf{s}}$ the number of positive tests in the observed sample $\mathbf{s}$. Since 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 3, 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 set of priors. This simple example shows that our previous theoretical results raise serious questions about the use of near-ignorance sets of priors also in very simple, common, and important situations. $\diamondsuit$

Example 4 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 5 Consider a sequence of IID categorical variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ with outcomes in $\mathcal{X}^n$ and unknown chances $\boldsymbol{\theta} \in \Theta$. 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

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

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

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

Therefore, according to Corollary 3, if we model our prior knowledge using a near-ignorance set of priors $\mathcal{M}_0$, 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 set of priors. $\diamondsuit$

Examples 4 and 5 raise, in general, serious criticisms about the use of near-ignorance sets of priors in real applications.

4.2 An important special case: predictive probabilities

We focus now on a very important special case: that of predictive inference.[12] 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 $\mathbf{x}'$.

[12]For a general presentation of predictive inference see Geisser [Gei93]; for a discussion of the imprecise probability approach to predictive inference see Walley and Bernard [WB99].

Given $\mathbf{x}' \in \mathcal{X}^{n'}$, our aim is to calculate $\underline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s})$ and $\overline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s})$. Modeling our prior ignorance about the parameters $\boldsymbol{\theta}$ with a near-ignorance set of priors $\mathcal{M}_0$ and denoting by $\mathbf{n}' := (n'_1, \ldots, n'_k)$ the frequencies of the dataset $\mathbf{x}'$, we have

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

where, according to Bayes' rule,

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

provided that $\int_\Theta \mathrm{P}(\mathbf{s} \,|\, \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \neq 0$. Analogously, substituting sup to inf in (4), we obtain

$$\overline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s}) = \overline{\mathbf{E}}\left(\prod_{i=1}^k \vartheta_i^{n'_i} \,|\, \mathbf{s}\right). \tag{4}$$

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 $f(\boldsymbol{\vartheta}) = \prod_{i=1}^k \vartheta_i^{n'_i}$.

It is easy to show that, in this case, the minimum of f is 0 and is reached in all the points $\boldsymbol{\vartheta} \in \Theta$ 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 the maximally imprecise probabilities regarding the dataset $\mathbf{x}'$, given that $\mathbf{x}'$ has been generated by a multinomial process, 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 4.1 hold also in the particular case of predictive probabilities. In particular, Corollary 3 can be rewritten as follows.

Corollary 6 *Consider a near-ignorance set of priors $\mathcal{M}_0$. Let $\mathbf{s}$ be given and let $\mathrm{P}(\mathbf{s} \,|\, \boldsymbol{\vartheta})$ be a continuous strictly positive function on Θ. Then, if $\mathcal{M}_0$ implies prior probabilities for a dataset $\mathbf{x}' \in \mathcal{X}^{n'}$ that are maximally imprecise, the predictive probabilities of $\mathbf{x}'$ are maximally imprecise also a posteriori, after having observed $\mathbf{s}$, i.e.,*

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

4.3 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 by $\mathcal{S}^i := \{s_1, \ldots, s_{n_i}\}$ 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 $x_j \in \mathcal{X}$.

A possible near-ignorance set of priors for this problem is the set $\mathcal{M}_0$ used in the IDM. We have seen, in Section 3, that this particular near-ignorance set of priors 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 $x_j \in \mathcal{X}$. For this particular choice, the following theorem[13] states necessary and sufficient conditions for learning.

Theorem 7 *Let $\Lambda(\mathrm{S}_i)$ be the emission matrix of S_i for $i = 1, \ldots, n$. Let $\mathcal{M}_0$ be the near-ignorance set of priors 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 $\Lambda(\mathrm{S}_i)$ 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 $\{1, \ldots, k\}$.*

[13]Theorem 7 is a slightly extended version of Theorem 1 in Piatti et al. [PZT05].

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.

Corollary 8 *Under the assumptions of Theorem 7, the following criteria hold:*

1. *If the j-th columns of matrices $\Lambda(\mathrm{S}_i)$ 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 $\Lambda(\mathrm{S}_i)$ have more than one nonzero element, then, for each* $\mathbf{s}$, $\underline{P}(\mathrm{X}_{n+1} = x_j \,|\, \mathbf{s}) = 0$.

Example 9 Consider again the medical test of Example 4. 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 set of priors $\mathcal{M}_0$ of the IDM, according to Theorem 7, we are unable to move away from the initial state of ignorance. This result confirms, in the case of the near-ignorance set of priors of the IDM, the general result of Example 4.

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. This point is particulary interesting when compared to what would be obtained using a model with a single non-informative prior. In this case, the difference between a model with perfect observations and a model that takes into account the probability or error would be very small and therefore the former model would be used instead of the latter. Our results show that this procedure, that is almost an automatism when using models with a single prior, may not be justified in models with sets of priors. The point here seems to be that the amount of imperfection of the observational process should not be evaluated in absolute terms; it should rather be evaluated in comparison with the weakness of the prior beliefs. $\diamondsuit$

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 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 4.1.

Example 10 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 7, if $\mathbf{x}$ contains only observations of the type x_j, then

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

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

$$\underline{\mathrm{P}}(\mathrm{X}_{n+1} = x_j \,|\, \mathbf{x}) > 0 \,, \overline{\mathrm{P}}(\mathrm{X}_{n+1} = x_j \,|\, \mathbf{x}) < 1 \,,$$

for each $j \in \{1, \ldots, k\}$. 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 4.1 thanks to its particular likelihood function. Having observed $\mathbf{S} = \mathbf{X} = \mathbf{x}$, we have

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

where n_i denotes the number of times that $x_i \in \mathcal{X}$ has been observed in $\mathbf{x}$. We have $\mathrm{P}(\mathbf{X} = \mathbf{x} \,|\, \boldsymbol{\vartheta}) = 0$ for all $\boldsymbol{\vartheta}$ such that $\vartheta_j = 0$ for at least one j such that $n_j > 0$ and $\mathrm{P}(\mathbf{X} = \mathbf{x} \,|\, \boldsymbol{\vartheta}) > 0$ for all the other $\boldsymbol{\vartheta} \in \Theta$, in particular for all $\boldsymbol{\vartheta}$ 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 $f(\boldsymbol{\vartheta}) = \vartheta_j$ for all $j \in \{1, \ldots, k\}$. This function reaches its minimum ($f_{\min} = 0$) if $\vartheta_j = 0$ and its maximum ($f_{\min} = 1$) if $\vartheta_j = 1$. Therefore, the points where the function $f(\boldsymbol{\vartheta}) = \vartheta_j$ 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 2 are not met. $\diamondsuit$

Example 10 shows that we are able to learn, using a near-ignorance set of priors, only if the likelihood function $\mathrm{P}(\mathbf{s} \,|\, \boldsymbol{\vartheta})$ 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 set of priors $\mathcal{M}_0$ that models in a satisfactory way a condition of prior (near-) ignorance.[14]

Yet, since the variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ are assumed to be observable, the successful application of a near-ignorance set of priors in the IDM is not helpful in addressing the doubts raised by our theoretical results about the applicability of near-ignorance set of priors in situations, where the variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ are latent, as shown in Example 9.

[14]See Walley [Wal96] and Bernard [Ber05] for an in-depth discussion on the properties of the IDM.

5 On modeling observable quantities

In this section, we discuss three alternative approaches that, at a first sight, might seem promising to overcome the problem of learning under prior near-ignorance. For the sake of simplicity, we consider the particular problem of calculating predictive probabilities for the next outcome and a very simple setting based on the IDM. The alternative approaches are based on trying to predict the manifest variable rather than the latent one, thus changing perspective with respect to the previous sections. This change of perspective is useful to consider also because on some occasions, e.g., when the imperfection of the observational process is considered to be low, one may deem sufficient to focus on predicting the manifest variable. We show, however, that the proposed approaches eventually do not solve the mentioned learning question, which remains therefore an open problem.

Let us introduce in detail the simple setting we are going to use. Consider a sequence of independent and identically distributed categorical binary latent variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ with unknown chances $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\theta_1, 1-\theta_1)$, and a sequence of IID binary manifest variables $(\mathrm{S}_i)_{i\in\mathbf{N}}$ with the same possible outcomes. Since the manifest variables are also IID, then they can be regarded as the product of an overall multinomial data-generating process (that includes the generation of the latent variables as well as the observational process) with unknown chances $\boldsymbol{\xi} := (\xi_1, \xi_2) = (\xi_1, 1-\xi_1)$. Suppose that the emission matrix Λ is known, constant for each i and strictly diagonally dominant, i.e.,

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

with $\varepsilon_1, \varepsilon_2 \neq 0$, $\varepsilon_1 < 0.5$ and $\varepsilon_2 < 0.5$. This simple matrix models the case in which, for each i, we are observing the outcomes of the random variable X_i but there is a positive probability of confounding the actual outcome of X_i with the other one. The random variable S_i represents our observation, while X_i represents the true value. A typical example for this kind of situation is the medical example discussed in Examples 4 and 9. Suppose that we have observed $\mathbf{S} = \mathbf{s}$ and our aim is to calculate $\underline{\mathrm{P}}(\mathrm{X}_{n+1} = x_1 \,|\, \mathbf{s})$ and $\overline{\mathrm{P}}(\mathrm{X}_{n+1} = x_1 \,|\, \mathbf{s})$.

In the previous sections we have dealt with this problem by modeling our ignorance about the chances of X_{n+1} with a near-ignorance set of priors and then calculating $\underline{\mathrm{P}}(\mathrm{X}_{n+1} = x_1 \,|\, \mathbf{s})$ and $\overline{\mathrm{P}}(\mathrm{X}_{n+1} = x_1 \,|\, \mathbf{s})$. But we already know from Example 4 that in this case we obtain vacuous predictive probabilities, i.e.,

$$\underline{\mathrm{P}}(\mathrm{X}_{n+1} = x_1 \,|\, \mathbf{s}) = 0, \qquad \overline{\mathrm{P}}(\mathrm{X}_{n+1} = x_1 \,|\, \mathbf{s}) = 1.$$

Because this approach does not produce any useful result, one could be tempted to modify it in order to obtain non-vacuous predictive probabilities for the next outcome. We have identified three possible alternative approaches that we discuss below. The basic structure of the three approaches is identical and is based on the idea of focusing on the manifest variables, that are observable, instead of the latent variables. The proposed structure is the following:

- specify a near-ignorance set of priors for the chances $\boldsymbol{\xi}$ of S_{n+1};
- construct predictive probabilities for the manifest variables, i.e.,

$$\underline{\mathrm{P}}(\mathrm{S}_{n+1} = x_1 \,|\, \mathbf{s}), \qquad \overline{\mathrm{P}}(\mathrm{S}_{n+1} = x_1 \,|\, \mathbf{s});$$

- use the predictive probabilities calculated in the previous point to say something about the predictive probabilities

$$\underline{\mathrm{P}}(\mathrm{X}_{n+1} = x_1 \,|\, \mathbf{s}), \qquad \overline{\mathrm{P}}(\mathrm{X}_{n+1} = x_1 \,|\, \mathbf{s}).$$

The three approaches differ in the specification of the near-ignorance set of priors for $\boldsymbol{\xi}$ and on the way $\underline{\mathrm{P}}(\mathrm{S}_{n+1} = x_1 \,|\, \mathbf{s})$ and $\overline{\mathrm{P}}(\mathrm{S}_{n+1} = x_1 \,|\, \mathbf{s})$ are used to reconstruct $\underline{\mathrm{P}}(\mathrm{X}_{n+1} = x_1 \,|\, \mathbf{s})$ and $\overline{\mathrm{P}}(\mathrm{X}_{n+1} = x_1 \,|\, \mathbf{s})$.

The first approach consists in specifying a near-ignorance set of priors for the chances $\boldsymbol{\xi}$ taking into consideration the fact that these chances are related to the chances $\boldsymbol{\theta}$ through the equation

$$\xi_1 = (1 - \varepsilon_2) \cdot \theta_1 + \varepsilon_1 \cdot (1 - \theta_1),$$

and therefore we have $\xi_1 \in [\varepsilon_1, 1 - \varepsilon_2]$. A possible way to specify correctly a near-ignorance set of priors in this case is to consider the near-ignorance set of priors $\mathcal{M}_0$ of the IDM on $\boldsymbol{\theta}$, consisting of standard $beta(s,t)$ distributions, and to substitute

$$\theta_1 = \frac{\xi_1 - \varepsilon_1}{1 - (\varepsilon_1 + \varepsilon_2)}, \qquad d\theta_1 = \frac{d\xi_1}{1 - (\varepsilon_1 + \varepsilon_2)},$$

into all the prior distributions in $\mathcal{M}_0$. We obtain thus a near-ignorance set of priors for $\boldsymbol{\xi}$ consisting of beta distributions scaled on the set $[\varepsilon_1, 1 - \varepsilon_2]$, i.e.,

$$\xi_1 \sim \frac{C}{1 - (\varepsilon_1 + \varepsilon_2)} \left(\frac{\xi_1 - \varepsilon_1}{1 - (\varepsilon_1 + \varepsilon_2)} \right)^{st_1 - 1} \left(\frac{(1 - \xi_1) - \varepsilon_2}{1 - (\varepsilon_1 + \varepsilon_2)} \right)^{st_2 - 1},$$

where $C := \frac{\Gamma(s)}{\Gamma(st_1)\Gamma(st_2)}$. But, scaling the distributions, we incur the same problem we have incurred with the IDM for the latent variable. Suppose that we have observed a dataset $\mathbf{s}$ containing n_1 times the outcome x_1 and $n - n_1$ times the outcome x_2. The likelihood function in this case is given by $L(\xi_1, \xi_2) = \xi_1^{n_1} \cdot (1 - \xi_1)^{(n - n_1)}$. Because $\xi_1 \in [\varepsilon_1, 1 - \varepsilon_2]$ the likelihood functions is always positive and therefore the extreme distributions that are present in the near-ignorance set of priors for $\boldsymbol{\xi}$ produce vacuous expectations for ξ_1, i.e., $\overline{E}(\xi_1 \,|\, \mathbf{s}) = 1 - \varepsilon_2$ and $\underline{E}(\xi_1 \,|\, \mathbf{s}) = \varepsilon_1$. It follows that this approach does not solve our theoretical problem. Moreover, it follows that the inability to learn is present under near-ignorance even when we focus on predicting the manifest variable!

The second, more naive, approach consists in using the near-ignorance set of priors $\mathcal{M}_0$ used in the standard IDM to model ignorance about $\boldsymbol{\xi}$. In this way we are assuming (wrongly) that $\xi_1 \in [0,1]$, ignoring thus the fact that $\xi_1 \in [\varepsilon_1, 1-\varepsilon_2]$ and therefore implicitly ignoring the emission matrix Λ. Applying the standard IDM on $\boldsymbol{\xi}$ we are able to produce non-vacuous probabilities $\underline{\mathrm{P}}(\mathrm{S}_{n+1} = x_1 \,|\, \mathbf{s})$ and $\overline{\mathrm{P}}(\mathrm{S}_{n+1} = x_1 \,|\, \mathbf{s})$. Now, because Λ is known, knowing the value of $P(\mathrm{S}_{n+1} = x_1 \,|\, \mathbf{s})$ it is possible to reconstruct $P(\mathrm{X}_{n+1} = x_1 \,|\, \mathbf{s})$. But this approach, that on one hand ignores Λ and on the other hand takes it into consideration, is clearly wrong. For example, it can be easily shown that it can produce probabilities outside $[0,1]$.

Finally, a third possible approach could be to neglect the existence of the latent level and consider S_{n+1} to be the variable of interest. Applying the standard IDM on the manifest variables we are clearly able to produce non vacuous probabilities $\underline{\mathrm{P}}(\mathrm{S}_{n+1} = x_1 \,|\, \mathbf{s})$ and $\overline{\mathrm{P}}(\mathrm{S}_{n+1} = x_1 \,|\, \mathbf{s})$ that are then simply used instead of the probabilities $\underline{\mathrm{P}}(\mathrm{X}_{n+1} = x_1 \,|\, \mathbf{s})$ and $\overline{\mathrm{P}}(\mathrm{X}_{n+1} = x_1 \,|\, \mathbf{s})$ in the problem of interest. This approach is the one typically followed by those who apply the IDM in practical problems.[15] This approach requires the user to assume perfect observability; an assumption that appears to be incorrect in most (if not all) real statistical problems. And yet this procedure, despite being wrong or hardly justifiable from a theoretical point of view, has produced in several applications of the IDM useful results, at least from an empirical point of view. This paradox between our theoretical results and the current practice is an open problem that deserves to be investigated in further research.

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 more common and simple 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. Given also that prior near-ignorance is arguably a privileged way to model a state of ignorance, our results appear to substantially reduce the hope to be able to adopt a form of prior ignorance to do objective-minded statistical inference.

With respect to future research, two possible research directions seem to be particularly important to investigate.

As reported by Bernard [Ber05], near-ignorance sets of priors, in the specific form of the IDM, have been successfully used in a number of applications. On the other hand, the theoretical results presented in this paper point to the impossibility of learning in real statistical problems when starting from a state of near-ignorance. This paradox between empirical and theoretical results should be investigated in

[15] See Bernard [Ber05] for a list of applications of the IDM.

order to better understand the practical relevance of the theoretical analysis presented here, and more generally to explain the mechanism behind such an apparent contradiction.

The proofs contained in this paper suggest that the impossibility of learning under prior near-ignorance with latent variables is mainly due to the presence, in the set of priors, of extreme distributions arbitrarily close to the deterministic ones. Some preliminary experimental analyses have shown that learning is possible as soon as one restricts the set of priors so as to rule out the extreme distributions. This can be realized by defining a notion of distance between priors and then by allowing a distribution to enter the prior set of probability distributions only if it is at least a certain positive distance away from the deterministic priors. The minimal distance can be chosen arbitrarily small (while remaining positive), and this allows one to model a state of very weak beliefs, close to near-ignorance. Such a weak state of beliefs could keep some of the advantages of near-ignorance (although it would clearly not be a model of ignorance) while permitting learning to take place. The main problem of this approach is the justification, i.e., the interpretation of the (arbitrary) restriction of the near-ignorance set of priors. A way to address this issue might be to identify a set of desirable principles, possibly similar to the symmetry and embedding principles, leading in a natural way to a suitably large set of priors describing a state close to near-ignorance.

Acknowledgements

This work was partially supported by Swiss NSF grants 200021-113820/1, 200020-116674/1, 100012-105745/1, and by the Hasler foundation grant 2233.

A 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 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

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

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

Theorem 11 *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 let $f_{\max} := \sup_{\Theta}(f)$ and $f_{\min} := \inf_{\Theta}(f)$. For this function define the measurable sets*

$$\Theta_\delta = \{\boldsymbol{\vartheta} \in \Theta \mid f(\boldsymbol{\vartheta}) \geq f_{\max} - \delta\}, \tag{5}$$

$$\widetilde{\Theta}_\delta = \{\boldsymbol{\vartheta} \in \Theta \,|\, f(\boldsymbol{\vartheta}) \leq f_{\min} + \delta\}. \tag{6}$$

1. *Assume that $(p_n)_{n\in\mathbf{N}}$ concentrates on a maximum of f for $n \to \infty$, in the sense that*
$$\mathbf{E}_n(f) \to f_{\max}, \tag{7}$$
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*
$$\mathbf{E}_n(f) \to f_{\min}, \tag{8}$$
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 (5) we know that on Θ_δ it holds $f(\boldsymbol{\vartheta}) \geq f_{\max} - \delta$, and therefore on $\bar{\Theta}_\delta$ we have $f(\boldsymbol{\vartheta}) \leq f_{\max} - \delta$, and thus

$$\frac{f_{\max} - f(\boldsymbol{\vartheta})}{\delta} \geq 1. \tag{9}$$

It follows that

$$\begin{aligned}
1 - \mathrm{P}_n(\Theta_\delta) &= \mathrm{P}_n(\bar{\Theta}_\delta) = \int_{\bar{\Theta}_\delta} p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \overset{(9)}{\leq} \int_{\bar{\Theta}_\delta} \frac{f_{\max} - f(\boldsymbol{\vartheta})}{\delta} p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \\
&\leq \int_{\Theta} \frac{f_{\max} - f(\boldsymbol{\vartheta})}{\delta} p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} = \frac{1}{\delta}(f_{\max} - \mathbf{E}_n(f)) \xrightarrow{(7)} 0,
\end{aligned}$$

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 (6) we know that on $\widetilde{\Theta}_\delta$ it holds $f(\boldsymbol{\vartheta}) \leq f_{\min} + \delta$, and therefore on $\widehat{\Theta}_\delta$ we have $f(\boldsymbol{\vartheta}) \geq f_{\min} + \delta$, and thus

$$\frac{f(\boldsymbol{\vartheta}) - f_{\min}}{\delta} \geq 1. \tag{10}$$

It follows that

$$\begin{aligned}
1 - \mathrm{P}_n(\widetilde{\Theta}_\delta) &= \mathrm{P}_n(\widehat{\Theta}_\delta) = \int_{\widehat{\Theta}_\delta} p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \overset{(10)}{\leq} \int_{\widehat{\Theta}_\delta} \frac{f(\boldsymbol{\vartheta}) - f_{\min}}{\delta} p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \\
&\leq \int_{\Theta} \frac{f(\boldsymbol{\vartheta}) - f_{\min}}{\delta} p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} = \frac{1}{\delta}(\mathbf{E}_n(f) - f_{\min}) \xrightarrow{(7)} 0,
\end{aligned}$$

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

Theorem 12 *Let $L(\boldsymbol{\vartheta}) \geq 0$ be a bounded measurable function and suppose that the assumptions of Theorem 11 hold. Then the following two statements hold.*

1. If the function $L(\boldsymbol{\vartheta})$ is such that

$$c := \lim_{\delta \to 0} \inf_{\boldsymbol{\vartheta} \in \Theta_\delta} L(\boldsymbol{\vartheta}) > 0, \tag{11}$$

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

$$\frac{\mathbf{E}_n(Lf)}{\mathbf{E}_n(L)} = \frac{\int_\Theta f(\boldsymbol{\vartheta}) L(\boldsymbol{\vartheta}) p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}}{\int_\Theta L(\boldsymbol{\vartheta}) p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}} \to f_{\max}. \tag{12}$$

2. If the function $L(\boldsymbol{\vartheta})$ is such that

$$c := \lim_{\delta \to 0} \inf_{\boldsymbol{\vartheta} \in \widetilde{\Theta}_\delta} L(\boldsymbol{\vartheta}) > 0, \tag{13}$$

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

$$\frac{\mathbf{E}_n(Lf)}{\mathbf{E}_n(L)} \longrightarrow f_{\min}. \tag{14}$$

Remark 13 *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 (11), resp. (13), 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 $\inf_{\boldsymbol{\vartheta} \in \Theta_\delta} L(\boldsymbol{\vartheta}) \geq \frac{c}{2}$. denote by $L_{\max}$ the supremum of the function $L(\boldsymbol{\vartheta})$ in Θ. From Theorem 11, we know that $\mathrm{P}_n(\Theta_\delta) \geq 1 - \varepsilon$, for n sufficiently large. This implies, for n sufficiently large,

$$\mathbf{E}_n(L) = \int_\Theta L(\boldsymbol{\vartheta}) p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \geq \int_{\Theta_\delta} L(\boldsymbol{\vartheta}) p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \geq \frac{c}{2}(1 - \varepsilon), \tag{15}$$

$$\mathbf{E}_n(Lf) \leq \mathbf{E}_n(Lf_{\max}) = f_{\max} \mathbf{E}_n(L), \tag{16}$$

$$\begin{aligned}
\mathbf{E}_n(L) &= \int_{\bar{\Theta}_\delta} L(\boldsymbol{\vartheta}) p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} + \int_{\Theta_\delta} L(\boldsymbol{\vartheta}) p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \\
&\leq L_{\max} \int_{\bar{\Theta}_\delta} p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} + \int_{\Theta_\delta} \underbrace{\frac{f(\boldsymbol{\vartheta})}{f_{\max} - \delta}}_{\geq 1 \text{ on } \Theta_\delta} L(\boldsymbol{\vartheta}) p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \\
&\leq L_{\max} \cdot \varepsilon + \frac{1}{f_{\max} - \delta} \mathbf{E}_n(Lf).
\end{aligned} \tag{17}$$

Combining (15), (16) and (17), 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 $\inf_{\boldsymbol{\vartheta} \in \widetilde{\Theta}_\delta} L(\boldsymbol{\vartheta}) \geq \frac{c}{2}$. From Theorem 11, 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,

$$\mathbf{E}_n(L) = \int_\Theta L(\boldsymbol{\vartheta}) p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \geq \int_{\widetilde{\Theta}_\delta} L(\boldsymbol{\vartheta}) p_n(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \geq \frac{c}{2}(1 - \varepsilon), \tag{18}$$

$$\mathbf{E}_n(Lf) \geq \mathbf{E}_n(Lf_{\min}) = f_{\min} \mathbf{E}_n(L) \Rightarrow f_{\min} \leq \frac{\mathbf{E}_n(Lf)}{\mathbf{E}_n(L)}. \tag{19}$$

Define the function

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

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

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

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 (18) and (19), 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

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

B Proofs of the main results

Proof of Theorem 2. Define, $f_{\min} := \inf_{\boldsymbol{\vartheta}\in\Theta} f(\boldsymbol{\vartheta})$, $f_{\max} := \sup_{\boldsymbol{\vartheta}\in\Theta} f(\boldsymbol{\vartheta})$, and define the bounded non-negative function $\tilde{f}(\boldsymbol{\vartheta}) := f(\boldsymbol{\vartheta}) - f_{\min} \geq 0$. 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 12, substituting $L(\boldsymbol{\vartheta})$ with $\mathrm{P}(\mathbf{s}\,|\,\boldsymbol{\vartheta})$ in (12), 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 $\underline{\mathbf{E}}(f\,|\,\mathbf{s}) = \underline{\mathbf{E}}(f) = f_{\min}$. $\blacksquare$

Corollary 3 is a direct consequence of Theorem 2.

Proof of Theorem 7 To prove Theorem 7 we need the following lemma.

Lemma 14 *Consider a dataset* $\mathbf{x}$ *with frequencies* $\mathbf{a} = (a_1^{\mathbf{x}}, \ldots, a_k^{\mathbf{x}})$. *Then, the following equality holds,*

$$\prod_{h=1}^{k} \vartheta_h^{a_h^{\mathbf{x}}} \cdot dir_{s,\mathbf{t}}(\boldsymbol{\vartheta}) = \frac{\prod_{h=1}^{k} \cdot \prod_{j=1}^{a_h^{\mathbf{x}}} (st_h + j - 1)}{\prod_{j=1}^{n} (s + j - 1)} \cdot dir_{s^{\mathbf{x}},\mathbf{t}^{\mathbf{x}}}(\boldsymbol{\vartheta}),$$

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 14 is in [PZT05]. Because $\mathrm{P}(\mathbf{x}\,|\,\boldsymbol{\vartheta}) = \prod_{h=1}^{k} \vartheta_h^{a_h^{\mathbf{x}}}$, according to Bayes' rule, we have $p(\boldsymbol{\vartheta}\,|\,\mathbf{x}) = dir_{s^{\mathbf{x}},\mathbf{t}^{\mathbf{x}}}(\boldsymbol{\vartheta})$ and

$$P(\mathbf{x}) = \frac{\prod_{h=1}^{k} \prod_{l=1}^{a_h^{\mathbf{x}}} (st_h + l - 1)}{\prod_{l=1}^{n} (s + l - 1)}. \tag{20}$$

Given a Dirichlet distribution $dir_{s,\mathbf{t}}(\boldsymbol{\vartheta})$, the expected value $\mathbf{E}(\vartheta_j)$ is given by $\mathbf{E}(\vartheta_j) = t_j$ (see [KBJ00]). It follows that

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

We are now ready to prove Theorem 7.

1. The first statement of Theorem 7 is a consequence of Corollary 6. Because S_i is independent of $\boldsymbol{\vartheta}$ given X_i for each $i \in \mathbf{N}$, we have

$$\mathrm{P}(\mathbf{s} \,|\, \mathbf{x}, \boldsymbol{\vartheta}) = \mathrm{P}(\mathbf{s} \,|\, \mathbf{x}), \tag{21}$$

   and therefore, using (21) and Bayes' rule, we obtain the likelihood function,

$$L(\boldsymbol{\vartheta}) = P(\mathbf{s} \,|\, \boldsymbol{\vartheta}) = \sum_{\mathbf{x} \in \mathcal{X}^n} P(\mathbf{s} \,|\, \mathbf{x}) \cdot P(\mathbf{x} \,|\, \boldsymbol{\vartheta}) = \sum_{\mathbf{x} \in \mathcal{X}^n} P(\mathbf{s} \,|\, \mathbf{x}) \cdot \prod_{h=1}^{k} \vartheta_h^{a_h^{\mathbf{x}}}. \tag{22}$$

   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 $\boldsymbol{\vartheta} \in \Theta$, 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,

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

   for each $\boldsymbol{\vartheta} \in \Theta$ and therefore, according to Corollary 6 with $n' = 1$, the predictive probabilities that are vacuous a priori remain vacuous also a posteriori.

2. We have $\mathrm{P}(\mathrm{X}_{n+1} = x_j \,|\, \mathbf{s}) = \mathbf{E}(\vartheta_j \,|\, \mathbf{s})$, and therefore, according to Lemma 14 and Bayes' rule,

$$\begin{aligned}
P(\mathrm{X}_{n+1} = x_j \,|\, \mathbf{s}) &= \frac{\int_\Theta \vartheta_j P(\mathbf{s} \,|\, \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}}{\int_\Theta P(\mathbf{s} \,|\, \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}} = \\
&\overset{(21)}{=} \frac{\sum_{\mathbf{x} \in \mathcal{X}^n} \int_\Theta \vartheta_j P(\mathbf{s} \,|\, \mathbf{x}) P(\mathbf{x} \,|\, \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}}{\sum_{\mathbf{x} \in \mathcal{X}^n} \int_\Theta P(\mathbf{s} \,|\, \mathbf{x}) P(\mathbf{x} \,|\, \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}} = \\
&= \frac{\sum_{\mathbf{x} \in \mathcal{X}^n} P(\mathbf{s} \,|\, \mathbf{x}) P(\mathbf{x}) \int_\Theta \vartheta_j p(\boldsymbol{\vartheta} \,|\, \mathbf{x}) d\boldsymbol{\vartheta}}{\sum_{\mathbf{x} \in \mathcal{X}^n} P(\mathbf{s} \,|\, \mathbf{x}) P(\mathbf{x})} = \\
&= \sum_{\mathbf{x} \in \mathcal{X}^n} \left(\frac{P(\mathbf{s} \,|\, \mathbf{x}) P(\mathbf{x})}{\sum_{\mathbf{x} \in \mathcal{X}^n} P(\mathbf{s} \,|\, \mathbf{x}) P(\mathbf{x})} \right) \cdot \mathbf{E}(\vartheta_j \,|\, \mathbf{x}), \\
&= \sum_{\mathbf{x} \in \mathcal{X}^n} \left(\frac{P(\mathbf{s} \,|\, \mathbf{x}) P(\mathbf{x})}{\sum_{\mathbf{x} \in \mathcal{X}^n} P(\mathbf{s} \,|\, \mathbf{x}) P(\mathbf{x})} \right) \cdot \frac{a_j^{\mathbf{x}} + st_j}{n+s}. \qquad (23)
\end{aligned}$$

   It can be checked that the denominator of (23) is positive and therefore conditioning on events with zero probability is not a problem in this setting. (23)

is a convex sum of fractions and is therefore a continuous function of t on $\mathcal{T}$. Denote by $\overline{\mathbf{x}}^j$ 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 (23) 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 (20), 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 (23),

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

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 (23) that

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

which is a convex combination of terms

$$\frac{a_j^{\mathbf{x}} + st_j}{n+s} \geq \frac{a_j^{\mathbf{x}}}{n+s} \geq \frac{1}{n+s},$$

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 manifest variable 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 (23) and letting $t_j \to 0$, we have, because of (20), 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,

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

Corollary 8 is a direct consequence of Theorem 7.

References

[Ber05] J.-M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(2–3):123–150, 2005.

[BMvH02] D. Boorsbom, G. J. Mellenbergh, and J. van Heerden. The theoretical status of latent variables. *Psychological Review*, 110(2):203–219, 2002.

[DCM06] G. De Cooman and E. Miranda. Symmetry of models versus models of symmetry. In W. Hofer and G. Wheeler, editors, *Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr.* King's College Publications, London, 2006.

[Gei93] S. Geisser. *Predictive Inference: An Introduction.* Monographs on Statistics and Applied Probability. Chapman and Hall, New York, 1993.

[Hut06] M. Hutter. On the foundations of universal sequence prediction. In *Proc. 3rd Annual Conference on Theory and Applications of Models of Computation (TAMC'06)*, volume 3959 of *LNCS*, pages 408–420. Springer, 2006.

[KBJ00] S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions, Volume 1: Models and Applications.* Wiley series in Probability and Statistics. Wiley, New York, 2000.

[KW96] R. Kass and L. Wassermann. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91:1343–1370, 1996.

[Lap51] P. S. Laplace. *Essai Philosophique sur les probabilités (1820). English translation: Philosophical Essays on Probabilities.* New York: Dover, 1951.

[PZT05] A. Piatti, M. Zaffalon, and F. Trojani. Limits of learning from imperfect observations under prior ignorance: the case of the imprecise Dirichlet model. In F. G. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA '05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, pages 276–286, Manno, Switzerland, 2005. SIPTA.

[SRH04] A. Skrondal and S. Rabe-Hasketh. *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models.* Chapman and Hall/CRC, Boca Raton, 2004.

[Wal91] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, New York, 1991.

[Wal96] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *J. R. Statist. Soc. B*, 58(1):3–57, 1996.

[WB99] P. Walley and J-M. Bernard. Imprecise probabilistic prediction for categorical data. Tech. rep. caf-9901, Laboratoire Cognition et Activités Finalisées, Université Paris 8, Saint-Denis, France, 1999.

[YB97] I. Yang and M. P. Becker. Latent variable modeling of diagnostic accuracy. *Biometrics*, 53:948–958, 1997.