Technical Report IDSIA-05-07

Learning about a Categorical Latent Variable under Prior Near-Ignorance

Alberto Piatti, IDSIA, Switzerland, alberto.piatti@idsia.ch
Marco Zaffalon, IDSIA, Switzerland, zaffalon@idsia.ch
Fabio Trojani, University of S. Gallen, Switzerland, fabio.trojani@unisg.ch
Marcus Hutter, ANU&NICTA, Australia ,marcus@hutter1.net

May 2007

Abstract

It is well known that 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 a state similar to full ignorance, that Walley calls *near-ignorance*, that permits learning to take place. In this paper we 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 show, for the case of categorical latent variables (and general *manifest* variables) that there is a sufficient condition that, if satisfied, prevents learning to take place under prior near-ignorance. This condition is shown to be easily satisfied in the most common statistical problems.

Contents

Keywords

Prior near-ignorance, latent and manifest variables, observational processes, vacuous beliefs, imprecise probabilities.

1 Introduction

Epistemic theories of statistics are often concerned 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]. But, at the same time, he proposes focusing on a slighlty different state of beliefs, called *near-ignorance*, that does enable learning to take place [Walley (1991), Section 4.6.9]. Loosely speaking, near-ignorant beliefs are beliefs close but not equal to vacuous (see Section 3). The possibility to learn under prior near-ignorance is shown, for instance, in the special case of the near-ignorance prior defining the *imprecise Dirichlet model* (IDM). This is a popular model used in the case of inference from categorical data generated by a discrete process ([Walley (1996), Bernard (2005)]).

In this paper, we also focus on a categorical random variable X, expressing the outcomes of a multinomial process, but 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 2).

In such a setting, we introduce a condition in Section 4, 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.

In 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.

2 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 modeled 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 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 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 $\theta \in \Theta := \{\theta = (\theta_1, \ldots, \theta_k) \mid \sum_{i=1}^{k} \theta_i = 1,\ 0 \leq \theta_i \leq 1\}$, without stressing any particular interpretation.

Suppose now that our aim is to predict, after N realizations of the variable X, the next outcome (or the next N' outcomes). Because the variable X is latent and therefore unobservable by definition, the only possible way to learn something about the probabilities of the next outcome 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 1 We 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[1] 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. $\diamondsuit$

We continue discussion about this example later on, in the light of our results, in Example 2 of Section 4.

3 Near-Ignorance Priors

Consider a categorical random variable X with outcomes in $\mathcal{X} = \{x_1, \ldots, x_k\}$ and unknown chances $\theta \in \Theta$. Suppose that we have no relevant prior information about θ and we are therefore in a situation of prior ignorance. 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 modeled 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[2] 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]For further details about the modeling of diagnostic accuracy with latent variables see [Yang and Becker (1997)].

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

In the case of finite possibility spaces, we agree with [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 principle* and the *embedding principle*. The *symmetry principle* states that, if we are completely 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 θ 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[3] that the only probability model consistent with coherence and 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 $\overline{\mathrm{P}}(A) = 1$. It is evident that this model cannot be expressed using a single probability distribution. It follows that, to model properly and in a coherent way a situation of prior ignorance, we need *imprecise probabilities*.[4]

Unfortunately, adopting the vacuous probability model for X is not a practical solution to our initial problem, because it produces only vacuous posterior probabilities. [Walley (1991)] suggests, as practical solution, the use of *near-ignorance priors*. A near-ignorance prior is a large closed convex set $\mathcal{M}_0$ of probability distributions for θ, very close to the vacuous probability model, which produces a priori *vacuous expectations* for various functions f on Θ, i.e., such that $\underline{\mathbf{E}}(f) = \inf_{\theta\in\Theta} f(\theta)$ and $\overline{\mathbf{E}}(f) = \sup_{\theta\in\Theta} f(\theta)$.

An example of near-ignorance prior that is particularly instructive is the set of priors $\mathcal{M}_0$ 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 in 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

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

and

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

[3]In Note 7, p. 526. See also Section 5.5.

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

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,

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

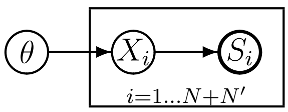
It follows that the particular near-ignorance prior $\mathcal{M}_0$ used in the IDM implies vacuous prior probabilities for each possible outcome of the variable X. It can be shown that this particular set of priors satisfies both the symmetry and embedding principles.

But what is the difference between the vacuous probability model and the the near-ignorance prior used in the IDM? In fact, although both models produce vacuous prior probabilities and both models satisfy the symmetry and embedding principles, the IDM yields posterior probabilities that are not vacuous, while the vacuous probability model produces only vacuous posterior probabilities. The answer to this question is the reason why we use the term *near-ignorance*: in the IDM, although we are completely ignorant about the possible outcomes of the variable X, we are not completely ignorant about the chances θ, because we assume a particular class of prior distributions, i.e., the Dirichlet distributions for a fixed value of s.

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 $\theta \in \Theta$, 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 an (unobservable) realization of the latent variable X_i and that the probability distribution of S_i given X_i is known for each $i \in \mathbf{N}$. Furthermore, we assume S_i to be independent of the chances θ of X_i given X_i. Define the random variables $\mathbf{X} := (\mathrm{X}_1, \ldots, \mathrm{X}_N)$, $\mathbf{S} := (\mathrm{S}_1, \ldots, \mathrm{S}_N)$ and $\mathbf{X}' := (\mathrm{X}_{N+1}, \ldots, \mathrm{X}_{N+N'})$.

We focus on the problem of predictive inference.[5] 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 $\mathrm{X}_1, \ldots, \mathrm{X}_N$. Using the notation defined above we have $\mathbf{S} = \mathbf{s}$ and $\mathbf{X} = \mathbf{x}$. Our aim is to predict the outcomes of the next N' variables $\mathrm{X}_{N+1}, \ldots, \mathrm{X}_{N+N'}$. In particular, given $\mathbf{x}' \in \mathcal{X}^{N'}$, our aim is to calculate $\underline{\mathrm{P}}(\mathbf{X}' = \mathbf{x}' \,|\, \mathbf{S} = \mathbf{s})$ and $\overline{\mathrm{P}}(\mathbf{X}' = \mathbf{x}' \,|\, \mathbf{S} = \mathbf{s})$. To simplify notation, when no confusion is possible, we denote in the rest of the paper $\mathbf{S} = \mathbf{s}$ with $\mathbf{s}$ and $\mathbf{X}' = \mathbf{x}'$ with $\mathbf{x}'$. The (in)dependence structure can be depicted graphically as follows:



[5]For a general presentation of predictive inference see [Geisser (1993)]; for a discussion of the imprecise probability approach to predictive inference see [Walley et al. (1999)].

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

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

where, according to Bayes theorem,

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

provided that $\int_\Theta \mathrm{P}(\mathbf{s} \,|\, \theta) p(\theta) d\theta \neq 0$. Analogously, substituting sup to inf in (3), we obtain

$$\overline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s}) = \overline{\mathbf{E}} \left(\prod_{i=1}^k \theta_i^{n'_i} \,|\, \mathbf{s} \right). \tag{3}$$

The central problem now is to choose $\mathcal{M}_0$ so as to be as ignorant as possible a priori and, at the same time, to be able to learn something from the observed dataset of manifest variables $\mathbf{s}$. Theorem 1 and the following corollaries yield a first partial solution to the above problem, stating several conditions for learning under prior near-ignorance.

Theorem 1 *Let* $\mathbf{s}$ *be given. Consider a bounded continuous function* f *defined on* Θ *and denote with* $f_{\max}$ *the Supremum of* f *on* Θ*. If the likelihood function* $\mathrm{P}(\mathbf{s} \,|\, \theta)$ *is strictly positive*[6] *in each point in which* f *reaches its maximum value* $f_{\max}$ *and it is continuous in an arbitrary small neighborhood of these points, and* $\mathcal{M}_0$ *is such that a priori* $\overline{\mathbf{E}}(f) = f_{\max}$*, then*

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

Many corollaries to Theorem 1 are listed in Section B of the Appendix. Here we discuss only the most important corollary. Consider, given a dataset $\mathbf{x}'$, the particular function $f(\theta) = \prod_{i=1}^k \theta_i^{n'_i}$. This function is particularly important for

[6]The Assumption about $\mathrm{P}(\mathbf{s} \,|\, \theta)$ in Theorem 1 can be substituted by the following weaker assumption. For a given arbitrary small $\delta > 0$, denote with Θ_δ the measurable set, $\Theta_\delta := \{\theta \in \Theta \,|\, f(\theta) \geq f_{\max} - \delta\}$. If $\mathrm{P}(\mathbf{s} \,|\, \theta)$ is such that, $\lim_{\delta \to 0} \inf_{\theta \in \Theta_\delta} \mathrm{P}(\mathbf{s} \,|\, \theta) = c > 0$, then Theorem 1 holds.

predictive inference, because its lower and upper expectations correspond to the lower and upper probabilities assigned to the dataset $\mathbf{x}'$. It is easy to show that, in this case, the minimum of f is 0 and is reached in all the points $\theta \in \Theta$ with $\theta_i = 0$ 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

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

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

Corollary 1 *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.,*

$$\underline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s}) = \underline{\mathrm{P}}(\mathbf{x}') = 0,$$

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

In other words, Corollary 1 states a sufficient condition that prevents learning to take place under prior near-ignorance: if the likelihood function $\mathrm{P}(\mathbf{s} \,|\, \theta)$ is continuous and strictly positive on Θ, then all the dataset $\mathbf{x}' \in \mathcal{X}^{N'}$ for which $\mathcal{M}_0$ implies vacuous probabilities have vacuous probabilities 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 for the same reason for which, in Section 3, we have excluded the vacuous probability model: because only vacuous posterior probabilities are produced.

The sufficient condition described above is satisfied very often in practice, as illustrated by the following striking examples.

Example 2 Consider the medical test introduced in Example 1 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 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 $(\theta, 1-\theta)$, where θ corresponds to the (unknown) proportion of diseased individuals in the population. Denote with $1 - \varepsilon_1$ the sensitivity and with $1 - \varepsilon_2$ the specificity of the test. Then it holds that

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

$$P(S_i = - \mid X_i = I) = \varepsilon_2 > 0,$$

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,

$$\begin{aligned} & P(S_i = + \mid \theta) = \\ =& P(S_i = + \mid X_i = I) P(X_i = I \mid \theta) + \\ +& P(S_i = + \mid X_i = H) P(X_i = H \mid \theta) = \\ =& \underbrace{(1 - \varepsilon_2)}_{>0} \cdot \theta + \underbrace{\varepsilon_1}_{>0} \cdot (1 - \theta) > 0. \end{aligned}$$

Analogously,

$$\begin{aligned} & P(S_i = - \mid \theta) = \\ =& P(S_i = - \mid X_i = I) P(X_i = I \mid \theta) + \\ +& P(S_i = - \mid X_i = H) P(X_i = H \mid \theta) = \\ =& \underbrace{\varepsilon_2}_{>0} \cdot \theta + \underbrace{(1 - \varepsilon_1)}_{>0} \cdot (1 - \theta) > 0. \end{aligned}$$

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

$$\begin{aligned} & P(\mathbf{S} = \mathbf{s} \mid \theta) = ((1 - \varepsilon_2) \cdot \theta + \varepsilon_1 \cdot (1 - \theta))^{n^{\mathbf{s}}} \cdot \\ & \cdot (\varepsilon_2 \cdot \theta + (1 - \varepsilon_1) \cdot (1 - \theta))^{N - n^{\mathbf{s}}} > 0 \end{aligned}$$

for each $\theta \in [0, 1]$ and each $\mathbf{s} \in \mathcal{X}^N$. Therefore, according to Corollary 1, all the predictive probabilities that, according to $\mathcal{M}_0$, are vacuous a priori remain vacuous a posteriori. It follows that, if we want to avoid vacuous posterior predictive probabilities, then we cannot model our prior knowledge (ignorance) using a near-ignorance prior implying some vacuous prior predictive probabilities. 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 the 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 2 focuses on discrete latent and manifest variables. In the next example, we show that our theoretical results have important implications also in models with discrete latent variables and continuous manifest variables.

Example 3 Consider the sequence of IID categorical variables $(\mathrm{X}_i)_{i\in\mathbf{N}}$ with outcomes in $\mathcal{X}^N$ and unknown chances $\theta \in \Theta$. 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

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

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,

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

Therefore, according to Corollary 1, if we model our prior knowledge using a near-ignorance prior $\mathcal{M}_0$, the vacuous prior predictive probabilities implied by $\mathcal{M}_0$ remain vacuous a posteriori. It follows that, if we want to avoid vacuous posterior predictive probabilities, we cannot model our prior knowledge using a near-ignorance prior implying some vacuous prior predictive probabilities. $\diamondsuit$

Examples 2 and 3 raise, in general, serious criticisms about the use of near-ignorance priors in practical applications.

The only predictive model in the literature, of which we are aware, where a near-ignorance prior is used successfully to obtain non-vacuous posterior predictive probabilities is the IDM. In the next example, we explain how the IDM avoids our theoretical limitations.

Example 4 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. 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 $\mathrm{P}(\mathbf{X} = \mathbf{x} \,|\, \theta) = 0$ for all θ such that $\theta_j = 0$ for at least one j such that $n_j > 0$ and $\mathrm{P}(\mathbf{X} = \mathbf{x} \,|\, \theta) > 0$ for all the other $\theta \in \Theta$, in particular for all θ in the interior of Θ.

The near-ignorance prior $\mathcal{M}_0$ used in the IDM consists in the set of all the Dirichlet densities $dir_{s,\mathbf{t}}(\theta)$ for a fixed $s > 0$ and all $\mathbf{t} \in \mathcal{T}$, where $dir_{s,\mathbf{t}}(\theta)$ and $\mathcal{T}$ have been defined in (1) and (2).

The particular choice of $\mathcal{M}_0$ in the IDM implies, for each $N' \geq 1$ and each $i \in \{1, \ldots, k\}$, that

$$\underline{\mathbf{E}}(\theta_i^{N'}) = 0, \quad \overline{\mathbf{E}}(\theta_i^{N'}) = 1.$$

Consequently, denoting with $\mathbf{d^i} \in \mathcal{X}^{N'}$ the dataset with $n'_i = N'$ and $n'_j = 0$ for each $j \neq i$, a priori we have,

$$\underline{\mathrm{P}}(\mathbf{X}' = \mathbf{d^i}) = 0, \quad \overline{\mathrm{P}}(\mathbf{X}' = \mathbf{d^i}) = 1,$$

and in particular

$$\underline{\mathrm{P}}(\mathrm{X}_1 = x_i) = 0, \quad \overline{\mathrm{P}}(\mathrm{X}_1 = x_i) = 1.$$

It can be shown that other prior predictive probabilities are not vacuous. For example, for $i \neq j$, we have

$$\overline{\mathbf{E}}(\theta_i \theta_j) = \frac{s}{4(s+1)} < \frac{1}{4} = \sup_{\theta \in \Theta} \theta_i \theta_j.$$

The IDM produces, for each possible observed data set $\mathbf{x}$, non-vacuous posterior predictive probabilities for each possible future data set (see [Walley (1996)]). This means that our previous theoretical limitations are avoided in some way. To explain this result we consider two cases. We consider firstly an observed data set $\mathbf{x}$ where we have observed at least two different outcomes. Secondly, we consider a data set $\mathbf{x}$ formed exclusively by outcomes of the same type, in other words, a data set of the type $\mathbf{d^i}$.

In the first case we have that $\mathrm{P}(\mathbf{x} \,|\, \theta) = \prod_{j=1}^{k} \theta_j^{n_j}$ is equal to zero for $\theta = \mathbf{e^i}$ for each $i \in \{1, \ldots, k\}$. In fact, $\theta_i = 1$ implies $\theta_j = 0$ for each $j \neq i$ and there is at least one j with $n_j > 0$. Therefore, the assumptions of Corollaries 4 and 5 are not satisfied. And in fact the IDM produces non-vacuous posterior predictive probabilities for each data set that, a priori, has vacuous predictive probabilities. On the other hand, all the datasets whose prior predictive probability reaches its maximum in a relative frequency $\mathbf{f} \in \Theta$ such that $\mathrm{P}(\mathbf{x} \,|\, \mathbf{f}) > 0$, are characterized by non-vacuous prior predictive probabilities.

The second case yields similar results. The only difference is that $\mathrm{P}(\mathbf{d^i} \,|\, \theta) = \theta_i^{N'}$ for a given $i \in \{1, \ldots, k\}$. In this case $\mathrm{P}(\mathbf{x} \,|\, \mathbf{e^i}) = 1 > 0$ and in fact, according to Corollaries 4 and 5, we obtain

$$\overline{\mathrm{P}}(x_i \,|\, \mathbf{x}) = \overline{\mathrm{P}}(x_i) = 1,$$

$$\overline{\mathrm{P}}(\mathrm{X}' = \mathbf{d^i} \,|\, \mathbf{x}) = \overline{\mathrm{P}}(\mathbf{d^i}) = 1,$$

and consequently, for each $j \neq i$ and each $\mathbf{y} \neq \mathbf{d^i}$,

$$\underline{\mathrm{P}}(x_j \,|\, \mathbf{x}) = \underline{\mathrm{P}}(x_j) = 0,$$

$$\underline{\mathrm{P}}(\mathrm{X}' = \mathbf{y} \,|\, \mathbf{x}) = \underline{\mathrm{P}}(\mathbf{y}) = 0.$$

But, on the other hand, we obtain

$$\underline{\mathrm{P}}(x_i \,|\, \mathbf{x}) > 0, \quad \underline{\mathrm{P}}(\mathrm{X}' = \mathbf{d^i} \,|\, \mathbf{x}) > 0,$$
$$\overline{\mathrm{P}}(x_j \,|\, \mathbf{x}) < 1, \quad \overline{\mathrm{P}}(\mathrm{X}' = \mathbf{y} \,|\, \mathbf{x}) < 1,$$

and therefore the posterior predictive probabilities are not vacuous for each possible future data set. ♢

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.

5 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 about 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 substantial evidence against the possibility to use prior near-ignorance in real statistical problems. Given that complete prior ignorance is not compatible with learning, as it is well known, we deduce that there is little hope to use any form of prior ignorance to do objective-minded statistical inference in practice.

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.

Acknowledgements

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

A Technical preliminaries

In this appendix we provide 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,

$$\mathbf{E}_n(f) := \int_{\Theta} f(\theta)p_n(\theta)d\theta,$$

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

In addition, for a given probability density p on Θ,

$$\mathbf{E}_p(f) := \int_{\Theta} f(\theta) p(\theta) d\theta,$$

$$\mathrm{P}_p(\widetilde{\Theta}) := \int_{\widetilde{\Theta}} p(\theta) d\theta, \quad \widetilde{\Theta} \subseteq \Theta.$$

Finally, with $\rightarrow$ we denote $\lim_{n\rightarrow\infty}$.

Theorem 2 *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. For this function define the measurable sets*

$$\Theta_\delta = \{\theta \in \Theta \,|\, f(\theta) \geq f_{\max} - \delta\}. \tag{4}$$

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

$$\mathbf{E}_n(f) \rightarrow f_{\max}, \tag{5}$$

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

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

Theorem 3 *Let $L(\theta) \geq 0$ be a bounded measurable function with*

$$\lim_{\delta\rightarrow 0} \inf_{\theta\in\Theta_\delta} L(\theta) =: c > 0, \tag{6}$$

under the same assumptions of Theorem 2. Then

$$\frac{\mathbf{E}_n(Lf)}{\mathbf{E}_n(L)} = \frac{\int_\Theta f(\theta) L(\theta) p_n(\theta) d\theta}{\int_\Theta L(\theta) p_n(\theta) d\theta} \rightarrow f_{\max}. \tag{7}$$

Remark 1 *If f has a unique maximum in $\theta = \theta_0$ and L is a function, continuous in an arbitrary small neighborhood of $\theta = \theta_0$, such that $L(\theta_0) > 0$, then (6) is satisfied.*

B Corollaries to Theorem 1

The following Corollaries to Theorem 1 are necessary to prove Corollary 1, and are useful to understand more deeply the limiting results implied by the use of near-ignorance priors with latent variables.

Corollary 2 *Let* $\mathbf{x}'$ *and* $\mathbf{s}$ *be given. Denote with* $\mathbf{f}' := \left(\frac{n'_1}{N'}, \ldots, \frac{n'_k}{N'}\right) \in \Theta$ *the vector of relative frequencies of the dataset* $\mathbf{x}'$*. If* $\mathrm{P}(\mathbf{s} \,|\, \theta)$ *is continuous in an arbitrary small neighborhood of* $\theta = \mathbf{f}'$*,* $\mathrm{P}(\mathbf{s} \,|\, \mathbf{f}') > 0$ *and* $\mathcal{M}_0$ *is such that*

$$\overline{\mathrm{P}}(\mathbf{x}') = \sup_{\theta \in \Theta} \left(\prod_{i=1}^{k} \theta_i^{n'_i} \right) = \prod_{i=1}^{k} \left(\frac{n'_i}{N'} \right)^{n'_i},$$

then

$$\overline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s}) = \overline{\mathrm{P}}(\mathbf{x}').$$

Corollary 3 *Let* $\mathbf{x}'$ *and* $\mathbf{s}$ *be given. If* $\mathrm{P}(\mathbf{s} \,|\, \theta) > 0$ *for each* $\theta \in \Theta$ *with* $\theta_i = 0$ *for at least one* i *with* $n'_i > 0$*, and* $\mathcal{M}_0$ *is such that* $\underline{\mathrm{P}}(\mathbf{x}') = 0$*, it follows that*

$$\underline{\mathrm{P}}(\mathbf{x}' \,|\, \mathbf{s}) = \underline{\mathrm{P}}(\mathbf{x}') = 0.$$

Corollary 4 *Let* $\mathbf{s}$ *be given. Consider an arbitrary* $x_i \in \mathcal{X}$ *and denote with* $\mathbf{e^i}$ *the particular vector of chances with* $\theta_i = 1$ *and* $\theta_j = 0$ *for each* $j \neq i$*. Suppose that* $\mathcal{M}_0$ *is such that, a priori,* $\overline{\mathrm{P}}(\mathrm{X}_1 = x_i) := \overline{\mathbf{E}}(\theta_i) = 1$*. Then, if* $\mathrm{P}(\mathbf{s} \,|\, \mathbf{e^i}) > 0$ *and* $\mathrm{P}(\mathbf{s} \,|\, \theta)$ *is continuous in a neighborhood of* $\theta = \mathbf{e^i}$*, we have*

$$\overline{\mathrm{P}}(\mathrm{X}_{N+1} = x_i \,|\, \mathbf{s}) = \overline{\mathrm{P}}(\mathrm{X}_1 = x_i) = 1, \tag{8}$$

and consequently,

$$\underline{\mathrm{P}}(\mathrm{X}_{N+1} = x_j \,|\, \mathbf{s}) = \underline{\mathrm{P}}(\mathrm{X}_j = x_i) = 0, \tag{9}$$

for each $j \neq i$*.*

Corollary 5 *Let* $\mathbf{s}$ *and* N' *be given and consider an arbitrary* $x_i \in \mathcal{X}$*. Suppose that* $\mathcal{M}_0$ *is such that, a priori,* $\overline{\mathrm{P}}(\mathrm{X}_1 = x_i) := \overline{\mathbf{E}}(\theta_i) = 1$*. Denote with* $\mathbf{d^i} \in \mathcal{X}^{N'}$ *the data set with* $n_i = N'$ *and* $n_j = 0$ *for each* $j \neq i$*. Then, if* $\mathrm{P}(\mathbf{s} \,|\, \mathbf{e^i}) > 0$ *and* $\mathrm{P}(\mathbf{s} \,|\, \theta)$ *is continuous in a neighborhood of* $\theta = \mathbf{e^i}$*, we have*

$$\overline{\mathrm{P}}(\mathbf{X}' = \mathbf{d^i} \,|\, \mathbf{s}) = 1,$$

and consequently,

$$\underline{\mathrm{P}}(\mathbf{X}' = \mathbf{y} \,|\, \mathbf{s}) = 0,$$

for each $\mathbf{y} \neq \mathbf{d^i}$*.*

References

[Bernard (2005)] 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 et al. (2002)] Boorsbom D., Mellenbergh G. J., van Heerden J. (2002) The theoretical status of latent variables. *Psychological Review*, 110 (2), 203–219.

[De Cooman and Miranda (2006)] 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.

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

[Hutter (2006)] 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 and Wassermann (1996)] Kass R., Wassermann L. (1996) The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91: 1343–1370.

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

[Piatti et al. (2005)] 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), 276–286.

[Skrondal and Rabe-Hesketh (2004)] Skrondal A., Rabe-Hasketh S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Chapman and Hall/CRC, Boca Raton.

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

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

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

[Walley et al. (1999)] 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.