INFORMES TÉCNICOS MIDE UC / TECHNICAL REPORTS MIDE UC
Centro de Medición MIDE UC / Measurement Center MIDE UC

IT 1003

Bayesian Identifiability: Contributions to an Inconclusive Debate

ERNESTO SAN MARTÍN AND JORGE GONZÁLEZ

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE PSICOLOGÍA

Bayesian Identifiability: Contributions to an Inconclusive Debate

ERNESTO SAN MARTÍN[a, b] AND JORGE GONZÁLEZ[b]
[a]*Department of Statistics, Pontificia Universidad Católica de Chile, Chile*
[b]*Measurement Center MIDE UC, Pontificia Universidad Católica de Chile, Chile*

April 12, 2010

Abstract

Using the concept of reduction by sufficiency of a Bayesian model, the issue of Bayesian identifiability is discussed. Various statements given in the literature on Bayesian identifiability as well as identifiability relationships in hierarchical structures, are revised. Examples and counter-examples are given to show that some of these statements are confusing or simply wrong. Most of these examples are developed in a fully discrete Bayesian model.

Keywords: Statistical Model; Hierarchical structure; Sufficient Parameter; Minimal Sufficient Parameter; Updating Process.

The author holds firmly to the view that
this contingent and cognitive universe of ours
is in reality only finite and, therefore, discrete.
[. . .] Infinite and continuous models [. . .] are to be
looked as mere approximations to the finite realities.
Basu (1975, p. 4)

Personal account

Pilar was an active supporter and developer of Bayesian statistics. She knew about the debate of whether identifiability is of interest or not in a Bayesian approach. The first author remember interesting discussions with Pilar in this debate. This paper represents an answer to our late discussions. It is couched following Florens, Mouchart and Rolin's "Elements of Bayesian Statistics" (1990) perspective, which in turn develops the concept of *reduction by sufficiency of the Bayesian model* formalized by Basu. It seems appropriate on this occasion mention that Pilar's advisor was Carlos Pereira (Univeridade de Sao Paulo, Brazil). Motivated by certain works of Michel Mouchart and Jean-Marie Rolin, Pereira did his Ph. D. thesis under the supervision of Basu. Basu met several times with Michel; and Michel was my advisor. Their contributions are the basis of this paper that we write in memory of Pilar.

1 Introduction

Identifiability is often treated as a necessary condition to have a rigorously well specified statistical model. When structural modeling is being considered, that is, when the model formalizes a certain phenomenon, the identification condition is more than a simple technical assumption, but covers a more fundamental aspect that is intuitively the adequacy of a theoretical statistical model for an observed process.

In the classical or sampling approach, a statistical model is defined as an indexed family of distributions on the sample space, whereas the Bayesian approach considers a unique probability measure on the product space "parameters $\times$ observations" (Florens et al., 1990; Gourieroux and Monfort, 1995). In both approaches the concept of identifiability has been extensively discussed in the literature (see, among others, Manski, 1995, 2007; Prakasa Rao, 1992, and the references there in).

Under the Bayesian approach, however, the concept of identifiability and, particularly, *nonidentifiability* (Dawid, 1979) has not been free from controversies, polemics and confusion. The famous sentence *In passing it might be noted that unidentifiability causes no real difficulty in the Bayesian approach* (Lindley, 1971) is a remarkable example.

Depending on whether the problem is in the prior, the likelihood or the posterior distribution, different views on the issue of identifiability have been given in the literature. Poirier (1998) argued that "A Bayesian analysis of a nonidentified model is always possible if a proper prior on all the parameters is specified". In the same line, Eberly and Carlin (2000) pointed out that "in some sense identifiability is a non-issue for Bayesian analyses, since given proper prior distributions the corresponding posteriors must be proper as well, hence every parameter can be well estimated". Gelfand and Sahu (1999) stated that too informative priors will dominate the inference and priors near to be improper will produce ill-behaved posteriors, yet on the other hand they argued that nonidentifiability would not depend on the nature of the prior specification but on lack of identifiability in the likelihood. In this sense, Dawid's (classical) definition of Bayesian nonidentifiability is equivalent with what a Bayesian would call likelihood identifiability (Eberly and Carlin, 2000). An unified thought is given by Kadane (1975) who stated that "identification is a property of the likelihood function and is the same whether considered classically or from the Bayesian approach." The issue has also been discussed when simulation-based techniques are used for model fitting and inferences. It is commonly argued that nonidentifiability does not preclude Bayesian inference as long as a suitable informative prior is specified. Kass et al. (1998) point out that provided the posterior is proper, there is no problem for MCMC methods-assuming that one has determined that the nonidentifiability "isn't due to a bug".

It is remarkable that all these authors refer to Bayesian unidentifiability or nonidentifiability rather than identifiability itself. The origin of the controversy in these two definitions seems to be the lack of distinction between the concepts of sufficient parameter and minimal sufficient parameter. These concepts are used to describe the information provided by the sampling process in a Bayesian model. Bayesian identifiability and nonidentifiability will then be related to minimal parameter sufficiency and parameter sufficiency, respectively.

Taking into the account the preceding differentiation, the main stream in Bayesian statistics can be qualified. Furthermore, in this paper we argue that Bayesian identifiabiliy defined through parametric minimal sufficiency, is not only a genuinely Bayesian concept but also related with other Bayesian and non-Bayesian statistical concepts.

In order to go in a deeper discussion, the previously mentioned concepts and controversy are discussed using a fully discrete bayesian model, that is, a model in which both the parameter and the sample space are finite. Using this approach, we believe it is possible to see in a more detailed way the origin of the possible confusions. Moreover, most of Bayesian models are specified using a hierarchical structure, as it is the case, for instance, in generalized linear mixed models. In this context, the applied literature suggests to relate the identifiability among the submodels composing the hierarchy. For instance, in a GLMMM it is typically questioned whether the variance of the random effect is still identified in the marginal observed model which is obtained after integrating out the random effects. We offer counter-examples to these relationships, when they are wrong, and review and illustrate them, when they are true.

The paper is organized as follows: First we introduce the Bayesian model together with the definitions of basics concepts necessary to define Bayesian identifiability. Next, the construction of a fully discrete Bayesian model is explained and its Bayesian identification discussed. Subsequently, we describe identifiability relationships in hierarchical structures using both counter-examples for relationships that seems to be true, and the fully discrete Bayesian approach to exemplify these relationship. The paper conclude with a discussion.

2 Bayesian model

2.1 General construction of a Bayesian model

In a (classical) sampling theory framework, a statistical model is formally defined as follows:

$$\mathcal{E} = \{(S, \mathcal{S}),\ P^a\ :\ a \in A\}, \tag{2.1}$$

where $(S, \mathcal{S})$ is a measurable space, the *sample space*, and $\{P^a\ :\ a \in A\}$ is a family of probability measures on the sample space indexed by a *parameter* a belonging to a *parameter space* A (see, e.g., Fischer, 1922; Barra, 1981; McCullagh, 2002). The probabilities $\{P^a\ :\ a \in A\}$ are called *sampling probabilities* as they describe the sampling process. The parameter space A might be either a Euclidean space, a functional space, or a product of both as it is the case in parametric, non-parametric and semi-parametric models, respectively. Note that the statistical model in (2.1) can be considered as an extension of a probability space $(S, \mathcal{S}, P)$ in the sense that a unique probability measure P is replaced by a *family* of probability measures P^a, $a \in A$.

A *Bayesian model* is defined as a *unique* probability measure Q defined on the product space "parameters $\times$ observations", denoted as $A \times S$. Taking as a starting point the statistical model given by (2.1), a probability measure Q on $A \times S$ is constructed by endowing the parameter space A with a probability measure μ on $(A, \mathcal{A})$, where the σ-field $\mathcal{A}$ of subsets of A makes $P^a(X)$ measurable for all $X \in \mathcal{S}$, and by extending to $\mathcal{A} \otimes \mathcal{S}$ (in a unique way) the function Q defined on $\mathcal{A} \times \mathcal{S}$ as follows:

$$Q(E \times X) \;=\; \int_E P^a(X)\, d\,\mu \qquad E \in \mathcal{A}, \quad X \in \mathcal{S}. \tag{2.2}$$

The measure constructed from (2.2) is denoted as the Markovian product $Q = \mu \otimes P^{\mathcal{A}}$. Thus, a Bayesian model is defined by the following probability space:

$$\mathcal{E} = (A \times S, \mathcal{A} \vee \mathcal{S}, Q = \mu \otimes P^{\mathcal{A}}). \tag{2.3}$$

Remark 1 In this paper, we shall systematically relate the sub-σ-field $\mathcal{B} \subset \mathcal{A}$ (resp., $\mathcal{T} \subset \mathcal{S}$) to the sub-$\sigma$-field of the corresponding cylinders $\mathcal{B} \times S$ (resp. $A \times \mathcal{T}$). Thus, in (2.3), we relate the product $\mathcal{A} \otimes \mathcal{S}$ to $\mathcal{A} \vee \mathcal{S}$, the σ-field generated by $(\mathcal{A} \times S) \cup (A \times \mathcal{S})$. This is to alleviate the notation.

By construction P^a in (2.2) becomes a transition of probability representing a regular version of $P^{\mathcal{A}}$, the restriction to $\mathcal{S}$ of the conditional probability Q given $\mathcal{A}$, and this is so for whatever probability μ on $(A, \mathcal{A})$. Moreover, the so-called *prior probability* μ corresponds to the marginal probability of Q on $(A, \mathcal{A})$, namely $\mu(E) = Q(E \times S)$ for $E \in \mathcal{A}$. Similarly, the marginal probability P on the sample space $(S, \mathcal{S})$ given by $P(X) = Q(A \times X)$ for $X \in \mathcal{S}$ is called the *predictive probability*.

Besides the decomposition $Q = \mu \otimes P^{\mathcal{A}}$, the probability Q is decomposed, under the usual hypotheses (e.g., Rao, 1993), into a marginal probability P, and a regular conditional probability given $\mathcal{S}$, represented by a probability transition denoted as $\mu^{\mathcal{S}}$; this is the so-called *posterior distribution*. When Q is decomposed as $Q = \mu \otimes P^{\mathcal{A}} = P \otimes \mu^{\mathcal{S}}$, the Bayesian model (2.3) is said to be *regular*. For more details, see Martin et al. (1973); Florens et al. (1990). In the remaining of this paper we assume that the Bayesian model (2.3) is regular.

The main difference between models (2.1) and (2.3) is that the first is a *family* of sampling distributions, whereas the second is a *unique* probability measure defined on the product space "parameters $\times$ observations". It should be emphasized that in a Bayesian model, the prior distribution μ is fixed. However, when the interest is focused on the sensitivity of Bayesian procedures with respect to changes on the prior distribution (see, e.g., Macci and Polettini, 2001), or on the Bayesian inference using intersubjective models (see, e.g., Dawid, 1979, section 9), we are dealing with a *family* of Bayesian models indexed by prior distributions defined on the parameter space $(A, \mathcal{A})$; that is,

$$\mathcal{E} = \{(A \times S, \mathcal{A} \vee \mathcal{S}),\; Q^{\mu} = \mu \otimes P^{\mathcal{A}},\;\; \mu \in \mathcal{P}(A, \mathcal{A})\}, \tag{2.4}$$

where $\mathcal{P}(A, \mathcal{A})$ denotes the space of probability measures defined on the parameter space. As can be seen, (2.1) and (2.4) share a common mathematical structure.

In the following subsections we define the concepts of statistics, parameter, sufficient parameter, minimal sufficient parameter, sampling information and Bayesian identification, based on the Bayesian model defined in (2.3).

2.2 Statistics and parameters in a Bayesian model

Taking as reference the Bayesian model (2.3), any $\mathcal{S}$-measurable function T defined on S with values in some measurable space $(U,\mathcal{U})$ (typically, a Borel space) is called a *statistics*. More precisely, the function $T : (S, \mathcal{S}) \to (U,\mathcal{U})$ is a statistics if and only if $\mathcal{T} \doteq T^{-1}(\mathcal{U}) \subset \mathcal{S}$. This σ-field, also denoted as $\sigma(T)$ –the σ–field generated by T–, is the smallest σ-field which makes measurable the function T. It should be remarked that $\mathcal{T}$ heuristically corresponds to the set of events that may be described in terms of that random variable (Florens and Mouchart, 1982, p. 588). Taking into account that the information of interest is in the probability space given by (2.3), we consider that $\mathcal{T}$ represents the information provided by T, which does not depend on the coordinate system chosen to represent the corresponding random variable. This is because $\sigma(T) = \sigma[h(T)]$ for all bi-measurable and bijective function h.

Taking into account these considerations, a statistics is a sub-σ-field of $\mathcal{S}$. Similarly, a *subparameter*, or more simply a *parameter*, is a sub-σ-field of $\mathcal{A}$.

2.3 Sufficient parameter

Let $\mathcal{T} \subset \mathcal{S}$ be a statistics and $\mathcal{B} \subset \mathcal{A}$ be a parameter. A statistics $\mathcal{T}$ is sufficient if, conditionally on it, the sampling process is independent of the parameter. Using properties of conditional independence (see Appendix A), the last statement can be written as $\mathcal{S} \perp\!\!\!\perp \mathcal{A} \mid \mathcal{T}$ which is read as "$\mathcal{S}$ is independent of $\mathcal{A}$ given $\mathcal{T}$". Thus, a statistics $\mathcal{T}$ is sufficient for the parameter $\mathcal{A}$ if, for all $\mathcal{S}$-measurable function s, $E(s \mid \mathcal{A} \vee \mathcal{T}) = E(s \mid \mathcal{T})$; that is, the process generating the observations conditionally on $\mathcal{A} \vee \mathcal{T}$ only depends on the statistics $\mathcal{T}$. Equivalently, for all $\mathcal{A}$-measurable function a, $E(a \mid \mathcal{S}) = E(a \mid \mathcal{T})$; that is, the posterior process is fully characterized by the sufficient statistics, being the observations $\mathcal{S}$ redundant once $\mathcal{T}$ is "given". Thus, the original Bayesian model (2.3) can be replaced, *without any loose of information*, by

$$\widetilde{\mathcal{E}} = (A \times S, \mathcal{A} \vee \mathcal{T}, Q_{\mathcal{T}}), \tag{2.5}$$

where $Q_{\mathcal{T}}$ is the restriction of Q on $\mathcal{T}$, that is, $Q_{\mathcal{T}}(E \times X) \doteq Q(E \times X)$ for $E \in \mathcal{A}$ and $X \in \mathcal{T}$. Following Basu (1975)'s and Florens et al. (1990)'s terminology, $\widetilde{\mathcal{E}}$ corresponds to *a reduction by sufficiency of the Bayesian model (2.3)*. The Bayesian learning process on the parameter $\mathcal{A}$ is unaffected by this reduction.

Taking advantage of the symmetric role of observations and parameters in a Bayesian model, it is possible to define a sufficient parameter in a way similar to that of sufficient statistics. As a matter of fact, a parameter $\mathcal{B}$ is sufficient with respect to $\mathcal{S}$ if and only if $\mathcal{A} \perp\!\!\!\perp \mathcal{S} \mid \mathcal{B}$, which means that, for all $\mathcal{S}$-measurable function s, $E(s \mid \mathcal{A}) = E(s \mid \mathcal{B})$; that is, the sufficiency of $\mathcal{B}$ means that $\mathcal{B}$ is "sufficient" to describe the sampling process and, consequently, that the observations $\mathcal{S}$ brings information on $\mathcal{B}$ only in the sense that, conditionally on $\mathcal{B}$, observation brings no information, that is, $E(a \mid \mathcal{S} \vee \mathcal{B}) = E(a \mid \mathcal{B})$ for all $\mathcal{A}$-measurable function a.

Using the properties of conditional independence, it can be verified that if $\mathcal{B} \subset \mathcal{A}$ is a sufficient parameter for $\mathcal{S}$ and $\mathcal{C} \subset \mathcal{A}$ is a parameter, then $\mathcal{A} \perp\!\!\!\perp \mathcal{S} \mid \mathcal{B} \vee \mathcal{C}$, that is, the parameter $\mathcal{B} \vee \mathcal{C}$ is also sufficient. This means that the sampling process is not only described by $\mathcal{B}$, but also by $\mathcal{B} \vee \mathcal{C}$ *for all* $\mathcal{C} \subset \mathcal{A}$. In other words, once we have a sufficient description of the sampling process, then such a process can be described using *redundant information at the parameter level*, which in turn is sufficient. Consequently, it makes sense to look for the *minimal* sufficient parameter describing the sampling process.

2.4 Minimal sufficient parameter

Let $\Sigma_{\mathcal{A}}$ be the class of sufficient parameters $\mathcal{B} \subset \mathcal{A}$ for $\mathcal{S}$, namely $\Sigma_{\mathcal{E}} = \{\mathcal{B} \subset \mathcal{A} : \mathcal{A} \perp\!\!\!\perp \mathcal{S} \mid \mathcal{B}\}$. It is clear that $\Sigma_{\mathcal{E}}$ is not empty since $\mathcal{A} \in \Sigma_{\mathcal{E}}$, hence $\Sigma_{\mathcal{E}} \neq \emptyset$. Take $\mathcal{B}_1, \mathcal{B}_2 \in \Sigma_{\mathcal{E}}$. Using both the definition of parametric sufficiency and the characterization of conditional independence in terms of a measurability condition (see Appendix A), it follows that the sufficiency of $\mathcal{B}_1$ implies that, for all $\mathcal{A}$-measurable function a, $E(a \mid \mathcal{A})$ is $\overline{\mathcal{B}}_1$-measurable; and the sufficiency of $\mathcal{B}_2$ implies that, for all $\mathcal{A}$-measurable function a, $E(a \mid \mathcal{A})$ is $\overline{\mathcal{B}}_2$-measurable. Here, $\overline{\mathcal{B}}_j$ ($j = 1, 2$) denotes the measurable completion $\overline{\mathcal{B}}_j = \mathcal{B}_j \vee \{E \in \mathcal{A} : \mu(E)^2 = \mu(E)\}$, where $\{E \in \mathcal{A} : \mu(E)^2 = \mu(E)\}$ is set of the prior null sets. It follows that, for all $\mathcal{A}$-measurable function a, $E(a \mid \mathcal{A})$ is $\overline{\mathcal{B}}_1 \cap \overline{\mathcal{B}}_2$-measurable, that is, $\mathcal{A} \perp\!\!\!\perp \mathcal{S} \mid \overline{\mathcal{B}}_1 \cap \overline{\mathcal{B}}_2$; thus, $\overline{\mathcal{B}}_1 \cap \overline{\mathcal{B}}_2 \in \Sigma_{\mathcal{A}}$. This argument is used to define a *minimal sufficient parameter* $\mathcal{B}_{\text{min}}$, as follows

$$\mathcal{B}_{\text{min}} = \bigcap_{\mathcal{B} \in \Sigma_{\mathcal{A}}} \overline{\mathcal{B}}. \tag{2.6}$$

Moreover, it shows that $\mathcal{B}_{\text{min}}$ *always* exists. By construction, the minimal sufficient parameter $\mathcal{B}_{\text{min}}$ contains all the prior null sets. Once the minimal sufficient parameter has been constructed, the original Bayesian model (2.3) should be replaced by

$$\mathcal{E}_{\text{min}} = (A \times S, \mathcal{B}_{\text{min}} \vee \mathcal{S}, Q_{\mathcal{B}_{\text{min}}}), \tag{2.7}$$

where $Q_{\mathcal{B}_{\text{min}}}$ is the restriction of Q on $\mathcal{B}_{\text{min}}$, that is, $Q_{\mathcal{B}_{\text{min}}}(E \times X) \doteq Q(E \times X)$ for $E \in \mathcal{B}_{\text{min}}$ and $X \in \mathcal{S}$.

The Bayesian model (2.7) does not contain redundant information at the parameter level because there is not a sufficient description of the sampling process (i.e. a $\mathcal{B} \in \Sigma_{\mathcal{A}}$) better than the description provided by the minimal sufficient parameter $\mathcal{B}_{\text{min}}$. Thus, the minimal sufficient parameter corresponds to the greatest possible parameter reduction for which the prior information is updated by the sample. Consequently, *the learning process underlying a Bayesian model is fully concentrated on the minimal sufficient parameter*.

2.5 Minimal sufficient parameter, sampling information and Bayesian identification

The minimal sufficient parameter $\mathcal{B}_{\text{min}}$ can be expressed in more operational terms. As a matter of fact, the σ-field generated by every version of the sampling expectations, namely $\sigma\{E(s \mid \mathcal{A}) : s \in [\mathcal{S}]^+\}$, is the smallest sub-$\sigma$-field of $\mathcal{A}$ that makes the sampling expectations measurable; here, $[\mathcal{S}]^+$ denotes the set of non-negative $\mathcal{S}$-measurable functions. This is equivalent to

$$\mathcal{A} \perp\!\!\!\perp \mathcal{S} \mid \sigma\{E(s \mid \mathcal{A}) \,:\, s \in [\mathcal{S}]^+\}.$$

Therefore, $\sigma\{E(s \mid \mathcal{A}) \,:\, s \in [\mathcal{S}]^+\} \in \Sigma_{\mathcal{A}}$ and, consequently,

$$\sigma\{E(s \mid \mathcal{A}) \,:\, s \in [\mathcal{S}]^+\} \supset \mathcal{B}_{\text{min}}.$$

On the other hand, $\mathcal{B}_{\text{min}}$ satisfies the condition that $\mathcal{A} \perp\!\!\!\perp \mathcal{S} \mid \mathcal{B}_{\text{min}}$, which is equivalent to say that, for all $\mathcal{S}$-measurable function s, $E(s \mid \mathcal{A})$ is $\mathcal{B}_{\text{min}}$-measurable. Therefore, by definition of $\sigma\{E(s \mid \mathcal{A}) \,:\, s \in [\mathcal{S}]^+\}$, it follows that

$$\sigma\{E(s \mid \mathcal{A}) \,:\, s \in [\mathcal{S}]^+\} \subset \mathcal{B}_{\text{min}}.$$

Thus, the minimal sufficient parameter $\mathcal{B}_{\text{min}}$ coincides with *all* the information provided by the sampling process, namely $\sigma\{E(s \mid \mathcal{A}) \,:\, s \in [\mathcal{S}]^+\}$. Additionally, this information contains *all* the prior null sets.

In summary, we have described the sampling process in terms of the information it provides. Using the concept of sufficiency, it has also been clarified how this information is updated by the learning-by-observation process. These considerations allow we to define Bayesian identifiabilty.

Definition 1 In the context of the Bayesian model (2.3), the parameter $\mathcal{A}$ is *Bayesian identified* by $\mathcal{S}$, which is denoted by $\mathcal{A} \prec \mathcal{S}$, if $\mathcal{A}$ is a minimal sufficient parameter; that is, $\mathcal{A} = \mathcal{B}_{\text{min}}$. More generally, let $\mathcal{M}_i \subset \mathcal{A} \vee \mathcal{S}$, for $i = 1, 2, 3$ –i.e. a function of the parameters, of the observations or of both. It is said that $\mathcal{M}_1$ *is Bayesian identified by* $\mathcal{M}_2$ *conditionally on* $\mathcal{M}_3$ if $\mathcal{M}_1 \vee \mathcal{M}_3$ is Bayesian identified by $\mathcal{M}_2 \vee \mathcal{M}_3$, *i.e.,*

$$\sigma\{E(f \mid \mathcal{M}_1 \vee \mathcal{M}_3) \,:\, f \in [\mathcal{M}_2 \vee \mathcal{M}_3]^+\} \;=\; \mathcal{M}_1 \vee \mathcal{M}_3.$$

This relationship is denoted as $\mathcal{M}_1 \prec \mathcal{M}_2 \mid \mathcal{M}_3$, which by definition is equivalent to $\mathcal{M}_1 \vee \mathcal{M}_3 \prec \mathcal{M}_2 \vee \mathcal{M}_3$.

In Bayesian statistics, this concept was introduced by Florens and Mouchart (1977) and thereafter further developed in, e.g., Florens and Rolin (1984); Mouchart and Rolin (1984) and, mainly, in Florens et al. (1990, chapter 4); see also Van Putten and Van Schuppen (1985) (for a table of correspondence between their results with that contained in Mouchart and Rolin (1984), see Mouchart and Rolin (1985)) and Gourieroux and Monfort (1995). In any case, this concept can be traced back to McKean (1963), although there is defined using the Lebesgue completion instead of the measurable completion.

From the definition of Bayesian identifiability given in Definition 1 there are four features that are worth to explain in more details: First, the learning-by-observing process in only based on the Bayesian identified parameter; second, the fact that we have a genuinely Bayesian concept of identifiability; third, the definition given yields a relation with Bayesian consistency; and fourth, the concept defined can be related to the classical parameter identifiability concept. We summarize this fourth issues in the followings comments:

Comment 1 The Bayesian identified parameter fully concentrates the learning-by-observing process. As a matter of fact, let us consider the Bayesian model (2.3); here, it is satisfied that

$$\mathcal{C} \perp\!\!\!\perp \mathcal{S} \mid \mathcal{B}_{\text{min}} \qquad \forall\, \mathcal{B}_{\text{min}} \subset \mathcal{C} \subset \mathcal{A}. \tag{2.8}$$

Here the parameter $\mathcal{C}$, although non-identified by $\mathcal{S}$, might be of interest and, consequently, it should be updated by the observations. As recognized by the Bayesian literature, this update can be done since, the posterior expectation $E(c \mid \mathcal{S})$ exists for all $\mathcal{C}$-measurable function c. However, the question is: are we actually learning about the unidentified parameter $\mathcal{C}$? The answer is *not*: although we compute the posterior distribution of $\mathcal{C}$, what we are actually updating is *always* the identified parameter, nothing more. Indeed, taking into account (2.8), it follows that

$$E(c \mid \mathcal{S}) = E[E(c \mid \mathcal{S} \vee \mathcal{B}_{\text{min}}) \mid \mathcal{S}] = E[E(c \mid \mathcal{B}_{\text{min}}) \mid \mathcal{S}]$$

for all $\mathcal{C}$-measurable function c. By definition of conditional expectation, $E(c \mid \mathcal{B}_{\text{min}})$ only depends on the identified parameter $\mathcal{B}_{\text{min}}$; this and only this is updated by the observations.

From a practical point of view, this means that if an unidentified parameter is estimated by its posterior distribution, the users should be warned that this estimates *does not provide any updating of the unidentified parameter, but only of the identified parameter*. This is more relevant when the unidentified parameter is a parameter of interest. In this case, if such a warning is not explicit, it will be providing an illusory result: the illusion that we obtain information about the unidentified parameter, when in reality we only get information about the identified parameter. This is illustrated in following example.

Example 1 Let us illustrate these considerations with the problem of estimating the prevalence, the sensitivity and the specificity of a diagnostic test in the absence of a gold-standard. Let $Z_i \in \{0,1\}$ be a binary random variable indicating the true state of a subject i, that is, $Z_i = 1$ if subject i is diseased, and $Z_i = 0$ otherwise. Let $Y_i \in \{0,1\}$ be a binary random variable indicating the classification of a subject i through a diagnostic test, that is, $Y_i = 1$ if subject i is classified as diseased by the diagnostic test, and $Y_i = 0$ otherwise. Note that Y_i is an observed variable, whereas Z_i is unobservable. Taking into account that this is a problem of misclassification, the parameters of interest are the sensitivity of the test, $\alpha = P(Y_i = 1 \mid Z_i = 1)$; the specificity of the test, $\beta = P(Y_i = 0 \mid Z_i = 0)$; and the true prevalence, $\omega = P(Z_i = 1)$. The sampling process is given by a sequence of mutually independent random variables Y_i conditionally on $p(\alpha,\beta,\omega)$, where $p(\alpha,\beta,\omega) = P(Y_i \mid \alpha,\beta,\omega) = \alpha\omega + (1-\beta)(1-\omega)$. If it is assumed that $\alpha + \beta > 1$, then $p(\alpha,\beta,\omega)$ is an increasing function of ω. The model is completed by specifying a prior probability distribution on (α,β,ω). The σ-field of the sample space is given by $\mathcal{S} = \sigma(Y_1,\ldots.Y_n)$, and the one of the parameters is given by $\mathcal{A} = \sigma(\alpha) \vee \sigma(\beta) \vee \sigma(\omega)$. Now, $p(\alpha,\beta,\omega) = E(Y_i \mid \mathcal{A})$, so

$$\sigma\{p(\alpha,\beta,\omega)\} \subset \sigma\{E(f \mid \mathcal{A}) \,:\, f \in [\mathcal{S}]^+\} \subset \mathcal{A}.$$

Moreover, $\sigma(\alpha) \not\subset \sigma\{p(\alpha,\beta,\omega)\}$, $\sigma(\beta) \not\subset \sigma\{p(\alpha,\beta,\omega)\}$, and $\sigma(\omega) \not\subset \sigma\{p(\alpha,\beta,\omega)\}$, because there not exist measurable functions such that α is a function of $p(\alpha,\beta,\omega)$, β is a function of $p(\alpha,\beta,\omega)$, and ω is a function of $p(\alpha,\beta,\omega)$. Therefore, neither $\mathcal{A}$ is Bayesian identified, nor α, nor β, nor ω. Furthermore, from the equality $E[Y_i \mid \mathcal{A}] = E[Y_i \mid p(\alpha,\beta,\omega)]$, and taking into account that Y_i is a binary random variable, it follows that

$$\sigma\{p(\alpha,\beta,\omega)\} \subset \sigma\{E(f \mid p(\alpha,\beta,\omega)) \,:\, f \in [\mathcal{S}]^+\} = \sigma\{E(f \mid \mathcal{A}) \,:\, f \in [\mathcal{S}]^+\},$$

which means that the sampling process is fully described by $p(\alpha,\beta,\omega)$. In other words, $p(\alpha,\beta,\omega)$ is Bayesian identified. Now, in this example, it is always possible to compute the posterior expectation of the parameters of interest. However, these expectations does not provide any information about them, but only about $p(\alpha,\beta,\omega)$ –which is typically called *apparent prevalence*.

□

Comment 2 The concept of identification as introduced in Definition 1 is genuinely Bayesian because it depends on the prior distribution through the prior null events. This means that if, additionally to the Bayesian model (2.3), a second one is specified in such a way that

$$\mathcal{E}' = (A \times S, \mathcal{A} \vee \mathcal{S}, Q' = \mu' \otimes P^{\mathcal{A}}),$$

then $\mathcal{B}_{\text{min}}$ is still the Bayesian identified parameter in the context of $\mathcal{E}'$ if and only if μ and μ' are equivalent probability measures (i.e., they have the *same* null sets). Thus, the Bayesian identifiability property (in the context of the Bayesian model (2.3)) can be lost if the prior null sets change; for more details, see Florens et al. (1990, Proposition 4.6.8). Contrary to what is typically accepted (Kass et al., 1998; Paulino and Pereira, 1994), this means that an unidentified parameter does not become identified once a proper prior distribution is specified on it, unless prior null sets are defined –which is equivalent to introduce dogmatic constraints. This is illustrated in the following example.

Example 2 Let us illustrate these considerations with the following example. Let $A = \mathbb{R}$ and $\mathcal{A}$ be the Borel sets of $\mathbb{R}$. Assume that the sampling process is specified as $(X_i \mid a) \sim \mathcal{N}(|a|, 1)$, where $X_1, \ldots . X_n$ are mutually independent conditionally on $\mathcal{A}$. Let μ be a prior probability distribution. Since $|a| = E(X_i \mid \mathcal{A})$, it follows that the corresponding minimal sufficient parameter is given by

$$\sigma\{|a|\} \vee \{E \in \mathcal{A} \,:\, \mu(E) = 0 \text{ or } 1\},$$

where $\sigma\{|a|\} = \{E \in \mathcal{A} \,:\, -E = E\}$. If μ is equivalent to the Lebesgue measure (i.e., μ has the same null sets as the Lebesgue measure on $\mathbb{R}$), then the parameter a is not identified because $\sigma\{a\} \not\subset \sigma\{|a|\}$. However, if μ is equivalent to the Lebesgue measure on $\mathbb{R}^+$ and $\mu(\mathbb{R}^-) = 0$, then the parameter a is identified.

□

Comment 3 Let us consider the Bayesian model (2.3) and let $\mathcal{B} \subset \mathcal{A}$ be a parameter of interest. It is said that $\mathcal{B}$ is exactly estimable if $\mathcal{B} \subset \overline{\mathcal{S}}$. This condition is equivalent to $Var(\mathbb{1}_B \mid \mathcal{S}) = 0$ for all $B \in \mathcal{B}$, which in turn is equivalent to $E(b \mid \mathcal{S}) = b$ Q-a.s. for all $\mathcal{B}$-measurable function b. Exact estimability means, therefore, that the posterior probability of a $\mathcal{B}$-measurable function b is a.s. 0 or 1. In other words, $\mathcal{B} \subset \overline{\mathcal{S}}$ formalizes the idea that the parameter $\mathcal{B}$ is "perfectly known" after the observation of the sample. For more details, see Florens et al. (1990, Section 4.7).

Bayesian identification is related with exact estimability, as in a pure sampling theory framework identification is related with consistency. As a matter of fact, let us consider the case of an iid process, namely $\{X_n : n \in \mathbb{N}\}$ be iid random variables conditionally on $\mathcal{A}$. In the context of the asymptotic Bayesian model, it can be shown that the minimal sufficient parameter $\mathcal{B}_{\text{min}} \doteq \sigma\{E(f \mid \mathcal{A}) : f \in [\mathcal{X}_1^\infty]^+\}$ satisfies the following condition:

$$\mathcal{B}_{\text{min}} \subset \overline{\mathcal{X}_1^\infty}; \tag{2.9}$$

here, $\mathcal{X}_1^\infty = \sigma(X_1, X_2, \ldots, X_n, \ldots)$; (for more details, see Florens et al., 1990, Theorem 9.3.2). This means that, in an iid process, the Bayesian identified parameter is exactly estimable. Furthermore, condition (2.9) ensures the convergence of the following posterior expectations: $E(b \mid \mathcal{X}_1^n)$ for all $\mathcal{B}_{\text{min}}$-measurable function b; and $E(a \mid \mathcal{X}_1^n)$ for all $\mathcal{A}$-measurable function a; here $\mathcal{X}_1^n = \sigma(X_1, \ldots, X_n)$. As a matter of fact, by the Martingale Theorem, $E(b \mid \mathcal{X}_1^n)$ converges a.s. and, moreover, it converges to $E(b \mid \mathcal{X}_1^\infty)$ in L^1 for all $\mathcal{B}_{\text{min}}$-measurable function b such that $E|b| < \infty$; being $L^1(A, \mathcal{A}, \mu) = \{f : A \to \mathbb{R} \mid f \text{is } \mathcal{A}\text{-measurable, and} \int_A \mid f \mid d\mu < \infty\}$. Using condition (2.9), it follows that $E(b \mid \mathcal{X}_1^\infty) = b$ a.s.

Similarly, let a be a $\mathcal{A}$-measurable function such that $E|a| < \infty$. Taking into account that $\mathcal{A} \perp\!\!\!\perp \mathcal{X}_1^n \mid \mathcal{B}_{\text{min}}$ for all $n \in \mathbb{N}$, it follows that

$$\begin{aligned} E(a \mid \mathcal{X}_1^n) &= E[E(a \mid \mathcal{X}_1^n \vee \mathcal{B}_{\text{min}}) \mid \mathcal{X}_1^n] \\ &= E[E(a \mid \mathcal{B}_{\text{min}}) \mid \mathcal{X}_1^n] \\ &\overset{n\longrightarrow\infty}{\longrightarrow} E[E(a \mid \mathcal{B}_{\text{min}}) \mid \mathcal{X}_1^\infty] = E(a \mid \mathcal{B}_{\text{min}}) \quad \text{a.s. and in } L^1. \end{aligned}$$

Thus, the Bayesian identified parameter is not only Bayesian consistent (in the sense that its posterior expectation converges to it), but also concentrates at the limit the posterior expectation of any other parameter.

Comment 4 In a pure sampling theory framework, parameter identifiability is defined as the injectivity of the mapping $a \longmapsto P^a$, where $\{P^a : a \in A\}$ is a family of sampling distributions. This identification concept (which we call s-identification) is related with Bayesian identification as introduced in Definition 1. More precisely if $\mathcal{A}$ is a Blackwell σ-field and $\mathcal{S}$ is separable, then s-identification implies Bayesian identification for all prior probability measure on $(A, \mathcal{A})$; for details and proofs, see Florens et al. (1985) and Florens et al. (1990, chapter 4). A σ-field $\mathcal{M}$ is a Blackwell σ-field if $\mathcal{M}$ is separable and if for all $\mathcal{M}$-measurable function m and for all $A \in \mathcal{M}$, $m(A)$ is an analytic set of $\mathbb{R}$. It should be emphasized that this relationship between s-identification and Bayesian identification depends on the separability of both the sample space and the parameter space. Since we use measurable completion (see Appendix A), we avoid the danger of loosing the separability and, consequently, we may use identification results established in a pure sampling theory approach.

3 Construction of a discrete Bayesian model

The preceding sections explain in a general way the concepts of Bayesian model and Bayesian identifiability. This generality means that the σ-fields $\mathcal{A}$ and $\mathcal{S}$ could have been either uncountable, or finite, or a mixture of both. In this section we illustrate these concepts for the particular case in which both $\mathcal{A}$ and $\mathcal{S}$ are σ fields of finite sub-sets, yielding a fully discrete Bayesian model.

3.1 Model construction

Let $S = \{s_1, \ldots, s_n\}$ be a finite set of observations and $A = \{a_1, \ldots, a_m\}$ be a finite set of parameters. The corresponding σ-fields are the the power sets of S and of A, respectively. Let μ be a prior probability defined on $(A, \mathcal{A})$ and let $A_\mu = \{a \in A : \mu(a) > 0\}$. The parametric support of an observation $s \in S$ is defined as $A_s = \{a \in A : p(s \mid a) > 0\}$. The sampling probabilities are defined as

$$\begin{aligned} p(s_i \mid a) &= \sum_{a_j \in A_\mu} p(s_i \mid a_j) \mathbb{1}_{\{a = a_j\}} \\ &= \sum_{a_j \in A_\mu \cap A_{s_i}} p(s_i \mid a_j) \mathbb{1}_{\{a = a_j\}}, \qquad i = 1, \ldots, n. \end{aligned}$$

When $a \notin A_\mu$, the sampling probabilities are arbitrary defined as $p(s \mid a) = c$ (for $s \in S$), with $c \neq 0$. Therefore, $A_\mu^c \subset A_s$ for each $s \in S$, and

$$q(s, a) = \begin{cases} p(s \mid a)\mu(a), & s \in S, \;\; a \in A_\mu \cap A_s, \\ \\ 0, & s \in S, \;\; a \in A_\mu^c \cup (A_\mu \cap A_s^c). \end{cases}$$

The predictive distribution is defined as

$$p(s) = \sum_{a_j \in A_\mu \cap A_s} p(s \mid a_j)\, \mu(a_j) = \sum_{a_j \in A} p(s_i \mid a_j) \mu(a_j), \qquad s \in S.$$

It should be remarked that for $s \in S$

$$p(s) > 0 \Longleftrightarrow A_\mu \cap A_s \neq \emptyset.$$

Therefore the joint probability $q(s, a)$ can be decomposed as

$$\begin{pmatrix} q(a_1,s_1) & \cdots & q(a_1,s_n) \\ \vdots & \ddots & \vdots \\ q(a_m,s_1) & \cdots & q(a_m,s_n) \end{pmatrix} = \begin{pmatrix} \mu(a_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mu(a_m) \end{pmatrix} \begin{pmatrix} p(s_1 \mid a_1) & \cdots & p(s_n \mid a_1) \\ \vdots & \ddots & \vdots \\ p(s_1 \mid a_m) & \cdots & p(s_n \mid a_m) \end{pmatrix} \tag{3.1}$$

$$= \begin{pmatrix} \mu(a_1 \mid s_1) & \cdots & \mu(a_1 \mid s_n) \\ \vdots & \ddots & \vdots \\ \mu(a_m \mid s_1) & \cdots & \mu(a_m \mid s_n) \end{pmatrix} \begin{pmatrix} p(s_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p(s_n) \end{pmatrix}.$$

These equalities illustrate the difference between a probability measure (the prior and the predictive) and a probability transition (the sampling and the posterior): the first one corresponds to a diagonal matrix, whereas the second one corresponds to a rectangular matrix. It can also be verified the dominance property of the posterior transition with respect to the prior distribution (Mouchart, 1976), that is, the prior null events are still posterior null events:

1. If $\mu(a_j) = 0$, then $\mu(a_j \mid s) = 0$ for all $s \in S$.
2. If $\mu(a_j) = 1$, then $\mu(a_k) = 0$ for each $k \neq j$ and, therefore, $\mu(a_k \mid s) = 0$ for each $k \neq j$ and for $s \in S$. Hence $\mu(a_j \mid s) = 1$ for each $s \in S$.

It can also be grasped that the converse relationship (namely, the dominance of the prior distribution with respect to the prior distribution) is not necessarily true. In fact, it could exists a $s \in S$ such that $\mu(a \mid s) = 0$ for $a \in A_s^c \cap A_\mu$; note that this intersection is equal to A_s^c since $A_\mu^c \subset A_s$.

3.2 Statistics and parameters in a discrete Bayesian model

In the discrete Bayesian model described in the previous section, a parameter $b(a)$ is characterized by the partition on the parameter space A induced by it, namely

$$A_b = \{a \in A \,:\, b(a) = b\}.$$

Similarly, a statistics $t(s)$ is characterized by the partition on the sample space induced by it, namely

$$S_t = \{s \in S \,:\, t(s) = t\}.$$

In this context, when using a random variable, it is important to keep in mind the partition generated by it, and not only the values taken by the random variable.

3.3 Bayesian identification in the discrete case

In this section, Bayesian identification is characterized for a fully discrete Bayesian model. More specifically, the idea is to characterize the relationship $X_1 \prec X_2 \mid X_3$, where X_1, X_2, X_3 are discrete random variables defined on a common probability space $(M, \mathcal{M}, P)$ such that $X_r : M \longrightarrow N_r$, for $r = 1, 2, 3$, where N_r ($r = 1, 2, 3$) are finite sets. These random variables can be interpreted differently. For example, X_1 can be interpreted as a parameter, X_2 as an observation and X_3 as a latent variable; or X_1 and X_3 as parameters, whereas X_2 as an observation; etc. In Section 4 different interpretations will be exploited.

Let us define

- $K = \{k \in N_3 \,:\, P(X_3 = k) > 0\}$.
- $N_1^k = \{i \in N_1 \,:\, P(X_1 = i \mid X_3 = k) > 0\}$ for $k \in K$.
- $N_2^k = \{j \in N_2 \,:\, P(X_2 = j \mid X_3 = k) > 0\}$ for $k \in K$.
- $\left(P^k\right)_{ij} = P(X_1 = i, X_2 = j \mid X_3 = k)$, a $|N_1^k| \times |N_2^k|$ matrix for $k \in K$.

By Definition 1, X_1 is Bayesian identified by X_2 conditionally on X_3 (i.e. $X_1 \prec X_2 \mid X_3$) if and only if

$$
\begin{aligned}
&\sigma\left\{P(X_2 = j, X_3 = k \mid X_1, X_3) \,:\, k \in K,\ j \in N_2^k\right\} = \\
&= \sigma\left\{\{X_1 = i\} \cap \{X_3 = k\} \,:\, k \in K,\ i \in N_1^k\right\} \ \vee \\
&\vee\ \sigma\left\{\{X_1 = i\} \cap \{X_3 = k\} \,:\, k \in N_3,\ i \in (N_1^k)^c\right\}.
\end{aligned} \tag{3.1}
$$

As a matter of fact, by definition of Bayesian identification, the σ-field of the left-side of (3.1) contains all the null sets of (X_1, X_3). This class corresponds to the σ-field

$$
\sigma\left\{\{X_1 = i\} \cap \{X_3 = k\} \,:\, k \in N_3,\ i \in (N_1^k)^c\right\}
$$

because

$$
P(X_1 = i, X_3 = k) \begin{cases} = P(X_1 = i \mid X_3 = k)\, P(X_3 = k), & i \in (N_1^k)^c,\ k \in K; \\ \leq P(X_3 = k), & k \in K^c. \end{cases}
$$

Thus, $P(X_1 = i \mid X_3 = k) = 0$ since $i \in (N_1^k)^c$ and $k \in K$, whereas $P(X_3 = k) = 0$ since $k \in K^c$. Therefore, $P(X_1 = i, X_3 = k) = 0$ for $i \in (N_1^k)^c$ and $k \in N_3$.

By this same fact –that all the null sets of (X_1, X_3) are contained into the σ-field of the left-side of (3.1)–, equality (3.1) is equivalent to the following relation:

$$\sigma\left\{\{X_1 = i\} \cap \{X_3 = k\} \,:\, k \in K,\ i \in N_1^k\right\} \subset$$
$$\subset \sigma\left\{P(X_2 = j, X_3 = k \mid X_1, X_3) \,:\, k \in K,\ j \in N_2^k\right\}. \tag{3.2}$$

Now, let us characterize the generators of $\sigma\left\{P(X_2 = j, X_3 = k \mid X_1, X_3) \,:\, k \in K,\ j \in N_2^k\right\}$. For each $k \in K$,

$$P(X_2 = j, X_3 = k \mid X_1, X_3) = \sum_{i \in N_1^k} p_{j|ik}\, \mathbb{1}_{\{X_1 = i, X_3 = k\}} \doteq Y_j.$$

Then , for each $j \in N_2^k$,

$$Y_j^{-1}\left[\{p_{j|ik}\}\right] = \{X_1 \in I_{ij},\, X_3 = k\},$$

where $I_{ij} \subset N_1^k$ is a set which depends on $(i, j) \in N_1^k \times N_2^k$. Moreover, $Y_j^{-1}\left[\{p_{j|ik}\}\right]$ contains $\{X_1 = i, X_3 = k\}$. It follows that

$$\{X_1 = i, X_3 = k\} \subset \bigcap_{j \in N_2^k} Y_j^{-1}\left[\{p_{j|ik}\}\right] = \{X_1 \in I_i, X_3 = k\},$$

where $I_i = \bigcap_{j \in N_2^k} I_{ij}$.

Thus, $\{X_1 \in I_i, X_3 = k\}$ is the smallest set in $\sigma\left\{P(X_2 = j, X_3 = k \mid X_1, X_3) \,:\, k \in K,\ j \in N_2^k\right\}$ containing $\{X_1 = i, X_3 = k\}$. Therefore, X_1 is Bayesian identified by X_2 conditionally on X_3 –which is equivalent to relation (3.2)– if and only if, for each $k \in K$,

$$I_i = \{i\} \quad \forall\, i \in N_1^k,$$

which is equivalent to, for each $k \in K$,

$$\bigcap_{j \in N_2^k} \left\{\omega \,:\, \sum_{l \in N_1^k} p_{j|lk} \mathbb{1}^{(\omega)}_{\{X_1 = l, X_3 = k\}} = p_{j|ik}\right\} = \{X_1 = i,\ X_3 = k\} \quad \forall\, i \in N_1^k.$$

This last condition is equivalent to the following one: for each $k \in K$,

$$\nexists\, i, i' \in N_1^k \quad \text{such that} \quad p_{j|ik} = p_{j|i'k} \quad \forall\, j \in N_2^k,$$

which in turn can equivalently be written as

$$\nexists\, i, i' \in N_1^k \quad \text{such that} \quad p_{ij|k} = c_{ii'}\, p_{i'j|k} \quad \forall\, j \in N_2^k.$$

Summarizing, we obtain the following theorem:

Theorem 1 *Let $(\Omega, \mathcal{M}, P)$ be a probability space and $X_r : \Omega \longrightarrow N_r$, with $r = 1, 2, 3$, be random variables, where N_r ($r = 1, 2, 3$) are finite sets. The following are equivalent:*

1. *$X_1 \prec X_2 \mid X_3$, that is, X_1 is Bayesian identified by X_2 conditionally on X_3.*
2. *For each $k \in K$, any two rows of P^k are linearly independent.*
3. *For each $k \in K$ and for each $i, i' \in N_1^k$, there not exists a $c_{ii'}$ such that*

$$P[X_1 = i, X_2 = j \mid X_3 = k] = c_{ii'} P[X_1 = i', X_2 = j \mid X_3 = k] \quad \forall\, j \in N_2^k.$$

3.4 Relationships between Bayesian and sampling identification

As pointed out in comment 4, Section 2.5, sampling identification implies Bayesian identification for all prior distribution provided the parameter space is Blackwell and the sample space is separable. This implication can be verified in the fully discrete Bayesian model. As a matter of fact, consider the fully discrete Bayesian model as specified in Section 3.1. By Theorem 1, the parameter a is Bayesian identified by s if and only if any two rows of the matrix

$$\begin{pmatrix} q(a_1, s_1) & \cdots & q(a_1, s_n) \\ \vdots & \ddots & \vdots \\ q(a_m, s_1) & \cdots & q(a_m, s_n) \end{pmatrix}$$

are linearly independent. Considering decomposition (3.1), this is equivalent to the following property: any two rows of the matrix representing the sampling transition are linearly independent; that is, the mapping $a \longmapsto p(\cdot \mid a)$ is injective, for all $a \in A_\mu \subset A$. Now, if $A_\mu = A$, then Bayesian identification and sampling identification are equivalent concepts. The property $A_\mu = A$ means that the prior distribution μ put positive mass on each element of A. Let us summarize this relationship in the following theorem:

Theorem 2 *Consider the fully discrete Bayesian model as specified in Section 3.1. If $A_\mu = A$ then Bayesian identification and sampling identification are equivalent.*

3.5 Updating unidentified parameters

As pointed out in comment 1, Section 2.5, the posterior distribution of an unidentified parameter can be computed. However, from a modeling point of view, the statistical meaning of this posterior distribution is of interest. Let us consider, therefore, a fully discrete Bayesian model defined by the following components: $S = \{s_1, s_2\}$, $A = \{a_1, a_2, a_3\}$, the sampling process characterized by

$$p(s_1 \mid a_1) = \frac{1}{2}, \quad p(s_1 \mid a_2) = p(s_1 \mid a_3) = \frac{1}{3},$$

and the prior distribution satisfying that $\mu(a_i) > 0$ for $i = 1, 2, 3$. All these components imply the following joint distribution of (a, s), with $a \in A$ and $s \in S$:

	s_1	s_2
a_1	$\frac{1}{2}\,\mu(a_1)$	$\frac{1}{2}\,\mu(a_1)$
a_2	$\frac{1}{3}\,\mu(a_2)$	$\frac{2}{3}\,\mu(a_2)$
a_3	$\frac{1}{3}\,\mu(a_3)$	$\frac{2}{3}\,\mu(a_3)$.

By Theorem 1, the Bayesian identified parameter is characterized by the partition

$$\{\{a_1\},\ \{a_2, a_3\}\}.$$

The posterior distribution of a_1 is given by

$$p(a_1 \mid s) = \frac{\frac{1}{2}\mu(a_1)}{\frac{1}{2}\mu(a_1) + \frac{1}{3}[\mu(a_2) + \mu(a_3)]}\mathbb{1}_{\{s=s_1\}} + \frac{\frac{1}{2}\mu(a_1)}{\frac{1}{2}\mu(a_1) + \frac{2}{3}[\mu(a_2) + \mu(a_3)]}\mathbb{1}_{\{s=s_2\}};$$

the posterior distribution of (a_1, a_2) is given by

$$p(a_2, a_3 \mid s) = \frac{\frac{1}{3}[\mu(a_2) + \mu(a_3)]}{\frac{1}{2}\mu(a_1) + \frac{1}{3}[\mu(a_2) + \mu(a_3)]}\mathbb{1}_{\{s=s_1\}} + \frac{\frac{2}{3}[\mu(a_2) + \mu(a_3)]}{\frac{1}{2}\mu(a_1) + \frac{1}{3}[\mu(a_2) + \mu(a_3)]}\mathbb{1}_{\{s=s_2\}}.$$

The posterior distribution of a_2 and of a_3 –which are unidentified parameters– are respectively given by

$$p(a_2 \mid s) = \frac{\frac{1}{3}\mu(a_2)}{\frac{1}{2}\mu(a_1) + \frac{1}{3}[\mu(a_2) + \mu(a_3)]}\mathbb{1}_{\{s=s_1\}} + \frac{\frac{2}{3}\mu(a_2)}{\frac{1}{2}\mu(a_1) + \frac{1}{3}[\mu(a_2) + \mu(a_3)]}\mathbb{1}_{\{s=s_2\}},$$

$$p(a_3 \mid s) = \frac{\frac{1}{3}\mu(a_3)}{\frac{1}{2}\mu(a_1) + \frac{1}{3}[\mu(a_2) + \mu(a_3)]}\mathbb{1}_{\{s=s_1\}} + \frac{\frac{2}{3}\mu(a_3)}{\frac{1}{2}\mu(a_1) + \frac{1}{3}[\mu(a_2) + \mu(a_3)]}\mathbb{1}_{\{s=s_2\}}.$$

It follows that

$$p(a_2 \mid s) = \frac{\mu(a_2)}{\mu(a_2)+\mu(a_3)}\, p(a_2, a_3 \mid s), \qquad p(a_3 \mid s) = \frac{\mu(a_3)}{\mu(a_2)+\mu(a_3)}\, p(a_2, a_3 \mid s).$$

That is, the posterior distribution of the unidentified parameter a_2 coincides with a function of the posterior distribution of the identified parameter (a_2, a_3). Similarly, for the posterior distribution of a_3. Therefore, it is not worth calculating the posterior probability of these unidentified parameters, but rather replace the original Bayesian model with the following:

	s_1	s_2
a_1	$\frac{1}{2}\,\mu(a_1)$	$\frac{1}{2}\,\mu(a_1)$
a_2, a_3	$\frac{1}{3}\,[\mu(a_2)+\mu(a_3)]$	$\frac{2}{3}\,[\mu(a_2)+\mu(a_3)].$

In this case, the σ-field $\mathcal{A}$ of the parameter space is given by

$$\{\emptyset, \{a_1\},\ \{a_2, a_3\},\ A\}.$$

It can be noted that the events $\{a_2\}$ and $\{a_3\}$ are *not* measurable, that is, *they are not events of interest*.

4 Identifiability relationships in a hierarchical structure

4.1 Motivation

A wide class of models typically used in psychometrics, biometrics and sociometrics, shares a common hierarchical structure composed of two submodels: (i) the *latent marginal model* $p(\eta \mid \theta_2)$ generating the latent variable η; and the *conditional model* or *measurement model* $p(Y \mid \eta, \theta_1)$ generating the observable variable Y given the latent variable η. This framework steams from the fact that, in these applied fields, the research interest is not just a population in the sense of a distribution of observable variables (Fischer, 1922), but a structure projected behind this distribution, by which the latter is thought to be generated; see Koopmans and Reiersøl (1950) and Manski (1995). Such a structure is not directly observable; consequently, it is specified through the latent marginal model. The conditional model corresponds to a measurement model: the observable variable Y measures, with error, the latent one η. The hierarchical specification is intended to explain an observed phenomenon described by a sampling distribution (or statistical model, or likelihood function) $p(Y \mid \omega)$. The relationship between the statistical model and the submodels of the hierarchy is given by

$$p(Y \mid \omega) = \int p(Y \mid \eta, \theta_2) p(\eta \mid \theta_1) d\eta, \tag{4.1}$$

where it is implicitly assumed that θ_1 and θ_2 are sufficient parameters of the conditional and latent marginal models, respectively; that is,

$$\text{(i) } \eta \perp\!\!\!\perp \theta \mid \theta_2, \qquad \text{(ii) } Y \perp\!\!\!\perp \theta \mid \eta, \theta_1, \tag{4.2}$$

with $\theta = (\theta_1, \theta_2)$. Two relevant examples of this hierarchical structure are the following:

Example 3 IRT-models: the basic idea of this class of models is that the probability of correctly answering an item depends on two factors: one characterizing the person, let η_i; the other one characterizing the item or task, let β_j; see Rasch (1960, 1966). More precisely, let Y_{ij} be a binary random variable such that $Y_{ij} = 1$ if person i correctly answers item j, and $Y_{ij} = 0$ otherwise. The probability of the event $\{Y_{ij} = 1\}$ is specified as

$$(Y_{ij} \mid \eta_i, \beta_j) \sim \text{Bern}\,[F(\eta_i - \beta_j)], \quad i = 1, \dots, n, \;\; j = 1, \dots, m, \tag{4.3}$$

where F is a known probability distribution, typically the logistic distribution or the normal distribution. Because of the incidental problem (see Neyman and Scott, 1948; Andersen, 1980; Ghosh, 1995; Lancaster, 2000), the individual characteristic η_i's are viewed as latent variables which satisfy the following condition:

$$(\eta_i \mid \varphi) \sim G^{\varphi}, \qquad \mathop{\perp\!\!\!\perp}_{1 \leq i \leq n} \eta_i \mid \varphi; \tag{4.4}$$

that is, $\{\eta_i \,:\, i = 1, \dots, n\}$ are mutually independent conditionally on φ, with a common distribution G^{φ} known up to a parameter φ. Typically, it is assumed that the distribution G^{φ} is a $\mathcal{N}(0, \varphi^2)$. The model is completed by assuming that, for each person i, the answers $\{Y_{ij} \,:\, j = 1, \dots, m\}$ are mutually independent conditionally on $(\eta_i, \beta_1, \dots, \beta_m)$; and that $\{\boldsymbol{Y}_i \,:\, i = 1, \dots, n\}$ are mutually independent conditionally on $(\eta_1, \dots, \eta_n, \beta_1, \dots, \beta_m)$, where $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{im})'$. For details, see Fischer and Molenaar (1995); van der Linden and Hambleton (1997); De Boeck and Wilson (2004).

The conditional model corresponds, therefore, to the distribution of $(\boldsymbol{Y}_i \mid \eta_i, \theta_1)$, where the sufficient parameter θ_1 is $(\beta_1, \dots, \beta_m)$; the latent marginal model corresponds to the distribution of $(\eta_i \mid \theta_2)$, where the sufficient parameter θ_2 is φ. The statistical model is obtained after integrating out the latent variables η_i's.

□

Example 4 Structural Equation Models: this type of models combine concepts of latent variables with the techniques of path analysis and simultaneous equations models, and represent the convergence of relatively independent research traditions in psychometrics, econometrics and biometrics. Traditionally, these models have been specified as a set of two submodels, namely a structural model and a measurement model; the first one specifies the relationship between the latent variables, and the second one specifies how the latent variables are related to the observed or measured variables. More specifically, for a sample of size n, the structural model may be written as

$$B\xi_i \,+\, C\zeta_i \,=\, \epsilon_i, \qquad i = 1, \dots, n, \tag{4.5}$$

where $\eta_i = ({\xi_i}', {\zeta_i}')' \in \mathbb{R}^p \times \mathbb{R}^l$, $i = 1, \dots, n$, are latent variables, B is a $p \times p$ matrix, C is a $p \times l$ matrix, and $\epsilon_i \in \mathbb{R}^p$, $i = 1, \dots, n$, are random vectors of residuals. It is assumed that the distribution of ϵ_i is known up to the variance-covariance matrix $\boldsymbol{\Sigma}_{\epsilon\epsilon}$; similarly, the distribution of ζ_i is known up to the variance-covariance matrix $\boldsymbol{\Sigma}_{\zeta\zeta}$. The measurement model is given by

$$x_i \,=\, \Lambda_x \xi_i \,+\, \epsilon_{x_i}, \qquad z_i \,=\, \Lambda_z \zeta_i \,+\, \epsilon_{z_i}, \qquad i = 1, \dots, n, \tag{4.6}$$

where $y_i = ({x_i}', {z_i}')' \in \mathbb{R}^r \times \mathbb{R}^s$, $i = 1, \dots, n$, are the observable (or manifest) variables, and ϵ_{x_i} and ϵ_{z_i}, $i = 1, \dots, n$, are vectors of errors of measurement in x_i and z_i, respectively. It is also assumed that, conditionally on the latent variables (ξ_i, ζ_i), the measurement errors ϵ_{x_i} and ϵ_{z_i} are mutually independent. Furthermore, the conditional distribution of ϵ_{x_i} given ξ_i, and the conditional distribution of ϵ_{z_i} given ζ_i, are known up to the variance-covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_x \epsilon_x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon_z \epsilon_z}}$, respectively. For details see, among others, Jöreskog (1981); Everitt (1984); Bollen (1989); Maruyama (1998).

The latent marginal model is, therefore, represented by the structural model; the corresponding sufficient parameter θ_2 is given by $\theta_2 = (B, C, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon\epsilon}}, \boldsymbol{\Sigma}_{\boldsymbol{\zeta\zeta}})$. The conditional model is represented by the measurement model; the corresponding sufficient parameter θ_1 is given by $\theta_1 = (\Lambda_x, \Lambda_y, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon_x \epsilon_x}}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon_z \epsilon_z}})$. The statistical model is obtained after integrating out the latent variables (ξ_i, ζ_i). For more details, see also San Martín (2003).

□

By construction, the parameter ω of the statistical model $p(Y \mid \omega)$ is a function of (θ_1, θ_2). Furthermore, ω can be viewed as the minimal sufficient parameter which characterizes the sampling process. In this sense, ω has a statistical meaning as far as it is a functional of the sampling process. However, from a modeling point of view, the problem is to know whether θ_1 and θ_2 have a statistical meaning, that is, whether (θ_1, θ_2) fully describes the sampling process. If it is the case, we say that the hierarchical specification has an empirical sense; otherwise, the explanation provided by the hierarchical specification is not unique.

These considerations lead seeking for identification relationships between the submodels composing a hierarchy. Thus, for instance, it has been proposed in the psychometric literature (without formal proof) that the identification of θ_1 in the conditional model $p(Y \mid \eta, \theta_1)$ and the identification of θ_2 in the marginal model $p(\eta \mid \theta_2)$ jointly imply the identifiability of (θ_1, θ_2) in the statistical model $p(Y \mid \theta_1, \theta_2)$; see, in the context of Structural Equation Models, Jöreskog (1981, pp. 89-90), Bollen (1989, p. 328) and Maruyama (1998, p. 191); and, in the context of IRT-models, see Adams et al. (1997) and Smits and Moore (2004).

It seems relevant, therefore, to explore possible identification relationships between the submodels composing a hierarchy, namely the latent marginal model, the conditional model and the statistical model. This section begins by reviewing a true identification relationship, which provides a necessary identification condition that can be used in *all* hierarchical models. Thereafter, counter-examples are provided to show that intuitively true identification relationships (as those above-mentioned) are false. These counter-examples are developed in a fully discrete Bayesian model; by so doing, it is possible to understand why such intuitions fail. The section ends by reviewing a second identification relationship, which provides a sufficient identification restriction; this one can be used in *some* hierarchical models.

4.2 A necessary identification relationship

Let us consider a general hierarchy defined by (4.2). The interest is to identify (θ_1, θ_2) by the observations in the statistical model $p(Y \mid \theta_1, \theta_2)$. The following theorem provided a *necessary* identification relationship:

Theorem 3 *The Bayesian identification of θ_2 by η in the latent marginal model is a necessary condition for the Bayesian identification of (θ_1, θ_2) by Y in the statistical model provided that*

$$\theta_1 \perp\!\!\!\perp \theta_2. \tag{4.7}$$

For a proof, see San Martín (2000) and Jara et al. (2008).

Condition (4.7) defines a *cut* between the latent marginal model and the conditional model, that is, the process generating $(Y, \eta, \theta_1, \theta_2)$ is decomposed into two factors:

$$p(Y, \eta, \theta_1, \theta_2) = p(Y \mid \eta, \theta_1)\pi(\theta_1) \times p(\eta \mid \theta_2)\pi(\theta_2).$$

This is a condition on the prior distribution of (θ_1, θ_2). In a pure sampling theory approach, condition (4.7) may be replaced by a condition of factorization of the parameter space (or a condition of *variation-free*), namely $(\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$, where Θ_i ($i = 1, 2$) are the parameter spaces; for details, see Barndorff-Nielsen (1978).

From a modeling point of view, Theorem 3 shows why it is necessary to specify the latent marginal model in such a way that θ_2 is Bayesian identified by the latent variable; otherwise, the parameter of interest (θ_1, θ_2) is not longer identified by the observations. In practice, when the latent marginal model is simple in the sense that θ_2 is Bayesian identified by η in a straightforward way (as, for instance, in an IRT model; see Example 3), this theorem can be "automatically applied". However, when the latent marginal model is complex in the sense of being over-parameterized (as, for instance, in a Structural Equation Model; see Example 4), this theorem requires identifying θ_2 by the latent variable η. For an example of this case, see Jara et al. (2008).

4.3 Sufficient identification relationships

As conjectured by the psychometric literature, the following relationship seems to be intuitively true: if θ_2 is Bayesian identified by η in the latent marginal model and if θ_1 is Bayesian identified by Y conditionally on η in the conditional model, then (θ_1, θ_2) is Bayesian identified by Y in the statistical model. It is, however, possible to offer counter-examples. In order to understand why the intuition fails, we provide one in a fully Bayesian discrete model.

4.3.1 Counter-example 1

Let $(Y, \eta, \theta_1, \theta_2) \in \{0, 1\}^4$ be four discrete random variables defined on a common probability space. Without additional restrictions, the corresponding Bayesian model is characterized by $2^4 - 1 = 15$ parameters, namely

$$\omega_{ijkl} \doteq P[Y = i, \eta = j, \theta_1 = k, \theta_2 = l].$$

Under the hierarchical structure (4.2), it follows that

$$\begin{aligned}\omega_{ijkl} &= P[Y=i \mid \eta=j,\theta_1=k,\theta_2=l]\, P[\eta=j \mid \theta_1=k,\theta_2=l]\, P[\theta_1=k,\theta_2=l] \\ &= P[Y=i \mid \eta=j,\theta_1=k]\, P[\eta=j \mid \theta_2=l]\, P[\theta_1=k,\theta_2=l] \\ &\doteq p_{i|jk}\cdot q_{j|l}\cdot o_{kl}.\end{aligned}$$

Therefore, the Bayesian model is characterized by 9 parameters, namely $p_{0|00},\ p_{0|01},\ p_{0|10},\ p_{0|11};\ q_{0|0},\ q_{0|1};\ o_{00},\ o_{01},\ o_{10}$. We assume that $o_{kl}>0$ for all k,l.

Using Theorem 1, we characterize (1) the Bayesian identification of θ_2 by η in the latent marginal model; (2) the Bayesian identification of θ_1 by Y conditionally on η in the conditional model; and (3) the Bayesian identification of (θ_1,θ_2) by Y in the statistical model. Thereafter, we study whether (1) and (2) jointly imply (3).

1. Bayesian identification of θ_2 by η. The relationship $\theta_2 \prec \eta$ requires to analyze the matrix whose entries are

$$s_{jl} \doteq P[\eta=j,\theta_2=l] = P[\eta=j \mid \theta_2=l]\, P[\theta_2=l] \doteq q_{j|l}\cdot n_l;$$

that is,

$$S=\begin{pmatrix} s_{00} & s_{01} \\ s_{10} & s_{11}\end{pmatrix}=\begin{pmatrix} q_{0|0} & q_{0|1} \\ q_{1|0} & q_{1|1}\end{pmatrix}\begin{pmatrix} n_0 & \\ & n_1\end{pmatrix}\doteq Q\ \cdot\ \mathrm{diag}\,(n_0,n_1),$$

where $n_l=o_{0l}+o_{1l}$ with $l=0,1$. Therefore, $\theta_2\prec\eta$ if and only if $(s_{00},s_{01})'$ and $(s_{10},s_{11})'$ are linearly independent. Since S is a 2×2 matrix, this last condition is equivalent to $r(S)=2$, which in turn is equivalent to $r(Q)=2$ because $o_{kl}>0$ for all k,l. Therefore, $\theta_2\prec\eta$ if and only if

A1. $r(Q)=2$.

□

2. Bayesian identification of θ_1 by Y conditionally on η. The relationship $\theta_1\prec Y\mid\eta$ requires to analyze the matrices whose entries are

$$t_{ik|0}\doteq P[Y=i,\theta_1=k\mid\eta=0],\qquad t_{ik|1}\doteq P[Y=i,\theta_1=k\mid\eta=1].$$

But

$$P[Y=i,\theta_1=k\mid\eta=j]=P[Y=i\mid\eta=j,\theta_1=k]\,P[\theta_1=k\mid\eta=j]\doteq p_{i|jk}\cdot\nu_{k|j}.$$

Therefore

$$\begin{aligned}T^{(0)}=\begin{pmatrix} t_{00|0} & t_{01|0} \\ t_{10|0} & t_{11|0}\end{pmatrix} &= \begin{pmatrix} p_{0|00}\nu_{0|0} & p_{0|01}\nu_{1|0} \\ p_{1|00}\nu_{0|0} & p_{1|01}\nu_{1|0}\end{pmatrix} \\ &= \begin{pmatrix} p_{0|00} & p_{0|01} \\ p_{1|00} & p_{1|01}\end{pmatrix}\begin{pmatrix} \nu_{0|0} & \\ & \nu_{1|0}\end{pmatrix},\end{aligned}$$

$$T^{(1)} = \begin{pmatrix} t_{00|1} & t_{01|1} \\ t_{10|1} & t_{11|1} \end{pmatrix} = \begin{pmatrix} p_{0|10} & p_{0|11} \\ p_{1|10} & p_{1|11} \end{pmatrix} \begin{pmatrix} \nu_{0|1} & \\ & \nu_{1|1} \end{pmatrix}.$$

Furthermore,

$$\begin{aligned} \nu_{k|j} &= \frac{P[\eta = j \mid \theta_2 = 0] P[\theta_1 = k, \theta_2 = 0] + P[\eta = j \mid \theta_2 = 1] P[\theta_1 = k, \theta_2 = 1]}{P[\eta = j \mid \theta_2 = 0] P[\theta_2 = 0] + P[\eta = j \mid \theta_2 = 1] P[\theta_2 = 1]} \\ &= \frac{q_{j|0}\, o_{k0} + q_{j|1}\, o_{k1}}{q_{j|0}\, n_0 + q_{j|1}\, n_1}. \end{aligned}$$

It follows that

$$T^{(0)} = \frac{1}{q_{0|0}\, n_0 + q_{0|1}\, n_1} \begin{pmatrix} p_{0|00} & p_{0|01} \\ p_{1|00} & p_{1|01} \end{pmatrix} \begin{pmatrix} q_{0|0} & q_{0|1} & 0 & 0 \\ 0 & 0 & q_{0|0} & q_{0|1} \end{pmatrix} \begin{pmatrix} o_{00} & 0 \\ o_{01} & 0 \\ 0 & o_{10} \\ 0 & o_{11} \end{pmatrix}.$$

$$T^{(1)} = \frac{1}{q_{1|0}\, n_0 + q_{1|1}\, n_1} \begin{pmatrix} p_{0|10} & p_{0|11} \\ p_{1|10} & p_{1|11} \end{pmatrix} \begin{pmatrix} q_{1|0} & q_{1|1} & 0 & 0 \\ 0 & 0 & q_{1|0} & q_{1|1} \end{pmatrix} \begin{pmatrix} o_{00} & 0 \\ o_{01} & 0 \\ 0 & o_{10} \\ 0 & o_{11} \end{pmatrix}.$$

Therefore, $\theta_1 \prec Y \mid \eta$ if and only if $r(T^{(0)}) = 2$ and $r(T^{(1)}) = 2$ (since both matrices are 2×2). Taking into account that $o_{kl} > 0$ for all k and l, these conditions are equivalent to the following:

B1. $q_{0|0} \cdot q_{0|1} \cdot q_{1|0} \cdot q_{1|1} > 0.$

B2. $r \begin{pmatrix} p_{0|00} & p_{0|01} \\ p_{1|00} & p_{1|01} \end{pmatrix} = 2$ and $r \begin{pmatrix} p_{0|10} & p_{0|11} \\ p_{1|10} & p_{1|11} \end{pmatrix} = 2.$

□

3. Bayesian identification of (θ_1, θ_2) by Y. The relationship $(\theta_1, \theta_2) \prec Y$ requires to analyze the matrix whose entries are the following:

$$\begin{aligned} f_{ikl} &\doteq P[Y = i, \theta_1 = k, \theta_2 = l] \\ &= P[Y = i \mid \eta = 0, \theta_1 = k]\, P[\eta = 0 \mid \theta_2 = l]\, P[\theta_1 = k, \theta_2 = l] + \\ &\quad + P[Y = i \mid \eta = 1, \theta_1 = k]\, P[\eta = 1 \mid \theta_2 = l]\, P[\theta_1 = k, \theta_2 = l] \\ &= p_{i|0k} \cdot q_{0|l} \cdot o_{kl} + p_{i|1k} \cdot q_{1|l} \cdot o_{kl}. \end{aligned}$$

It follows that

$$F = \begin{pmatrix} f_{000} & f_{001} \\ f_{010} & f_{011} \\ f_{100} & f_{101} \\ f_{110} & f_{111} \end{pmatrix} = \begin{pmatrix} (p_{0|00}q_{0|0} + p_{0|10}q_{1|0})\, o_{00} & (p_{1|00}q_{0|0} + p_{1|10}q_{1|0})\, o_{00} \\ (p_{0|00}q_{0|1} + p_{0|10}q_{1|1})\, o_{01} & (p_{1|00}q_{0|1} + p_{1|10}q_{1|1})\, o_{01} \\ (p_{0|01}q_{0|0} + p_{0|11}q_{1|0})\, o_{10} & (p_{1|01}q_{0|0} + p_{1|11}q_{1|0})\, o_{10} \\ (p_{0|01}q_{0|1} + p_{0|11}q_{1|1})\, o_{11} & (p_{1|01}q_{0|1} + p_{1|11}q_{1|1})\, o_{11} \end{pmatrix}$$

$$= \Delta \begin{pmatrix} Q' & \\ & Q' \end{pmatrix} \begin{pmatrix} p_{0|00} & p_{1|00} \\ p_{0|10} & p_{1|10} \\ p_{0|01} & p_{1|01} \\ p_{0|11} & p_{1|11} \end{pmatrix} \doteq \Delta \begin{pmatrix} Q' & \\ & Q' \end{pmatrix} P,$$

where $\Delta = \text{diag}\,(o_{00}, o_{01}, o_{10}, o_{11})$ and Q as defined in step **1**. Taking into account that $o_{kl} > 0$ for all k and l, $(\theta_1, \theta_2) \prec Y$ if and only if

C1. $r(Q) = 2$.

C2. All the two pairs of rows of matrix P are linearly independent.

□

Is it true that conditions A1, B1 and B2 jointly imply conditions C1 and C2? The answer is *not*. As a matter of fact, matrix P is composed of the entries corresponding to the two matrices of condition **B2**. The entries corresponding to the first rank condition of **B2** are in boldface in P, namely

$$\begin{pmatrix} \boldsymbol{p_{0|00}} & \boldsymbol{p_{1|00}} \\ p_{0|10} & p_{1|10} \\ \boldsymbol{p_{0|01}} & \boldsymbol{p_{1|01}} \\ p_{0|11} & p_{1|11} \end{pmatrix}.$$

It can be verified that two rows of F are linearly dependent, although **B2** is verified. Take, for instance, $p_{0|00} = p_{0|10}$; in this case, the Bayesian identified parameter in the statistical model would be characterized by the partition

$$\{\{(0,0),(1,0)\},\ \ \{(0,1),(1,1)\}\}.$$

If (θ_1, θ_2) were the Bayesian identified parameter in the statistical model, then it were characterized by the partition

$$\{\{(0,0)\},\ \ \{(0,1)\},\ \ \{(1,0)\},\ \ \{(1,1)\}\}.$$

4.3.2 Counter-example 2

The previous counter-example shows that both the identification of the latent marginal model and the identification of the conditional model are not enough to implying the identification of the statistical model. However, it would be possible to obtain the identification of the statistical model if the identification relationship $\theta_1 \prec Y \mid \eta$ is replaced by a stronger identification relationship, namely $(\theta_1, \eta) \prec Y$. That is, to study whether $\theta_2 \prec \eta$ and $(\theta_1, \eta) \prec Y$ jointly imply $(\theta_1, \theta_2) \prec Y$.

Let $(Y, \eta, \theta_1, \theta_2) \in \{1, 2, 3, 4, 5\} \times \{1, 2\} \times \{1, 2, 3\} \times \{1, 2\}$. Using Theorem 1, we begin characterizing the Bayesian identification of (θ_1, η) by Y.

2'. Bayesian identification of (θ_1, η) by Y. The relationship $(\theta_1, \eta) \prec Y$ requires to analyze the matrix whose entries are

$$c_{jk\,i} \doteq P[Y = i, \theta_1 = k, \eta = j] = P[Y = i \mid \eta = j, \theta_1 = k]\, P[\theta_1 = k, \eta = j] \doteq p_{i|jk}\, r_{jk},$$

where

$$r_{jk} = q_{j|1} o_{k1} + q_{j|2} o_{k2}.$$

The matrix C is, therefore, given by

$$C = \begin{pmatrix} r_{11}\boldsymbol{p_{1|11}} & r_{11}\boldsymbol{p_{2|11}} & r_{11}\boldsymbol{p_{3|11}} & r_{11}\boldsymbol{p_{4|11}} & r_{11}\boldsymbol{p_{5|11}} \\ r_{12}\boldsymbol{p_{1|12}} & r_{12}\boldsymbol{p_{2|12}} & r_{12}\boldsymbol{p_{3|12}} & r_{12}\boldsymbol{p_{4|12}} & r_{12}\boldsymbol{p_{5|12}} \\ r_{13}\boldsymbol{p_{1|13}} & r_{13}\boldsymbol{p_{2|13}} & r_{13}\boldsymbol{p_{3|13}} & r_{13}\boldsymbol{p_{4|13}} & r_{13}\boldsymbol{p_{5|13}} \\ r_{21}p_{1|21} & r_{21}p_{2|21} & r_{21}p_{3|21} & r_{21}p_{4|21} & r_{21}p_{5|21} \\ r_{22}p_{1|22} & r_{22}p_{2|22} & r_{22}p_{3|22} & r_{22}p_{4|22} & r_{22}p_{5|22} \\ r_{23}p_{1|23} & r_{23}p_{2|23} & r_{23}p_{3|23} & r_{23}p_{4|23} & r_{23}p_{5|23} \end{pmatrix}$$

□

3'. Bayesian identification of (θ_1, θ_2) by Y. Similarly to step **3** in Section 4.3.1, $(\theta_1, \theta_2) \prec Y$ requires to analyze the rows of matrix

$$F = \Delta \cdot \text{diag} \cdot (Q, Q, Q) \begin{pmatrix} \boldsymbol{p_{1|11}} & \boldsymbol{p_{2|11}} & \boldsymbol{p_{3|11}} & \boldsymbol{p_{4|11}} & \boldsymbol{p_{5|11}} \\ p_{1|21} & p_{2|21} & p_{3|21} & p_{4|21} & p_{5|21} \\ \boldsymbol{p_{1|12}} & \boldsymbol{p_{2|12}} & \boldsymbol{p_{3|12}} & \boldsymbol{p_{4|12}} & \boldsymbol{p_{5|12}} \\ p_{1|22} & p_{2|22} & p_{3|22} & p_{4|22} & p_{5|22} \\ \boldsymbol{p_{1|13}} & \boldsymbol{p_{2|13}} & \boldsymbol{p_{3|13}} & \boldsymbol{p_{4|13}} & \boldsymbol{p_{5|13}} \\ p_{1|23} & p_{2|23} & p_{3|23} & p_{4|23} & p_{5|23} \end{pmatrix} \doteq \Delta \cdot \text{diag} \cdot (Q, Q, Q) \cdot P,$$

where $\Delta = \text{diag}\,(o_{11}, o_{12}, o_{21}, o_{22}, o_{31}, o_{32})$ and

$$Q = \begin{pmatrix} q_{1|1} & q_{2|1} \\ q_{1|2} & q_{2|2} \end{pmatrix}.$$

□

In boldface, we distinguished how the elements of matrix C –which are related with the process generating $(Y \mid \eta, \theta_1)$– are distributed in matrix F -which is related with the statistical model. Now, θ_2 is Bayesian identified by η if $r(Q) = 2$. Furthermore, we assume that $o_{kl} > 0$ for all (k, l). Therefore, $r_{jk} > 0$ for each pair (j, k). Let us suppose that

$$p_{i|11} = p_{i|12} = \frac{1}{5} \quad \forall\, i = 1, 2, 3, 4, 5,$$

and that the matrices Q and Δ are chosen such that $r_{11} \neq r_{21}$. These conditions imply that the first two columns of the matrix C are linearly independent. If we assume furthermore that any two rows of C are linearly independent, then $(\theta_1, \eta) \prec Y$, but (θ_1, θ_2) is not Bayesian identified by Y because the first and third row of P (and therefore of F) are linearly dependent.

Summarizing, the Bayesian identification of θ_2 by η in the marginal model, and the Bayesian identification of (θ_1, η) by Y in the conditional model, are *not* enough to ensure the Bayesian identification of (θ_1, θ_2) by Y in the statistical model.

4.4 Solutions to the given counter-examples

However, matrix C can be written as $C = R \cdot \Pi$, where

$$R = \text{diag}\,(r_{11}, \cdots, r_{1K}, \cdots, r_{J1}, \ldots, r_{JK}),$$

$$\Pi = \left(\begin{array}{ccc} p_{1|11} & \cdots & p_{I|11} \\ \vdots & \ddots & \vdots \\ p_{1|1K} & \cdots & p_{I|1K} \\ \hline p_{1|21} & \cdots & p_{I|21} \\ \vdots & \ddots & \vdots \\ p_{1|2K} & \cdots & p_{I|2K} \\ \hline \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ \hline p_{1|J1} & \cdots & p_{I|J1} \\ \vdots & \ddots & \vdots \\ p_{1|JK} & \cdots & p_{I|JK} \end{array}\right),$$

where $Y \in \{1, \ldots, I\}$, $\eta \in \{1, \ldots, J\}$, $\theta_1 \in \{1, \ldots, K\}$ and $\theta_2 \in \{1, \ldots, L\}$. Note that Π is a $JK \times I$ matrix.

The Bayesian identifiability of (θ_1, η) by Y can be ensured by a stronger condition than that established in Theorem 1, namely that $r(\Pi) = JK$. As a matter of fact, this last condition implies that all the two rows of Π are linearly independent, whereas the converse is in general false. Taking into account that $r_{jk} > 0$ for each (j, k), it follows that $\theta_2 \prec \eta$ and $r(\Pi) = KL$ jointly imply the Bayesian identification of (θ_1, θ_2) by Y in the statistical model. It should be noted that the rank condition $r(\Pi) = KL$ imposes a restriction, namely $KL \leq I$.

4.4.1 Bayesian completeness on the parameter space

What statistical concept can be associated to this rank condition? A rank condition can be understood as a condition ensuring the injectivity of a linear transformation. In a sampling theory framework, completeness is precisely defined through the injectivity of a linear operator –namely, a sampling expectation. In a Bayesian approach, completeness is defined as follows:

Definition 2 *Let $(M, \mathcal{M}, P)$ be a probability space and, for $i = 1, 2, 3$, let X_i be a random variable defined from $(M, \mathcal{M})$ to a measurable space $(S_i, \mathcal{S}_i)$. Let $q \in [1, \infty]$. X_1 is said to be q-complete w.r.t. X_2, denoted as $X_1 \ll_q X_2$, if for all Borel function $h : (S_1, \mathcal{S}_1) \longrightarrow (\mathbb{R}, \mathcal{B})$ such that $E(|h(X_1)|^q) < \infty$ (when $q = \infty$, it is taken the essential supremum), the following implication follows:*

$$E[h(X_1) \mid X_2] = 0 \;\; P\text{-a.s.} \Longrightarrow h(X_1) = 0 \;\; P\text{-a.s.} \tag{4.8}$$

Moreover, conditionally on X_3, X_1 is q-complete w.r.t X_2, denoted as $X_1 \ll_q X_2 \mid X_3$, if and only if $(X_1, X_3) \ll_q (X_2, X_3)$.
When this relationship is valid for all $q \in [1, \infty]$, q is omitted.

Since $E[\cdot \mid X_2]$ is a linear operator on the q-integrable functions, expression (4.8) means that the null space of $E[\cdot \mid X_2]$ reduces to $\{0\}$. This is equivalent to the injectivity of the conditional expectation $E[\cdot \mid X_2]$ (see Conway, 1985, pp. 376). In other words, the q-completeness of X_1 w.r.t. X_2 means that the expectation conditional on X_2 is an injective operator defined on $L^q(\Omega, \mathcal{M}_1, P)$. Thus, if X_1 plays the role of a "statistic" and X_2 of a "parameter", then $X_1 \ll_q X_2$ corresponds to the Bayesian counterpart of the classical definition of a complete statistic and can be viewed as the injectivity of the sampling expectation on the q-integrable functions of the statistic. Similarly, if X_1 plays the role of a "parameter" and X_2 of a "statistic", $X_1 \ll_q X_2$ corresponds to the injectivity of the posterior expectation on the q-integrable functions of the parameter. We refer to Chapter 5 in Florens et al. (1990) for details and properties. It can be proved that if X_1 is q-complete w.r.t. X_2, then X_1 is Bayesian identified by X_2; see Florens et al. (1990, Theorem 5.4.12). For a comparison between the classical and the Bayesian completeness, see San Martín and Mouchart (2007).

4.4.2 Bayesian completeness in a fully discrete Bayesian model

Using the notation introduced in Section 3.3, let us characterize the completeness of X_1 w.r.t. X_2 conditionally on X_3. By Definition 2, $X_1 \ll X_2 \mid X_3$ is equivalent to the following implication:

$$E[f(X_1, X_3) \mid X_2, X_3] = 0 \text{ a.s.} \Longrightarrow f(X_1, X_3) = 0 \text{ a.s.}$$

This condition is, therefore, equivalent to

$$E[f(X_1,X_3) \mid X_2=j, X_3=k]=0 \quad \forall\, k\in K,\;\; \forall\, j\in N_2^k \Longrightarrow f(i,k)=0 \quad \forall\, k\in K,\;\; \forall\, i\in N_1^k.$$

Using the definition of a conditional expectation, this condition is equivalent to

$$\sum_{i\in N_1^k} f(i,k)\, p_{i|jk}=0 \quad \forall\, k\in K,\;\; \forall\, j\in N_2^k \Longrightarrow f(i,k)=0 \quad \forall\, k\in K,\;\; \forall\, i\in N_1^k;$$

which in turn is equivalent to

$$\forall\, k\in K \qquad (P^k)' f_k = 0 \Longrightarrow f_k=0,$$

where $(f_k)_i = f(i,k)$ with $i\in N_1^k$. This implication is equivalent to say that, for each $k\in K$, $\text{Ker}\,[(P^k)']=\{0\}$, or equivalently that, for each $k\in K$, $\text{Im}\,(P^k)=\mathbb{R}^{|N_1^k|}$. Since P^k is a linear transformation, this last condition is equivalent to, for each $k\in K$, $r(P^k)=|N_1^k|$. Summarizing, we obtain the following theorem:

Theorem 4 *Let $(\Omega,\mathcal{M},P)$ be a probability space and $X_r:\Omega\longrightarrow N_r$, with $r=1,2,3$, be random variables, where N_r ($r=1,2,3$) are finite sets. The following are equivalent:*

1. $X_1 \ll X_2 \mid X_3$.
2. $\forall\, k\in K\; r(P^k)=|N_1^k|$.

This theorem deserves the following comments:

Comment 5 In the discrete case, the relationship $X_1 \ll X_2 \mid X_3$ imposes a dimensional restriction between X_1 and X_2, namely $\forall\, k\in K$, $|N_1^k|\leq|N_1^k|$. It should be mentioned that the Bayesian identification restriction $X_1 \prec X_2 \mid X_3$ does not imply any dimensional restriction.

Comment 6 In the discrete case, it is easily verified the following general theorem: $X_1 \ll X_2 \mid X_3 \Rightarrow X_1 \prec X_2 \mid X_3$.

Comment 7 If, for each $k\in K$, P^k is a bijective linear transformation (so $|N_1^k|=|N_2^k|$), then $X_1 \ll X_2 \mid X_3$ and $X_2 \ll X_1 \mid X_3$, and conversely.

4.4.3 Coming back to identification relationships in a hierarchical structure

Let us coming back to the (counter-)example developed in Sections 4.3.2 and 4.4. The rank condition $r(\Pi) = KL$ is equivalent to the completeness relationship $(\theta_1, \eta) \ll Y$. Therefore, what is it true is the following implication: if θ_2 is Bayesian identified by η in the marginal latent model and (θ_1, η) is complete w.r.t. Y in te conditional model, then (θ_1, θ_2) is Bayesian identified by Y.

How general is this relationship? Mouchart and San Martin (2003, Theorem 1) and San Martín and Mouchart (2007, Section 7.2) established the following theorem:

Theorem 5 *Consider the general hierarchical structure defined by (4.2). If θ_2 is Bayesian identified by η in the latent marginal model and if (θ_1, η) is 2-complete w.r.t. Y in the conditional model, then (θ_1, θ_2) is Bayesian identified by Y in the statistical model.*

This theorem can be applied to *some* hierarchical structure, being one example the fully discrete Bayesian model above-discussed. Other example is offered by San Martín and Mouchart (2007), where the theorem is applied to obtain the identifiability of a semi-parametric Rasch Poisson Count Model. The applicability of Theorem 5 depends on the possibility to establish the 2-completeness of (θ_1, η) in the conditional model $p(Y \mid \theta_1, \eta)$. According to our recent experience, the 2-completeness depends on the support of the conditional distribution: when it is a finite set, the 2-completeness fails except in a fully discrete model as discussed above; when it is a countable set, it is possible to verify it.

5 Discussion

The identification problem arose as a consequence of a reformulation of the specification problem as stated by Fischer (1922). Such a reformulation establishes that "the investigator's inquisitiveness is not just a population in the sense of a distribution of observable variables, but a physical structure projected behind this distribution, by which the latter is thought to be generated"; Koopmans and Reiersøl (1950, p.165); see also Hurwicz (1950). The identification problem consists, therefore, in investigating whether one and only one structure explains the observed phenomenon. From a probabilistic point of view, this type of specification is modeled through a hierarchical structure; see the examples in Section 4. Thus, the identification problem leads to ensure whether such a hierarchy has an empirical sense.

These considerations explain why identifiability was considered as a pre-statistical problem (McHugh, 1956) and, consequently, more related to statistical modeling than statistical inference. In spite of that, identifiability has been traditionally considered as a necessary condition to ensure a coherent inference –that is, the existence of unbiased estimators and consistent estimators; see San Martín and Quintana (2002) and the references therein. In a pure sampling theory approach, this constitutes a limitation in the sense that it is not possible to compute estimators of unidentified parameters.

Bayesian statistics seems to be a solution. As a matter of fact, it is always possible (if not, some hypotheses can be introduced) to compute the posterior distribution of unidentified parameters. However, from a modeling point of view, the question is to know what is the statistical meaning of these estimators. To answer it, an identifiability analysis is unavoidable: once a parameter of interest has been identified, its statistical meaning becomes explicit. In practice, this is *the* issue which matters; for an example in psychometrics, see San Martín et al. (2009). Therefore, in fields where parameters have a substantive meaning, Bayesian statistics will be useful if it is warranted about identifiability. Otherwise, it will give illusory solutions; see comment 1.

With respect to identification relationships in a hierarchical structure, this paper offers counter-examples showing that some results are false. Furthermore, it provides identification strategies which can be used in some hierarchical models; it depends on the 2-completeness of a parameter with respect to a statistics in a conditional model or, equivalently, to study the injectivity of a posterior expectation. This result is an implicit invitation to study this problem for particular sampling distributions.

Last, but not least, Bayesian identification is always implied by sampling identification. This means that (some) identification results established in a pure sampling theory framework could be useful when models are specified under a Bayesian approach.

Acknowledgements Part of the material included in Section 4 was developed while the first author was visiting the Department of Psychology, K. U. Leuven, Belgium, during the winter of 2008. The first author also acknowledges the partial financial support of the PUENTE Grant 08/2009 from the Pontificia Universidad Católica de Chile.

References

Adams, R. J., M. Wilson, and W. Wang (1997). Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement 21*, 1–23.

Andersen, E. B. (1980). *Discrete Statistical Models with Social Sciences Applications*. North-Holland.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families*. New York, USA: Wiley.

Barra, J. R. (1981). *Mathematical Basis of Statistics*. New York, USA: Academic Press.

Basu, D. (1975). Statistical information and likelihood (with discussion). *Sankhyā 37*, 1–71.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley.

Chow, Y. S. and H. Teicher (1988). *Probability Theory. Independence, Interchangeability, Martingales. Second Edition*. New York, USA: Springer.

Conway, J. B. (1985). *A course in functional analysis*. Springer.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B 41*, 1–31.

De Boeck, P. and M. Wilson (2004). *Explanatory item response models. A generalized linear and non-linear approach*. Springer.

Dellacherie, C. and P. A. Meyer (1975). *Probabilité et Potentiel I.* Paris: Hermann.

Döhler, R. (1980). On The Conditional Independence of Random Variables. *Theory of Probability and Its Applications 25*, 628–634.

Eberly, L. and B. Carlin (2000). Identifiability and convergence issues for markov chain monte carlo fitting of spatial models. *Statistics in Medicine 19*, 2279–2294.

Everitt, B. S. (1984). *An Introduction to Latent Variable Models*. Chapman and Hall.

Fischer, G. H. and I. W. Molenaar (1995). *Rasch Models Foundations, Recent Developments and Applications*. Springer.

Fischer, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London, Series A 222*, 309–368.

Florens, J.-P. and M. Mouchart (1977). Reduction of Bayesian Experiments. Technical report, CORE Discussion Paper 7737, Université Catholique de Louvain, Belgium.

Florens, J.-P. and M. Mouchart (1982). A note on non causality. *Econometrica 50*, 583–591.

Florens, J.-P., M. Mouchart, and J.-M. Rolin (1985). On two definitions of identification. *Statistics 16*, 213–218.

Florens, J.-P., M. Mouchart, and J.-M. Rolin (1990). *Elements of Bayesian Statistics*. Marcel Dekker.

Florens, J.-P. and J.-M. Rolin (1984). Asymptotic sufficiency and exact estimability. In J.-P. Florens, M. Mouchart, J.-P. Raoult, and L. Simar (Eds.), *Alternative Approaches to Time Series Analysis*, pp. 121–142. Publications des Facultés Universitaires Saint-Louis, Bruxelles.

Gelfand, A. and S. Sahu (1999). Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models. *Journal of the American Statistical Association 94*, 247–253.

Ghosh, M. (1995). Inconsistent maximum likelihood estimators for the Rasch model. *Statistical and Probability Letters 23*, 165–170.

Gourieroux, C. and A. Monfort (1995). *Statistics and Econometric Models*. Cambridge University Press.

Hurwicz, L. (1950). Generalization of the Concept of Identification. In T. C. Koopmans (Ed.), *Statistical Inference in Dynamic Economic Models, Cowles Commision Research Monograph 11*, pp. 238–257. New York: Wiley.

Jara, A., F. Quintana, and E. San Martín (2008). Linear mixed models with skew-elliptical distrbutions: A Bayesian approach. *Computational Statistics and Data Analysis 52*, 5033–5045.

Jöreskog, K. G. (1981). Analysis of covariances structures (with Discussion). *Scandinavian Journal of Statistics 8*, 65–92.

Kadane, J. (1975). The role of identification in bayesian theory. In S. Fienberg and A. Zellner (Eds.), *Studies in Bayesian Econometrics and Statistics*, pp. 175–191. Amsterdam: North-Holland.

Kass, R., B. Carlin, A. Gelman, and R. Neal (1998). Markov Chain Monte Carlo in Practice: A Roundtable Discussion. *The American Statistician 52*, 93–100.

Koopmans, T. C. and O. Reiersøl (1950). The identification of structural characteristics. *The Annals of Mathematical Statistics 21*, 165–181.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics 95*, 391–413.

Lindley, D. V. (1971). *Bayesin Statistics: A Review*. Philadelphia: Society for Industrial and Applied Mathematics.

Macci, C. and S. Polettini (2001). Bayes factor for non-dominated statistical models. *Statistics and Probability Letters 53*, 79–89.

Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Harvard University Press.

Manski, C. F. (2007). *Identification for Prediction and Decision*. Harvard University Press.

Martin, F., J. L. Petit, and M. Littaye (1973). Indépendance conditionelle dans le modèle statistique bayésien. *Annales de l'Institut Henri Poincaré - Section B 9*, 19–40.

Maruyama, G. M. (1998). *Basic of Structural Equation Modeling*. Sage Publications.

McCullagh, P. (2002). What is a statistical model? (with discussion). *The Annals of Statistics 30*, 1225–1310.

McHugh, R. B. (1956). Efficient Estimation and Local Identification in Latent Class Analysis. *Psychometrika 21*, 331–347.

McKean, H. (1963). Brownian motion with a several-dimensional time. *Theory of Probability and Its Applications 8*, 335–354.

Mouchart, M. (1976). A Note on Bayes Theorem. *Statistica 36*, 349–357.

Mouchart, M. and J.-M. Rolin (1984). A note on conditional independence. *Statistica 44*, 557–584.

Mouchart, M. and J.-M. Rolin (1985). Letter to the Editor. *Statistica 45*, 427–430.

Mouchart, M. and E. San Martin (2003). Specification and identification issues in models involving a latent hierarchical structure. *Journal of Statistical Planning and Inference 111*, 143–163.

Neyman, J. and E. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*, 1–32.

Paulino, C. and C. Pereira (1994). On identifiability of parametric models. *Journal of the Italian Statistical Society 3*, 125–151.

Poirier, D. J. (1998). Revising Beliefs in Nonidentified Models. *Econometric Theory 14*, 483–509.

Prakasa Rao, B. L. S. (1992). *Identifiability in Stochastic Models. Characterizations of Probability Distributions*. Academic Press.

Rao, M. M. (1984). *Probability Theory with Applications*. New York, USA: Academic Press.

Rao, M. M. (1993). *Conditional Measures and Applications*. Marcel Dekker.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. The Danish Institute for Educational Research (Expanded Edition, 1980, Chicago: The University Chicago Press).

Rasch, G. (1966). An Individualistic Approach to Item Analysis. In P. F. Lazarsfeld and N. W. Henry (Eds.), *Readings in Mathematical Social Sciences*, pp. 89–107. Massachusetts: MIT Press.

San Martín, E. (2000). *Latent Structural Models: Specification and Identification Problems*. Belgium: Ph. D. Dissertation, Institute of Statistics, Université Catholique de Louvain.

San Martín, E. (2003). Modeling problems motivated by the specification of latent linear structures. *Journal of Mathematical Psychology 47*, 572–579.

San Martín, E., J. González, and F. Tuerlinckx (2009). Identified Parameters, Parameters of Interest and Their Relationships. *Measurement: Interdisciplinary Research and Perspective 7*, 97–105.

San Martín, E. and M. Mouchart (2007). On joint completness: Sampling and Bayesian versions, and their connections. *Sankhyā 69*, 780–807.

San Martín, E. and F. Quintana (2002). Consistency and identifiability revisited. *Brazilian Journal of Probability and Statistics 16*, 99–106.

Smits, D. J. M. and S. Moore (2004). Latent item predictors with fixed effects. In P. De Boeck and M. Wilson (Eds.), *Explanatory Item Response Models. A Generalized Liner and Nonlinear Approach*, pp. 267–287. New York: Springer.

van der Linden, W. and R. K. Hambleton (1997). *Handbook of Modern Item Response Theory*. Springer.

Van Putten, C. and J. H. Van Schuppen (1985). Invariance properties of the conditional independence relation. *The Annals of Probability 13*, 934–945.

Appendix

A Conditional Independence

A.1 General definition

The concept of conditional independence, well known and very useful in probability theory, becomes more and more interesting in statistical theory, where it can be used as a basic tool to express many of the important concepts of statistics (such as sufficiency, ancillarity, identifiability, etc.), unifying many areas that are, at first sight, different.

Let $(M, \mathcal{M}, P)$ be a probability space and $\mathcal{N}$ be a sub-σ-algebra of $\mathcal{M}$. The completed σ-field $\overline{\mathcal{N}}$ of $\mathcal{N}$ is defined as

$$\overline{\mathcal{N}} = \mathcal{N} \vee \{E \in \mathcal{M} \,:\, P(A) = P^2(A)\},$$

that is, the σ-algebra generated by the union between $\mathcal{N}$ and the completed trivial σ-field. Following Florens et al. (1990) (see also Chow and Teicher, 1988), the σ-fields are completed by *measurable sets* only and not by subssets of measurable sets as is usually done in Lebesgue completion. In this way, it is avoided the danger of losing the separability of σ-field. As mentioned in the main text, the separability of σ-fields is essential to relate Bayesian and sampling identification.

Definition 3 Let $\mathcal{M}_1, \mathcal{M}_2$ and $\mathcal{M}_3$ be sub-σ-algebras of $\mathcal{M}$. It is said that $\mathcal{M}_1$ *is independent of* $\mathcal{M}_2$ *conditionally on* $\mathcal{M}_3$, denoted as $\mathcal{M}_1 \perp\!\!\!\perp \mathcal{M}_2 \mid \mathcal{M}_3$, if and only if one of the following equivalent conditions hold:

(i) $E[f_1\, f_2 \mid \mathcal{M}_3] = E[f_1 \mid \mathcal{M}_3]\, E[f_2 \mid \mathcal{M}_3]$ a.s. for all positive function f_i measurable with respect to $\mathcal{M}_i$ (for $i = 1, 2$).

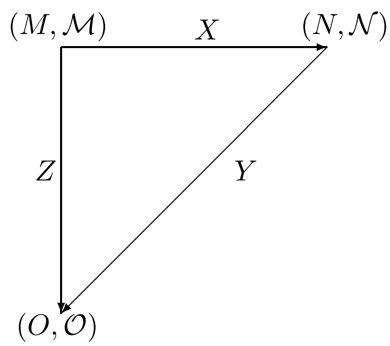
(ii) $E[f_1 \mid \mathcal{M}_2 \vee \mathcal{M}_3] = E[f_1 \mid \mathcal{M}_3]$ a.s for all positive function f_1 measurable with respect to $\mathcal{M}_1$.

For a proof on the equivalence between (i) and (ii), see Florens et al. (1990, Theorem 2.2.1). When $\mathcal{M}_3$ is equal to the trivial σ-field $\{\emptyset, M\}$, this definition reduces to the usual independence between σ-fields; in such a case, we write $\mathcal{M}_1 \perp\!\!\!\perp \mathcal{M}_2$.

It is clear from condition (i) that the concept of conditional independence is *symmetric* in $\mathcal{M}_1$ and $\mathcal{M}_2$, namely $\mathcal{M}_1 \perp\!\!\!\perp \mathcal{M}_2 \mid \mathcal{M}_3$ is equivalent to $\mathcal{M}_2 \perp\!\!\!\perp \mathcal{M}_1 \mid \mathcal{M}_3$. Condition (ii) provides an heuristic meaning of conditional independence: $\mathcal{M}_1 \perp\!\!\!\perp \mathcal{M}_2 \mid \mathcal{M}_3$ means that the process generating $\mathcal{M}_1$ conditionally on $\mathcal{M}_2 \vee \mathcal{M}_3$ depends on $\mathcal{M}_3$ only; or, equivalently, the process generating $\mathcal{M}_2$ conditionally on $\mathcal{M}_1 \vee \mathcal{M}_3$ depends on $\mathcal{M}_3$ only. This heuristic meaning actually corresponds to a measurability property of conditional independence: $\mathcal{M}_1 \perp\!\!\!\perp \mathcal{M}_2 \mid \mathcal{M}_3$ if and only if, for all $\mathcal{M}_2$-measurable function f_2, $E[f_2 \mid \mathcal{M}_1 \vee \mathcal{M}_3]$ is measurable with respect to $\overline{\mathcal{M}}_3$; for a proof, see Florens et al. (1990, Theorem 2.2.6). For details and properties on conditional independence, the reader is referred, among many other, to Martin et al. (1973); Dawid (1979); Döhler (1980); Florens et al. (1990).

A.2 Conditional independence: its meaning in terms of random variables

For a better understanding of the abstract concept of conditional independence we now present its definition in terms of random variables. this presentation is based on the Lemma of Dynkin-Doob. This lemma establishes the following: let $(M, \mathcal{M})$, $(N, \mathcal{N})$ and $(O, \mathcal{O})$ be three measurable spaces and let $X : M \longrightarrow N$ be a measurable function (that is, $\sigma(X) \doteq X^{-1}(\mathcal{N}) \subset \mathcal{M}$) and let $Z : M \longrightarrow O$ be measurable function with respect to $\sigma(X)$ (that is, $\sigma(Z) \doteq Z^{-1}(\mathcal{O}) \subset \sigma(X)$). Then there exists a function $Y : N \longrightarrow O$ measurable with respect to $\mathcal{N}$ (that is, $\sigma(Y) \doteq Y^{-1}(\mathcal{O}) \subset \mathcal{N}$) such that $Z = Y \circ X$; for a proof, see Dellacherie and Meyer (1975); Rao (1984). The following diagram summarizes these relationships.



As pointed out in the main text (see Section 2.2), the σ-field generated by a random variable corresponds to the set of events that may be described in terms of that random variable. In this sense, the Lemma of Dynkin-Doob establishes that when the information provided by a random variable Z is strictly contained in the information provided by X (that is, the events described by Z are contained into the events described by X), then Z is a measurable transformation of X –i.e., Z is a reduction of X.

Let $m_i : (M, \mathcal{M}) \longrightarrow (\mathbb{R}, \mathcal{B})$, $i = 1, 2, 3$, be three random variables and let $\mathcal{M}_i = \sigma(m_i)$. Now, $\mathcal{M}_1 \perp\!\!\!\perp \mathcal{M}_2 \mid \mathcal{M}_3$ if and only if, for all $\mathcal{M}_2$-measurable function f_2, $E[f_2 \mid \mathcal{M}_2 \vee \mathcal{M}_3]$ is $\overline{\mathcal{M}}_3$-measurable. Using the Lemma of Dynkin-Doob, this is equivalent to say that for all measurable function h, there exists a measurable function g such that

$$E[h(m_1) \mid \mathcal{M}_2 \vee \mathcal{M}_3] = g(m_3) \quad \text{a.s.}$$

That is, the conditional expectation of all measurable transformations of m_1 given (m_2, m_3) are a.s. a function of m_3.