Identifiability of structural characteristics: How relevant is it for the Bayesian approach?

Ernesto San Martín[a,b]

[a]*Faculty of Mathematics, Pontificia Universidad Católica de Chile*
[b]*Center for Operations Research and Econometric CORE*

Abstract. The role of identification in the Bayesian approach is still debatable. Since Lindley [Bayesian Statistics. A Review (1971) Philadelphia], most Bayesian statisticians pretend that unidentifiabiity causes no real difficulty in their approach. Recently, Wechsler, Izbicki and Esteves [*Amer. Statist.* **67** (2013) 90–93] provide a simple example illustrating this perspective. By critically reading Wechsler, Izbicki and Esteves [*Amer. Statist.* **67** (2013) 90–93], we intend to show that the Bayesian approach is far from being free of the identification problems, provided that the interest is focused on the interpretation of the parameters. It is written using a rather ancient style, the so-called Platonic dialogues. In modern times, there are beautiful examples of that, particularly in Foundations of Mathematics, where debatable subjects are discussed: let us refer Heyting [Intuitionism. An Introduction (1971) North-Holland Publishing Company], where the debate between a formalist and an intuitionist is presented as a dialogue; or Lakatos [Proofs and Refutations. The Logic of Mathematical Discovery (1976) Cambridge University Press], where the relationship between proofs and conjectures is magnificently illustrated. We hope that this style will help to understand why identifiability really matters in the Bayesian approach.

1 Introduction

The dialogue takes place in an imaginary classroom where some basic concepts related to statistical model construction were being discussed. One of those concepts is the concept of *identification*. The previous session was focused on understanding how identifiability became a statistical problem and, particularly, related to the statistical meaning of a parameter. At the beginning of the current session, the TEACHER does the following summary.

In 1922, R. A. Fisher pointed out that, in spite of the large amount of fruitful applications of statistics, its basic principles were still in a "state of obscurity" and that "during the recent rapid development of practical methods, fundamental problems have been ignored" (Fisher (1922), p. 310). This evaluation led Fisher to define the scope of statistical methods as the reduction of data. "No human mind

is capable of grasping in its entirety the meaning of any considerable quantity of numerical data" (Fisher (1973), p. 6). Consequently, data should be replaced by a relatively few quantity of values which shall contain the *relevant information* provided by the original data.

Conceptually speaking, the reduction of data is accomplished by constructing a hypothetical infinite population, of which the actual data are considered a random sample. This hypothetical population is fully characterized by a probability distribution, which in turn is specified by relatively few parameters. The parameters "are sufficient to describe it exhaustively in respect of all qualities under discussion". Thus, any information provided by a sample, which is used to estimate those parameters, is relevant information. This conceptual scheme leads to distinguish three problems of statistics (Fisher (1922, 1973), Chapter 1): *Problems of Specification*, *Problems of Estimation* and *Problems of Distribution*. Problems of Specification arise in the choice of the mathematical form of the probability distribution which characterizes the observed population. Its choice "is not arbitrary, but requires an understanding of the way in which the data are supposed to, or did in fact, originate" (Fisher (1973), p. 8).

How to make precise the way in which the data are generated? A quarter of century later, Koopmans and Reiersøl suggested a response, which leads to reformulate the specification problem as defined by Fisher. Such a reformulation, "appropriate to many applications of statistical methods", is based on a distinction between population and structure:

> In many fields the objective of 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. The word "physical" is used merely to convey that the structure concept is based on the investigator's ideas as to the "explanation" or "formation" of the phenomena studied, briefly, on his theory of these phenomena, whether they are classified as physical in the literal sense, biological, psychological, sociological, economic or otherwise (Koopmans and Reiersøl (1950), p. 165).

A structure is determined on the basis of a substantive theory, combined with systematically collected data for the relevant variables. The collected data are represented by observed variables, whereas (some) aspects of a substantive theory are represented by unobservable or latent ones. Thus, a structure is composed of a particular probability distribution of the latent variables, combined with a relationship between the observed and unobserved variables. Koopmans and Reiersøl's reformulation of the specification problem is accordingly defined as follows:

> The specification is therefore concerned with the mathematical forms of both the distribution of the latent variables and the relationships connecting observed and latent variables (p. 168).

The distribution of the observed variables is regarded as produced by an operation performed on the distribution of the unobserved variables. The specification

problem is, therefore, concerned with specifying a model which, by hypothesis, contains the structure generating the distribution of the observed variables. The result of this reformulation is the emergence of the identification problem: can the distribution of the observations, generated by a given structure, be generated by only one structure in that model? When the answer is affirmative, it is said that the structure is identified by the observations. In such a case, the structure has a statistical meaning in the sense that it is supported by the observations through the probability distribution of the observations (Koopmans (1949), Hurwicz (1950), Clifford (1982), Manski (1995)). The main characteristic of the so-called structural modeling is that it considers structures and/or parameters as endowed with a meaning that should be supported by the observations.

TEACHER: How can we make precise the last statement, namely that once a parameter is identified, it is then endowed with a statistical meaning? By using a concept of Bayesian identification, we will show that identified parameters can be expressed as functions of the data generating process. In this way, the statistical meaning of a parameter can actually be based on empirical information only. However, our discussion should take into account Lindley' statement "unidentifiability causes no real difficulty in the Bayesian approach" (Lindley (1971)). Our challenge is to show that this statement intends to summarize Bayesian inferential aspects only, but does not prevent on the ambiguousness on the statistical meaning of unidentified parameters.

2 Identifiability in a sampling theory framework approach

TEACHER: In order to answer this question, let us recall the definition of identifiability. This definition depends on a more fundamental concept: what is it?

PUPIL TAU: It's the concept of *statistical model*; in the last session, we learnt it. It come back to Fisher (1922).

TEACHER: It's correct. Now the question is, why identifiability depends on the concept of *statistical model*?

PUPIL ALPHA: Because identifiability deals with the relationship between parameters and sampling probabilities (or the corresponding likelihood functions when a dominated statistical model is considered; see Barra (1981), Chapter 1). The concept of *statistical model* emphasizes the fact that a statistical model is a family of sampling distributions indexed by a parameter.

TEACHER: This is an exact statement. Let us formalize it. Who can do that?

ALPHA: The components of a statistical models are the following:

- *Sample space* $(S, \mathcal{S})$, where S is the set of all possible outcomes, and $\mathcal{S}$ is the σ-field of subsets of S: the elements of $\mathcal{S}$ are considered as the events of interest.
- A probability distribution P^θ defined on the sample space $(S, \mathcal{S})$, where θ is a parameter. P^θ is accordingly called *sampling distribution* or *sampling probability*.

- $\theta \in \Theta$, where the set Θ is called *parameter space*.

Thus, a statistical model can compactly be written as

$$\mathcal{E} = \left\{ (S, \mathcal{S}) : P^{\theta} : \theta \in \Theta \right\}.$$

When the statistical model is dominated (that is, when for all $\theta \in \Theta$, P^{θ} is absolutely continuous with respect to a σ-finite measure λ defined on $(S, \mathcal{S})$, the statistical model is equivalently described through the likelihood function p^{θ}, where $p^{\theta} = \frac{dP^{\theta}}{d\lambda}$:

$$\mathcal{E} = \left\{ (S, \mathcal{S}) : p^{\theta} : \theta \in \Theta \right\}.$$

TAU: It is important to remark that this concept covers what it is commonly called parametric, semi-parametric and non-parametric models: the distinction depends on the structure of Θ. Thus, the parameter space Θ of a parametric model is contained in a finite-dimensional space; for non-parametric models, the parameter space Θ is an infinite-dimensional space; semi-parametric models can be described by saying that Θ is a cartesian product of the form $A \times \Psi$, where A is a finite-dimensional space and Ψ is an infinite-dimensional space.

TEACHER: The definition of statistical model implicitly involves a relationship between the parameter space and the set of sampling distributions defined on $(S, \mathcal{S})$. If we denote this set as $\mathcal{P}(S, \mathcal{S})$, such a relationship corresponds to the mapping $\Phi : \Theta \longrightarrow \mathcal{P}(S, \mathcal{S})$ such that $\Phi(\theta) = P^{\theta}$ for $\theta \in \Theta$.

Our interest is focused on the sample space $(S, \mathcal{S})$: there we collect specific features of interest characterizing a particular population. For instance, the number of correct responses in a test of 40 items. A sampling probability P^{θ} is supposed to describe the behavior of those features. This behavior could be interpreted as the variability of such features across the population. The relevant point here is that P^{θ} fully characterizes the population that is analyzed with respect to those features of interest. If we accept that, then we focus our interest on the family $\{P^{\theta} : \theta \in \Theta\}$, assuming that θ is a characteristic of P^{θ} which in turn describes the units of the population in terms of those features of interest. As Fisher said, the parameters "are sufficient to describe [the population of interest] exhaustively in respect of all qualities under discussion" (Fisher (1973)).

PUPIL SIGMA: That explains why we are interested in the mapping Φ! Thus, it is not desirable that two different parameters be related to the same sampling probability: if not, a same feature of interest can be explained by two different set of data each of them generated by different sampling probabilities.

TAU: Now I understand the formal definition of identifiability; that is,

$$\theta_1 \neq \theta_2 \quad \Longrightarrow \quad P^{\theta_1} \neq P^{\theta_2}.$$

That is, the mapping Φ is injective.

SIGMA: This definition also shows why it is not possible to make inferences on unidentified parameters. In fact, if $\theta_1 \neq \theta_2$, but $P^{\theta_1} = P^{\theta_2} \doteq P^{\theta}$, then the data we

use to make inference on the parameter are supposed to be generated by P^θ. But P^θ is related to θ_1 as well as to θ_2! That is, observations do not provide information on a unique parameter. Econometric as well as psychometric jargon is, in this respect, palatable: θ_1 and θ_2 are observationally equivalent.

TEACHER: The previous discussion deserves several remarks:

1. The concept of identifiability supposes a hypothetical knowledge of the sampling probabilities $\{P^\theta : \theta \in \Theta\}$. An exact knowledge of them cannot be derived from any finite number of observations. Consequently, "by hypothesizing nevertheless the full availability of such knowledge, we obtain a clear separation between problems of statistical inference arising from the variability of finite samples, and problems of identification in which we explore the limits to which inference even from an infinite number of observations is suspect" (Koopmans (1949)); see also Manski (1995).

2. The eventual limitations of statistical inference due to identification problems are typically described through the following results:

   - In a given statistical model $\mathcal{E}$, if θ is unidentified, then there not exists an unbiased estimator of it.
   - In a given statistical model $\mathcal{E}$, if θ is unidentified, then there not exists a consistent estimator of it.

   Proofs of these results can be found in Gabrielsen (1978), with a correction in San Martín and Quintana (2002) and Paulino and Pereira (1994).

3 Revising unidentified parameters

PUPIL KAPPA: These considerations help to show the advantage of the Bayesian approach over the sampling theory framework approach because, according to Lindley, unidentifiability causes no real difficulty in the Bayesian framework.

TEACHER: Could you be more precise? Can you explain why the Bayesian approach seems to be stronger than the sampling theory framework approach? I'm just asking to explain in which sense Lindley's statement is related to inferential aspects rather than to structural modeling ones.

KAPPA: Yes, I can explain Lindley's statement using a simple example recently published by Wechsler, Izbicki and Esteves (2013).

TEACHER: Let us download the paper, thus we can follow the details.

.
.
.

Some minutes after, all the students have access to Wechsler, Izbicki and Esteves (2013), and realize that its merit is to provide a simple example of nonidentifiability. In fact, most of the examples of nonidentifiability "occur in areas of statistics that require further statistical background such as linear models, time series, and survival analysis".

TEACHER: This paper is suitable for us because, as the authors state, "students in statistics can be introduced to the subject through straightforward examples".

SIGMA: Moreover, the following statement is crucial for our future work as statisticians: "Understanding nonidentifiability and its consequences can be crucial when designing new experiments and analyzing data".

KAPPA: Yes, the example on which we will focus our attention is at page 91, taken from an exercise on probability (Ross (2009), p. 111):

"A ball is in any one of n boxes and is in the ith box with probability p_i. If the ball is in box i, a search of that box will uncover it with probability α_i. Show that the conditional probability that the ball is in box j, given that a search of box i did not uncover it, is

$$\frac{p_j}{1 - \alpha_i p_i} \quad \text{if } j \neq i, \quad \text{and} \quad \frac{(1 - \alpha_i)p_i}{1 - \alpha_i p_i} \quad \text{if } j = i."$$

TAU: How this example will help us to understand why under a Bayesian approach inferences on unidentified parameters can be performed?

SIGMA: The objective of Wechsler, Izbicki and Esteves (2013) is precisely to answer this question!

TEACHER: Could you explain us their answer?

SIGMA: Let me try! Kappa, help me!

KAPPA: The first key consideration is the following comment on Ross's exercise: "This problem exemplifies the use of Bayes theorem, to evaluate how an opinion on the occurrence of an event is updated by the evidence of experimental data". That is, Ross's example illustrates a basic aspect of Bayesian *inference*.

TAU: Good remark! Bayesian approach involves a specific inferential paradigm: we update our prior opinions by taking into account the observations. But, where is the prior information in this example? Where the experimental data are?

SIGMA: Your questions are actually answered in the paper. The authors rephrase Ross's problem in the following terms:

"A missing plane has crashed in one of n possible regions. A search for the plane is then performed in region i; p_i is the probability that the plane has gone down in region i and α_i the probability that the plane will be found in a search of the ith region when it is, in fact, in that region. We consider $0 < p_i < 1$ and $0 < \alpha_i < 1$ to make the problem reasonable and avoid trivialities."

TAU: Considering that rephrasing, we note that the parameter of interest is the region in which the plane has gone down; let us denote it as θ. The parameter space can be represented by a set of integers because each region is labeled by a number. Thus, the parameter space Θ corresponds to the set $\{1, 2, \ldots, n\}$. Therefore, the prior specification is the following:

$$p_j = P[\theta = j], \qquad j = 1, 2, \ldots, n.$$

SIGMA: The experimental data is represented by the random variable X: $X = 1$ if the plain is found in region i; 0 otherwise. Thus, $X \in \{0, 1\}$ and, according to the conditions of the problem,

$$P[X = 0 \mid \theta = j] = 1 \quad \text{if } j \neq i, \qquad P[X = 1 \mid \theta = i] = \alpha_i.$$

Therefore, a routine application of Bayes theorem gives us (for $j \neq i$)

$$P[\theta = i \mid X = 1] = 1, \qquad P[\theta = i \mid X = 0] = \frac{(1 - \alpha_i)\,p_i}{1 - \alpha_i\,p_i},$$

$$P[\theta = j \mid X = 1] = 0, \qquad P[\theta = j \mid X = 0] = \frac{p_j}{1 - \alpha_i\,p_i}.$$

ALPHA: Before continuing, it seems relevant to rewrite the example in terms of the formal concept of statistical model. We start from the observations—or experimental data—which are actually represented by a statistical model. In our case, its components are the following:

- Sample space: $(\{0, 1\}, \{\varnothing, \{0\}, \{1\}, \{0, 1\}\})$.
- Sampling probabilities: for simplicity, let me suppose that the label i of the original problem corresponds to 1; then the sampling probabilities are the following:

$$P[X = 1 \mid \theta = 1] = \alpha_1, \qquad P[X = 1 \mid \theta = j] = 0 \qquad \text{for all } j \geq 2.$$

  It is clear how these probabilities are indexed by the parameter θ.
- Parameter space: $\Theta = \{1, 2, \ldots, n\}$.

Thus, compactly written, the statistical model corresponding to the problem at hand is given by

$$\mathcal{E} = \{(\{0, 1\}, \{\varnothing, \{0\}, \{1\}, \{0, 1\}\}), P[\bullet \mid \theta] : \theta \in \Theta\},$$

where

$$P[X = 0 \mid \theta] = (1 - \alpha_1)\mathbb{1}_{\{\theta=1\}} + \sum_{j=2}^{n} \mathbb{1}_{\{\theta=j\}}.$$

TAU: How the prior information is considered in this structure?
ALPHA: We learnt that for defining a probability measure, we need a measurable space, that is, a set M of possible outcomes or results, and a σ-field of subsets of M. Now, we want to define a probability on the parameter space Θ. Because Θ is a finite set, we use the class of all the subsets of Θ, denoted as $\mathcal{P}(\Theta)$. Thus, our measurable space is given by $(\Theta, \mathcal{P}(\Theta))$. In order to define a probability on it, it is enough to take n real numbers $p_1, p_2, \ldots, p_n$ such that

$$0 \leq p_j \leq 1, \qquad \sum_{j=1}^{n} p_j = 1.$$

We interpret p_j as $P[\theta = j]$. The prior distribution corresponds, therefore, to the vector $(p_1, p_2, \ldots, p_n)'$: there are as many prior distributions as numbers p_j's we choose. Each person—as Bayesian statisticians say—chooses as its initial opinion.
TAU: Summarizing, then, we have, on the one hand, the statistical model

$$\mathcal{E} = \{(\{0, 1\}, \mathcal{P}(\{0, 1\}), P(\bullet \mid \theta) : \theta \in \Theta\},$$

Table 1 *Joint probability distribution of (X, θ)*

	$\{X = 0\}$	$\{X = 1\}$
$\{\theta = 1\}$	$(1 - \alpha_1)p_1$	$\alpha_1 p_1$
$\{\theta = 2\}$	p_2	0
$\{\theta = 3\}$	p_3	0
$\vdots$	$\vdots$	$\vdots$
$\{\theta = n\}$	p_n	0

where

$$P[X = 0 \mid \theta] = (1 - \alpha_1)\mathbb{1}_{\{\theta=1\}} + \sum_{j=2}^{n} \mathbb{1}_{\{\theta=j\}};$$

and, on the other hand, the prior specification $(\Theta, \mathcal{P}(\Theta), \boldsymbol{p})$, where $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)'$ with $p_j = P[\theta = j]$.

ALPHA: These elements allow us to construct a Bayesian model, which is characterized by a *unique* probability defined on the product space "parameters × observations". In our case, the Bayesian model is given by

$$\big(\Theta \times \{0, 1\}, \mathcal{P}(\Theta) \otimes \mathcal{P}(\{0, 1\}), \Pi\big),$$

where Π can be fully described by the Table 1.

TEACHER: It is time to get two important conclusions:

- The starting point to perform an identification analysis is the specification of the statistical model, or likelihood function: this is a component common to statisticians following either the Bayesian approach, or the sampling theory framework approach.
- The structural difference between a Bayesian model and a statistical model is that a Bayesian model corresponds to a *unique* probability measure defined on the product space "parameters × observations", whereas a statistical model corresponds to a *family* of sampling distributions indexed by a parameter. Thus, the Bayesian reasoning is done using one probability; the reasoning in a sampling theory framework approach is done with respect to a family of probabilities.

PUPIL DELTA: What are the consequences of such a structural difference?

TEACHER: Before answering this relevant question, let us continue our reading of Wechsler, Izbicki and Esteves (2013). In fact, we just constructed the Bayesian model . . . ; we want to understand how unidentified parameters are treated in this set-up.

KAPPA: Let us consider the sampling distributions. It is easy to verify that the parameter θ is unidentified. In fact,

$$P[X = 1 \mid \theta = 2] = P[X = 1 \mid \theta = 3] = 0,$$
$$P[X = 0 \mid \theta = 2] = P[X = 0 \mid \theta = 3] = 1.$$

That is, we found at least two different parameters such that the respective sampling probabilities are equal.

SIGMA: In spite of this identification problem, it is still possible to make inferences on all the parameters θ, in particular on those that are unidentified. Thus, for instance,

$$P[\theta = 2 \mid X = 0] = \frac{p_2}{1 - \alpha_1 p_1}, \qquad P[\theta = 2 \mid X = 1] = 0.$$

KAPPA: Those computations show the advantage of the Bayesian approach when compared to the sampling theory framework approach:

> "We reinforce that even a nonidentifiable model may bring information about the parameters of interest. Bayesian inference is easily performed in cases of nonidentifiability [...] Nonidentifiability may bring undesirable nonuniqueness to the ML estimation method" (Wechsler, Izbicki and Esteves (2013), p. 93).

DELTA: Wow! These considerations seem to illustrate Lindley's statement: it is always possible to learn about unidentified parameters, while in a sampling theory framework it is impossible to say anything about them.

4 Constructing a Bayesian identification concept

TAU: Yes ... but let us consider your previous question, Delta: how the structural difference between both approaches have an impact on the construction of statistical concepts. I mention that because we learnt that to construct a Bayesian model, we endow the parameter space with a probabilistic structure. However, the concept of identifiability is defined in terms of the injectivity of a mapping Φ (previously introduced by the Teacher): this property does NOT require any probabilistic structure on the parameter space!

TEACHER: This is a remarkable comment. The concept of identification couched in a sampling theory framework needs to be either translated, or reframed in Bayesian terms. In other words, we need to define a pertinent Bayesian concept of identification: once this is done, the second task will be to study if such a concept is related to the identification concept previously introduced, or not.

DELTA: By the way, Wechsler, Izbicki and Esteves (2013) use the identification concept couched in a pure sampling framework approach! If I understood correctly your remark, Teacher, we need to interpret the, let say, standard concept in terms of the Bayesian paradigm: is my intuition correct?

TEACHER: Yes, it is.

DELTA: We have to start with a Bayesian model, therefore, where parameters and observations are endowed with a probability distribution.

TEACHER: Let us denote it as $\Pi(X, \theta)$, where X denotes the observations and θ the parameters. This probability is defined on the product space $(\Theta \times S, \mathcal{A} \otimes \mathcal{S})$,

where $\mathcal{A}$ is a σ-field of subset of Θ and $\mathcal{S}$ is a σ-field of subsets of S. Now, in a Bayesian model the probability Π is decomposed in two complementary ways:

$$\Pi(X, \theta) = p(X \mid \theta)\mu(\theta)$$
$$= \mu(\theta \mid X)p(X),$$

where $p(X \mid \theta)$ is the sampling distribution, $\mu(\theta \mid X)$ is the posterior distribution, $\mu(\theta)$ is the prior distribution, and $p(X)$ is the predictive distribution. However, $p(X \mid \theta)$ and $\mu(\theta \mid X)$ satisfy additional properties, in particular that $p(X \mid \bullet)$ as a function of the parameter is measurable with respect to $\mathcal{A}$, and $\mu(\theta \mid \bullet)$ as a function of the observations is measurable with respect to $\mathcal{S}$.

   Additional remark: when the sampling probability P^{θ} is represented through a density $l(x \mid \theta)$, the previous decomposition of $\Pi(X, \theta)$ can be performed only if $l(\bullet \mid \bullet)$ is a bi-measurable function with respect to $\mathcal{A} \otimes \mathcal{S}$. This condition is NOT related to (un-)identifiability and, therefore, the existence of posterior distribution is just a matter of measurability. Doob (1953) provides a sufficient condition to ensure such a bi-measurability property. Under this condition, Mouchart (1976) deduces the Bayes theorem: such a deduction helps us to understand the "troubles" with improper priors. But let discuss these issues in a next session (in Appendix, we summarize the arguments).

ALPHA: In a Bayesian model, how can we define a parameter of interest?

TEACHER: A parameter (of interest) is a measurable function of θ, say $\psi \doteq g(\theta)$. Following Basu and Pereira (1983) and Florens and Mouchart (1985), we denote it as $\psi \subset \theta$.

KAPPA: If ψ is of interest, then I focus my inference on it: I update it by computing $\mu(\psi \mid X)$.

TEACHER: How do you do that?

KAPPA: First, I obtain $\Pi(X, \psi)$ from $\Pi(X, \theta)$ by marginalization with respect to $\mu(\theta \mid \psi)$, the conditional prior distribution of θ given ψ. Thereafter, I apply the Bayes theorem.

TEACHER: Could you be more explicit, please?

KAPPA: Well, let me suppose that the parameter θ is decomposed as $\theta = (\psi, \lambda)$ [as done by Poirier (1998)]. Then the two complementary decompositions of $\Pi(X, \psi)$ are the following:

$$\Pi(X, \theta) = p(X \mid \psi)\mu(\psi)$$
$$= \mu(\psi \mid X)p(X),$$

where the likelihood $p(X \mid \psi)$ is obtained by integrating out with respect to $\mu(\lambda \mid \psi)$:

$$p(X \mid \psi) = \int p(X \mid \theta)\mu(\lambda \mid \psi)\, d\lambda. \tag{4.1}$$

TEACHER: This shows that a Bayesian model can be *reduced*—to use the terminology introduced by Basu (1975), who borrowed it from Fisher (1922). In fact,

a probability distribution can be reduced either by marginalization or by conditionalization. Now, the reduction performed by you, Kappa, *modifies* the original data generating process because $p(X \mid \theta)$ is a function of λ and therefore $p(X \mid \theta) \neq p(X \mid \psi)$.

TAU: So, each time we focus our attention on a parameter $g(\theta)$, it is enough to perform a reduction of the original Bayesian model.

TEACHER: No; your statement is incorrect. If the data generating process is initially described by the sampling distribution $p(X \mid \theta)$, the Bayesian inference should be performed with respecto to such a distribution. When you perform a reduction in order to obtain $p(X \mid \psi)$, you change the likelihood and, therefore, you not only change the original problem, but also the original data. However, the concept of reduction of a Bayesian model is relevant to define identifiability in Bayesian terms.

TAU: Why?

TEACHER: Because such a tool allows us to define the concept of *sufficient parameter*: in the Bayesian model $\Pi(X, \theta)$, a subparameter $\psi \subset \theta$ is said to be sufficient if ψ describes the sampling process, being θ redundant.

TAU: Do you mean that $p(X \mid \theta)$ can be "replaced" by $p(X \mid \psi)$ *without losing relevant information*? I say that because I remember a similar statement when we study sufficient statistics.

TEACHER: Yes, it is correct.

ALPHA: Taking into account that $\psi \subset \theta$, we can use a probabilistic tool to formalize the definition of sufficient parameter, the tool of conditional independence. We know that it is very efficient to use it as a language to represent statistical concepts!

TEACHER: I agree ... it were just my following remark ... but go ahead, shows us the formalization.

ALPHA: If I take Tau's statement, I mean $p(X \mid \theta) = p(X \mid \psi)$, then I can write

$$X \perp\!\!\!\perp \theta \mid \psi, \qquad \text{where } \psi \subset \theta.$$

TEACHER: That is correct. Just remark that such a condition is well defined with respect to the probability Π.

TAU: To what extend a reduction performed under the condition $X \perp\!\!\!\perp \theta \mid \psi$ is different from a reduction performed without assuming such a condition?

ALPHA: Well, let consider your computation of the likelihood $p(X \mid \psi)$ (see equality (4.1)). Under the condition $X \perp\!\!\!\perp \theta \mid \psi$, $p(X \mid \theta) = p(X \mid \psi)$ and, therefore,

$$p(X \mid \psi) = \int p(X \mid \theta) \mu(\lambda \mid \psi) \, d\lambda$$

$$= p(X \mid \psi) \int \mu(\lambda \mid \psi) \, d\lambda$$

$$= p(X \mid \psi) \times 1.$$

That is, the original likelihood was not modified and, consequently, we can say that the condition $X \perp\!\!\!\perp \theta \mid \psi$ ensures that the reduction of the original Bayesian

model is performed *without loosing any relevant information*. Consequently, the conditional prior $\mu(\lambda \mid \psi)$ is not useful to performing an update.

DELTA: Is it possible to interpret the definition of sufficient parameter in Bayesian terms? I make this question because the equality $p(X \mid \theta) = p(X \mid \psi)$ is easily understood in a sampling theory approach.

ALPHA: I guess it is possible if we recall the symmetry of the conditional independence, namely

$$X \perp\!\!\!\perp \lambda \mid \psi \quad \Longleftrightarrow \quad \lambda \perp\!\!\!\perp X \mid \psi.$$

Thus, $\mu(\theta \mid X, \psi) = \mu(\lambda \mid \psi)$.

DELTA: That is interesting because $\mu(\lambda \mid \psi)$ is the conditional prior distribution used by Kappa to reduce $\Pi(X, \theta)$ into $\Pi(X, \psi)$. It means that, once we condition on ψ, the observation X does not bring information on $\theta = (\lambda, \psi)$: this is exactly Alpha's statement on the uselessness of $\mu(\lambda \mid \psi)$! This is a notion already introduced by Dawid (1979), and widely used by Poirier (1998) in a paper dealing with revising belief in nonidentified models.

TEACHER: Your remarks are correct. Let me add a last one: a conditional independence condition can be defined in terms of conditional expectations, as shown by Dellacherie and Meyer (1975) and widely pursued by Florens, Mouchart and Rolin (1990). Thus, $X \perp\!\!\!\perp \theta \mid \psi$ means that (because $\psi \subset \theta$)

$$E\big[f(X) \mid \theta\big] = E\big[f(X) \mid \psi\big]$$

for all positive measurable function f; or, equivalently,

$$E\big[h(\theta) \mid X, \psi\big] = E\big[h(\theta) \mid \psi\big]$$

for all positive measurable function h.

TAU: I guess I understood this definition. My question now is, how can we use this concept to define identifiability in Bayesian terms?

TEACHER: First, remark the following fact:

$$X \perp\!\!\!\perp \theta \mid \psi \quad \text{and} \quad \omega \subset \theta \quad \Longrightarrow \quad X \perp\!\!\!\perp \theta \mid (\psi, \omega).$$

That is, once we have a sufficient parameter, we can find out as many sufficient parameters as subparameters ω we get!

ALPHA: This means that if the sampling process is described by ψ, at the same time is described by (ψ, ω) because

$$E\big[f(X) \mid \psi\big] = E\big[f(X) \mid \theta\big] = E\big[f(X) \mid \psi, \omega\big]$$

for all positive measurable function f. Therefore, in the description of the sampling process there is a kind of "indeterminacy"!

TEACHER: That is correct. So, the question is, how this "indeterminacy" can be avoided?

TAU: I guess to answer this question, it is relevant to note that

$$\psi \subset (\psi, \omega) \subset \theta.$$

Is it helpful?

TEACHER: Yes! It shows that, once we have a sufficient parameter, it is always possible to find out a "bigger" one.

TAU: So, it is enough to find out the "smallest" sufficient parameter.

TEACHER: Could you formalize your statement?

TAU: Well, let me denote the "smallest" sufficient parameter as $\theta_{\min}$. Then, it should satisfy the definition of sufficient parameter, so

$$X \perp\!\!\!\perp \theta \mid \theta_{\min}.$$

Now $\theta_{\min}$ is the smallest one, hence $\theta_{\min}$ is a function of all sufficient parameter, but there not exists a sufficient parameter that is a function of $\theta_{\min}$.

TEACHER: This is a good definition. Let me summarize it: let $\Pi(X, \theta)$ be a Bayesian model. We said that $\theta_{\min} \subset \theta$ is a *minimal sufficient parameter* if the following two conditions hold:

- $X \perp\!\!\!\perp \theta \mid \theta_{\min}$.
- $X \perp\!\!\!\perp \theta \mid \psi$ with $\psi \subset \theta$, then $\theta_{\min} \subset \psi$ a.s.

This last condition depends on the prior distribution.

TAU: Therefore, $\theta_{\min}$ fully describes the sampling process in the sense that the sampling process is parameterized unambiguously. Moreover, $\theta_{\min}$ is the smallest parameter conditionally on which all other parameter is not updated by the data.

TEACHER: It is time to define identifiability in a Bayesian framework. We said that θ is Bayesian identified by X if and only if $\theta = \theta_{\min}$ with prior probability 1.

DELTA: The way we follow to define Bayesian identifiability is quite clear. But how can we ensure the existence of the minimal sufficient parameter?

TEACHER: This is an excellent question: by answering, we will obtain an operational characterization of $\theta_{\min}$.

KAPPA: Show us the existence of $\theta_{\min}$!

TEACHER: To do that, we will use two concepts we learnt in probability theory:

- How we characterize the smallest σ-field that makes a function measurable? (this corresponds to the σ-field generated by a function).
- How we characterize the measurability of a function with respect to a given σ-field? (this corresponds to the Lemma of Dynkin–Doob).

KAPPA: Yes, we recall that. So, go ahead!

TEACHER: First, we begin with our Bayesian model $\Pi(X, \theta)$, where Π is defined on the σ-field $\mathcal{A} \otimes \mathcal{S}$. Second, let us define the class of sufficient parameters. To do that, let us consider as a subparameter, a sub-σ-field of $\mathcal{A}$—which can be viewed as the σ-field generated by $g(\theta)$. Thus, the class

$$\Sigma = \{\mathcal{B} \subset \mathcal{A} : \mathcal{S} \perp\!\!\!\perp \mathcal{A} \mid \mathcal{B}\}$$

is the class of sufficient parameters. Why this class is non-empty?

KAPPA: Because $\mathcal{A} \in \Sigma$: it is easy to verify that $\mathcal{S} \perp\!\!\!\perp \mathcal{A} \mid \mathcal{A}$.

TEACHER: That it's correct. Now, we need to recall a characterization of conditional independence in terms of measurability.

DELTA: The property is in terms of σ-fields, and it is the following: let $(M, \mathcal{M}, P)$ be a probability space and let $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ be sub-σ-fields of $\mathcal{M}$. Then $\mathcal{M}_1 \perp\!\!\!\perp \mathcal{M}_2 \mid \mathcal{M}_3$ if and only if $E(m_1 \mid \mathcal{M}_2 \vee \mathcal{M}_3)$ is $\overline{\mathcal{M}}_3$-measurable positive for all m_1 $\mathcal{M}_1$-measurable positive function. Here $\mathcal{M}_1 \vee \mathcal{M}_2 = \sigma(\mathcal{M}_1 \cup \mathcal{M}_2)$ and $\overline{\mathcal{M}}_3$ is the completed σ-field, where the completion is in terms of measurable null events, that is,

$$\overline{\mathcal{M}}_3 = \mathcal{M}_3 \vee \{A \in \mathcal{M} : P(A) \in \{0, 1\}\}.$$

TEACHER: Excellent! This property is very useful to ensure the existence of $\theta_{\min}$. In fact, because $\Sigma \neq \varnothing$, if we take $\mathcal{B}_1, \mathcal{B}_2 \in \Sigma$ then

- $\mathcal{S} \perp\!\!\!\perp \mathcal{A} \mid \mathcal{B}_1$ if and only if $E[s \mid \mathcal{A}]$ is $\overline{\mathcal{B}}_1$-measurable positive for all s $\mathcal{S}$-measurable positive function.
- $\mathcal{S} \perp\!\!\!\perp \mathcal{A} \mid \mathcal{B}_2$ if and only if $E[s \mid \mathcal{A}]$ is $\overline{\mathcal{B}}_2$-measurable positive for all s $\mathcal{S}$-measurable positive function.

It follows that $E[s \mid \mathcal{A}]$ is $\overline{\mathcal{B}}_1 \cap \overline{\mathcal{B}}_2$-measurable positive function for all s $\mathcal{S}$-measurable positive function.; that is,

$$\mathcal{S} \perp\!\!\!\perp \mathcal{A} \mid \overline{\mathcal{B}}_1 \cap \overline{\mathcal{B}}_2,$$

so $\overline{\mathcal{B}}_1 \cap \overline{\mathcal{B}}_2 \in \Sigma$.

DELTA: Your argument shows how a minimal sufficient parameter can be obtained: it is enough to take the intersection of all the sufficient parameters in Σ. It is, by construction, the minimal one.

TEACHER: It is correct. The minimal sufficient parameter is given by

$$\mathcal{B}_{\min} = \bigcap_{\mathcal{B} \in \Sigma} \overline{\mathcal{B}}.$$

Note that $\overline{\mathcal{B}}_{\min} = \mathcal{B}_{\min}$, so the events of prior probability equal to 0 or 1 belongs to $\mathcal{B}_{\min}$. This is actually a pertinent Bayesian concept as far as it depends on the prior distribution through the null events. Moreover, the existence of the minimal sufficient parameter is ensured by construction.

KAPPA: I agree with the argument. Now, is it possible to offer an operational characterization?

TEACHER: Yes, it is. Let us define the following subparameter:

$$\sigma\{E[s \mid \mathcal{A}] : s \in [\mathcal{S}]^+\},$$

where $[\mathcal{S}]^+$ is the set of positive $\mathcal{S}$-measurable functions. It is a subparameter because it is contained in $\mathcal{A}$ by definition of conditional expectation. Somebody can provide an interpretation?

ALPHA: First, $E[s \mid \mathcal{A}]$ for $s \in [\mathcal{S}]^+$ is a sampling expectation, which is a measurable function of $\mathcal{A}$. If we take the σ-field generated by the set $\{E[s \mid \mathcal{A}] : s \in [\mathcal{S}]^+\}$, we obtain *all* the information provided by the sampling process ... What it is nice is that this σ-field is a sub-parameter!

TEACHER: Nice interpretation. Now, if we take into account that it is a σ-field generated by a function, what can we state?

ALPHA: Well, that $\sigma\{E[s \mid \mathcal{A}] : s \in [\mathcal{S}]^+\}$ is the smallest σ-field that makes the sampling expectations measurable.

TEACHER: That is correct. Now, if we focus on this last characterization, could you write it in terms of conditional independence?

ALPHA: Let see ... $E[s \mid \mathcal{A}]$ is $\sigma\{E[s \mid \mathcal{A}] : s \in [\mathcal{S}]^+\}$-measurable for all $s \in [\mathcal{S}]^+$. This is equivalent to

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

if $\sigma\{E[s \mid \mathcal{A}] : s \in [\mathcal{S}]^+\}$ contains the prior null sets.

TEACHER: It's correct, and in fact the last statement is true because

$$\mathcal{A} \cap \overline{\mathcal{S}} \subset \sigma\{E[s \mid \mathcal{A}] : s \in [\mathcal{S}]^+\}.$$

To see that, it is enough to note that

$$A \in \mathcal{A} \cap \overline{\mathcal{S}} \quad \Longleftrightarrow \quad A \in \mathcal{A} \quad \text{and} \quad \mathbb{1}_A = E[\mathbb{1}_A \mid \mathcal{S}].$$

Remark that $\mathcal{A} \cap \overline{\mathcal{S}}$ represents the parameters that are a.s. equal to observations. In particular, this σ-field includes the events in $\mathcal{A}$ of prior probability equal to 0 or 1.

ALPHA: Therefore, we have that

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

which implies that $\sigma\{E[s \mid \mathcal{A}] : s \in [\mathcal{S}]^+\} \in \Sigma$. We conclude that

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

TEACHER: Correct! Who can prove the converse relationship?

KAPPA: This is straightforward. $\sigma\{E[s \mid \mathcal{A}] : s \in [\mathcal{S}]^+\}$ is the smallest σ-field that makes $E[s \mid \mathcal{A}]$ measurable for all $s \in [\mathcal{S}]^+$. But $\mathcal{S} \perp\!\!\!\perp \mathcal{A} \mid \mathcal{B}_{\min}$, hence $\mathcal{B}_{\min}$ makes sampling expectations measurable. Therefore,

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

TEACHER: Correct! Thus, we have proved that, in a Bayesian model, the minimal sufficient parameter always exists and corresponds to $\sigma\{E[s \mid \mathcal{A}] : s \in [\mathcal{S}]^+\}$. Moreover, the Bayesian model $\Pi(X, \theta)$ is Bayesian identified if and only if

$$\sigma\{E[s \mid \mathcal{A}] : s \in [\mathcal{S}]^+\} = \mathcal{A}.$$

SIGMA: I have a question: why we say that $\mathcal{B}_{\min}$ is the Bayesian *identified* parameter? I mean, our starting point was to look for a Bayesian concept that can

be considered the Bayesian counterpart of the sampling theory framework concept of identification. In which sense the concept developed up to now is such a counterpart?

TEACHER: Good question! The answer is based on a theorem relating both concepts. Without entering in details, if for instance both $\mathcal{A}$ and $\mathcal{S}$ are Borel σ-fields, then the injectivity of the mapping $\theta \longmapsto P^\theta$ implies that θ is Bayesian identified for all prior distribution. This important result was proved by Florens, Mouchart and Rolin (1985). A systematic treatment can be found in Florens, Mouchart and Rolin (1990), Chapter 4. This concept of identification has been used to analyze the identifiability of semi-parametric Item Response Theory models (San Martín et al. (2011), San Martín, Rolin and Castro (2013)) and to analyze the identifiability of linear mixed-models with skew-elliptical distributions (Jara, Quintana and San Martín (2008)).

SIGMA: This relationship actually allows us to say that the Bayesian identification concept already developed is the counterpart of the sampling theory framework concept. Now we have all the necessary concepts to revisit Wechsler, Izbicki and Esteves (2013)'s example.

5 Revisiting Lindley's (1971) claim

TEACHER: Yes, it is time to revisit Lindley's (1971) claim. We already have specified the discrete Bayesian model used by Wechsler, Izbicki and Esteves (2013). Now, let us focus our attention on the sampling distributions, which can be written in the following terms:

$$P[X = 0 \mid \theta] = (1 - \alpha_1)\mathbb{1}_{\{\theta=1\}} + \mathbb{1}_{\{\theta \neq 1\}},$$

where $\{\theta \neq 1\} = \{\theta \in \{2, 3, \ldots, n\}\}$. Who can describe the corresponding minimal sufficient parameter?

ALPHA: Well, taking into account that X is a binary random variable, it is enough to characterize the following σ-field:

$$\sigma\{P[X = x \mid \theta] : x \in \{0, 1\}\}.$$

Now, the sampling distribution is a simple function which takes only two different values: $(1 - \alpha_1)$ and 1. Therefore,

$$\sigma\{P[X = x \mid \theta] : x \in \{0, 1\}\} = \sigma\{\{\theta = 1\}, \{\theta \neq 1\}\} = \{\varnothing, \{\theta = 1\}, \{\theta \neq 1\}, \Theta\}. \tag{5.1}$$

But $\varnothing$ and Θ are prior null events. Therefore, the Bayesian identified parameter is characterized by the partition

$$\{\{\theta = 1\}, \{\theta \neq 1\}\}.$$

TEACHER: This is correct. Taking into account this finding, look at the probability distribution describing the discrete Bayesian model (see Table 1).

Rows $2, 3, \ldots, n$ are linearly dependent. This is a necessary and sufficient condition to state that θ is not Bayesian identified by X. In fact, San Martín and González (2010) show that if both θ and X are discrete random variables, and P represents the joint probability distribution $(P)_{ij} \doteq P[\theta = i, X = j]$, then θ is Bayesian identified by X if and only if any two rows of P are linearly independent.

DELTA: I'm surprising because the lack of identifiability is, for instance, due to the parameters $\{\theta = 2\}$ and $\{\theta = 3\}$, as in the sampling framework!

TEACHER: It is not a surprise because both concepts are related! Moreover, when $P[\theta = j] > 0$ for all $j = 1, \ldots, n$, San Martín and González (2010) show that both concepts are equivalent.

KAPPA: Let go to the updating process. I can update $\{\theta = 1\}$. This is easy. But I can also update $\{\theta = j\}$ for $j \geq 2$, in spite that

$$\{\theta = j\} \notin \big\{\{\theta = 1\}, \{\theta \neq 1\}\big\}.$$

TEACHER: Your statement need to be carefully considered. In fact, when we compute the posterior probability of $\{\theta = j\}$ with $j = 1, 2, \ldots, n$, we consider the measurability structure given by $\mathcal{P}(\{1, 2, \ldots, n\})$. In this context, we can compute the posterior distribution not only of the identified parameters $\{\theta = 1\}$ and $\{\theta \neq 1\}$, but also of $\{\theta = j\}$ for $j = 2, 3, \ldots, n$. The corresponding posterior probabilities are given by

$$P\big[\{\theta = 1\} \mid X\big] = \frac{(1 - \alpha_1)p_1}{1 - \alpha_1 p_1}\mathbb{1}_{\{X=0\}} + \mathbb{1}_{\{X=1\}};$$

$$P\big[\{\theta \neq 1\} \mid X\big] = \frac{p_2 + \cdots + p_n}{1 - \alpha_1 p_1}\mathbb{1}_{\{X=0\}} + 0 \cdot \mathbb{1}_{\{X=1\}};$$

$$P\big[\{\theta = j\} \mid X\big] = \frac{p_j}{1 - \alpha_1 p_1}\mathbb{1}_{\{X=0\}} + 0 \cdot \mathbb{1}_{\{X=1\}} \qquad \text{for } j \geq 2.$$

SIGMA: It's clear: identified as well as unidentified parameters are updated!

TEACHER: Yes, but the question is the following: when we update unidentified parameters, do we learn something different from what we learn from the updating of the identified parameter?

ALPHA: If I correctly understand your question, you suggest that what we learn about unidentified parameters in *not* more than what we learn about the identified one: is it correct?

TEACHER: Yes, it is. Let us consider the general Bayesian model $\Pi(X, \theta)$. We know that the Bayesian identified parameter $\theta_{\min} \subset \theta$ always exists and satisfies $X \perp\!\!\!\perp \theta \mid \theta_{\min}$. Hence, if $\psi \subset \theta$, and we suppose that we are interested in ψ, then

$$E\big[f(\psi) \mid X\big] = E\big[E\big[f(\psi) \mid X, \theta_{\min}\big] \mid X\big]$$
$$= E\big[E\big[f(\psi) \mid \theta_{\min}\big] \mid X\big]$$

for all measurable positive function f. Thus, when we update ψ, we *only* update a function of $\theta_{\min} \subset \theta$! By the way, this general result is also remarked by Wechsler, Izbicki and Esteves (2013):

Table 2 *Reduced Bayesian model*

	$\{X = 0\}$	$\{X = 1\}$
$\{\theta = 1\}$	$(1 - \alpha_1)p_1$	$\alpha_1 p_1$
$\{\theta \neq 1\}$	$p_2 + \cdots + p_n$	0

"Information about θ is partial and comes only through the medium of learning about $g(\theta)$" (p. 91);

here, $g(\theta)$ corresponds to the minimal sufficient parameter. Before discussing this statement, let see the details in the example under analysis.

ALPHA: We can apply the general argument to our simple example: for $j = 2, 3, \ldots, n$,

$$
\begin{aligned}
P[\theta = j \mid X = 0] &= E\big[\mathbb{1}_{\{\theta = j\}} \mid \{X = 0\}\big] \\
&= E\big\{E\big[\mathbb{1}_{\{\theta = j\}} \mid \{X = 0\}, \{\theta = 1\}, \{\theta \neq 1\}]\{X = 0\}\big\} \\
&= E\big\{E\big[\mathbb{1}_{\{\theta = j\}} \mid \{\theta = 1\}, \{\theta \neq 1\}] \mid \{X = 0\}\big\}.
\end{aligned}
$$

But

$$
P\big[\theta = j \mid \{\theta = 1\}, \{\theta \neq 1\}\big] = P[\theta = j \mid \theta = 1]\mathbb{1}_{\{\theta = 1\}} + P[\theta = j \mid \theta \neq 1]\mathbb{1}_{\{\theta \neq 1\}}
$$

$$
= 0 \cdot \mathbb{1}_{\{\theta = 1\}} + \frac{p_j}{1 - p_1}\mathbb{1}_{\{\theta \neq 1\}}.
$$

Therefore,

$$
P[\theta = j \mid X = 0] = E\left[\frac{p_j}{1 - p_1}\mathbb{1}_{\{\theta \neq 1\}} \mid X = 0\right] = \frac{p_j}{1 - p_1}P[\theta \neq 1 \mid X = 0].
$$

TAU: This type of statements should be complemented by a remark previously done by the Teacher, namely that the original Bayesian model can be replaced by a *reduced* model without losing relevant information. Accordingly, I can reduce our original Bayesian model to one which only includes the identified parameters. By doing so, we avoid redundant information! The original Bayesian model can, therefore, be reduced to the Table 2.

Let me emphasize: this reduction can be done *without losing relevant information*, as widely explained earlier (see page 356) … because the Bayesian identified parameter is the *minimal sufficient parameter*. Consequently, the likelihood is *exhaustively* described by the minimal sufficient parameter, and the learning-by-observing process is fully concentrated on the minimal sufficient parameter in the sense that, conditionally on it, *there is nothing more to learn on $\{\theta = j\}$ for $j = 2, \ldots, n$*. Moreover, the posterior probability of the unidentified parameters are less than the posterior probability of the identified ones: for each $j = 2, \ldots, n$,

$$
P[\theta = j \mid X = 0] < P[\theta \neq 1 \mid X = 0], \qquad P[\theta = j \mid X = 1] < P[\theta \neq 1 \mid X = 1].
$$

In other words, it is more informative the posterior distribution of an identified parameter than that of the unidentified parameter.

ALPHA: The reduced Bayesian model is now defined on the product space

$$(\{1, \ldots, n\} \times \{0, 1\}, \mathcal{A} \otimes \mathcal{S}, \Pi),$$

where

- $\mathcal{A} = \{\varnothing, \{\theta = 1\}, \{\theta \neq 1\}, \Theta\}$;
- $\mathcal{S} = \{\varnothing, \{X = 0\}, \{X = 1\}, \{0, 1\}\}$;
- and Π is the matrix that Alpha computed.

In this model, the events

$$\{\theta = j\}, \qquad j = 2, 3, \ldots, n$$

are *not* measurable and, consequently, they don't provide information on the observations are not informative: all the relevant information regarding the data is in $\theta_{\min}$.

6 The statistical meaning of identified parameters

TEACHER: It is time to discuss about the statistical meaning of the identified parameters. The objective of this session was to formalize the claims according to which identified parameters can only be supported by the observations through the data generating process. The concept of Bayesian identification actually allows us to precise this statement. Who can explain it?

TAU: In a Bayesian model, the minimal sufficient parameter is given by $\sigma\{E[s \mid \mathcal{A}] : s \in [\mathcal{S}]^+\}$. Now, a Bayesian model is identified if

$$\mathcal{A} = \sigma\{E[s \mid \mathcal{A}] : s \in [\mathcal{S}]^+\},$$

which means that the parameters are measurable functions of sampling expectations. These expectations fully describe the data generating process, so the statistical meaning of the identified parameter is supported by the observations.

ALPHA: This was just what we done when we characterized the minimal sufficient parameter of Wechsler, Izbicki and Esteves's (2013) example:

$$\sigma\{P[X = x \mid \theta] : x \in \{0, 1\}\} = \{\varnothing, \{\theta = 1\}, \{\theta \neq 1\}, \Theta\}.$$

This equality means that

1. The sampling probabilities are measurable with respect to $\{\varnothing, \{\theta = 1\}, \{\theta \neq 1\}, \Theta\}$.
2. The parameters $\{\theta = 1\}$ and $\{\theta \neq 1\}$ can therefore be expressed as functions of the sampling distributions—in other words, the sampling distributions provide information on $\{\theta = 1\}$ and $\{\theta \neq 1\}$. This is valid in the original Bayesian model (Table 1) as well as in the reduced Bayesian model (Table 2).

3. The parameters $\{\theta = j\}$ for $j = 2, 3, \ldots, n$ can not be supported by the observations because such parameters *are not measurable* with respect to the σ-field of the parameter space in the reduced Bayesian model.

TEACHER: Who can explicitly write the functions linking the identified parameters to the sampling distributions?

SIGMA: Well, let me define the functions Y_0 and Y_1 as follows:

$$P[X = 0 \mid \theta] = (1 - \alpha_1)\mathbb{1}_{\{\theta=1\}} + \mathbb{1}_{\{\theta\neq1\}} \doteq Y_0,$$

$$P[X = 1 \mid \theta] = \alpha_1\mathbb{1}_{\{\theta=1\}} + 0 \cdot \mathbb{1}_{\{\theta\neq1\}} \doteq Y_1.$$

Therefore,

$$\{\theta = 1\} = Y_0^{-1}[\{1 - \alpha_1\}] \cap Y_1^{-1}[\{\alpha_1\}], \qquad \{\theta \neq 1\} = Y_0^{-1}[\{1\}] \cap Y_1^{-1}[\{0\}].$$

TEACHER: These fundamental equalities show in which sense the data generating process provides information leading to state the statistical meaning of the parameters. As it was pointed out by Alpha, these equalities can also be derived from the original Bayesian model and, therefore, the parameters $\{\theta = j\}$ for $j = 2, \ldots, n$ can no be expressed as functions of the sampling distributions. Using these relationships, who can exploit the statistical meaning of the parameters?

ALPHA: The parameter $\{\theta = 1\}$ that the plain is in region 1, provides the same information as the probability of finding it and of not finding it there. The parameter $\{\theta \neq 1\}$ that the plain is not in region 1 provides the same information as the probability of finding and of no finding the plain in a region different from 1.

TEACHER: These considerations show that, although it seems possible to make inferences on unidentified parameters, such inferences do not provide any new insight because the statistical meaning of unidentified parameters is *not* supported by the observations. Thus, Lindley's (1971) claim is not correct: identifiability matters for Bayesian model construction.

7 Where does the problem of identification come from?

At the end of the session, the Teacher considers the role of identifiability in the formulation of scientific statements, which leads him to consider the source of an identification problem: In 1949, the great psychologist Piaget published a treatise on logic with a double objective: to provide a basic text on modern logic to non specialist readers; and to develop specific piagetian theses dealing with structures of sets and with reversibility of deductive operations. These tools were widely used by Piaget to explain the cognitive development of child and adolescents. One year later, the logician Beth published a book review on Piaget's treatise: the review was extremely critical as Beth (1950b) pointed out the failures and gaps in Piaget's treatise.

7.1 Definitions by abstraction

Among the gaps, Beth (1950b) mentions the *definitions by abstraction*. This type of definitions play an important role in the formation of concepts in modern mathematics. Let us start by discussing them [adapted from Beth (1950a, 1965)]. Suppose that a class $\mathfrak{S}$ and a relation R, defined for the elements of $\mathfrak{S}$, are given. We assume that the relation R is an equivalence relation; that is, R is reflexive, symmetric and transitive. Let $u \in \mathfrak{S}$ be an arbitrary element. Then we let correspond to u the class $\mathcal{C}(u)$ (typically called *equivalence class*), which is composed of all elements $v \in \mathfrak{S}$ such that vRu. The equivalence relation R accordingly determines an exhaustive classification of the elements of $\mathfrak{S}$ in the sense that the family $\{\mathcal{C}(u) : u \in \mathfrak{S}\}$ defines a partition of $\mathfrak{S}$, that is,

  (i) Each element $x \in \mathfrak{S}$ belong to exactly one class $\mathcal{C}(x)$.
 (ii) For all $x, y \in \mathfrak{S}$, $\mathcal{C}(x) \cap \mathcal{C}(y) = \varnothing$ if and only if $x \neq y$.
(iii) $\bigcup_{x \in \mathfrak{S}} \mathcal{C}(x) = \mathfrak{S}$.

The equivalence classes $\{\mathcal{C}(u) : u \in \mathfrak{S}\}$ constitute a new type of objects in $\mathfrak{S}$ that were derived from the original objects u in $\mathfrak{S}$ by means of a definition by abstraction related to the equivalence relation R. This procedure, that is related to Aristotle's theory of abstraction, allows us to say that a *proper name* has a *property* (Lorenzen (1969), Chapter 1) and, consequently, to formulate specific scientific statements. For more details on definitions by abstraction, see also Lorenzen (1987), Chapters 5 and Kim (2011).

7.2 Examples of definitions by abstraction

The most significant physical and mathematical concepts are likewise obtained in accordance with the scheme of definitions by abstraction, mainly those related to the process of measurement (Weyl (1949), pp. 10–11). In order to understand the scope of definitions by abstraction, let us consider a concrete example. Let $\mathfrak{S}$ be the class of all sticks, slats, staffs, beams, masts and similar elongated objects; we wish to introduce the concept of *length* into this class.

 (I) In accordance with Aristotle's theory of abstraction, one would say: let u be an arbitrary element of the class $\mathfrak{S}$ (thus a stick, slat, staff, ...). Now we disregard the thickness, color, weight, material ... of u. Then we obtain an abstract object (u) that has but one characteristic property remaining. This property is designated as the *length* of u.
(II) On the basis of the method of definition by abstraction we say: let R be the relation *equally long*; this relation is defined in the class $\mathfrak{S}$ and it fulfils the properties of an equivalence relation. Let u again be an arbitrary object of $\mathfrak{S}$. Then the above class $\mathcal{C}(u)$ will consist of all objects which are just as long as u.

In this way we define the *length* of u as the class $\mathcal{C}(u)$; it follows that the length of u is the same as the length of v if and only if u is just as long as v.

As it was mentioned above, several mathematical objects are constructed through definitions by abstraction: for the integers and rational numbers, see Landau (1976) and Waismann (2003), Chapters 4 and 5; for examples in physics, see Weyl (1949), Part II, Chapter II.

7.3 Definitions by abstraction and identifiability

Let us consider a statistical model $\mathcal{E}$ given by $\mathcal{E} = \{(S, \mathcal{S}), P^\theta : \theta \in \Theta\}$. In the parameter space Θ, we say that θ_1 and θ_2 are *observationally equivalent*, which is denoted as $\theta_1 \sim \theta_2$, if and only if $P^{\theta_1} = P^{\theta_2}$. The binary relation $\sim$ is actually an equivalence relation and, therefore, induces a partition on the parameter space Θ. The elements of the partition are the equivalent classes $\mathcal{C}(\theta) = \{\theta' \in \Theta : P^\theta = P^{\theta'}\}$ of observationally equivalent parameters.

$\mathcal{C}(\theta)$ are the new objects defined by abstraction: we disregard the parameters associated to the sampling probabilities, obtaining thus a fully description of the observations through *a* probability distribution. Taking into account the equivalence between observations and sampling probabilities (see 350), each object $\mathcal{C}(\theta)$ corresponds to a specific set of data, nothing more. However, when we impose identifiability, these abstract objects—sampling probabilities—reduce to parameters because in such a case $\mathcal{C}(\theta) = \{\theta\}$ for all $\theta \in \Theta$. In other words, the corresponding quotient space $\Theta / \sim$ reduces to the trivial partition of Θ. Thus, a set of data is fully characterized by those parameters through functionals of the sampling distributions. In this way, the parameters acquire a new status: the parameters are specific characteristics of the data.

7.4 The identification problem: A consequence of the construction of a Bayesian model

The construction of a Bayesian model inherently involves an identification problem. As a matter of fact, in Section 2, we introduce the notion of statistical model as the triplet $\mathcal{E} = \{(S, \mathcal{S}), P^\theta : \theta \in \Theta\}$, where the sample space $(S, \mathcal{S})$ is a measurable space, P^θ is the so-called sampling probability defined on the sample space, θ is a parameter and Θ is the parameter space. In order to construct a Bayesian model, we first endow the parameter space Θ with a σ-field $\mathcal{A}$ in such a way that P^θ becomes a transition probability, that is, $P^\theta(\bullet)$ is a probability measure on $(S, \mathcal{S})$ for all $\theta \in \Theta$, and $P^\bullet(X)$ is $\mathcal{A}$-measurable for all $X \in \mathcal{S}$ as a function of θ. If on the measurable space $(\Theta, \mathcal{A})$, we define a probability measure μ (the so-called *prior distribution*), then, by Proposition III-2-1 in Neveu (1964), there exists a unique probability measure Π defined on the product space $(\Theta \times S, \mathcal{A} \otimes \mathcal{S})$ such that

$$\Pi(E \times X) = \int_{\theta \in E} P^\theta(X) \mu(d\theta), \qquad E \in \mathcal{A}, X \in \mathcal{S}. \qquad (7.1)$$

In this way, we obtain a Bayesian model given by the probability space $(\Theta \times S, \mathcal{A} \otimes \mathcal{S}, \Pi)$.

From a modeling perspective, once $\mathcal{A}$ is defined, it is defined on what parameters the interest is focused. Nevertheless, the choice of $\mathcal{A}$ is not unique: it is enough that $\mathcal{A}$ contains the smallest σ-field that makes the mapping $\theta \longmapsto P^\theta(X)$, $X \in \mathcal{S}$, measurable. Such a smallest σ-field is characterized as follows: for each $X \in \mathcal{S}$ fixed, $[P^\bullet(X)]^{-1}(\mathcal{B}_{[0,1]})$, where $\mathcal{B}_{[0,1]}$ is the Borel σ-field on $[0, 1]$, is the smallest σ-field that makes measurable the mapping $\theta \longmapsto P^\theta(X)$. Therefore, the smallest σ-field that makes the mapping $\theta \longmapsto P^\theta(X)$ measurable for all $X \in \mathcal{S}$ is given by

$$\bigvee_{X \in \mathcal{S}} [P^\bullet(X)]^{-1}(\mathcal{B}_{[0,1]}). \tag{7.2}$$

Now, when both the parameter space and the sample space are Borel spaces, it can be proved that

$$\bigvee_{X \in \mathcal{S}} [P^\bullet(X)]^{-1}(\mathcal{B}_{[0,1]}) = \{\{\theta\} : \theta \in \Theta\}; \tag{7.3}$$

for a proof and a more general statement, see Florens, Mouchart and Rolin (1990), Theorem 4.6.18. The reader can recognize that the right side component of equality (7.3) corresponds to the quotient space $\Theta/\sim$ under identifiability. Thus, *there exists one and only one choice of the σ-field $\mathcal{A}$ that is compatible with identifiability*. Furthermore, it can be verified that the σ-field (7.2) is related to the minimal sufficient parameter up to the prior null sets; that is,

$$\overline{\bigvee_{X \in \mathcal{S}} [P^\bullet(X)]^{-1}(\mathcal{B}_{[0,1]})}^\mu = \sigma\{E[s \mid \mathcal{A}] : s \in [\mathcal{S}]^+\}, \tag{7.4}$$

where the upper line denotes the completion of a σ-field with the null sets of the parameter space with respect to the prior distribution μ; for a proof, see Florens, Mouchart and Rolin (1990), Theorem 4.4.12. This last equality shows that the role of the prior distribution in obtaining identifiability is through the 0/1-prior null sets only. In the context of Wechsler, Izbicki and Esteves's (2013) example, the prior null sets are $\{\varnothing, \Theta\}$ only. Therefore, equality (5.1) corresponds to both (7.2) and (7.4), showing that the arguments developed above are examples of a large theory.

Summarizing, the identification problem is related to the specification of the sampling transitions [as already pointed out by Kadane (1974)], in particular to the information provided by the parameters of the model. The problem is overcame when the parameters are endowed with a data-based meaning: this is an effort to be performed by statisticians in collaboration with scientists of different fields. In this respect, Koopmans and Reiersøl (1950)'s words seem to be a relevant ethical requirement:

> Scientific honesty demands that the specification of a model be based on prior knowledge of the phenomenon studied and possibly on criteria of simplicity, but not on the desire for identifiability of characteristics in which the researcher happens to be interested (pp. 169–170).

Appendix: The measurabiliy structure underlying the Bayes' theorem

The objective of this Appendix (adapted from Mouchart (1976)) is twofold: On the one hand, to show that the existence of the posterior probability distribution is not related to (un-)identifiability problems, but to measurability conditions only; on the other hand, to show under which conditions it is still possible to obtain a posterior probability distribution when the prior measure is improper.

A.1 Proper prior case

Let us consider the statistical model $\mathcal{E}$ introduced in Section 7.4. This model is said to be *dominated* if there exists a σ-finite measure m defined on $(S, \mathcal{S})$ such that $P^\theta \ll m$ for all $\theta \in \Theta$—that is, for each $\theta \in \Theta$, P^θ is absolutely continuous with respect to m. By the Radon–Nikodym theorem, this condition is equivalent to the following one: for each $\theta \in \Theta$, P^θ can be represented by a density with respect to m, i.e., for all $\theta \in \Theta$, there exists a $\mathcal{S}$-measurable function $l(\bullet \mid \theta)$ unique up to an m-equivalence such that

$$(\forall X \in \mathcal{S}) \qquad P^\theta(X) = \int_{\omega \in X} l(\omega \mid \theta) m(d\omega). \tag{A.1}$$

We choose a particular version $l(\omega \mid \theta) \in \frac{dP^\theta}{dm}$ and $l(\omega \mid \theta)$ corresponds to the *likelihood function*.

As it was discussed in Section 7.4, the parameter space Θ is endowed with a σ-field $\mathcal{A}$ in such a way that P^θ becomes a transition probability. Once a prior probability μ is defined on the measurable space $(\Theta, \mathcal{A})$, we obtained a probability measure Π defined on the product space $(\Theta \times S, \mathcal{A} \otimes \mathcal{S})$, which is given by (7.1). We write $\Pi = P^\theta \otimes \mu$.

The probability Π allows us to define a *predictive probability P* on $(S, \mathcal{S})$ in the following way:

$$P(X) = \Pi(X \times \Theta) = \int_{\theta \in \Theta} P^\theta(X) \mu(d\theta), \qquad X \in \mathcal{S}. \tag{A.2}$$

If we want to endow the Bayesian model $(\Theta \times S, \mathcal{A} \otimes \mathcal{S}, \Pi)$ with a dominated structure, it seems to be enough to use (A.1). However, as we pointed out in Section 4, it should be ensured that the likelihood function $l(\omega \mid \theta)$ is measurable not only with respect to $\mathcal{S}$ (which is guaranteed by the dominance of P^θ with respect to m for all $\theta \in \Theta$), but also with respect to $\mathcal{A}$:

$$l(\omega \mid \theta) \text{ is } (\mathcal{A} \otimes \mathcal{S})\text{-measurable.} \tag{A.3}$$

If the sample space $(S, \mathcal{S})$ is separable, then condition (A.3) holds; for a proof, see Doob (1953) and Rolin (2007). Under condition (A.3), (A.1) and (7.1) can be combined, obtaining thus

$$\Pi(E \times X) = \int_{\theta \in E} \int_{\omega \in X} l(\omega \mid \theta) m(d\omega) \mu(d\theta), \qquad E \in \mathcal{A}, X \in \mathcal{S}. \tag{A.4}$$

In fact, condition (A.3) ensures that the likelihood function $l(\omega \mid \theta)$ can be integrated with respect to both ω and θ. Moreover, by Fubini's theorem and condition (A.3), we may reverse the order of integration in (A.4), in particular (A.2) may be rewritten as

$$P(X) = \int_{\omega \in X} \int_{\theta \in \Theta} l(\omega \mid \theta) \mu(d\theta) m(d\omega) \qquad X \in \mathcal{S}. \tag{A.5}$$

It follows that $P \ll m$ and, therefore, we may now define a Radon–Nikodym derivative of P with respect to m and choose the following version:

$$g(\omega) \doteq \int_{\theta \in \Theta} l(\omega \mid \theta) \mu(d\theta) \in \frac{dP}{dm}. \tag{A.6}$$

This derivative is called a *predictive density*. Note that for any fixed $C \in \mathcal{A}$, $\Pi(C \times \bullet)$ is a finite measure on the sample space $(S, \mathcal{S})$. Moreover, $\Pi(C \times \bullet) \ll P$ and, therefore, we can define for every $C \in \mathcal{A}$ a Radon–Nikodym derivative $\mu(C \mid \omega)$:

$$\mu(C \mid \omega) \in \frac{d\Pi(C \times \bullet)}{dP}, \qquad C \in \mathcal{A}. \tag{A.7}$$

But from (A.4) it follows that

$$\int_{\theta \in C} l(\omega \mid \theta) \mu(d\theta) \in \frac{d\Pi(C \times \bullet)}{dm}, \qquad C \in \mathcal{A}. \tag{A.8}$$

We can therefore choose the following version of (A.7):

$$\begin{aligned}
\mu(C \mid \omega) &\in \frac{d\Pi(C \times \bullet)}{dm} \Big/ \frac{dP}{dm} \\
&= \frac{\int_{\theta \in C} l(\omega \mid \theta) \mu(d\theta)}{\int_{\theta \in \Theta} l(\omega \mid \theta) \mu(d\theta)} = \frac{1}{g(\omega)} \int_{\theta \in C} l(\omega \mid \theta) \mu(d\theta).
\end{aligned} \tag{A.9}$$

We conclude that for ω fixed, $\mu(\bullet \mid \omega)$ is a probability on $(\Theta, \mathcal{A})$; it is called *posterior probability*. When $g(\omega) = 0$, $\mu(C \mid \omega)$ will arbitrarily be putted equal to 0 for all $C \in \mathcal{A}$. Finally, note that from (A.7), we deduce that $\mu(\bullet \mid \omega) \ll \mu$: a priori null sets are still a posteriori null sets.

Note that, in (A.9), $\mu(\bullet \mid \omega)$ is chosen so as to be a transition probability on $(\mathcal{A} \times S)$. The probability Π defined on $(\Theta \times S, \mathcal{A} \otimes \mathcal{S})$ and constructed as $\Pi = P^\theta \otimes \mu$ has therefore been decomposed as

$$\Pi(E \times X) = \int_{\omega \in X} \mu(E \mid \omega) P(d\omega), \quad E \in \mathcal{A}, X \in \mathcal{S}. \tag{A.10}$$

When the prior probability μ is defined through a *prior density* $p(\theta)$ with respect to a σ-finite measure ν on $(\Theta, \mathcal{A})$, we may write the *posterior density* $p(\theta \mid \omega)$ with respect to ν as

$$p(\theta \mid \omega) = \frac{l(\omega \mid \theta) p(\theta)}{\int_{\theta \in \Theta} l(\omega \mid \theta) p(\theta) \nu(d\theta)} = \frac{l(\omega \mid \theta) p(\theta)}{g(\omega)} \in \frac{d\mu(\bullet \mid \omega)}{d\nu}. \tag{A.11}$$

This is essentially the well-known form of Bayes' theorem.

A.2 Improper prior case

The previous analysis can be extended to the case where the prior μ is an unbounded measure (or improper distribution) instead of a probability. More specifically, let μ be a positive σ-finite measure on $(\Theta, \mathcal{A})$ instead of a probability measure; it is called a *prior measure*.

Repeating Neveu's argument on each element of a countable partition of Θ whose elements have finite μ-measure, we ensure the existence of a unique measure Π defined on $(\Theta \times S, \mathcal{A} \otimes \mathcal{S})$ such that (7.1) holds. This measure is positive and σ-finite. Furthermore, under the measurability condition (A.3), the measure Π can be written as in (A.4). Again a *predictive measure* can be defined as in (A.2). However, this measure will not necessarily be σ-finite. We have then to distinguish two cases:

1. *Case* 1: The predictive measure is σ-finite. In this case, the Radon–Nikodym derivative at the right hand-side of (A.7) is well defined and the predictive density, defined by (A.6), is m-a.e. finite so that the right-hand of (A.9) is again well defined and $\mu(\Theta \mid \omega)$ will always be equal to 1. Summarizing, when the predictive measure is σ-finite, the Bayesian inference can still be constructed in the same way as when the prior measure is, in fact, a probability. In this case, the posterior result will be a posterior probability.
2. *Case* 2: The predictive measure is not σ-finite. In this case, the derivation in (A.7) is not warranted (because the Radon–Nikodym theorem can not be applied) and the predictive density defined in (A.6) will not be m-a.e. finite. It is therefore impossible to decompose Π as in (A.10). However, (A.7) remains valid and the right-hand side of (A.7) defines a transition measure on $S \times \mathcal{A}$: this measure will be infinite when $g(\omega)$ is infinite. However, its utility in a Bayesian framework is questionable because it is not a decomposition with respect to P. Moreover, it is determined only u to a multiplicative constant which corresponds to the choice of a specific dominating measure m on the family of sampling probabilities; this constant depends on the observed ω.

Acknowledgments

References

Barra, J. R. (1981). *Mathematical Basis of Statistics*. New York: Academic Press. MR0628863

Basu, D. (1975). Statistical information and likelihood (with discussion). *Sankyā: The India Journal of Statistics* **37**, 1–71. MR0440747

Basu, D. and Pereira, C. A. B. (1983). Conditional independence in statistics. *Sankyā: The India Journal of Statistics* **45**, 324–337. MR0747602

Beth, E. W. (1950a). *Les Fondements Logiques des Mathématiques*. Paris: Gauthier-Villars. MR0036197

Beth, E. W. (1950b). A propos d'un "Traité de logique". *Methodos* **2**, 258–264.

Beth, E. W. (1965). *Mathematical Thought. An Introduction to the Philosophy of Mathematics*. Dordrecht: D. Reidel Publishing Company. MR0252158

Clifford, P. (1982). Some general comments on nonidentifiability. In *Probability Models and Cancer* (L. Le Cam and J. Neyman, eds.) 81–83. Amsterdam: Noth Holland.

Dawid, P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B* **41**, 1–31. MR0535541

Dellacherie, C. and Meyer, P. A. (1975). *Probabilité et Potentiel I*. Paris: Herman. MR0488194

Doob, J. L. (1953). *Stochastic Process*. New York: Wiley. MR0058896

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–3681.

Fisher, R. A. (1973). *Statistical Methods for Research Workers*. New York: Hafner Publishhing. MR0346954

Florens, J.-P. and Mouchart, M. (1985). A linear theory for noncausality. *Econometrica* **53**, 157–175. MR0770885

Florens, J.-P., Mouchart, M. and Rolin, J.-M. (1985). On two definitions of identification. *Statistics* **16**, 213–218. MR0789064

Florens, J. P., Mouchart, M. and Rolin, J.-M. (1990). *Elements of Bayesian Statistics*. New York: Marcel Dekker. MR1051656

Gabrielsen, A. (1978). Consistency and identifiability. *Journal of Econometrics* **8**, 261–263. MR0547145

Heyting, A. (1971). *Intuitionism. An Introduction*. Amsterdam: North-Holland Publishing Company. MR0075147

Hurwicz, L. (1950). Generalization of the concept of identification. In *Statistical Inference in Dynamic Economic Models, Cowles Commision Research Monograph 11* (T. C. Koopmans, ed.) 238–257. New York: Wiley. MR0038640

Jara, A., Quintana, F. and San Martín, E. (2008). Linear mixed models with skew-elliptical distributions: A Bayesian approach. *Computational Statistics & Data Analysis* **52**, 5033–5045. MR2526212

Kadane, J. (1974). The role of identification in Bayesian theory. *Studies in Bayesian econometrics and statistics* **1**, 175–191. MR0483124

Kim, K. (2011). A strengthening of the caesar problem. *Erkenntnis* **74**, 123–136. MR2824138

Koopmans, T. C. (1949). Identification problems in economic model construction. *Econometrica* **17**, 125–144. MR0031703

Koopmans, T. C. and Reiersøl, O. (1950). The identification of structural characteristics. *The Annals of Mathematical Statistics* **21**, 165–181. MR0039967

Lakatos, I. (1976). *Proofs and Refutations. The Logic of Mathematical Discovery*. Cambridge: Cambridge University Press. MR0479916

Landau, E. (1976). *Foundations of Analysis. The Arithmetic of Whole, Rational, Irrational and Complex Numbers*. Providence, Rhode Island: AMS Chelsea Publishing. MR0038404

Lindley, D. V. (1971). *Bayesian Statistics. A Review*. *CBMS-NSF Regional Conference Series in Applied Mathematics*. Philadelphia. MR0329081

Lorenzen, P. (1969). *Normative Logic and Ethics*. Mannheim-Zürich: Bibliograhisches Institut. MR0290929

Lorenzen, P. (1987). *Constructive Philosophy*. Amherst: The University of Massachusetts Press.

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

Mouchart, M. (1976). A note on Bayes theorem. *Statistica* **36**, 349–357. MR0455164

Neveu, J. (1964). *Bases Mathématiques du Calcul des Probabilités*. Paris: Masson. MR0198504

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

Piaget, J. (1949). *Traité de Logique. Essai de Logistique Opératoire*. Paris: Librairie Armand Colin.

Poirier, D. J. (1998). Revising beliefs in nonidentified models. *Econometric Theory* **14**, 483–509. MR1650041

Rolin, J.-M. (2007). Bayesian Consistency of Density Estimators. Belgium. Discussion Paper DP0705, Institut de Statistique, Université catholique de Louvain, Belgium.

Ross, S. (2009). *A First Course in Probability*, 8th ed. Upper Saddle River: NJ.

San Martín, E. and González, J. (2010). Bayesian identifiability: Contributions to an inconclusive debate. *Chilean Journal of Statistics* **1**, 69–91. MR2756120

San Martín, E., Jara, A., Rolin, J.-M. and Mouchart, M. (2011). On the Bayesian nonparametric generalization of IRT-type models. *Psychometrika* **76**, 385–409. MR2823006

San Martín, E. and Quintana, F. (2002). Consistency and identifiability revisited. *Brazilian Journal of Probability and Statistics* **16**, 99–106. MR1962497

San Martín, E., Rolin, J.-M. and Castro, M. (2013). Identification of the 1PL model with guessing parameter: Parametric and semi-parametric results. *Psychometrika* **78**, 341–379. MR3046223

Waismann, F. (2003). *Introduction to Mathematical Thinking. The Formation of Concepts in Modern Mathematics*. New York: Dover Publications. MR1972880

Wechsler, S., Izbicki, R. and Esteves, L. G. (2013). A Bayesian look at nonidentifiability: A simple example. *The American Statistician* **67**, 90–93. MR3303596

Weyl, H. (1949). *Philosophy of Mathematics and Natural Sciences*. Princeton: Princeton University Press. MR2489603

Ernesto San Martín
Faculty of Mathematics
Pontificia Universidad Católica de Chile
Vucuña Mackenna 4860
Macul, Santiago
Chile
E-mail: esanmart@mat.puc.cl