Bayesian Identification and Partial Identification

Jean-Pierre Florens†  Anna Simoni‡

Toulouse School of Economics  Department of Decision Sciences
Université de Toulouse 1 - Capitole  Università Bocconi and IGIER

First draft: February 2011 - Last changes: November 2011

Abstract

The problem of identification and partial identification in econometrics is considered from a Bayesian point of view. We carry out a comprehensive analysis of this topic and stress the links existing with classical identification. We first analyze how the prior distribution has to be specified in order to identify a parameter that is not identified in the sampling model. The focus is on infinite-dimensional parameters. Then, for a parameter that is partially identified, we introduce the concept of prior and posterior capacity functional of the set parameter and propose a completely nonparametric Bayesian approach. Finally, we show that parameters that are non-identified or partially-identified in the complete model can be identified in the marginal model, where the marginalization is over the identified parameter. The paper provides several examples that show how to implement in practise our techniques.

Key words: Sufficiency, Minimal Sufficiency, Exact Estimability, Set identification, Dirichlet Process, Capacity functional.

JEL code: C11; C10; C14

1 Introduction

This paper studies Bayesian identification and partial identification in econometric models. A Bayesian approach is appealing when a non-identified parameter can be identified through the specification of a suitable prior distribution. We give new results about identification by the prior distribution of non-identified models with infinite-dimensional parameters. On top of that, Bayesian theory can be convenient when the model’s parameter of interest is partially-identified – which means that the parameter is a set. For that case, we build up a new Bayesian theory for set-parameters and define their prior and posterior distribution in terms of capacity functionals. Furthermore, a Bayesian approach provides a distribution

---

*We acknowledge helpful comments from Andrew Chesher, Frank Kleibergen and participants to the 2011-Econometric Summer Conference in Yale and ESEM-2011 in Oslo. The usual disclaimer applies and all errors remain ours.

†Toulouse School of Economics - 21, allée de Brienne - 31000 Toulouse (France). Email: florens@cict.fr

‡Corresponding author. Università Bocconi - Department of Decision Sciences - via Röntgen, 1 - 20136 Milano (Italy). Email: anna.simoni@unibocconi.it
over the identified-set.

The identification issue is a very important topic in econometrics and is at the core of recent developments in the econometric literature - see Manski [1990], Manski and Tamer [2002], Chesher [2003], Severini and Tripathi [2006], Chernozhukov et al. [2007], Liao and Jiang [2010], D'Haultfoeuille [2011], Bontemps et al. [2011], Beresteanu et al. [2011], Galichon and Henry [2011], Kitagawa [2011a,b], Kleibergen and Mavroeidis [2011] to name only a few. A Bayesian analysis is motivated by the extensive use of Bayesian methodology in many fields of econometric - e.g. in macroeconomics, see Primiceri [2005], Sims and Zha [2006] and An and Schorfheide [2007], in simultaneous equation models, see e.g. Drèze [1974], Drèze [1976], Maddala [1976], Kleibergen and van Dijk [1998], Kleibergen and Zivot [2003] and Hoogerheide et al. [2007], or in the GMM literature, see e.g. Florens and Rolin [1994] and Kitamura and Otsu [2011] and references therein. In this Bayesian context our paper recalls the links existing between Bayesian and frequentist identification. This is an old question in econometrics, see e.g. Lindley [1971]. Kadane [1974] observed that identification is a property of the likelihood function and that is the same whether considered from Bayesian or classical perspectives. However, there are situations where the introduction of a prior distribution can make a parameter which is non-identified in frequentist theory identified in Bayesian theory. We explain further this idea in Section 3.

Roughly speaking, a parameter is non-identified in Bayesian theory if its prior distribution is not revised through the information brought by the data so that the conditional posterior and conditional prior distributions are the same. In other words, in a Bayesian experiment, the observed sample bears information only on a minimal sufficient parameter; conditionally on this parameter, the Bayesian experiment is completely non-informative.

The minimal sufficient parameter is the smallest $\sigma$-field on the parameter space that makes the sampling probabilities measurable and is the identified $\sigma$-field. It is then natural to ask why should one introduce a $\sigma$-field larger than the minimal sufficient parameter. In nonexperimental fields redundant parametrization is usually introduced either as an early stage of model building or as a support for relevant prior information or because the parameter of interest (making e.g. the loss function measurable) is larger than the minimal sufficient parameter. In experimental fields, it may be the case that the experimental design will not provide information on all the parameters of a theoretically relevant model, see Florens et al. [1990]. Summarizing, we say that a model is non-identified if the parametrization is redundant. Otherwise, the model is identified.

We define in this paper three concepts of identification: sampling or frequentist identification, measurable identification and Bayesian identification. The difference among these

---

1 General theoretical concepts of identification in sampling and Bayesian framework were presented in some previous works of the first author jointly with M. Mouchart and J-M. Rolin - see in particular Florens et al. [1985], Florens and Mouchart [1986] and Florens et al. [1990]. These concepts have been applied in different other frameworks jointly with E. Scheihing and E. San Martin, see e.g. Mouchart et al. [1999], San Martín and González [2010]
concepts relies on the different levels of specification of the model. The first concept of identification is defined without the introduction of a \(\sigma\)-field associated with the parameter space. Such a measurable structure is necessary for the other two concepts of identification. In addition, the Bayesian identification concept requires the definition of a unique joint probability measure over the sample and the parameters. Examples are provided in the paper in order to show how these three concepts can differ in practice.

As Lindley [1971] remarked, the problem of non-identification causes no real difficulty in the Bayesian approach. As long as a proper prior distribution can be defined, the posterior distribution will be well-defined. On the contrary, in a non-identified model if the prior is non-informative the posterior distribution of the non-identified parameter may not exist because, roughly speaking, the denominator in the Bayes theorem is unbounded.

A Bayesian approach has several advantages even when the model results to be non-identified from a Bayesian point of view. (i) The posterior distribution always exists (under mild assumptions on the spaces of reference and if the prior distribution is proper) even in a non-identified model. Therefore, in a “decision problem setting” we can always take decisions as, by using a proper prior distribution, the resulting posterior distribution may not be flat within the identified region. (ii) When the parameter of the econometric model is multidimensional, with some components that are identified and some others are not, we can learn from the data something also about the non-identified parameters. In fact, the information contained in the data about the non-identified parameters will pass through the identified parameters so that the conditional prior distribution of the non-identified parameters, given the identified ones, will change. (iii) If some prior information about the parameters of the model is available, e.g. in the forms of constraints, this can be incorporated in a proper prior distribution and it will help in identification. For instance, if we have set identification, the prior density can be informative only about a subset of the identified set. (iv) Methods like MCMC allow to easily draw from the posterior distribution even when this is not available in closed-form. Thus, we can analyze e.g. the marginal posterior distribution of a partially-identified parameter; we can approximate moments of the posterior distribution of a parameter and/or of the identified region.

Our paper contributes to the literature on identification in three ways. First, we provide the general theory about how to specify the prior distribution in order to identify, in a Bayesian sense, finite- and infinite-dimensional parameters that are not identified in the frequentist sense. In particular, we analyze Gaussian processes and Dirichlet processes priors. Second, we analyze partially-identified models where the identified parameter is a set. For these models we build up a new Bayesian nonparametric approach for constructing prior and posterior distributions for set-parameters. These distributions are defined in terms of prior and posterior capacity functionals and are based on Dirichlet process priors.
They are easy to compute either by simulation or in closed-form. The posterior capacity functional results to be an appealing tool to build estimators and credible set for the set-parameter. Third, we show that, when the parameter of interest lacks identification in the full original model, it can be identified in the marginal model where the marginalization is carried out with respect to the identified parameter. The marginal model depends on the specification of the prior distribution of the identified parameter. To complement our theoretical approach we develop many examples and simulations that show how our method can be directly applied.

The paper is organized as follows. In Section 2 we introduce and recall preliminary definitions and explain the difference among the three concepts of identification: sampling identification, measurable identification and Bayesian identification. In Section 3 we explore situations where parameters that are non-identified in the sampling model can be Bayesian identified if the prior distribution is suitably specified. We propose in Section 4 a Bayesian nonparametric approach for computing the posterior distribution of the identified set. Finally, in Section 5 we analyze the marginal model.

2 Some General Definition

2.1 Sampling Theory

We first recall some basic definitions in the non-Bayesian framework, called in the paper sampling theory approach. A statistical model is usually defined by a sampling space \( X \) provided with a \( \sigma \)-field \( \mathcal{X} \) and by a collection of probabilities on this space. In a small sample approach we observe the realization \( x \in X \) of a random variable – for example an \( iid \) sample \( x = (x_1, \ldots, x_n) \) – taking values in a measurable space where the sample size is not made explicit or is kept fixed. In an asymptotic approach, \( \mathcal{X} \) is provided by a filtration \( (\mathcal{X}_n)_{n\geq 1} \), where \( \mathcal{X}_n \) represents the information contained by a sample of size \( n \), we denote by \( \mathcal{X}_\infty = \bigvee_{n\geq 1} \mathcal{X}_n \) the \( \sigma \)-field generated by \( \bigcup_{n\geq 1} \mathcal{X}_n \) and write \( \mathcal{X}_n \uparrow \mathcal{X}_\infty \).

The collection of probabilities defined on \( (X, \mathcal{X}) \) is in general indexed by a parameter \( \theta \in \Theta \), which may be a functional parameter (but we use the name parameter even in that case) and this collection is noted \( (P^\theta)_{\theta \in \Theta} \). In summary, a statistical model \( \mathcal{E}_s \) is defined by the elements \( \{\Theta, (X, \mathcal{X}), (P^\theta)_{\theta \in \Theta}\} \). The notation \( \mathcal{E}_s \) is for the sampling statistical model.

In that framework two parameters \( \theta_1 \) and \( \theta_2 \) are said to be observationally equivalent, in symbol \( \theta_1 \sim \theta_2 \), if \( P^{\theta_1} = P^{\theta_2} \). This relation defines an equivalence relation and we denote by \( \hat{\Theta} \) the quotient space \( \Theta / \sim \), i.e. the elements of \( \hat{\Theta} \) are the equivalence classes over \( \Theta \) by \( \sim \). We now define the concept of sampling identification.

**Definition 1.** A real valued function \( a(\theta) \) is identified if \( \theta_1 \sim \theta_2 \) implies \( a(\theta_1) = a(\theta_2) \). The sampling model \( \mathcal{E}_s \) is identified if and only if any real valued function defined on \( \Theta \) is
This definition is equivalent to say that $\theta_1 \sim \theta_2$ implies $\theta_1 = \theta_2$ or to say that any equivalence class is reduced to a singleton. This turns to be equivalent to saying that the sampling model $E_s$ is identified if the mapping $\theta \to P^\theta$ is injective.

Let us underline that for any sampling statistical model $E_s$ there exists a canonical identified model $\tilde{E}_s$ defined as the sampling statistical model with parameter space $\tilde{\Theta}$:

$$\tilde{E}_s = \{\tilde{\Theta}, (X, X), (P^\theta)_{\tilde{\theta} \in \tilde{\Theta}}\}$$

where $P^\tilde{\theta} = P^\theta$ for any $\theta \sim \tilde{\theta}$. Remember that $\tilde{\theta}$ is a set and $\tilde{\Theta}$ is a set of sets of parameters. $\tilde{E}_s$ is also the set identified statistical model associated with $E_s$.

As it is sometimes difficult to work with models where the "parameters" are sets, one may construct an identified sampling model by selecting a single element in each equivalence class. Let us call section a function $\sigma$ such that

$$\sigma : \tilde{\Theta} \to \Theta$$

such that $\sigma(\tilde{\theta}) \in \tilde{\theta}$ and $\sigma(\tilde{\theta}) \in \Theta$.

In that case one may define the identified sampling model

$$E_{s, \sigma} = \{\Theta_\sigma, (X, X), (P^\theta)_{\theta \in \Theta_\sigma}\}$$

where $\Theta_\sigma$ is the image of $\tilde{\Theta}$ by $\sigma$ and $P_\sigma = P_{\sigma(\tilde{\theta})}$. In general, it is not possible to devise constructive rules for selecting elements in $\Theta$ for each $\tilde{\theta} \in \tilde{\Theta}$. The existence of $\Theta_\sigma$ cannot be proved by using axioms from set theory if we do not have some structure on the parameter space but must be asserted as additional axiom called axiom of choice, see e.g. Kolmogorov and Fomin [1975].

In the sampling theory statistics, a topological structure is needed in the parameter space in order to define statistical decision rules based on convergence, risk or loss function. This canonical topological structure is defined on $\Theta$ and then may be carried on $\tilde{\Theta}$. Let $\rho$ be the canonical application $\theta \to \rho(\theta) = \tilde{\theta}$ mapping $\Theta$ onto $\tilde{\Theta}$. The natural topological structure is the smallest one for which $\rho$ is continuous, i.e. $\rho^{-1}(\tilde{O})$ is open in $\Theta$ whenever $\tilde{O} \subset \tilde{\Theta}$ is open. We will not detail in this paper the topological aspects of set identification.

Note however that in that context it is natural to look for continuous sections $\sigma$ (or to bicontinuous bijections between $\tilde{\Theta}$ and $\Theta_\sigma$) and the existence of such sections is a difficult topological question (see Husmoller [1994] and the appendix to chapter 3 in Dellacherie and Meyer [1975]).

**Remark 2.1.** A natural construction of sampling model which may lead to identification questions is the following. The econometrician first considers a sampling model perfectly identified $(\Theta, (X, X), (P^\theta)_{\theta \in \Theta})$ but she completes her specification by considering another parameter $\gamma \in \Gamma$ (also possibly functional). This parameter $\gamma$ is the natural parameter arising in economic models and is related to $\theta$. The relation between $\theta$ and $\gamma$ may be of the form $A(\theta, \gamma) = 0$ or $A(\theta, \gamma) \geq 0$ or more generally $A(\theta, \gamma) \in A_0$, $A_0 \in \Phi$ where
Φ is a suitable defined space. The parameter γ may be identified or not depending on the relation $A$.

For example, the GMM model in an iid sampling case is characterized by a sampling space $X^n$ and a parameter $\theta = F$ where $F$ is the cumulative distribution function (cdf in the following) of any observation $x_i$. The parameter $\gamma$ is then characterized through a relation

$$E^F(h(x, \gamma)) = 0$$

(1)

where $h$ is a known function. The parameter $\gamma$ is identified if a unique solution to (1) exists.

Condition (1) may be extended into

$$E^F(h(x, \gamma)) \geq 0$$

which in general defines a set of $\gamma$ solutions of this inequality.

### 2.2 Measurable Statistical Model

A preliminary step for the construction of a Bayesian model consists in defining a measurable statistical model. This is the sampling model $E_m = \{(\Theta, A), (X, \mathcal{X}), (P^\theta)_{\theta \in \Theta}\}$ where we use the notation previously introduced and where $A$ is a σ-field on $\Theta$ such that $P^\theta$ is a transition probability, i.e. (i) $\forall \theta \in \Theta$, $P^\theta$ is a probability measure on $(X, \mathcal{X})$, (ii) $\forall E \in \mathcal{X}$, $P^\theta(E)$ is a measurable function on $(\Theta, A)$. A σ-field represents an information structure and the parameter of interest is naturally introduced as a σ-field $A$ that makes both the transition probability $P^\theta(E)$, $E \in \mathcal{X}$, and the loss function of the underlying decision model $A$-measurable. This structure is motivated by the possibility to introduce a given distribution on $\Theta$. We start here to discuss identification issues in that framework without the specification of this prior distribution, which will be specified in the next section.

**Definition 2.** A sub-σ-field $B$ of $A$ is said to be sufficient if $P^\theta(E)$ is $B$-measurable for any $E \in \mathcal{X}$.

The following proposition introduces another concept of identification of $E_m$, called measurable identification.

**Proposition 1.** There exists a minimal sufficient σ-field $A_*$ equal to the intersection of all the sufficient σ-fields. The measurable model $E_m$ is identified if and only if $A_* = A$.

**Definition 3.** A real valued function $a$ defined on $(\Theta, A)$ is measurably identified if it is $A_*$-measurable.

The notion of sampling and measurable identification are identical if $\Theta$ is a measurable subset of $\mathbb{R}^k$ provided with the Borelian σ-field and if $X$ is included in $\mathbb{R}^n$ and also provided with the Borelian σ-field. Actually, this equivalence is true more generally and requires
that $\Theta$ “looks like” a Borelian of $\mathbb{R}$ and that the $\sigma$-field on $X$ is separable or equivalently generated by a countable family of subsets. The general results have been given in Theorems 0.2.16 and 4.6.18 in Florens et al. [1990] and follow from the Blackwell theorem (see Dellacherie and Meyer [1975]). The link between sampling and measurable identification requires some mathematical technicalities. It may be proven the following result:

**Theorem 1.** Let us assume that $(\Theta, \mathcal{A})$ is a Souslin space, i.e. $\exists B \subset \mathbb{R}$ analytic such that there exists a bimeasurable bijection between $(\Theta, \mathcal{A})$ and $(B, B \cap \mathcal{B})$, where $B \cap \mathcal{B}$ denotes the restriction to $B$ of the Borel $\sigma$-fields of $\mathbb{R}$. Then,

- a real valued function $a$ defined on $(\Theta, \mathcal{A})$ is measurably identified if and only if $a$ is constant on the equivalence class $\tilde{\Theta}$;
- if $X$ is separable, a model is sampling identified if and only if it is measurably identified.

Let us recall that an analytic set on $\mathbb{R}$ is the projection on $\mathbb{R}$ of a Borelian set in $\mathbb{R}^2$. In particular, all Borelian sets are analytic. The property that $(\Theta, \mathcal{A})$ is a Souslin space is clearly true for finite dimensional parameter spaces or more generally for Polish spaces, which include the $L^2$ spaces on a real measure space. However, this property is not necessarily satisfied when $\theta$ is a functional parameter in a “big” space.

### 2.3 Bayesian Identification

A Bayesian model is defined by a measurable model provided with a measure on $(\Theta, \mathcal{A})$. This measure is informative if it is a probability or non-informative if the measure of $\Theta$ is infinite. For simplicity we will essentially consider a probability measure denoted by $\mu$ which is called prior distribution. Then, $\mu$ and $P^\theta$ generate a unique measure on $(\Theta \times X, \mathcal{A} \otimes \mathcal{X})$ denoted by $\Pi$. The marginal probability measure on $(X, \mathcal{X})$ is denoted by $P$ and is called the predictive probability. We assume that there exists a regular version of the conditional probability on $\Theta$ given $X$, i.e. a transition probability $\mu^x(\cdot)$ which is a probability on $(\Theta, \mathcal{A})$ for any $x$ and such that $\mu^x(A)$ is $\mathcal{X}$-measurable for any $A \in \mathcal{A}$. This probability measure is called the posterior probability. In the following, the notation $\mu(\cdot|X)$ can be used instead of $\mu^x(\cdot)$; we use the same notation $\mu(\theta|x)$ (resp. $\mu(\theta)$) indistinctly for the posterior (resp. the prior) distribution and the posterior (resp. the prior) density function with respect to the Lebesgue measure. The sampling (resp. predictive) density function with respect to the Lebesgue measure is denoted by $p(x|\theta)$ (resp. $p(x)$). The product of $P$ and $\mu^x$ is also equal to $\Pi$. In summary:

$$\Pi(A \times E) = \int_A P^\theta(E) \mu(d\theta) = \int_E \mu^x(A) P(dx), \quad A \in \mathcal{A}, \ E \in \mathcal{X}$$

or

$$\Pi = \mu \otimes P^\theta = P \otimes \mu^x$$
and the Bayesian model is defined by the following probability space

\[ \mathcal{E}_b = \{ \Theta \times X, \mathcal{A} \otimes \mathcal{X}, \Pi \} \]

**Definition 4.** A sub \( \sigma \)-field \( \mathcal{B} \) of \( \mathcal{A} \) is sufficient in the Bayesian model if and only if \( \mathcal{A} \| \mathcal{X} | \mathcal{B} \).

The conditional independence of the previous definition has two equivalent characterizations:

(i) \( \forall t : \mathcal{X} \to \mathbb{R} \),

\[ E(\mathcal{E}(t(x)|\mathcal{A})) = E(\mathcal{E}(t(x)|\mathcal{B})) \text{ a.s.} \quad (2) \]

provided that the conditional expectations exist;

(ii) \( \forall a : \mathcal{A} \to \mathbb{R} \),

\[ E(\mathcal{E}(a(\theta)|\mathcal{X} \vee \mathcal{B})) = E(\mathcal{E}(a(\theta)|\mathcal{B})) \text{ a.s.} \quad (3) \]

We point out that all the a.s. are related to the joint probability \( \Pi \). The first characterization (2) weakens the concept of sufficient \( \sigma \)-field because the property is only required almost surely with respect to the prior probability. The second characterization (3) says that the conditional prior and posterior distributions are a.s.-identical given a sufficient \( \sigma \)-field. If \( \mathcal{A} \) is generated by a function \( a \) and \( \mathcal{B} \) by a function \( b \) then (2) means that the likelihood functions \( p(x|a) \) and \( p(x|b) \) are a.s. equal. Equivalently, (3) means that the posterior distribution \( \mu(a|x, b) \) of \( a \) is a.s. equal to \( \mu(a|b) \).

It may be proven that there exists a minimal Bayesian sufficient \( \sigma \)-field in \( \mathcal{A} \), denoted by \( \mathcal{A}_\mu^\ast \) and defined below.

**Definition 5.** A minimal Bayesian sufficient \( \sigma \)-field \( \mathcal{A}_\mu^\ast \) (or \( \mu \)-a.s. minimal sufficient \( \sigma \)-field) is the \( \sigma \)-field generated by all the versions of \( E(t(x)|\mathcal{A}) \) for any positive \( \mathcal{X} \)-measurable function \( t : \mathcal{X} \to \mathbb{R}_+ \), provided the expectation exists. This \( \sigma \)-field \( \mathcal{A}_\mu^\ast \) is called the projection of \( \mathcal{X} \) on \( \mathcal{A} \) and is also denoted \( \mathcal{A}_\mathcal{X} \).

In an equivalent way, we can say that \( \mathcal{A}_\mu^\ast \) is the smallest \( \sigma \)-field which makes the sampling probabilities \( \mathcal{P}(\mathcal{E}(t(x)|\mathcal{A}) \text{ measurable}, \forall E \in \mathcal{X} \), completed by the null sets of \( \mathcal{A} \) with respect to \( \mu \). It may be easily verified that

\[ \overline{\mathcal{A}} \cap \mathcal{A} = \mathcal{A}_\mathcal{X} \equiv \mathcal{A}_\mu^\ast \]

where \( \overline{\mathcal{A}} \) is the \( \sigma \)-field generated by \( \mathcal{A} \), and all the null sets of \( \mathcal{A} \vee \mathcal{X} \) and \( \overline{\mathcal{A}} \cap \mathcal{A} \) is the \( \sigma \)-field generated by \( \mathcal{A} \), and all the null sets of \( \mathcal{A} \) with respect to \( \mu \).

**Definition 6.** The Bayesian model \( \mathcal{E}_b \) is Bayesian identified if and only if \( \mathcal{A}_\mathcal{X} = \mathcal{A} \).
This definition means that \( \mathcal{A}_* \) is almost surely equal to \( \mathcal{A} \), i.e. \( \mathcal{A}_* \cap \mathcal{A} = \mathcal{A} \) or: \( \forall A \in \mathcal{A}, \exists B \in \mathcal{A}_* \) such that \( \mu(A \triangle B) = 0 \). This property shows that Bayesian identification is actually an almost sure measurable identification with respect to the prior. We also have the following characterization of Bayesian identification, see Theorem 4.6.21 in Florens et al. [1990].

**Theorem 2.** If \((\Theta, \mathcal{A})\) is a Souslin space and if \(\mathcal{X}\) is separable the Bayesian model is identified if and only if \(\exists \Theta_0 \subset \mathcal{A}\) such that (i) \(\mu(\Theta - \Theta_0) = 1\) and (ii) the sampling model restricted to \(\Theta - \Theta_0\) is identified, i.e. \(\theta \rightarrow P^\theta\) is injective on \(\Theta - \Theta_0\).

Before concluding this section we want to stress more links between the three concepts of identification that we have seen. (i) Measurable identification implies Bayesian identification for any \(\mu\). (ii) Measurable identification implies sampling identification if and only if \(\mathcal{A}\) is separating, i.e. all the atoms of \(\mathcal{A}\) are singletons.

**Remark 2.2.** There are two basic cases which lead to non-identified models. (i) We have seen in Remark 2.1 that we may start from an identified model parameterized by \(\theta\) and add a supplementary parameter \(\gamma\) which is the parameter of interest and is linked to \(\theta\). This parameter \(\gamma\) may be non-identified. In this case the prior distribution is naturally specified by the marginal prior on \(\theta\) and the conditional prior on \(\gamma\) given \(\theta\) (which incorporates the link between \(\theta\) and \(\gamma\), see section 5): \(x|\theta, \gamma \sim P^\theta, \theta \sim \mu\) and \(\gamma|\theta \sim \nu^\theta\). (ii) Another canonical structure arises when we start with an identified model but we specify the prior on \(\theta\) as a function of an hyperparameter (also called latent variable) \(\gamma\). In that case the prior is naturally on \(\theta|\gamma\) completed by a prior on \(\gamma\). Remark that this natural prior decomposition is not identical to the decomposition in the previous case. The parameter \(\gamma\) is non-identified in the sampling model but it can be Bayesian identified by the prior. This second case will be considered in Section 3.

### 2.4 Identification and Bayesian consistency

There exists an important link between the concepts of identification and the concept of exact estimability and we want to present briefly the relations between these two important ideas. We reduce our presentation to the Bayesian models.

**Definition 7.** In a Bayesian model \(\mathcal{E}_b\), let \(\mathcal{B}\) be a sub-\(\sigma\)-field of \(\mathcal{A}\). Then, \(\mathcal{B}\) is exactly estimable if and only if \(\mathcal{B} \subset \overline{\mathcal{X}}\) where \(\overline{\mathcal{X}}\) is the \(\sigma\)-field generated by \(\mathcal{X}\) and all the null sets of the product space \(\Theta \times \mathcal{X}\) with respect to \(\Pi\).

The inclusion \(\mathcal{B} \subset \overline{\mathcal{X}}\) means that for any \(\mathcal{B}\)-measurable positive function \(a\) of \(\theta\), the posterior expectation \(\mathbf{E}(a|\mathcal{X}) = a\) \(\Pi\)-a.s., provided that the conditional expectation exists. This
means that we know \( a \) a-posteriori almost surely, \textit{i.e.} after the observation of the sample we know \( a \) perfectly. Thus, any sub-\( \sigma \)-field of \( \mathcal{A} \cap \overline{X} \) is exactly estimable. Equivalently, we can say that \( \mathcal{B} \) is exactly estimable if \( \mathcal{B} \parallel \mathcal{A}\mid \overline{X} \).

**Proposition 2.** Any exactly estimable parameter \( \mathcal{B} \subset \mathcal{A} \cap \overline{X} \) is identified.

The concept of exact estimability has a particular interest in asymptotic models because, by using the martingale convergence theorem, we have \( \mathbb{E}(a|X_n) \to \mathbb{E}(a|X_\infty) \) a.s. for any function \( a \) of \( \theta \). Then, if \( \mathcal{B} \subset \mathcal{X}_\infty \), we have \( \mathbb{E}(a|X_n) \to a(\theta) \) \( \mathbb{P}^\theta \)-a.s., is not obvious and may not be verified, see Diaconis and Freedman \[1986\] and Florens and Simoni \[2010\]. We introduce now the concept of \textit{asymptotically exact estimability}:

**Definition 8.** Let \( \mathcal{E}_\infty \) be the sequential Bayesian experiment defined as \( \mathcal{E}_\infty = (\Theta \times X, \mathcal{A} \otimes \mathcal{X}, \Pi, X_n \uparrow X_\infty) \) and \( \mathcal{B} \) be a sub-\( \sigma \)-field of \( \mathcal{A} \) such that \( \mathcal{B} \subset X_\infty \). Then, \( \mathcal{B} \) is said to be asymptotically exactly estimable in \( \mathcal{E}_\infty \).

This definition means that \( \mathcal{B} \subset \mathcal{A} \) is asymptotically exactly estimable if, for every integrable real random variable \( a \) defined on \( \mathcal{B} \), \( \mathbb{E}(a|X_n) \to a \) \( \Pi \)-a.s. This is the concept of Bayesian consistency discussed above, for which the convergence must be taken with respect the joint probability distribution \( \Pi \).

In a Souslin space, asymptotic exact estimability means that the posterior distribution is asymptotically a Dirac measure on a function of the sample, see Theorem 4.8.3 in Florens \textit{et al.} \[1990\]. The intuition is that, if \( \mathbb{E}(a|X_n) \to a(\theta) \) \( \Pi \)-a.s., then also \( \mathbb{E}(a^2|X_n) \to a^2 \) \( \Pi \)-a.s. which implies that the posterior variance converges to 0. This clearly explains that the posterior distribution concentrates around the \( \Pi \)-a.s. limit of its mean. More generally, a integrable real random variable \( a \) defined on \( \mathcal{B} \) is asymptotically exactly estimable if there exists a strongly consistent sequence of estimators of \( a \), \textit{i.e.} if there exists a set of random variables \( t_n \) defined on \( \mathcal{X} \), \( n \in \mathbb{N} \), such that \( t_n \to a \) \( \Pi \)-a.s. In sampling theory, a necessary condition for the existence of such a sequence is the identifiability of the parameter.

The main result is the Doob theorem, see Doob \[1949\] or Florens \textit{et al.} \[1990\]:

**Theorem 3.** In an i.i.d. model, the minimal sufficient \( \sigma \)-field of \( \mathcal{A} \) is exactly estimable.

Reciprocally, we have the following theorem.

**Theorem 4.** Let \( \{(\Theta, \mathcal{A}), (X, \mathcal{X}), \Pi\} \) be a Bayesian model such that the Bayesian minimal sufficient \( \sigma \)-field \( \mathcal{A}\mathcal{X} \) is exactly estimable, \textit{i.e.} \( \mathcal{A}\mathcal{X} \subset \overline{X} \). Then, \( \mathcal{B} \subset \mathcal{A} \) is exactly estimable if and only if \( \mathcal{B} \) is a.s. identified, \textit{i.e.} \( \mathcal{B} \subset \mathcal{A} \mathcal{X} \).
Proof. Remember that \( \mathcal{AX} = \overline{A} \cap A \). First, assume \( B \) identified, then \( B \subset \mathcal{AX} \) and, as \( \mathcal{AX} \subset \overline{X} \), we have \( B \subset \overline{X} \) and \( B \) is exactly estimable.

Reciprocally, let \( B \) be exactly estimable. As \( \mathcal{AX} \) is the minimal sufficient \( \sigma \)-field we have

\[
\mathcal{X} \parallel B | \mathcal{AX}.
\]

This conditional independence implies

\[
\overline{X} \cap B \subset \overline{AX}
\]

by Corollary 2.2.9 in Florens et al. [1990]. In particular, \( B \subset \overline{AX} \) and \( B \) is identified.

\[\square\]

3 Identification by the prior distribution

We want to illustrate in this section the possibility that a model satisfies Bayesian identification even if it is not sampling or measurable identified. This situation may occur intuitively when the function \( \theta \mapsto P^\theta \) is not-injective but the prior put a mass zero on the non-identified component of the model (as it has been shown in Theorem 2). Therefore, identification by the prior distribution refers to the fact that a suitable choice of the prior distribution may solve the problem of lack of injectivity of the likelihood function. This situation is fully artificial in the finite-dimensional parameter case but is not so trivial in non-parametric cases or in infinite-dimensional sample.

Consider the case where the parameter space \( (\Theta, \mathcal{A}) \) contains two sub-\( \sigma \)-fields \( \mathcal{B} \) and \( \mathcal{G} \), where \( \mathcal{B} \) is the \( \mu \)-a.s. minimal sufficient \( \sigma \)-field in the parameter space. Intuitively, the sampling distribution only depends on \( \mathcal{B} \). The \( \sigma \)-field \( \mathcal{G} \) is not identified in the sense that \( \mathcal{G} \) is not included in \( \mathcal{B} \).

In that case, recall that in terms of probability density properties one has that \( p(x|b, \gamma) = p(x|b) \) a.s. where \( p \) is the data density and \( b \in (B, \mathcal{B}), \gamma \in (\Gamma, \mathcal{G}) \). This is equivalent to \( \mu(\gamma|x, b) = \mu(\gamma|b) \) a.s. Note that the existence of the densities is absolutely not necessary and that a density does not exist in general in the case of functional data and/or parameters. An interesting situation arises when the \( \sigma \)-field \( \mathcal{G} \) is not identified only for particular values taken on by the parameter \( \mathcal{B} \), that is, \( p(x|b = \bar{b}, \gamma) = p(x|b = \bar{b}) \) for some value \( \bar{b} \). This is known as local non-identification and has been widely addressed in Drèze [1974], Kleibergen and van Dijk [1998], Hoogerheide et al. [2007] and Kleibergen and Mavroeidis [2011]. We do not analyze this situation further in the paper.

In this section we are interested essentially in models where the prior specification is naturally given as the product of the conditional distribution of \( \mathcal{B} \) given \( \mathcal{G} \) and the marginal on \( \mathcal{G} \). In this case, \( \mathcal{G} \) is interpreted as a parameter of the prior on \( \mathcal{B} \), usually called an hyperparameter, see e.g. Berger [1985], or latent variable. By construction, the hyperparameters are not identified. However, we may look at the marginal model on \( \Gamma \times X \) by integrating out
the identified parameters using the prior on $\mathcal{B}$ given $\mathcal{G}$. This approach will be considered in section 5.

The main argument which is useful in order to analyze Bayesian identification of this type of models is the following.

**Proposition 3.** Let us consider the Bayesian model $\mathcal{E}_b = \{\Theta \times X, \mathcal{A} \otimes \mathcal{X}, \Pi\}$ and a sub-$\sigma$-field $\mathcal{B} \subset \mathcal{A}$ such that $\mathcal{B} = AX$. Then, any sub-$\sigma$-field $\mathcal{G} \subset \mathcal{A}$ is identified if and only if $\mathcal{G} \subset \mathcal{B}$, $\mu$-a.s. or, equivalently, if any $\mathcal{G}$-measurable function is $\mu$-a.s. equal to a $\mathcal{B}$-measurable function.

Notice that the a.s. in the proposition are related to the prior and that $\mathcal{B}$ is the minimal sufficient $\sigma$-field. The proposition states that any sub-$\sigma$-field $\mathcal{G} \subset \mathcal{A}$ is identified if and only if it is almost surely included in $\mathcal{B}$. Intuitively, the result of the proposition holds if the prior distribution on $\mathcal{G}$ given $\mathcal{B}$ is degenerate into a Dirac measure on a $\mathcal{B}$-measurable function.

We now illustrate these concepts by some examples.

### 3.1 Examples

**Example 3.1.** Unobserved heterogeneity or incidental parameters. The incidental parameter problem typically arises with panel data models when a regression model admits an agent specific intercept $\xi_i$. If we take the fixed effect approach then the model is written conditionally on $\{\xi_i\}$. The general model can be simplified to be

$$x_i | \xi_i, \gamma \sim \text{i.i.d.} \mathcal{N}(\xi_i, \sigma^2), \quad i = 1, \ldots, \infty$$

with $\sigma^2$ known. Let $\xi = (\xi_1, \xi_2, \ldots)$. The $\mu$-a.s. minimal sufficient $\sigma$-field $\mathcal{B}$ is generated by $\xi$, that is, $\xi$ is the identified parameter. The parameter $\gamma$ is not measurably identified. However, it can be identified in the Bayesian model. To see this, let us specify a prior probability for $\xi$ and $\gamma$ as follows:

$$\xi_i | \gamma \sim \text{i.i.d.} \mathcal{N}(\gamma, \sigma_0^2), \quad i = 1, \ldots, \infty$$

$$\gamma \sim \pi$$

where $\sigma_0^2$ is known and $\pi$ is a probability measure. By the strong law of large numbers, $\gamma = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \xi_i$ a.s. with respect to the conditional distribution of $\{\xi_i\}$ given $\gamma$ (and clearly with respect to the joint distribution of $\{\xi_i, \gamma\}$). Therefore, the conditional distribution of $\gamma | \{\xi_i\}_{i=1}^{\infty}$ puts all its mass on $\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \xi_i$ and the model is fully Bayesian identified. This property does not depend on the marginal prior on $\gamma$ but is crucially dependent on the prior on the $\{\xi_i\}$ given $\gamma$. 

**Example 3.2. Gaussian process and hyperparameter in the mean.** Let us consider the sample space \( X = L^2[0,1] \), that is, the space of square integrable functions defined on \([0,1]\) with respect to the uniform distribution, endowed with the \( L^2\)-norm \( \| \cdot \| \), the \( L^2\)-inner product \( < \cdot , \cdot > \) and the associated Borel \( \sigma \)-field \( \mathcal{X} \). Let \( B = L^2[0,1] \), we consider two parameters \((\beta , \gamma )\) such that \( \beta \in (B , B) \) and \( \gamma \in (\mathbb{R}_+ , \mathcal{R}) \), where \( B \) and \( \mathcal{R} \) are the Borel \( \sigma \)-field associated with \( B \) and \( \mathbb{R}_+ \), respectively.

The sample is represented by a single observation of a trajectory \( x \) from a Gaussian process in \( X \): \( x | \beta , \gamma \sim \mathcal{N}(\beta , \Sigma) \) where \( \beta \in L^2([0,1]) \) and \( \Sigma : X \rightarrow X \). By definition of Gaussian process in Hilbert spaces, it follows that \( \mathbb{E} \| x \|^2 < \infty , \Sigma \) is a bounded, linear, positive definite, self-adjoint and trace-class covariance operator. While the parameter \( \beta \) is measurably identified, the parameter \( \gamma \) is not measurably identified.

Suppose now that \( \beta \) and \( \gamma \) are two random variables with values in \((B , B)\) and \((\mathbb{R}_+ , \mathcal{R})\), respectively, so that they induce a prior distribution on \((B \times \mathbb{R}_+ , B \otimes \mathcal{R})\):

\[
\beta | \gamma \sim \mathcal{N}(\beta_0 , \sigma_0^2 \Omega), \quad \beta_0 \in L^2[0,1], \quad \Omega : L^2[0,1] \rightarrow L^2[0,1]
\]

\[
\gamma \sim \mathcal{N}(\gamma_0 , \sigma_0^2), \quad \gamma_0 , \sigma_0^2 \in \mathbb{R}_+
\]

where \( \beta_0 , \gamma_0 , \sigma_0^2 \) and \( \Omega \) are known. \( \Omega \) is a bounded, linear, positive definite, self-adjoint and trace-class covariance operator. Let \((\lambda_j^\Omega , \psi_j^\Omega)\) be the eigensystem associated with \( \Omega \). Then, \( < \beta , \varphi_j > | \gamma \sim \mathcal{N}(< \beta_0 , \varphi_j > , \sigma_0^2 \lambda_j^\Omega) \) and

\[
< \beta , \varphi_j > \sim \mathcal{N}(< \beta_0 , \varphi_j > / \sigma_0 \sqrt{\lambda_j^\Omega} , 1).
\]

Therefore, we can write \( \gamma \) as a function of \( \beta \):

\[
\gamma = \hat{\gamma}(\beta) - \frac{\sum_{j=1}^{\infty} \xi_j < \beta_0 , \varphi_j > \sqrt{\lambda_j^\Omega}}{\sum_{j=1}^{\infty} \frac{< \beta_0 , \varphi_j >^2}{\sigma_0^2 \lambda_j^\Omega}} \equiv \frac{\sum_{j=1}^{\infty} \xi_j < \beta_0 , \varphi_j > \sqrt{\lambda_j^\Omega}}{\sum_{j=1}^{\infty} \frac{< \beta_0 , \varphi_j >^2}{\sigma_0^2 \lambda_j^\Omega}} - \frac{\sum_{j=1}^{\infty} \xi_j < \beta_0 , \varphi_j > \sqrt{\lambda_j^\Omega}}{\sum_{j=1}^{\infty} \frac{< \beta_0 , \varphi_j >^2}{\sigma_0^2 \lambda_j^\Omega}}
\]

where \( \hat{\gamma}(\beta) \) denotes the least squares estimate of \( \gamma \) and \( \xi_j \sim iid \mathcal{N}(0,1) \), \( j \geq 1 \). In order to have exact estimability and Bayesian identification of \( \gamma \), i.e. \( \gamma \in \overline{B} \), the series \( \sum_{j=1}^{\infty} \frac{< \beta_0 , \varphi_j >^2}{\lambda_j^\Omega} \) must diverge so that \( \gamma = \hat{\gamma}(\beta) \).

We show now that if \( \sum_{j=1}^{\infty} \frac{< \beta_0 , \varphi_j >^2}{\lambda_j^\Omega} \) diverges then the specification of the prior distribution for \((\gamma , \beta)\) is such that the conditional distribution of \( \gamma | \beta \) puts all its mass on \( \hat{\gamma}(\beta) \).

We define the following operators:

\[
K : \mathbb{R} \rightarrow L^2[0,1], \quad K^* : L^2[0,1] \rightarrow \mathbb{R}
\]

\[
r \mapsto r \beta_0, \quad \psi \mapsto < \beta_0 , \psi >
\]

where \( < \cdot , \cdot > \) denotes the inner product in \( L^2[0,1] \). Then, \( KK^* = \beta_0 < \beta_0 , \cdot > \) and we can write the joint distribution of \((\gamma , \beta)\) as

\[
\left( \begin{array}{c} \gamma \\ \beta \end{array} \right) \sim \mathcal{N}\left( \begin{array}{c} \gamma_0 \\ K \gamma_0 \end{array} , \begin{array}{cc} 1 & K^* \\ K & (KK^* + \Omega) \end{array} \sigma_0^2 \right)
\]

13
and the conditional distribution of $\gamma$ given $\beta$ as:

$$\gamma|\beta \sim \mathcal{N}(\gamma_0 + K^*(KK^* + \Omega)^{-1}(\beta - K\gamma_0), (1 - K^*(KK^* + \Omega)^{-1}K)\sigma_0^2).$$

Let $\mathcal{R.K.H.S.}(\Omega)$ denote the Reproducing Kernel Hilbert Space associated with $\Omega$. The convergence of the series $\sum_{j=1}^{\infty} \frac{|<\beta_0, \gamma_j>|^2}{\lambda_j^2}$ is equivalent to the source condition: $\beta_0 \in \mathcal{R.K.H.S.}(\Omega) \equiv \mathcal{R}(\Omega^{1/2})$, that is, there exists a $\delta_0 \in L^2[0,1]$ such that $\beta_0 = \Omega^{1/2} \delta_0$. In this example the parameter $\gamma$ may be Bayesian identified if and only if the conditional distribution of $\gamma$ given $\beta$ is a degenerated Dirac measure on a deterministic function of $\beta$. This happens when $\beta_0 \notin \mathcal{R}(\Omega^{1/2})$ as we are going to explain. Define $T = \Omega^{-\frac{1}{2}} K$, which is well defined if $\beta_0 \in \mathcal{R}(\Omega^{1/2})$, and $T^*T = \langle \Omega^{-\frac{1}{2}} \beta_0, \Omega^{-\frac{1}{2}} \beta_0 \rangle >$. The conditional mean $\mathbf{E}(\gamma|\beta)$ becomes

$$\mathbf{E}(\gamma|\beta) = \gamma_0(1 - T^*(TT^* + I)^{-1}T) + K^*(KK^* + \Omega)^{-1}\beta$$

and the conditional variance is:

$$\text{Var}(\gamma|\beta) = (1 - K^*(KK^* + \Omega)^{-1}K)\sigma_0^2 = (1 - T^*(TT^* + I)^{-1}T)\sigma_0^2$$

$$= (1 - (1 + T^*T)^{-1}T^*T)\sigma_0^2.$$

The $\text{Var}(\gamma|\beta)$ is equal to 0 if and only if $(1 + T^*T)^{-1}T^*T = 1$ which is impossible if $\beta_0 \notin \mathcal{R}(\Omega^{1/2})$ since in this case $T^*T = ||\delta_0||^2 < \infty$. On the contrary, when $\beta_0 \notin \mathcal{R}(\Omega^{1/2})$ the variance is equal to 0 since $(1 + T^*T)^{-1}T^*T = \left(\langle \Omega^{-\frac{1}{2}} \beta_0, \Omega^{-\frac{1}{2}} \beta_0 \rangle^{-1} + 1\right)^{-1} = 1$ because $\langle \Omega^{-\frac{1}{2}} \beta_0, \Omega^{-\frac{1}{2}} \beta_0 \rangle >$ is unbounded. In the same way, when $\beta_0 \notin \mathcal{R}(\Omega^{1/2})$ then, $\mathbf{E}(\gamma|\beta) = K^*(KK^* + \Omega)^{-1}\beta$ and it is equal to the function $\hat{\gamma}(\beta)$ defined above.

This shows that if the source condition does not hold, i.e. $\beta_0 \notin \mathcal{R}(\Omega^{1/2})$ then, $\gamma$ is Bayesian identified by the prior distribution and exactly estimable since its conditional distribution is degenerated at the conditional mean which is function of $\beta$. On the contrary, if the source condition is verified then, $\gamma$ is not identified by the prior distribution. However, in the measurable model, $\gamma$ is never identified.

The condition $\beta_0 \notin \mathcal{R}(\Omega^{1/2})$ is quite natural since it is well-known that, with $\mu$-probability 1, $(\beta - \gamma_0) \notin \mathcal{R}(\Omega^{1/2})$. For this reason, in the opposite case where we assumed $\beta_0 \in \mathcal{R}(\Omega^{1/2})$ then the trajectories of $\beta$ generated by the Gaussian process $\mathcal{N}(\gamma_0, \sigma^2\Omega)$ would be very irregular.

**Example 3.3. Gaussian process and hyperparameter in the variance.** Let $X$ be an infinite dimensional Polish space. We observe a trajectory $x$ from a Gaussian process:

$$x|\beta, \sigma^2 \sim \text{ind}\mathcal{N}(\beta, \Sigma)$$

where $\Sigma : X \to X$ is a bounded, linear, positive definite, self-adjoint and trace-class covariance operator. The parameter $\beta$ is identified but we are interested in the parameter
\( \sigma^2 \) which is not measurably identified. However, we show that it can be identifiable in the Bayesian sense.

We endow \( \beta \) and \( \sigma^2 \) with a prior distribution:

\[
\beta | \sigma^2 \sim \mathcal{N}(0, \sigma^2 \Omega)
\]

\[
\sigma^2 \sim \mathcal{I}(\nu_0, \sigma^2_0)
\]

where \( \mathcal{I} \) denotes an inverse-gamma distribution. The covariance operator \( \Omega : X \to X \) is known, bounded, linear, positive definite, self-adjoint and trace-class. The parameters \( \nu_0 \) and \( \sigma^2_0 \) are supposed to be known.

Knowledge of \( \Omega \) implies knowledge of its eigensystem \( (\lambda_j^0, \varphi_j^0) \) The random elements \( \left\{ \frac{\langle \beta, \varphi_j^0 \rangle \sqrt{\lambda_j^0}}{\lambda_j^0} \right\} \) are independently and identically distributed as \( \mathcal{N}(0, \sigma^2) \). By the strong law of large numbers we have

\[
\lim_{J \to \infty} \frac{1}{J} \sum_{j=1}^{J} \frac{\langle \beta, \varphi_j^0 \rangle^2}{\lambda_j^0} = \sigma^2, \quad \mu - a.s.
\]

This shows that \( \sigma^2 \) is a \( \mu - a.s. \) function of \( \beta \), then it is identified.

\textbf{Example 3.4. Bayesian nonparametric and Dirichlet process.} We observe an \( n \)-sample from the following sampling model

\[
x_i | F, G \sim \text{iid } F, \quad i = 1, \ldots, n
\]

where \( F \) and \( G \) are two probability distributions –for instance on \( \mathbb{R} \). We can interpret \( F \) as a nuisance parameter and \( G \) is the only parameter of interest. The parameter \( G \) is not measurably identified in the original sampling model but it may be Bayesian identified under some prior assumptions. To see this, we assume that \( F \) is a Dirichlet process with parameters \( n_0 \) and \( G \), i.e. \( F \sim \text{Dir}(n_0, G) \). Denote by \( F \) the \( \sigma \)-field generated by \( F \). The prior distribution is

\[
F|G \sim \text{Dir}(n_0, G)
\]

\[
G \sim \pi
\]

where \( \pi \) denotes a probability measure on the space of distributions such that \( G \) is \( a.s. \) a diffuse distribution, i.e. \( G(x) = 0, \forall x \). By using the Sethuraman [1994] - Rolin [1992] decomposition, \( F = \sum_j \alpha_j \delta_{\xi_j} \), where \( \{\xi_j\}_{j \geq 1} \) are independent drawings from \( G \), \( \alpha_j = v_j \prod_{k=1}^{j-1} (1 - v_k) \) with \( \{v_j\}_{j \geq 1} \) independent drawings from a beta distribution \( \text{Be}(1, n_0) \) and \( \{v_j\}_{j \geq 1} \parallel \{\xi_j\}_{j \geq 1} \). Then,

\[
\lim_{J \to \infty} \frac{1}{J} \sum_{j=1}^{J} \delta_{\xi_j} = G, \quad \mu - a.s.
\]

which proves that \( G \) is a \( \mu \)-almost sure function of \( F \) then, it is identified.
4 Bayesian set estimation

Let us come back to Remark 2.1. We consider here a model where the parameter $\theta$ is identified but the model is completed by a theoretical condition

$$A(\theta, \gamma) \in A_0$$

(4)

where $\gamma \in \Gamma$, $A : \Theta \times \Gamma \rightarrow \Phi$, for some finite or infinite dimensional space $\Phi$ and $A_0$ is a subset of $\Phi$. This parameter $\gamma$ is for instance the natural parameter arising in economic models and it is not necessarily identified. Depending on the relation $A$ it may be only partially-identified which means that there exists more than one solution of (4) in $\gamma$. Define $\Gamma_I$ as $\Gamma_I \equiv \Gamma_I(\theta) = \{ \gamma; A(\theta, \gamma) \in A_0 \}$ and call it the identified-set. Then, the model is identified if $\Gamma_I$ is a singleton and is partially-identified if not. Remarks that $\Gamma_I$ depends on $\theta$.

In this section, $\gamma$ is not considered as a parameter in the sense that it is not provided with a prior measure. We will introduce a prior specification on $\gamma$ in the next section. Our characterization of (4) differs from the frequentist characterization of it since the quantity $A(\theta, \gamma)$ has to be understood in our analysis as a random element (where the randomness comes from $\theta$) and the theoretical condition (4) has to hold a.s. with respect to the prior distribution of $\theta$. Therefore, $\Gamma_I(\theta)$ is a random set and condition (4) may be rewritten as a function:

$$g_\theta(\gamma) = 1\{A(\theta, \gamma) \in A_0 \in \Phi\}$$

(5)

where $1\{\cdot\}$ is the indicator function which is equal to 1 if the condition inside the curly brackets is satisfied and 0 otherwise. Therefore, we can focus either on (4) or on the random function $g_\theta(\cdot) : \Gamma \rightarrow \{0, 1\}$. In that construction, $\gamma$ is just the index of the stochastic process $g_\theta(\cdot)$.

A Bayesian analysis requires the specification of a prior distribution for the random set $\Gamma_I$. We propose to construct such a prior by first specifying a prior distribution for $\theta$ and then recovering the prior for $\Gamma_I$ through a transformation. Let $F$ denote the distribution which generates the data. The random parameter $\theta$ may be rewritten as a functional of the distribution $F$: $\theta = \phi(F)$, for some functional $\phi$. For instance, $\theta = E_F(\cdot)$. In our analysis, we are not going to restrict $F$ to belong to some parametric class. Thus, we specify a Dirichlet process prior for $F$ and write $F \sim Dir(n_0, F_0)$ where $n_0 \in \mathbb{R}_+$ and $F_0$ is a diffuse probability measure on $\mathbb{R}$, i.e. $F_0(y) = 0, \forall y \in \mathbb{R}$, see e.g. Ferguson [1973] and Florens [2002]. The prior distribution for $\Gamma_I$ is going to be determined by the prior capacity functional. Define $\mathcal{K}$ as the family of compact subsets of $\Gamma$. The prior capacity functional $T_{\Gamma_I} : \mathcal{K} \mapsto [0, 1]$ is given by

$$T_{\Gamma_I}(K) = P\{K \cap \Gamma_I \neq \emptyset\}, \quad K \in \mathcal{K}$$
where the probability $P$ is determined by the prior distribution of $\theta$ as a functional of the Dirichlet process prior $\phi(F)$ where $F \sim \text{Dir}(n_0, F_0)$.

When the prior capacity functional is defined on singletons instead of on $\mathcal{K}$, i.e. $K = \{\gamma\}$, $\gamma \in \Gamma$, then it is called prior coverage function of $\Gamma_I$ and denoted by $p_{\Gamma_I}(\gamma)$. In particular,

$$p_{\Gamma_I}(\gamma) = E(g_\theta(\gamma)) = \text{Prob}(g_\theta(\gamma) = 1) = P(A(\theta, \gamma) \in A_0) \in [0, 1], \ \forall \gamma \in \Gamma \quad (6)$$

where $g_\theta(\cdot)$ is the stochastic process defined above and $P$ and $E$ are the probability and expectation, respectively, taken with respect to the prior of $\theta$. The appealing fact of the prior coverage function with respect to the prior capacity functional is that $p_{\Gamma_I}(\gamma)$ can be represented graphically in an easier way than $T_{\Gamma_I}(K)$ (at least if $\Gamma \subset \mathbb{R}$), see for instance Figures 1, 3 and 5 where we have represented the functions $p_{\Gamma_I}(\cdot)$ and $p_{\Gamma_I}(\cdot|\mathcal{X})$ against $\gamma$.

We detail now the computation of prior and posterior capacity functionals. The intuition for this computation is the following. Consider the example where $A(\theta, \gamma) \in A_0$ writes as $\gamma \in [\theta_1, \theta_2]$, for $\theta = (\theta_1, \theta_2)^t$, $\theta_1, \theta_2 \in \mathbb{R}$ so that $\Gamma_I = [\theta_1, \theta_2]$ and we specify a parametric prior for $\theta$. For instance, $\theta_1 \sim \mathcal{U}[0, 1]$ and $\theta_2 \sim \mathcal{U}[1, 2]$. For any $K \in \mathcal{K}$, write $K = [K_1, K_2]$, $\bar{K} = [\bar{K}_1, \bar{K}_2] := K \cap [0, 1]$ and $\bar{K} = [\bar{K}_1, \bar{K}_2] := K \cap [1, 2]$. The prior capacity functional then is:

$$T_{\Gamma_I}(K) = \begin{cases} 
0 & \text{if } K \cap [0, 2] = \emptyset \\
\bar{K}_2 - \bar{K}_1 & \text{if } K \cap [0, 1] \neq \emptyset \text{ and } K \cap [1, 2] = \emptyset \\
\bar{K}_2 - \bar{K}_1 + \bar{K}_2 - \bar{K}_1 & \text{if } K \cap [0, 1] \neq \emptyset \text{ and } K \cap [1, 2] \neq \emptyset \\
\bar{K}_2 - \bar{K}_1 & \text{if } K \cap [0, 1] = \emptyset \text{ and } K \cap [1, 2] \neq \emptyset 
\end{cases}$$

The expression for the prior coverage function is simpler and is given by

$$p_{\Gamma_I}(\gamma) = P(\gamma \in [\theta_1, \theta_2]) = \gamma 1\{\gamma \in [0, 1]\} + (2 - \gamma) 1\{\gamma \in [1, 2]\}.$$ 

With this intuition in mind, let us move to the nonparametric Bayesian approach. This approach is based on a Dirichlet process prior and requires to write $\theta = (\theta_1, \theta_2)^t$ as $\theta_i = \phi_i(F)$, $i = 1, 2$. If the data are observations of the random vector $Y$ then, the Bayesian model writes

$$Y|F \sim F$$

$$F \sim \text{Dir}(n_0, F_0).$$

The probability measure $F_0$ should be chosen such that $\phi_2(F) > \phi_1(F)$, $\mu$-a.s. The prior capacity functional of $\Gamma_I = [\phi_1(F), \phi_2(F)]$ is given by

$$T_{\Gamma_I}(K) = P\{K \cap \left[ \phi_1\left(\sum_{j=1}^{\infty} \alpha_j \delta_{\xi_j}\right), \phi_2\left(\sum_{j=1}^{\infty} \alpha_j \delta_{\xi_j}\right) \right] \neq \emptyset\}, \ \ K \in \mathcal{K} \quad (7)$$

where $\{\xi_j\}_{j \geq 1}$ are independent drawings from $F_0$, $\delta_{\xi_j}$ denotes the Dirac mass in $\xi_j$, $\alpha_j = v_j \prod_{l=1}^{j-1} (1 - v_l)$ with $\{v_l\}_{l \geq 1}$ independent drawings from a beta distribution $\text{Be}(1, n_0)$ and
\{v_j\}_{j \geq 1} are independent of \{\xi_j\}_{j \geq 1}. This representation of the Dirichlet process is called “stick-breaking” and is due to Rolin [1992] and Sethuraman [1994]. In Appendix A we recall how to simulate \( \phi_1(F) \) from the prior and posterior distribution by using this representation. The prior coverage function of \( \Gamma_I \) is

\[
p_{\Gamma_I}(\gamma) = P\left( \gamma \in [\phi_1(F), \phi_2(F)] \right) = P\left( \phi_1\left( \sum_{j=1}^{\infty} \alpha_j \delta_{\xi_j} \right) \leq \gamma \leq \phi_2\left( \sum_{j=1}^{\infty} \alpha_j \delta_{\xi_j} \right) \right)
\]

(8)

where we have taken \( K = \{ \gamma \} \). In general we do not have an analytical form for \( T_{\Gamma_I} \) and \( p_{\Gamma_I} \) but we have a perfect knowledge of them since we can easily simulate from them by using the “stick-breaking” representation of the Dirichlet process.

Suppose now to observe an \( n \)-sample of \( Y: (y_1, \ldots, y_n) \). The posterior distributions of \( F \) is

\[
F|y_1, \ldots, y_n \sim D \left( n_0 + n, \frac{n_0}{n_0 + n} F_0 + \frac{n}{n_0 + n} F_n \right)
\]

where \( F_n \) denotes the empirical cumulative distribution. Denote \( F_\ast = \frac{n_0}{n_0 + n} F_0 + \frac{n}{n_0 + n} F_n \) and \( n_\ast = n_0 + n \). If the true distribution \( F \) which generates the data is such that \( \phi_2(F) > \phi_1(F) \) then the same is true for the distribution \( F \) generated by the posterior. The posterior capacity functional is denoted by \( T_{\Gamma_I}(K|\{y_i\}_{i=1}^n) \) and given by, for \( \forall K \in \mathcal{K} \)

\[
T_{\Gamma_I}(K|\{y_i\}_{i=1}^n) = P\{ K \cap \left[ \phi_1\left( \rho \sum_{j=1}^{n} \beta_j \delta_{y_j} + (1-\rho) \sum_{j=1}^{\infty} \alpha_j \delta_{\xi_j} \right) , \phi_2\left( \rho \sum_{j=1}^{n} \beta_j \delta_{y_j} + (1-\rho) \sum_{j=1}^{\infty} \alpha_j \delta_{\xi_j} \right) \right] \neq \emptyset \}
\]

where, \{\alpha_j\}, \{\delta_{\xi_j}\}, \{\beta_j\} and \{\xi_j\}, are as above, \( \rho \) is drawn form a beta distribution \( \text{Be}(n, n_0) \) independently of the other quantities and \( (\beta_1, \ldots, \beta_n) \) has been drawn from a Dirichlet distribution of parameters \( (1, \ldots, 1) \) on the simplex \( S_{n-1} \) of dimension \( (n - 1) \).

The posterior coverage function \( p_{\Gamma_I}(\gamma|\{y_i\}_{i=1}^n) = P\left( \gamma \in [\phi_1(F), \phi_2(F)] \right| \{y_i\}_{i=1}^n \) is:

\[
P\left( \phi_1\left( \rho \sum_{j=1}^{n} \beta_j \delta_{y_j} + (1-\rho) \sum_{j=1}^{\infty} \alpha_j \delta_{\xi_j} \right) \leq \gamma \leq \phi_2\left( \rho \sum_{j=1}^{n} \beta_j \delta_{y_j} + (1-\rho) \sum_{j=1}^{\infty} \alpha_j \delta_{\xi_j} \right) \right).
\]

(9)

An estimator of the interval \( \Gamma_I \) then can be proposed based either on the posterior capacity functional of \( \Gamma_I \) or on the posterior coverage function.

For simplicity we have presented only the case where \( A(\theta, \gamma) \in A_0 \) writes as \( [\phi_1(F), \phi_2(F)] \) but our nonparametric method can be generalized to the case where \( \Gamma_I \) is not an interval. In that case, if a Dirichlet process prior is specified for \( F \), the prior capacity functional of \( \Gamma_I \) is given by, \( \forall K \in \mathcal{K} \)

\[
T_{\Gamma_I}(K) = P(K \cap \Gamma_I \neq \emptyset) = P \left( K \cap \left\{ \gamma \in \Gamma; A(\phi\left( \sum_{j=1}^{\infty} \alpha_j \delta_{\xi_j} \right), \gamma) \in A_0 \right\} \neq \emptyset \right)
\]

(10)

18
and the posterior is, \( \forall K \in \mathcal{K} \)

\[
T_{1_f}(K|\{y_i\}_{i=1}^n) = P(K \cap \Gamma_f \neq \emptyset | \{y_i\}_{i=1}^n) = \\
\left( K \cap \left\{ \gamma \in \Gamma; A(\phi \left( \sum_{j=1}^n \beta_j \delta_{y_j} + (1 - \rho) \sum_{j=1}^{\infty} \alpha_j \delta_{\xi_j} \right), \gamma) \in A_0 \right\} \neq \emptyset | \{y_i\}_{i=1}^n \right)
\]

where, \( \{\alpha_j\}, \{\delta_{\xi_j}\}, \{\delta_{y_j}\}, \{\xi_j\} \), \( \rho \) and \( (\beta_1, \ldots, \beta_n) \) are as above. We provide in the next section some examples for which the analysis is carried out by our proposed Bayesian nonparametric method. For simplicity, we only focus on the prior and posterior coverage functions, which are easy to represent graphically.

### 4.1 Examples

**Example 4.1. Interval Censored Data.** This is example 1 in Chernozhukov et al. [2007] and is motivated by missing data. The real random variable \( Y \) is unobserved but is known to be almost surely (with respect to the sampling distribution) in the interval \([Y_1, Y_2]\), where \( Y_1 \) and \( Y_2 \) are two real random variables with unknown probability distribution. Denote by \( \gamma \) the parameter \( \gamma := \mathbf{E}(Y) \) and by \( \mathbf{E}_F(\cdot) \) the expectation taken with respect to a distribution \( F \), i.e. \( \mathbf{E}_F(\cdot) = \mathbf{E}(\cdot|F) \). Since \( Y \in [Y_1, Y_2] \) a.s., the condition \( A(\theta, \gamma) \in A_0 \) takes the form

\[
\mathbf{E}_{F_1}(Y_1) \leq \gamma \leq \mathbf{E}_{F_2}(Y_2)
\]

a.s. with respect to the prior distribution of \( \theta \), where \( \theta = \phi(F) = (\mathbf{E}_{F_1}(Y_1), \mathbf{E}_{F_2}(Y_2))' \), \( F = (F_1, F_2)' \) is the joints distribution of \((Y_1, Y_2)\) and \( F_i \) denotes the probability distribution of \( Y_i \), \( i = 1, 2 \). We compute the prior and posterior coverage function of the set \( \Gamma_f = [\mathbf{E}_{F_1}(Y_1), \mathbf{E}_{F_2}(Y_2)] \) by using a Dirichlet process prior for \( \theta \). We assume that \( Y_1 \parallel Y_2|F \).

The Bayesian hierarchical model is the following

\[
(y_{11}, \ldots, y_{1n}) | F \sim iid \quad F_1 \\
(y_{21}, \ldots, y_{2n}) | F \sim iid \quad F_2
\]

\[
F_1 \sim \text{Dir}(n_0^1, F_0^1) \\
F_2 \sim \text{Dir}(n_0^2, F_0^2)
\]

If the probability measures \( F_0^1 \) and \( F_0^2 \) have disjoint supports, that is, \( \max \text{Supp}(F_0^1) \leq \min \text{Supp}(F_0^2) \) then, \( \mathbf{E}_{F_2}(Y_2) > \mathbf{E}_{F_1}(Y_1) \) \( \mu \)-a.s. Let \( g_\theta(\gamma) := 1\{ \gamma \in [\mathbf{E}_{F_1}(Y_1), \mathbf{E}_{F_2}(Y_2)] \} \). The prior coverage function is \( \mathbf{E}(g_\theta(\gamma)), \forall \gamma \in \Gamma \), and can be represented by

\[
p_{F_1}(\gamma) = P\left( \gamma \in [\mathbf{E}_{F_1}(Y_1), \mathbf{E}_{F_2}(Y_2)] \right) = P\left( \left( \sum_j \alpha_j^1 \xi_j^1 \leq \gamma \leq \sum_j \alpha_j^2 \xi_j^2 \right) \right)
\]

19
where, for \( i = 1, 2 \), \( \{\xi_j^i\}_{j \geq 1} \) are independent drawings from \( F_0^i \), \( \alpha_j^i = v_j^i \prod_{j=1}^i (1 - v_j^i) \) with \( \{v_j^i\}_{j \geq 1} \) independent drawings from a beta distribution \( Be(1, n_0^i) \) and \( \{v_j^i\}_{j \geq 1} \) are independent of \( \{\xi_j^i\}_{j \geq 1} \). In practice, we have that: 

(i) \( \forall \gamma \notin [\min \text{Supp}(F_0^1), \max \text{Supp}(F_0^2)] \), \( p_{r_1}(\gamma) = 0 \); 
(ii) \( \forall \gamma \in [\max \text{Supp}(F_0^1), \min \text{Supp}(F_0^2)] \), \( p_{r_1}(\gamma) = 1 \); 
(iii) \( \forall \gamma \in \text{Supp}(F_0^1) \), \( p_{r_1}(\gamma) = P\left(\sum_j \alpha_j^1 \xi_j^1 \leq \gamma\right) = F_1(\gamma) \); 
(iv) \( \forall \gamma \in \text{Supp}(F_0^2) \), \( p_{r_1}(\gamma) = P\left(\gamma \leq \sum_j \alpha_j^2 \xi_j^2\right) = 1 - F_2(\gamma) \), where for \( i = 1, 2 \), \( F_i(\cdot) \) denotes the cumulative distribution function of \( E_{F_i}(Y_i) \).

The posterior distributions of \( F_1 \) and \( F_2 \) are

\[
F_i(y_1, \ldots, y_m) \sim \text{Dir}(n_i^1, F_i^1), \quad i = 1, 2
\]

\[
n_i^1 = n_0^1 + n \quad F_i^1 = \frac{n_i^0}{n_0^0 + n} + \frac{n}{n_0^1 + n} F_n^i, \quad i = 1, 2
\]

where \( F_n^i(\cdot) = \frac{1}{n} \sum_j \delta_{y_j(\cdot)}(\cdot) \) is the empirical distribution of the sample \( (y_{i1}, \ldots, y_{im}) \), \( i = 1, 2 \). The posterior coverage function of \( \Gamma_I \) is

\[
p_{r_1}(\gamma|\{y_{i1}\}, \{y_{2i}\}) = P\left(\gamma \in [E_{F_1}(Y_1), E_{F_2}(Y_2)]\big|\{y_{i1}\}, \{y_{2i}\}\right)
\]

and can be represented as:

\[
P\left(\rho_1 \sum_{j=1}^n \beta_{1j} y_{1j} + (1 - \rho_1) \sum_{j=1}^n \alpha_j^1 \xi_j^1 \leq \gamma \leq \rho_2 \sum_{j=1}^n \beta_{2j} y_{2j} + (1 - \rho_2) \sum_{j=1}^n \alpha_j^2 \xi_j^2\right)
\]

where, for \( i = 1, 2 \), \( \alpha_j^i \) and \( \xi_j^i \) are as above, \( \rho_i \) is drawn from a Beta distribution \( Be(n, n_0^i) \) independently of the other quantities and \( (\beta_{1i}, \ldots, \beta_{im}) \) has been drawn from a Dirichlet distribution of parameters \( (1, \ldots, 1) \) on the simplex \( S_{n-1} \) of dimension \( (n-1) \).

A simulation exercise allows to visualize the prior and posterior coverage functions of \( \Gamma_I \). We generate an \( n \)-sample of observations of \( Y_1 \) and \( Y_2 \) from the following distributions:

\[
Y_1 \sim N(0, 0.1), \quad Y_2 \sim N(5, 0.1).
\]

The parameters are fixed as follows: \( n = 1000 \), \( n_1^1 = 10 \), \( n_0^1 = 20 \), \( F_0^1 = N(0, 1) \) and \( F_0^2 = N(10, 1) \). The supports of \( F_0^1 \) and \( F_0^2 \) are not disjoint, however, since the tails of a normal density function are very thin, the prior probability that \( E_{F_2}(Y_2) < E_{F_1}(Y_1) \) is very small. The true identified set in our simulation is \( [E_{F_1}(Y_1), E_{F_2}(Y_2)] = [0, 5] \).

Then, we draw 1000 times from the prior and posterior distribution of \( (F_1, F_2) \) and for every \( \gamma \in [-3, 12] \) we count the number of times that the simulated \( \Gamma_I \) contains \( \gamma \). We show in Figure 1 the prior and posterior coverage functions for each value of \( \gamma \in [-3, 12] \). In Figure 2 we represent the intervals \( \Gamma_I \) drawn from the prior and posterior distributions (on the vertical axis) against the true interval \( [0, 5] \) (on the horizontal axis).

**Example 4.2. Interval Regression Model.** This is example 2 in Chernozhukov et al. [2007] and it is a generalization of Example 4.1. Let \( X \) be a vector of regressors

\[\text{This data generating process is the same used in Liao and Jiang [2010] in their Example 5.1.}\]
and the conditional expectation of an endogenous variable $Y$ be supposed to be linear: $E(Y|X) = X'\gamma$, where $\gamma$ is a vector of parameters for which we do not specify a prior distribution. The random variable $Y$ is unobservable but it is known that with probability 1

$$E(Y_1|X) \leq X'\gamma \leq E(Y_2|X)$$

where $Y_1$ and $Y_2$ are two observable real random variables with unknown probability distribution. These conditional moment inequalities are then transformed in unconditional moment inequalities by using a vector $Z$ of positive transformations of $X$ or of positive instrumental variables:

$$E_F(Y_1Z) \leq E_F[ZX]'\gamma \leq E_F(Y_2Z), \quad \mu - a.s. \quad (13)$$

$E_F(\cdot)$ denotes the expectation taken with respect to the joint distribution $F$ of $(Y_1, Y_2, X, Z)$
and $E_F(\cdot) \equiv E(\cdot|F)$. The nonparametric Bayesian model is

$$(y_{1i}, y_{2i}, x_i, z_i)|F \sim iid \quad F \sim Dir(n_0, F_0)$$

where $(y_{1i}, y_{2i}, x_i, z_i)$ is a realization of $(Y_1, Y_2, X, Z)$ and we simplify the notation as $W := (Y_1, Y_2, X, Z)$, and $w_i := (y_{1i}, y_{2i}, x_i, z_i)$. The identified set is $\Gamma(F) = \{ \gamma \in \Gamma; \ E_F(Y_1Z) \leq E_F(Y_2Z) \}$ which simplifies to a (multidimensional) interval when $E_F(ZX)'$ is full-column rank. Its prior coverage function $p_{\Gamma, i}(\gamma)$ is given by

$$P\left( E_F(Y_1Z) \leq E_F(ZX)' \gamma \leq E_F(Y_2Z) \right) = P\left( \sum_j \alpha_j \xi_j \leq \sum_j \alpha_j \xi_j' \leq \sum_j \alpha_j^2 \xi_j \right)$$

where $\xi_j = (\xi_j^y, \xi_j^z, \xi_j^x, \xi_j^w)' \sim iid \ F_0$ and $\{\alpha_k\}_{k \geq 1}$ are computed as in (7). The posterior of $F$ is

$$F|(w_1, \ldots, w_n) \sim Dir(n_s, F_s), \quad n_s = n_0 + n \quad F_s = \frac{n_0}{n_0 + n} F_0 + \frac{n}{n_0 + n} F_n, \quad i = 1, 2$$

where $F_n(\cdot) = \frac{1}{n} \sum_j \delta w_j(\cdot)$ is the empirical distribution of the sample $(w_1, \ldots, w_n)$. The posterior coverage function of $\Gamma$ is given by

$$P\left( E_F(Y_1Z) \leq E_F(ZX)' \gamma \leq E_F(Y_2Z) \right) = P\left( \sum_j \beta_j y_j z_j + (1 - \rho) \sum_j \alpha_j \xi_j \leq \sum_j \beta_j y_j' z_j + (1 - \rho) \sum_j \alpha_j \xi_j' \leq \sum_j \beta_j y_j z_j + (1 - \rho) \sum_j \alpha_j^2 \xi_j \right)$$

where $\rho$ and $(\beta_1, \ldots, \beta_n)$ are defined as in (11).

A simulation exercise allows to visualize the prior and posterior coverage function of $\Gamma$. We generate an $n$-sample of observations of $(Y_1, Y_2, X, Z)$ in the following way:

$$z_i \sim iid \ U[0, 1], \quad x_i = z_i + u_i, \quad u_i \sim iid \ N(0, 1)$$

$$y_{1i} = \theta_1 x_i + \varepsilon_{1i}, \quad \theta_1 = 2, \quad \varepsilon_{1i} \sim iid \ N(0, 0.1)$$

$$y_{2i} = \theta_2 x_i + \varepsilon_{2i}, \quad \theta_2 = 6, \quad \varepsilon_{2i} \sim iid \ N(0, 0.1).$$

Therefore, the true identified-set is $\Gamma_0 = \left[ \frac{E_F(Y_1Z)}{E_F(ZX)'}, \frac{E_F(Y_2Z)}{E_F(ZX)'} \right] = [2, 6]$. The parameters of the Dirichlet process are fixed as $n_0 = 20$ and

$$F_0 = N_4\left( \begin{pmatrix} 0 \\ 4 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 0.1 & 0 & 0.2 & \frac{3}{2} \\ 0 & 0.1 & 0.2 & 3 \\ 0.2 & 0.2 & 0.1 & 0.5 \\ \frac{3}{2} & 3 & 0.5 & 0.1 \end{pmatrix} \right).$$
Moreover, we set: \( n = 1000 \). We draw 1000 intervals \( \left[ \frac{E_F(Y_1 Z)}{E_F(Z X)}, \frac{E_F(Y_2 Z)}{E_F(Z X)} \right] \) from the prior and posterior distributions of \( F \). In Figure 3 we represent the prior and posterior coverage functions of the random interval \( \left[ \frac{E_F(Y_1 Z)}{E_F(Z X)}, \frac{E_F(Y_2 Z)}{E_F(Z X)} \right] \) for each value of \( \gamma \in [-3, 12] \). Figure 4 displays the intervals drawn from the prior and posterior distributions (on the vertical axis) against the true interval \([2, 6]\) (on the horizontal axis).

**Example 4.3. Linear Regression with errors in Regressors.** This is the well-known linear errors-in-variables structural model considered in Frisch [1934] and Klepper and Leamer [1984]. For simplicity, we focus here on the univariate linear regression model. Let \( Y \) be an observable random variable satisfying the model \( Y = \gamma \xi + u \) where \( \xi \) is an unobservable random variable such that \( E(\xi) = 0, \) \( Var(\xi) = \tau^2 \) and for which only realizations affected by an error are available: \( Z = \xi + v \). The error terms \((u, v)\) are zero-
mean jointly distributed random variables, independent of $\xi$, and with a variance-covariance matrix $\Sigma$ which may be diagonal.

Let $\Sigma = \text{diag}(\sigma_u^2, \sigma_v^2)$, the structural parameters are $\gamma$, $\sigma_u^2$, $\sigma_v^2$ and $\tau^2$ while $\xi$ is the incidental parameter. Denote $\sigma_{yz} = \text{Cov}(Y, Z)$, $\sigma_{zz} = \text{Var}(Z)$ and $\sigma_{yy} = \text{Var}(Y)$. Due to the endogeneity of $Z$, the parameter $\gamma$ lies in the identified set $\Gamma_I = [\theta_1, \theta_2]$

$$[\theta_1, \theta_2] = \left[ \min \left( \frac{\sigma_{yz}}{\sigma_{zz}}, \frac{\sigma_{yy}}{\sigma_{yz}} \right), \max \left( \frac{\sigma_{yz}}{\sigma_{zz}}, \frac{\sigma_{yy}}{\sigma_{yz}} \right) \right],$$

where $\frac{\sigma_{yz}}{\sigma_{zz}}$ is the regression line of $Y$ onto $Z$ and $\frac{\sigma_{yy}}{\sigma_{yz}}$ is the reverse regression line of $Z$ onto $Y$, both without intercept.

We specify a Dirichlet process prior on the joint probability distributions $F_{yz}$ of $(Y, Z)$. In the simulation exercise that we have performed the data have been generated as $\xi_i \sim \text{iid } \mathcal{N}(0, 1)$, $(u_i, v_i)' \sim \text{iid } \mathcal{N}_2(0, I_2)$

$$Z_i = \xi_i + v_i, \quad Y_i = \gamma \xi_i + u_i, \quad \gamma = 1, \quad i = 1, \ldots, n$$

where $I_2$ is the 2-dimensional identity matrix. As $\gamma > 0$ the identified set is $[\frac{\sigma_{yz}}{\sigma_{zz}}, \frac{\sigma_{yy}}{\sigma_{yz}}] = [1/2, 1]$. The prior is $F_{yz} \sim \text{Dir}(n_0, F_0)$ with $n_0 = 20$ and

$$F_0 = \mathcal{N}_2\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0.9 \\ 0.9 & 2 \end{pmatrix} \right).$$

The sample size is $n = 1000$ and we have drawn 1000 intervals from the prior and posterior coverage functions. The results are shown in Figures 5 and 6.

**Figure 5:** Prior and Posterior coverage functions $p_{\gamma}$ of $\Gamma_I$.

**Example 4.4.** Binary outcome and missing data. This example has been considered very often in the literature, see Manski [1990], Sebastiani and Ramoni [2000] and Kitagawa [2011a]. We consider a random binary variable $Y$ taking values in $\{0, 1\}$ and
a \( n \)-random sample of it. This random sample is incomplete in the sense that, for some units, the outcome \( Y \) is not observed due to unknown reasons. Therefore, the observed sample is \( x_i = (y_i d_i, d_i), \; i = 1, \ldots, n \), where \( y_i \) is the realization of \( Y \) for the \( i \)-th unit and \( d_i \) is a dummy variable that takes the value 1 if the corresponding \( y_i \) is observed and 0 otherwise. We denote \( p_{kj} = \mathbb{P}(y_i = k, d_i = j), \; j, k \in \{0, 1\} \) and the parameter of interest is \( p_1 := \mathbb{E}(Y) = \mathbb{P}(Y = 1) = p_{10} + p_{11} \). The likelihood of \( p = (p_{11}, p_{10}, p_{01}, p_{00}) \) is equal to

\[
l(x_1, \ldots, x_n | p) = \frac{n!}{n_1!n_0!m!} p_{11}^{n_1} p_{01}^{n_0} (p_{10} + p_{00})^m, \tag{15}
\]

where \( n_1 = \sum_{i=1}^n y_i d_i \) is the number of observed units with \( y_i = 1 \), \( n_0 = \sum_{i=1}^n (1 - y_i) d_i \) is the number of observed units with \( y_i = 0 \) and \( m = \sum_{i=1}^n (1 - d_i) \) is the number of unobserved units. The identified parameter, i.e. the minimal sufficient parameter, is \( (p_{11}, p_{01}, \tilde{p}_0) \), where \( \tilde{p}_0 = p_{10} + p_{00} \) and the probability of interest \( p_1 \) is not identified. Condition (4) writes

\[
p_1 \in [p_{11}, p_{11} + \tilde{p}_0]
\]

with \( \gamma = p_1, \; \theta = (p_{11}, p_{11} + \tilde{p}_0) \) and \( \Gamma_I \equiv [p_{11}, p_{11} + \tilde{p}_0] \subset [0, 1] \). We use a Bayesian nonparametric approach where we specify a Dirichlet prior distribution on \( (p_{11}, p_{01}, \tilde{p}_0) \):

\[
(p_{11}, p_{01}, \tilde{p}_0) \sim \text{Dir}(\alpha),
\]

where \( \text{Dir} \) denotes a Dirichlet distribution and \( \alpha = (\alpha_1, \alpha_2, \alpha_3) \) is a vector of positive numbers. The Bayesian model is the following

\[
x_1, \ldots, x_n | p_{11}, p_{01}, \tilde{p}_0 \sim \mathcal{M}u(n, (p_{11}, p_{01}, \tilde{p}_0))
\]

\[
(p_{11}, p_{01}, \tilde{p}_0) \sim \text{Dir}(\alpha)
\]

where \( \mathcal{M}u \) denotes a joint Multinomial distribution. The prior coverage function is \( p_{r_I} = P(p_{01} \leq \gamma \leq p_{01} + \tilde{p}_0) \) where \( p_{01} \sim \text{Be}(\alpha_2, \alpha_1 + \alpha_3), \; (p_{01} + \tilde{p}_0) \sim \text{Be}(\alpha_2 + \alpha_3, \alpha_1) \) and \( \text{Be} \) denotes a Beta distribution. The posterior distribution is

\[
(p_{11}, p_{01}, \tilde{p}_0) | x_1, \ldots, x_n \sim \text{Dir}(\alpha_*), \; \alpha_* = (\alpha_1 + n_1, \alpha_2 + n_0, \alpha_3 + m) \tag{16}
\]
and the posterior coverage functional is defined accordingly. In our simulation exercise we generate an $n$-sample of observations in the following way:

$$y_i \sim B(p_y), \quad p_y = 0.8$$

$$d_i \sim B(p_d), \quad p_d = 0.5$$

where $B(q)$ denotes a Bernoulli distribution with probability $q$. Thus, $x_i = (y_i, d_i)$ and $p_{11} = p_y p_d$, $p_{10} = p_y (1 - p_d)$, $p_{01} = (1 - p_y) p_d$, $p_{00} = (1 - p_y) (1 - p_d)$ and the true set $[p_{11}, \tilde{p}_0]$ is $[0.4, 0.9]$. The parameter $\alpha$ of the Dirichlet distribution is set equal to $\alpha = (2, 3, 1)$. We draw 1000 times $p_{11}$ and $\tilde{p}_0$ from the prior and posterior distributions. The results are shown in Figures 7 and 8. Figure 8 has been obtained by taking the first 100 drawings from the prior and posterior distributions and represents the corresponding intervals $\Gamma_I$ (on the vertical axes) against the true interval $[0.4, 0.9]$. Once the posterior distribution is found we could take as an estimator for $[p_{11}, p_{11} + \tilde{p}_0]$ the interval

$$[\tilde{p}_{11}, \tilde{p}_{11} + \tilde{p}_0] = \left[\frac{\alpha_1 + n_1}{\bar{\alpha} + n}, \frac{\alpha_1 + \alpha_2 + n_1 + m}{\bar{\alpha} + n}\right],$$

where $\bar{\alpha} = \alpha_1 + \alpha_2 + \alpha_3$, as proposed in Sebastiani and Ramoni [2000]. The lower and upper bounds of the interval on the right hand side are the posterior means of $p_{11}$ and $p_{11} + \tilde{p}_0$, respectively.

![Prior coverage function](image1.png)

![Posterior coverage function](image2.png)

Figure 7: Prior and Posterior coverage functions $p_{11}$ of $\Gamma_I$.

### 5 Marginal Identification

In Sections 3 and 4 we have considered the statistical model containing all the parameters and eventually augmented by an additional parameter. When the parameter of interest is $\gamma$, that is, a sub-parameter, we might want to perform the analysis by getting rid of the other parameters of the model that are not of interest. Therefore, it is natural to examine the marginal model in this sub-parameter.
5.1 Marginal Identification of hyperparameters

We start by considering the marginalization of models analyzed in Section 3. Let $\theta = (\beta, \gamma)$ denotes the whole parameter of the model and $\gamma$ be the parameter of interest. In terms of probability distributions the marginal model writes

$$P^\gamma(E) = \int P^{\beta}(E) \mu_\beta(d\beta|\gamma), \quad \forall E \in \mathcal{X}$$

where $\mu_\beta(\beta|\gamma)$ is the conditional prior measure of $\beta$ given $\gamma$. This model keeps unchanged the marginal prior on $\gamma$. In terms of the conditional probability density function of $\beta|\gamma$ the marginal likelihood writes

$$p(x|\gamma) = \int p(x|\beta, \gamma) \mu_\beta(\beta|\gamma) d\beta.$$

The predictive distribution $p(x)$ is identical as in the full model and the posterior $\mu_\gamma(\gamma|x)$ is the marginal posterior on $\gamma$ of the joint posterior $\mu(\beta, \gamma|x)$. This framework has a particular interest in the case where $\gamma$ is not identified in the full model but may be identified in the marginal model as it is shown by the following example.

**Example 5.1. Classical model of hyperparameter.** Let us consider the following Bayesian model

$$x_i|\beta, \gamma \sim iid \mathcal{N}(\beta, \sigma^2), \quad i = 1, \ldots, n$$
$$\beta|\gamma \sim \mathcal{N}(\gamma, \sigma_\beta^2)$$
$$\gamma \sim \mathcal{N}(\gamma_0, \tau_0^2)$$

where $\gamma$ is the hyperparameter of the prior distribution and $\sigma^2, \sigma_\beta^2, \gamma_0, \tau_0^2$ are known parameters. The parameter $\gamma$ is not identified in the sampling model. The minimal sufficient
\( \sigma \)-field \( \mathcal{A} \mathcal{X} \) is almost surely equal to the \( \sigma \)-field generated by \( \beta \).

The marginal model is
\[
\begin{pmatrix}
  x_1 \\
  \vdots \\
  x_n
\end{pmatrix} | \gamma \sim \mathcal{N} \left( \gamma_\iota, \sigma^2 \mathcal{I}_n + \sigma_0^2 \mathcal{I}' \right)
\]
where \( \iota = (1, \ldots, 1)' \) is \( n \times 1 \) and \( \mathcal{I}_n \) is the \( n \)-dimensional identity matrix. Therefore, \( \gamma \) is identified in the marginal model.

Note that, in the previous example, even if \( \gamma \) is identified in the marginal model it is not exactly estimable: the Doob theorem does not apply because the marginal sampling distribution is not \( i.i.d. \). The only case of exact estimability would be the case where the conditional distribution of \( \gamma \) given \( \beta \) is a degenerated Dirac measure on a deterministic function of \( \beta \).

More generally, we may consider the asymptotic behavior of a sub-parameter, possibly non-identified, in the global model.

**Theorem 5.** Let us consider a Bayesian model \( \{ \Theta \times \mathcal{X}, \mathcal{A} \otimes \mathcal{X}, \Pi \} \) with a filtration \( \mathcal{X}_n \uparrow \mathcal{X}_\infty \) and where the identified parameter \( \mathcal{A} \mathcal{X}_\infty \) is asymptotically exactly estimable, i.e. \( \mathcal{A} \mathcal{X}_\infty \subset \overline{\mathcal{X}}_\infty \). Let \( c \) be an integrable function defined on \( \mathcal{A} \), then:
\[
E(c|\mathcal{A}_n) \rightarrow E(c|\mathcal{A}\mathcal{X}_\infty), \quad \Pi - a.s.
\]

**Proof.** By the martingale convergence theorem
\[
E(c|\mathcal{A}_n) \rightarrow E(c|\mathcal{A}_\infty), \quad \Pi - a.s.
\]
Moreover, as \( \mathcal{A} \parallel \mathcal{X}_\infty|\mathcal{A}\mathcal{X}_\infty \) we have, see Florens et al. [1990]:
\[
E(c|\mathcal{X}_\infty) = E(E(c|\mathcal{A}\mathcal{X}_\infty)|\mathcal{X}_\infty) = E(c|\mathcal{A}\mathcal{X}_\infty), \quad \Pi - a.s.
\]
due to the exact estimability of \( \mathcal{A}\mathcal{X}_\infty \).

This means that the posterior mean of \( c \) converges \( \Pi \)-a.s. to the conditional prior mean given the identified parameter. Remark that the \( a.s. \) is with respect to the joint distribution.

**5.2 Marginal models in Bayesian set estimation**

Consider now the semiparametric model of Section 4. Let \( X \) be the sample generated given \( F \), which is an identified parameter. The parameter of the model can be written as \( \theta = \phi(F) \), for some functional \( \phi \); the parameter of interest is denoted by \( \gamma \) and is related to \( \theta \) by relation (4): \( A(\theta, \gamma) \in \mathcal{A}_0 \). Let \( \mathcal{G} \) be the \( \sigma \)-field on \( \Gamma \), by using the structural relation
we can provide $\gamma \in (\Gamma, \mathcal{G})$ with a prior distribution conditional on $\theta$. This prior will put all its mass on the set of the $\gamma$s that satisfy the constraint $A(\theta, \gamma) \in A_0$. In alternative, we may relax this constraint into a constraint on the hyperparameter of the distribution of $\gamma$, as illustrated by the examples 4.2 and 4.3 below.

The marginal posterior model, in terms of probability distributions, writes

$$
\forall \Gamma_1 \in \mathcal{G}, \quad \mu_\gamma(\gamma \in \Gamma_1 | \mathcal{X}) = \int_{\Gamma_1} \int_\Theta \mu_\theta(\theta | \mathcal{X}) \mu_\gamma(\gamma | \theta) d\theta d\gamma,
$$

where $\mu_\gamma(\cdot | \theta)$ denotes the conditional prior distribution of $\gamma$ given $\theta$, $\mu_\theta(\cdot | \mathcal{X})$ denotes the posterior distribution of $\theta$ and $\mu_F(\cdot | \mathcal{X})$ denotes the posterior distribution of $F$. In the second equality we rewrite the integral in terms of the posterior distribution of $F$; this aims to stress the fact that the prior distribution of $\theta$ is recovered from the Dirichlet process prior specified for $F$. Theorem 5 applies also to this case.

While the marginal posterior density (or distribution) of $\gamma$ is not usually known in closed-form – in particular if $\theta$ is infinite dimensional – we can have perfect knowledge of it because we can easily simulate from this distribution with the help of a random generator. In general, the model is specified in a way such that the marginal prior distribution $\mu_\theta$ of $\theta$ and the posterior $\mu_\theta(\cdot | \mathcal{X})$ of $\theta | x$ are easy to compute or to simulate from. In the same way, is usually easy to simulate from $\mu_\gamma(\cdot | \theta)$. Thus, the simulations from $\mu_\gamma(\cdot | \mathcal{X})$, i.e. from the posterior of the marginal model, are very easy: we first simulate $\theta$ given $x$ from $\mu_\theta(\cdot | \mathcal{X})$ and then, for each drawing of $\theta$, we simulate $\gamma$ given $\theta$ from $\mu_\gamma(\cdot | \theta)$. This simulation scheme produces in the end simulations from $\mu_\gamma(\cdot | \mathcal{X})$ and this is due to the lack of identification of $\gamma$ which implies that $\gamma \parallel x | \theta$, see section 2.3. On the other side, computation of the prior distribution of $\theta$ given $\gamma$ and of the marginal likelihood may be very difficult, but as already stressed these computations are not necessary for the simulation of the marginal posterior of $\gamma$ $\mu_\gamma(\cdot | \mathcal{X})$.

Having a marginal posterior distribution $\mu_\gamma(\cdot | \mathcal{X})$ of the parameter of interest $\gamma$ is important for instance in a “decision problem setting”. In fact, knowledge of $\mu_\gamma(\cdot | \mathcal{X})$ allows to select the most likely value of $\gamma$ or the region inside $\Gamma_I$ with the highest posterior probability. This choice of a value or a region inside $\Gamma_I$ is clearly affected by the choice of the prior. A study of this phenomenon is beyond the scope of this paper and has been addressed by Kitagawa [2011a,b].

In the next section we develop further the semiparametric examples of Section 4.1 and we provide the parameter $\gamma$ is provided with a conditional prior distribution given $\theta = \phi(F)$. To simplify notation we denote this distribution by $\mu_\gamma^F$: $\mu_\gamma^F := \mu_\gamma(\cdot | \phi(F))$.
5.3 Examples

Example 4.1 (continued). Interval Censored Data. Suppose that we are not only interested in the random set \( \Gamma_I = [\theta_1, \theta_2] \equiv [E_{F_1}(Y_1), E_{F_2}(Y_2)] \) but also in the posterior distribution of \( \gamma \) on this interval. Knowledge of the posterior distribution of \( \gamma \) is useful because it is informative about the areas of the identified region \( \Gamma_I \) where the parameter is more likely. The distribution of \( \theta = (\theta_1, \theta_2) \) is specified as in Section 4 through the specification of a prior Dirichlet process for \( F \). Under the assumption that \( Y_1 \parallel Y_2 | F \), the Bayesian hierarchical model is

\[
\begin{align*}
(y_{11}, \ldots, y_{1n}) | F, \gamma & \sim iid \ F_1 \\
(y_{21}, \ldots, y_{2n}) | F, \gamma & \sim iid \ F_2 \\
F_1 & \sim Dir(n_0^1, F_0^1) \\
F_2 & \sim Dir(n_0^2, F_0^2) \\
\gamma | F & \sim \mu_0^F.
\end{align*}
\]

We consider four different specifications for \( \mu_0^F \) in our simulation exercise:

1. \( \gamma | F \sim \mathcal{N}(\gamma_0, \sigma_0^2) \), \( \gamma_0 = \frac{\bar{y}_1^i + \bar{y}_2^i}{2} \) and \( \bar{y}_0^i := \mathbf{E}_{F_i}(Y_i) = \int y F_i(dy), \ i = 1, 2 \) and we discard the drawings of \( \gamma \) that do not belong to the interval \([\mathbf{E}_{F_1}(Y_1), \mathbf{E}_{F_2}(Y_2)]\);

2. \( \gamma | F \sim \mathcal{N}(0, \sigma_0^2) \) truncated to the interval \([\mathbf{E}_{F_1}(Y_1), \mathbf{E}_{F_2}(Y_2)]\);

3. \( \gamma | F \sim U[\mathbf{E}_{F_1}(Y_1), \mathbf{E}_{F_2}(Y_2)] \) (flat prior);

4. \( \gamma | F \sim \mathcal{B}(\mathbf{E}_{F_1}(Y_1), \mathbf{E}_{F_2}(Y_2), p, q) \), i.e. \( \gamma \) has a Beta prior distribution with support \([\mathbf{E}_{F_1}(Y_1), \mathbf{E}_{F_2}(Y_2)]\) and shape parameters \( p \) and \( q \). The corresponding probability density function is:

\[
f(\gamma | F) = \frac{\left( \gamma - \mathbf{E}_{F_1}(Y_1) \right)^{p-1} \left( \mathbf{E}_{F_2}(Y_2) - \gamma \right)^{q-1}}{B(p, q) \left( \mathbf{E}_{F_2}(Y_2) - \mathbf{E}_{F_1}(Y_1) \right)^{p+q-1}}
\]

where \( B(p, q) \) is the beta function.

The parameter \( \gamma \) is not identified by the data since its posterior distribution depends on the data only through \( F \). This means that the posterior distribution of \( \gamma \) is affected by the data only indirectly through \( F \) by replacing the moments \( \mathbf{E}_{F_i}(Y_i), \ i = 1, 2 \), in the prior distribution with \( \mathbf{E}(Y_i | F_1, y_{i1}, \ldots, y_{im}) \), for \( i = 1, 2 \). The notation \( \mathbf{E}(Y_i | F_1, y_{i1}, \ldots, y_{im}) \) is for the mean of \( Y_i \) drawn from the posterior distribution of \( \phi(F) \).

Our simulation exercise allows to visualize the marginal prior and posterior probability distributions of \( \gamma \). We generate an \( n \)-sample of observations of \( (Y_1, Y_2) \) as in section 4.1:

\[
Y_1 \sim \mathcal{N}(0, 0.1), \quad Y_2 \sim \mathcal{N}(5, 0.1).
\]
The parameters are fixed as follows: \( n = 1000, n_1^0 = 10, n_2^0 = 20, F_0^1 = \mathcal{N}(0,1), F_0^2 = \mathcal{N}(10,1) \) and \( \tau_0^2 = 1, \sigma_0^2 = 2, p = 1 \) and \( q = 0.5 \). The true identified set is \( \Gamma_I = [0, 5] \).

We draw 1000 times from the marginal prior and posterior distribution of \( \gamma \). The simulation scheme is the following: for each \( 1 \leq j \leq 1000 \), draw \( F^{(j)} \) from \( \mu_F \) (resp. \( \mu_F\cdot|X) \), compute \( \theta^{(j)}_i = \mathbb{E}_{F^{(j)}}(Y_i), i = 1, 2 \) and draw \( \gamma^{(j)} \) from \( \mu^{F^{(j)}}_\gamma \) (resp. \( \mu^{F^{(j)}}_\gamma(\cdot|F^{(j)}, X) \)). We show in Figures 9 and 10 the kernel estimate of the marginal prior (in blue) and posterior (in red) density functions of \( \gamma \) computed by using the \( \gamma \)'s drawn from the marginal prior and posterior distributions, respectively. Each panel corresponds to one of the four specifications for \( \mu_F \). We see how the areas with higher probability strongly depend on the prior specification.

![Figure 9](image1.png)

**Figure 9:** Kernel estimation of the prior (in blue) and posterior (in red) probability density functions. Normal kernel and bandwidth equal to \( h = 0.1 \).

![Figure 10](image2.png)

**Figure 10:** Kernel estimation of the prior (in blue) and posterior (in red) probability density functions. Normal kernel and bandwidth equal to \( h = 0.1 \) for the uniform and \( h = 0.15 \) for the Beta distribution.
Example 4.2 (continued). Interval Regression Model. Suppose that the parameter \( \gamma \) is the parameter of interest and we provide it with a prior probability distribution conditioned on \( F \). The Bayesian hierarchical model of Section 4 becomes

\[
(y_{1i}, y_{2i}, x_i, z_i) | F, \gamma \sim iid \quad F, \quad i = 1, \ldots, n
\]

\[
F \sim Dir(n_0, F_0)
\]

\[
\gamma | F \sim \mu^F_\gamma.
\]

The prior \( \mu^F_\gamma \) is constructed so that the constraint (13) on \( \gamma \) is satisfied. In particular, we consider four different prior specifications in our simulation exercise (where \( X \) and \( Z \) are univariate random variables):

1. \( \gamma | F \sim N(\gamma_0, \sigma_0^2) \), \( \gamma_0 = \frac{\gamma_0^1 + \gamma_0^2}{2} \) and \( \gamma_0^j := E_F(Y_j Z) = \int y_j z F(dy, dz), \; j = 1, 2 \) and we discard the drawings of \( \gamma \) that do not belong to the interval \( \left[ \frac{E_P(Y_1 Z)}{E_F(ZX)}, \frac{E_P(Y_2 Z)}{E_F(ZX)} \right] \);

2. \( \gamma | F \sim N(0, \sigma^2_0) \) truncated to the interval \( \left[ \frac{E_P(Y_1 Z)}{E_F(ZX)}, \frac{E_P(Y_2 Z)}{E_F(ZX)} \right] \);

3. \( \gamma | F \sim U \left[ \frac{E_P(Y_1 Z)}{E_F(ZX)}, \frac{E_P(Y_2 Z)}{E_F(ZX)} \right] \) (flat prior);

4. \( \gamma | F \sim Be \left( \frac{E_P(Y_1 Z)}{E_F(ZX)}, \frac{E_P(Y_2 Z)}{E_F(ZX)}, p, q \right) \), i.e. \( \gamma \) has a Beta prior distribution with support \( \left[ \frac{E_P(Y_1 Z)}{E_F(ZX)}, \frac{E_P(Y_2 Z)}{E_F(ZX)} \right] \) and shape parameters \( p \) and \( q \).

The parameter \( \gamma \) is not identified by the data since its posterior distribution depends on the data only through \( F \). This means that the prior-to-posterior transformation of \( \gamma \) is not directly affected by the data but it is affected through \( F \) by replacing the moments \( E_F(Y_j Z) \) and \( E_F(ZX) \) with \( E(Y_j Z | F, \{w_i\}_{i=1}^n) \) and \( E(ZX | F, \{w_i\}_{i=1}^n) \), for \( i = 1, 2 \), respectively. The notations \( E(Y_j Z | F, \{w_i\}_{i=1}^n) \) and \( E(ZX | F, \{w_i\}_{i=1}^n) \) refer to the mean of \( Y_j Z, \; i = 1, 2 \), and of \( ZX \), respectively, computed with a \( F \) drawn from the posterior of \( F \).

Our simulation exercise allows to see how the support of the marginal posterior and posterior distributions overlaps with the true identified set. We generate an \( n \)-sample of observations of \( (Y_1, Y_2, X, Z) \) as in section 4.1:

\[
\begin{align*}
    z_i & \sim iid \ U[0, 1], \quad x_i = z_i + u_i, \quad u_i \sim iid \ N(0, 1) \\
    y_{1i} & = \theta_1 x_i + \varepsilon_{1i}, \quad \theta_1 = 2, \quad \varepsilon_{1i} \sim iid \ N(0, 0.1) \\
    y_{2i} & = \theta_2 x_i + \varepsilon_{2i}, \quad \theta_2 = 6, \quad \varepsilon_{2i} \sim iid \ N(0, 0.1).
\end{align*}
\]

Then, the true identified set is \([2, 6]\).

The parameters of the Dirichlet process are fixed as \( n_0 = 20 \) and

\[
F_0 = N_4 \left( \begin{pmatrix} 0 \\ 4 \\ 0 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 0.1 & 0 & 0.2 & \frac{3}{2} \\ 0 & 0.1 & 0.2 & 3 \\ 0 & 0.2 & 0.1 & 0.5 \\ 0.2 & 0.3 & 0.5 & 0.1 \end{pmatrix} \right).
\]
Moreover, we set: \( n = 1000, \tau_0^2 = 1, \sigma_0^2 = 2, p = 1 \) and \( q = 0.5 \).

We draw 1000 times from the marginal prior and posterior distributions of \( \gamma \). The simulation scheme is the following: for each \( 1 \leq j \leq 1000 \), draw \( F^{(j)} \) from \( \mu_F \) (resp. \( \mu_F(\cdot|X) \)), compute \( E_{F^{(j)}}(Y_iZ), i = 1, 2 \) and \( E_{F^{(j)}}(ZX) \), and draw \( \gamma^{(j)} \) from \( \mu_{\gamma}^{F^{(j)}} \) (resp. \( \mu_{\gamma}^{F^{(j)}}(\cdot|F^{(j)}, X) \)). Figures 11 and 12 show the marginal prior (in blue) and posterior (in red) probability density functions corresponding to the four prior specifications (obtained with a kernel density estimation and the 1000 drawings from the marginal prior and posterior respectively). Each panel corresponds to one of the four specifications for \( \mu_{\gamma}^F \).

**Figure 11:** Kernel estimation of the prior (in blue) and posterior (in red) probability density functions. Normal kernel and bandwidth equal to \( h = 0.2 \).

**Figure 12:** Kernel estimation of the prior (in blue) and posterior (in red) probability density functions. Normal kernel and bandwidth equal to \( h = 0.2 \).

**Example 4.4 (continued).** Binary outcome and missing data. The parameter of interest is now the probability \( p_1 = p_{11} + p_{10} \) which is not identified but is known to belong
to the interval $\Gamma_I = [p_{11}, p_{11} + \tilde{p}_0]$. We provide $p_1$ with a prior distribution conditioned on $p_{11}$ and $\tilde{p}_0$ and the Bayesian hierarchical model is

$$x_1, \ldots, x_n|p_{11}, p_{01}, \tilde{p}_0 \sim \mathcal{M}u(n, (p_{11}, p_{01}, \tilde{p}_0))$$
$$\langle p_{11}, p_{01}, \tilde{p}_0 \rangle \sim \text{Dir}(\alpha)$$
$$p_1|p_{11}, \tilde{p}_0 \sim \mu_{p_1}(p_1|p_{11}, \tilde{p}_0)$$

where $\mathcal{M}u$ denotes a Multinomial distribution and $\text{Be}$ a Beta distribution. We have considered four specifications for $\mu_{p_1}(p_1|p_{11}, \tilde{p}_0)$:

1. $p_1|p_{11}, \tilde{p}_0 \sim \mathcal{N}(\gamma_0, 1)$, $\gamma_0 = p_{11} + \frac{\tilde{p}_0}{2}$ and we discard the drawings of $p_1$ that do not belong to the interval $[p_{11}, p_{11} + \tilde{p}_0]$;
2. $p_1|p_{11}, \tilde{p}_0 \sim \mathcal{N}(0, 2)$ truncated to the interval $[p_{11}, p_{11} + \tilde{p}_0]$;
3. $p_1|p_{11}, \tilde{p}_0 \sim \text{U}[p_{11}, p_{11} + \tilde{p}_0]$ (flat prior);
4. $p_1|p_{11}, \tilde{p}_0 \sim p_{11} + \tilde{p}_0 \text{Be} \left( \frac{1}{\tilde{p}_0}, \frac{2}{\tilde{p}_0} \right)$.

We implement the same simulation scheme as in section 4: $y_i \sim B(p_y), \ p_y = 0.8$
$d_i \sim B(p_d), \ p_d = 0.5$

so that the true identified set is $[0.4, 0.9]$ and the parameter $\alpha$ of the Dirichlet distribution is set equal to $\alpha = (2, 3, 1)$. The simulation scheme is the following: for each $1 \leq j \leq 1000$, draw $\langle p_{11}^{(j)}, p_{01}^{(j)}, \tilde{p}_0^{(j)} \rangle$ from $\text{Dir}(\alpha)$ (resp. $\text{Dir}(\alpha_*)$) and draw $p_1^{(j)}$ from $\mu_{p_1}(p_1|p_{11}, \tilde{p}_0)$ (resp. from $\mu_{p_1}(p_1|p_{11}, \tilde{p}_0, \mathcal{X})$). The results are shown in Figures 13 and 14 where each panel corresponds to one of the four specifications for $\mu_{p_1}(p_1|p_{11}, \tilde{p}_0)$.

**Appendix**

**A Representation of the Dirichlet Process**

In this appendix we recall the main results concerning the representation of the Dirichlet process used for simulations and we refer to Florens and Rolin [1994] and Florens [2002] for more details.

1. Sethuraman [1994] - Rolin [1992] decomposition. Consider the Dirichlet process $F \sim \text{Dir}(n_0, F_0)$ and suppose for simplicity that $F_0$ is a diffuse probability, i.e.
\( \gamma \sim \mathcal{N}(\gamma_\ast, 1) \) by discarding all the drawings that are not inside \([p_{11}, p_{11} + \tilde{p}_0]\).

\( \gamma \sim \mathcal{N}(0, 2) \) truncated to \([p_{11}, p_{11} + \tilde{p}_0]\).

Figure 13: Kernel estimation of the prior (in blue) and posterior (in red) probability density functions. Normal kernel and bandwidth equal to \(h = 0.1\).

\( p_{1} \sim \mathcal{U}[p_{11}, p_{11} + \tilde{p}_0] \)

\( \gamma \sim \text{Beta}(p_{11}, p_{11} + \tilde{p}_0, 1, 0.5) \)

Figure 14: Kernel estimation of the prior (in blue) and posterior (in red) probability density functions. Normal kernel and bandwidth equal to \(h = 0.1\).

\( F_0(x) = 0, \forall x \). The probability distribution \( F \) is generated by a \( \text{Dir}(n_0, F_0) \) process if and only if \( F \) can be written

\[
F = \sum_{j=1}^{\infty} \alpha_j \delta_{\xi_j}
\]

(17)

where \( \delta_{\xi_j} \) is a Dirac mass at the value \( \xi_j \) and the sequences \( \{\alpha_j\} \) and \( \{\xi_j\} \) are generated as

- \( \xi_j \sim \text{i.i.d. } F_0 \);
- \( \alpha_j = v_j \prod_{k=1}^{j-1} v_k \), with \( v_k \sim \text{i.i.d. } \text{Be}(1, n_0) \) for \( k = 1, \ldots, j \);
- \( \{\xi_j\} \) and \( \{v_j\} \) are independent.

Suppose now that we have observed an \( n \)-sample \( x_1, \ldots, x_n \) of realizations of \( X \), where \( X | F \sim F \), and denote by \( F_n \) the empirical cumulative distribution function. The rep-
representation of the posterior Dirichlet process $F \sim \text{Dir}(n_*, F_*)$, with $n_* = n_0 + n$ and $F_* = n_0 F_0 + n F_n$, is

$$F = (1 - \gamma) \sum_{j=1}^{\infty} \alpha_j \delta_{\xi_j} + \gamma \sum_{j=1}^{n} \beta_j \delta_{x_j}.$$  

(18)

where the sequences $\{\alpha_j\}$ and $\{\xi_j\}$ are generated as above, $\gamma \sim \text{Be}(n, n_0)$ and $(\beta_1, \ldots, \beta_n)$ follows a Dirichlet distribution with parameter $(1, \ldots, 1)$ over the simplex $S_{n-1}$.

2. **Draw from the Dirichlet process.** To draw a trajectory from a Dirichlet process we use the Sethuraman - Rolin representation described in 1. In practice, (17) cannot be implemented since it requires an infinity of drawings. Therefore, we truncate the sum at $K < \infty$ and we re-normalize:

$$F_K = \frac{1}{\sum_{k=1}^{K} \alpha_k} \sum_{k=1}^{K} \alpha_k \delta_{\xi_k}.$$  

Once we have drawn $\{\xi_k\}$ and $\{\alpha_k\}$ for the Dirichlet process, we can compute any functional of the Dirichlet, for instance $E_F(Y)$:

$$E_F(Y) = \frac{1}{\sum_{k=1}^{K} \alpha_k} \sum_{k=1}^{K} \alpha_k \xi_k.$$  

The error that we make by approximating $F$ by $F_K$ is measured by $\varepsilon = \sup_B |F(B) - F_K(B)|$, where $B$ is a Borel set of $\mathbb{R}^m$ (if $F$ is a distribution on $\mathbb{R}^m$), and is equal to $\sum_{j=K+1}^{\infty} \alpha_j$. The probability distribution of $\varepsilon$ is known and has a density function equal to

$$\frac{n_0^K}{\Gamma(K)} \left( \ln \frac{1}{\varepsilon} \right)^{K-1} \varepsilon^{n_0-1}$$  

over $[0, 1]$.

**References**


