Semi-parametric Bayesian Partially Identified Models based on Support Function*

Yuan Liao† Anna Simoni‡

University of Maryland CNRS and THEMA

November 2013

Abstract

We provide a comprehensive semi-parametric study of Bayesian partially identified econometric models. While the existing literature on Bayesian partial identification has mostly focused on the structural parameter, our primary focus is on Bayesian credible sets (BCS's) of the unknown identified set and the posterior distribution of its support function. We construct a (two-sided) BCS based on the support function of the identified set. We prove the Bernstein-von Mises theorem for the posterior distribution of the support function. This powerful result in turn infers that, while the BCS and the frequentist confidence set for the partially identified parameter are asymptotically different, our constructed BCS for the identified set has an asymptotically correct frequentist coverage probability. Importantly, we illustrate that the constructed BCS for the identified set does not require a prior on the structural parameter. It can be computed efficiently for subset inference, especially when the target of interest is a sub-vector of the partially identified parameter, where projecting to a low-dimensional subset is often required. Hence, the proposed methods are useful in many applications.

The Bayesian partial identification literature has been assuming a known parametric likelihood function. However, econometric models usually only identify a set of moment

*The authors are grateful to Federico Bugni, Ivan Canay, Joel Horowitz, Enno Mammen, Francesca Molinari, Andriy Norets, Adam Rosen, Frank Schorfheide, Jörg Stoye and seminar participants at CREST (LS and LMI), Luxembourg, Mannheim, Tsinghua, THEMA, University of Illinois at Urbana-Champaign, 4^{th} French Econometrics Conference, Bayes in Paris workshop, Oberwolfach workshop, 2013 SBIES meeting, CMES 2013, EMS 2013 in Budapest and ESEM 2013 in Gothenburg for useful comments. Anna Simoni gratefully acknowledges financial support from the University of Mannheim through the DFG-SNF Research Group FOR916, ANR-13-BSH1-0004, labex MMEDII (ANR11-LBX-0023-01) and hospitality from University of Mannheim and CREST.

†Department of Mathematics, University of Maryland at College Park, College Park, MD 20742 (USA). Email: yuanliao@umd.edu

‡CNRS and THEMA, Université de Cergy-Pontoise - 33, boulevard du Port, 95011 Cergy-Pontoise (France). Email: simoni.anna@gmail.com

inequalities, and therefore using an incorrect likelihood function may result in misleading inferences. In contrast, with a nonparametric prior on the unknown likelihood function, our proposed Bayesian procedure only requires a set of moment conditions, and can efficiently make inference about both the partially identified parameter and its identified set. This makes it widely applicable in general moment inequality models. Finally, the proposed method is illustrated in a financial asset pricing problem.

Key words: partial identification, posterior consistency, concentration rate, support function, two-sided Bayesian credible sets, identified set, coverage probability, moment inequality models

JEL code: C110, C140, C58

1 Introduction

Partially identified models have been receiving extensive attentions in recent years due to their broad applications in econometrics. Partial identification of a structural parameter arises when the data available and the constraints coming from economic theory only allow to place the parameter inside a proper subset of the parameter space. Due to the limitation of the data generating process, the data cannot provide any information within the set where the structural parameter is partially identified (called *identified set*).

This paper aims at developing a semi-parametric Bayesian inference for partially identified models. A Bayesian approach may be appealing for several reasons. First, Bayesian procedures often conduct inference through Bayesian credible sets (BCS's), which are often relatively easy to construct thanks to the use of Markov Chain Monte Carlo (MCMC) methods. This is particularly useful when we are concerned about marginalizing the BCS to a low-dimensional space. In some situations, we are interested only in a projection of the identified region for the subset inference. We demonstrate that our proposed approach provides tractable computational tools for projecting a high-dimensional identified region to a low-dimensional space. This has important implications for practical implementations.

Secondly, our Bayesian procedures also have comprehensive frequentist validations. In particular, our constructed BCS of the identified set also has a correct asymptotic frequentist coverage probability. We construct credible sets based on the support function; the latter completely characterizes convex and closed identified sets. We also show the Bernstein-von Mises theorem for the posterior distribution of the support function. At the best of our knowledge this has not been studied in the literature yet. This powerful result in turn allows us to establish the (asymptotic) equivalence between BCS's and frequentist confidence sets (FCS's) for the identified set. The literature on partial identification distinguishes between credible/confidence sets for the partially identified parameter and for the identified set. Credible sets for the identified set play an important role not only when the target of interest is the partially identified parameter but even when the identified set is itself the object of interest. While focusing on the study of BCS for the identified set, we also extend Moon and Schorfheide (2010)'s analysis for the partially identified parameter to a semi-parametric setup, which is relevant in more general *moment inequality models* where the likelihood function may be unknown. Moreover, if we admit the existence of a true value of the structural parameter and the identified set, the corresponding posterior distributions concentrate asymptotically in a neighborhood of the true value (set). This property is known as the *posterior consistency*. It is important because it guarantees that, with a sufficiently large amount of data, we can recover the truth accurately with large probabilities.

Third, putting a prior on the partially identified parameter can be viewed as a way of incorporating researchers' beliefs. A Bayesian approach conveniently combines the information from both the observed data and other sources of prior information. The prior information is, for instance, information coming from historical data, information based on experience or on previous survey data. In some applications this information is largely available, *e.g.* in macroeconomics, central banking and finance. We stress that the prior information will

not affect the boundary of the identified set, but will only play a role in determining which areas inside the identified set are a priori "more likely" than others. On the other hand, when specifying a prior distribution on the partially identified parameter is either difficult or conflicting with the philosophy of partial identification, a researcher can still use our procedure and either specify a uniform prior or just construct the BCS for the identified set for inference. The latter is due to an important feature of our procedure that the Bayesian analysis of the identified set does not require to specify a prior on the partially identified parameter. Therefore, we accommodate both situations where a researcher does have prior information as well as situations where she does not.

From the posterior perspective, the Bayesian partial identification produces a posterior distribution of the partially identified parameter whose support will asymptotically concentrate around the true identified set. When informative priors are available, the shape of the posterior density may not be flat inside the identified set, and will ground on the prior distribution even asymptotically. Therefore, the asymptotic behavior of the posterior distribution is different from that of the traditional point identified case where (in the latter case) the information from the prior is often washed away by the data asymptotically. Thus, the Bayesian approach to partial identification links conclusions and inferences to various information sources – data, prior, experience, etc.– in a transparent way.

Finally, when the identified set depends on a point identified nuisance parameter, say ϕ, and this is integrated out with respect to its posterior, then the prior information on the partially identified parameter is completely revised by the data. Hence, the proposed procedure also learns about the partially identified parameter based on the whole posterior distribution of ϕ, which is potentially useful in finite samples. Consequently, there is a strong motivation for us to conduct a comprehensive Bayesian study for the partially identified econometric models.

There are in general two approaches in the literature on Bayesian partial identification. The first approach specifies a parametric likelihood function and assumes it is known up to a finite-dimensional parameter. This approach has been used frequently in the literature, see e.g., Moon and Schorfheide (2012), Poirier (1998), Bollinger and Hasselt (2009), Norets and Tang (2012) among many others. In many applications, however, econometric models usually only identify a set of moment inequalities instead of the full likelihood function. Examples are: interval-censored data, interval instrumental regression, asset pricing (Chernozhukov et al. 2008), incomplete structural models (Menzel 2011), etc. Assuming a parametric form of the likelihood function is ad-hoc in these applications. Once the likelihood is mis-specified, the posterior can be misleading. The second approach starts from a set of moment inequalities, and uses a moment-condition-based likelihood such as the limited information likelihood (Kim 2002) and the exponential tilted empirical likelihood (Schennach 2005). Further references may be found in Liao and Jiang (2010), Chernozhukov and Hong (2003) and Wan (2011). This approach avoids assuming the knowledge of the true likelihood function. However, it only studies the structural parameter, and it is hard to construct posteriors and credible sets for the identified set. Moreover, it does not have a

Bayesian probabilistic interpretation.

This paper proposes a pure Bayesian procedure without assuming a parametric form of the true likelihood function. We place a nonparametric prior on the likelihood and obtain the marginal posterior distribution for the partially identified parameter and the identified set. A similar Bayesian procedure was recently used in Florens and Simoni (2011). As a result, our procedure is semi-parametric Bayesian and can make inference about both the partially identified parameter and its identified set easily. It only requires a set of moment conditions and then it can be completely nonparametric on the data generating process. This is an appealing feature in general moment inequality models. On the other hand, if the likelihood function is known, our procedure continues to work and this paper is still well-motivated. In fact, many contributions of this paper, e.g., Bayesian inference of the support function, construction of BCS for the identified set, subset inferences, etc., are relevant and original also for the case with a known likelihood.

There is a growing literature on Bayesian partially identified models. Besides those mentioned above, the list also includes Gelfand and Sahu (1999), Neath and Samaniego (1997), Gustafson (2012), Epstein and Seo (2011), Stoye (2012), Kitagawa (2012), Kline (2011), etc. There is also an extensive literature that analyzes partially identified models from a frequentist point of view. A partial list includes Andrews and Guggenberger (2009), Andrews and Soares (2010), Andrews and Shi (2013), Beresteanu, Molchanov and Molinari (2011), Bugni (2010), Canay (2010), Chernozhukov, Hong and Tamer (2007), Chiburis (2009), Imbens and Manski (2004), Romano and Shaikh (2010), Rosen (2008), Stoye (2009), among others. See Tamer (2010) for a review.

When the identified set is closed and convex, the support function becomes one of the useful tools to characterize its properties. The literature on this perspective has been growing rapidly, see for example, Bontemps, Magnac and Maurin (2012), Beresteanu and Molinari (2008), Beresteanu et al. (2012), Kaido and Santos (2013), Kaido (2012) and Chandrasekhar et al. (2012). This paper is also closely related to the asymptotic nonparametric Bayesian literature: Wu and Ghosal (2008), Ghosh and Ramamoorthi (2003), Ghosal and van der Vaart (2001), Shen and Wasserman (2001), Ghosal et al. (1999), Amewou-Atisso et al. (2003), Walker et al. (2007), van der Vaart and van Zanten (2008), Bickel and Kleijn (2012), Jiang (2007), Choi and Ramamoorthi (2008), Castillo (2008), Freedman (1999), Rivoirard and Rousseau (2012), among others.

The paper is organized as follows. Section 2 outlines our main results and contributions. Section 3 sets up the model and discusses the prior specification on the underlying likelihood function. Section 4 studies the (marginal) posterior distribution of the structural parameter. Section 5 studies the posterior of the support function in moment inequality models. In particular, the Bernstein-von Mises theorem and a linear representation for the support function are obtained. Section 6 constructs the Bayesian credible sets for both the structural parameter and its identified set. In addition, the frequentist coverages of these credible sets are studied. Section 7 addresses the subset inference when the target of interest is only a component of the full parameter. Section 8 shows the posterior consistency for the identified set and provides the concentration rate. Section 9 addresses the uniformity. In particular,

it discusses the case when point identification is actually achieved. Section 10 applies the support function approach to a financial asset pricing study. Finally, Section 11 concludes with further discussions. All the proofs are given in the appendix to this paper and in a supplementary appendix.

2 Highlights of Our Contributions

This section provides a global vision of our main contributions of this paper. Formal setup of the model starts from Section 3.

Semi-parametric Bayesian partial identification

We focus on semi-parametric models where the true likelihood function may be unknown, which is more relevant in moment inequality models. Then there are three types of parameters in the Bayesian setup: θ, which is the partially identified structural parameter; ϕ, a point-identified parameter that characterizes the identified set, and the unknown likelihood F. The identified set can be written as $\Theta(\phi)$. According to the Bayesian philosophy, we treat the identified set as random, and construct its posterior distribution.

Without assuming any parametric form for the likelihood, we place a nonparametric prior $\pi(F)$ on it. The posteriors of ϕ and of the identified set can then be constructed via the posterior of F. Such a semi-parametric posterior requires only a set of moment inequalities, and therefore is robust to the likelihood specification. Moreover, to make inference about the partially identified θ, we place a conditional prior $\pi(\theta|\phi)$ supported only on $\Theta(\phi)$. Note that Bayesian inference for the identified set may be carried out based on the posterior of $\Theta(\phi)$ which does not depend on $\pi(\theta|\phi)$. So the prior specification for θ plays a role only in the inference about θ.

For these posteriors, we show that asymptotically $p(\theta|Data)$ will be supported within an arbitrarily small neighborhood of the true identified set, and the posterior of $\Theta(\phi)$ also concentrates around the true set in the Hausdorff distance. These are the notion of *posterior consistency* under partial identification.

Support function

To make inference on $\Theta(\phi)$ we can take advantage of the fact that when $\Theta(\phi)$ is closed and convex it is completely characterized by its *support function* $S_\phi(\cdot)$ defined as:

$$S_\phi(\nu) = \sup_{\theta \in \Theta(\phi)} \theta^T \nu$$

where $\nu \in \mathbb{S}^{\dim(\theta)}$, the unit sphere. Therefore, inference on $\Theta(\phi)$ may be conveniently carried out through inference on its support function. The posterior distribution of $S_\phi(\cdot)$ is also determined by that of ϕ. We show that in a general moment inequality model, the support function has an asymptotic linear representation in a neighborhood of the true value of ϕ,

which potentially extends the inference in Bontemps et al. (2012) to nonlinear models. Our paper also establishes the Bernstein-von Mises theorem for the support function, that is, the posterior distribution of $S_\phi(\cdot)$ converges weakly to a Gaussian process. We also calculate the support function for a number of interesting examples, including interval censored data, missing data, interval instrumental regression and asset pricing models.

Two-sided Bayesian credible sets for the identified set

We construct two types of Bayesian credible sets (BCS's): one for the identified set $\Theta(\phi)$ and the other for the partially identified parameter θ. In particular, the BCS for the identified set is constructed based on the support function, is two-sided, and has an asymptotically correct frequentist coverage probability. Specifically, we find sets $\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}}$ and $\Theta(\hat{\phi})^{q_\tau/\sqrt{n}}$, satisfying: for level $1-\tau$ where $\tau \in (0,1)$,

Bayesian coverage:

$$P(\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}} \subset \Theta(\phi) \subset \Theta(\hat{\phi})^{q_\tau/\sqrt{n}} | Data) = 1-\tau; \tag{2.1}$$

Frequentist coverage:

$$P_{D_n}(\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}} \subset \Theta(\phi_0) \subset \Theta(\hat{\phi})^{q_\tau/\sqrt{n}}) \geq 1-\tau, \tag{2.2}$$

where P_{D_n} denotes the sampling probability, and $\Theta(\phi_0)$ is the true identified set. In (2.1) the random set is $\Theta(\phi)$ while in (2.2) the random sets are $\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}}$ and $\Theta(\hat{\phi})^{q_\tau/\sqrt{n}}$. One of the important features is that the BCS for the identified set does not require specifying a prior on the partially identified parameter. The notation $\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}}$, $\Theta(\hat{\phi})^{q_\tau/\sqrt{n}}$, $\hat{\phi}$ and q_τ are to be formally defined in the paper. Therefore, the constructed two-sided BCS can also be used as frequentist confidence set for the identified set.

Furthermore, we find that in the semi-parametric Bayesian model, Moon and Schorfheide (2012)'s conclusion about the BCS for the partially identified parameter θ still holds: it is smaller than frequentist confidence sets in large samples. Hence, while the BCS for the partially identified parameter does not have a correct frequentist coverage, the asymptotic equivalence between BCS and FCS for the identified set holds. Intuitively, this is because the prior information still plays an important role in the posterior of the partially identified parameter even asymptotically; on the other hand, as the identified set is "point identified", whose BCS is independent of the prior on θ, then its prior information is "washed away" asymptotically. Thus, the proposed inference for the identified set and the support function is asymptotically robust to their prior specification.

Projection and subset inference

We show that with our approach it is easy to project (marginalize) onto low-dimensional subspaces for subset inferences. This computation is fast. Suppose the dimension of θ is relatively large, but we are interested in only a few components of θ, and aim to make inference

about these components and their marginal identified set. In our approach, constructing the identified set and BCS for the marginal components simply requires the marginalization of a joint distribution and can be carried out efficiently thanks to the use of MCMC methods. It is also computationally convenient to calculate the BCS for the marginal identified set. Hence, the proposed procedure has large potentiality in many empirical applications.

Uniformity

The proposed Bayesian inference for the identified set is valid uniformly over a class of data generating process. In particular, using specific examples, we illustrate that as the identified set shrinks to a singleton, so that point identification is (nearly) achieved, our Bayesian inference for the identified set carries over.

Applications

We develop a detailed application of Bayesian partial identification to financial asset pricing, which is an example where the identified set is of direct interest. Estimation and inference for the support function as well as for the identified set are conducted. Moreover, throughout the paper, we study in detail other typical examples including the interval censoring, interval regression and missing data problems.

3 General Setup of Bayesian Partially Identified Model

3.1 The Model

Econometric models often involve a structural parameter $\theta \in \Theta$ that is only partially identified by the data generating process (DGP) on a non-singleton set, which we call *identified set*. The model also contains two parameters that are point identified by the DGP: a finite-dimensional parameter $\phi \in \Phi \subset \mathbb{R}^{d_\phi}$ and the distribution function F of the observed data, which is infinite-dimensional. Here, Φ denotes the parameter space for ϕ and d_ϕ its dimension. The point identified parameter ϕ often arises naturally as it characterizes the data distribution. In most of partially identified models, the identified set is also characterized by ϕ, hence we denote it by $\Theta(\phi)$ to indicate that once ϕ is determined, so is the identified set. Let $d = \dim(\theta)$ and $\Theta \subset \mathbb{R}^d$ denote the parameter space for θ; we assume $\Theta(\phi) \subseteq \Theta$.

In a parametric Bayesian partially identified model as in Poirier (1998), Gustafson (2012) and Moon and Schorfheide (2012), F is linked with a known likelihood function to ϕ. However, as in the usual point identified models, in some applications assuming a known likelihood function may suffer from a model specification problem, and may lead to misleading conclusions. Instead, econometric applications often involve only a set of moment conditions as in (3.1) below. This gives rise to the *moment inequality models*. A parametric form of the likelihood function and of F can be unavailable in these models. A robust approach is to

proceed without assuming a parametric form for the likelihood function, but to put a prior on (θ, ϕ, F) instead. This yields the semi-parametric Bayesian setup.

We specify a nonparametric prior on data's cumulated distribution function (CDF) F, which can deduce a prior for ϕ through a transformation $\phi = \phi(F)$, as ϕ often is a functional of F. Moreover, the prior on the identified set $\Theta(\phi)$ is determined through that of ϕ. Due to the identification feature, for any given $\phi \in \Phi$, we specify a conditional prior $\pi(\theta|\phi)$ such that

$$\pi(\theta \in \Theta(\phi)|\phi) = 1.$$

By construction, this prior for θ puts all its mass on $\Theta(\phi)$ for any $\phi \in \Phi$. So it takes the form:

$$\pi(\theta|\phi) \propto I_{\theta \in \Theta(\phi)} g(\theta),$$

where $g(\cdot)$ is some probability density function and $I_{\theta \in \Theta(\phi)}$ is the indicator function of $\Theta(\phi)$. In Section 4.1 we discuss the philosophy of specifying the prior on θ.

Our analysis focuses on the situation where $\Theta(\phi)$ is a closed and convex set for each ϕ. Therefore, $\Theta(\phi)$ can be uniquely characterized by its *support function*. For any fixed ϕ, the support function for $\Theta(\phi)$ is a function $S_\phi(\cdot) : \mathbb{S}^d \to \mathbb{R}$ such that

$$S_\phi(\nu) = \sup_{\theta \in \Theta(\phi)} \theta^T \nu.$$

where $\mathbb{S}^d$ denotes the unit sphere in $\mathbb{R}^d$. The support function plays a central role in convex analysis since it determines all the characteristics of a convex set. Hence, it is one of the essential objects for our Bayesian inference. In a similar way as for $\Theta(\phi)$, we put a prior on $S_\phi(\cdot)$ via the prior on ϕ.

Suppose $p(\phi|D_n)$ denotes the posterior of ϕ, given the data D_n and a prior $\pi(\phi)$. It is readily seen that (see e.g., Poirier 1998) the joint posterior of (θ, ϕ) is given by

$$p(\theta, \phi|D_n) \propto \pi(\theta|\phi) p(\phi|D_n).$$

By integrating out ϕ, we obtain the marginal posterior for θ. On the other hand, the posteriors of $\Theta(\phi)$ and $S_\phi(\cdot)$ are also determined through the marginal posterior $p(\phi|D_n)$. This also highlights an important feature of this paper: our results on $\Theta(\phi)$ and the support function do not require placing a prior on the partially identified parameter θ, because as far as $p(\phi|D_n)$ is concerned, the prior for θ is not needed at all. Furthermore, as the identified set and support function are "point identified", their posteriors are asymptotically robust to the prior specifications on ϕ.

Let us present a few examples that have received much attention in partially identified econometric models literature. In the rest of the paper, we denote by X the observable random variable for which we have n i.i.d. observations $D_n = \{X_i\}_{i=1}^n$. Let $(\mathcal{X}, \mathcal{B}_x, F)$ denote a probability space in which X takes values and $\mathcal{F}$ denote the parameter space of F.

Example 3.1 (Interval censored data)**.** Let (Y, Y_1, Y_2) be a 3-dimensional random vector such that $Y \in [Y_1, Y_2]$ with probability one. The random variables Y_1 and Y_2 are observed

while Y is unobservable (see, e.g., Moon and Schorfheide 2012). We denote: $\theta = E(Y)$ and $\phi = (\phi_1, \phi_2)' \equiv (E(Y_1), E(Y_2))'$. Therefore, we have the following identified set for θ: $\Theta(\phi) = [\phi_1, \phi_2]$. The support function for $\Theta(\phi)$ is easy to derive:

$$S_\phi(1) = \phi_2, \quad S_\phi(-1) = -\phi_1.$$

The non-parametric prior specification on the likelihood is to be discussed in Section 3.2. □

Example 3.2 (Interval regression model)**.** The regression model with interval censoring has been studied by, for example, Haile and Tamer (2003). Let (Y, Y_1, Y_2) be a 3-dimensional random vector such that $Y \in [Y_1, Y_2]$ with probability one. The random variables Y_1 and Y_2 are observed while Y is unobservable. Assume that

$$Y = x^T\theta + \epsilon$$

where x is a vector of observable regressors. In addition, assume there is a d-dimensional vector of nonnegative exogenous variables Z such that $E(Z\epsilon) = 0$. Here Z can be either a vector of instrumental variables when X is endogenous, or a nonnegative transformation of x when x is exogenous. It follows that

$$E(ZY_1) \leq E(ZY) = E(Zx^T)\theta \leq E(ZY_2). \tag{3.1}$$

We denote $\phi = (\phi_1, \phi_2, \phi_3)$ where $(\phi_1^T, \phi_3^T) = (E(ZY_1)^T, E(ZY_2)^T)$ and $\phi_2 = E(Zx^T)$. Then the identified set for θ is given by $\Theta(\phi) = \{\theta \in \Theta : \phi_1 \leq \phi_2\theta \leq \phi_3\}$. Suppose ϕ_2^{-1} exists. The support function for $\Theta(\phi)$ is given by (denote $(x)_i$ as the ith component of x)[1]:

$$S_\phi(\nu) = \nu^T\phi_2^{-1}\left(\frac{\phi_1 + \phi_3}{2}\right) + \alpha_\nu^T\left(\frac{\phi_3 - \phi_1}{2}\right), \quad \nu \in \mathbb{S}^d$$

where $\alpha_\nu = (|(\nu^T\phi_2^{-1})_1|, ..., |(\nu^T\phi_2^{-1})_d|)^T$.

□

Example 3.3 (Missing data)**.** Consider a bivariate random vector (Y, M) where M is a binary random variable which takes the value $M = 0$ when Y is missing and 1 otherwise. Here Y represents whether a treatment is successful ($Y = 1$) or not ($Y = 0$). The parameter of interest is the probability $\theta = P(Y = 1)$. This problem without the missing-at-random assumption has been extensively studied in the literature, see for example, Manski and Tamer (2002), Manski (2003), etc. Since $P(Y = 1|M = 0)$ cannot be recovered from the data, the empirical evidence partially identifies θ and θ is characterized by the following moment restrictions:

$$P(Y = 1|M = 1)P(M = 1) \leq \theta \leq P(Y = 1|M = 1)P(M = 1) + P(M = 0).$$

[1]See Appendix C.1 in the supplementary material for detailed derivations of the support function in this example. Similar results but in a slightly different form are presented in Bontemps et al. (2012).

Here, $\phi = (P(M = 1), P(Y = 1|M = 1)) = (\phi_1, \phi_2)$. The identified set is $\Theta(\phi) = [\phi_1\phi_2, \phi_1\phi_2 + 1 - \phi_1]$, and its support function is: $S_\phi(1) = \phi_1\phi_2 + 1 - \phi_1$, $S_\phi(-1) = -\phi_1\phi_2$.

3.2 Nonparametric prior scheme for (ϕ, F)

When the model only specifies a set of moment inequalities, we can place a non-parametric prior on the likelihood function through F, e.g., a Dirichlet process prior. Since ϕ is point identified, we assume it can be rewritten as a measurable function of F as $\phi = \phi(F)$. The prior distribution for ϕ is then deduced from that of F via $\phi(F)$. The Bayesian experiment is (we use the notation "$\sim$" to mean "distributed as")

$$X|F \sim F, \qquad F \sim \pi(F), \qquad \theta|\phi \sim \pi(\theta|\phi(F))$$

For instance, in the interval censored data example 3.1, let F be the joint CDF of (Y_1, Y_2), then $(\phi_1, \phi_2) = \phi(F) = (E(Y_1|F), E(Y_2|F))$, and the identified set is modeled as $\Theta(\phi) = [\phi_1(F), \phi_2(F)]$, which is a set-valued function of F.

The prior distribution $\pi(F)$ is a distribution on $\mathcal{F}$. Examples of such a prior include Dirichlet process priors (Ferguson 1973) and Polya tree (Lavine 1992). The case where $\pi(F)$ is a Dirichlet process prior in partially identified models is proposed by Florens and Simoni (2011).

Let $p(F|D_n)$ denote the marginal posterior of F which, by abuse of notation, can be written $p(F|D_n) \propto \pi(F) \prod_{i=1}^n F(X_i)$. The posterior distributions of ϕ, $\Theta(\phi)$, and the support function $S_\phi(\cdot)$ are deduced from the posterior of F, but do not depend on the prior on θ. Moreover, it can be shown that $p(\theta|\phi(F), D_n) = \pi(\theta|\phi(F))$. Then, for any measurable set $B \subset \Theta$, the marginal posterior probability of θ is given by, averaging over F:

$$\begin{aligned} P(\theta \in B|D_n) &= \int_{\mathcal{F}} P(\theta \in B|\phi(F), D_n) p(F|D_n) dF \\ &= \int_{\mathcal{F}} \pi(\theta \in B|\phi(F)) p(F|D_n) dF = E[\pi(\theta \in B|\phi(F))|D_n] \end{aligned}$$

where the conditional expectation is taken with respect to the posterior of F. The above posterior is easy to calculate via simulation when F has a Dirichlet process prior.

An alternative prior scheme for (ϕ, F) consists in putting a prior on ϕ directly. This is particularly useful when there is informative prior information for ϕ. It models the unknown likelihood function semi-parametrically through reformulating F as $F = F_{\phi,\eta}$ where η is an infinite-dimensional nuisance parameter (often a density function) that is *a priori* independent of ϕ. The prior on (ϕ, F) is then deduced from the prior on (ϕ, η). We describe this alternative semi-parametric prior in Appendix A.

4 Bayesian Inference for θ

4.1 Putting priors on partially identified θ

In this section we briefly discuss the meaning of the prior $\pi(\theta|\phi)$. As stated in Tamer (2010): "(Partial identification) links conclusions drawn from various empirical models to sets of assumptions made in a transparent way. It allows researchers to examine the informational content of their assumptions and their impacts on the inferences made."

By imposing a prior on the partially identified parameter θ, we reflect how prior beliefs and/or assumptions can impact the associated statistical inference. To illustrate the rationale of imposing such a prior, let us consider the missing data example (Example 3.3). Writing $\alpha = P(Y = 1|M = 0)$, we then link θ with α by

$$\theta = \phi_2\phi_1 + \alpha(1 - \phi_1). \tag{4.1}$$

As ϕ is point identified, statistical inferences about θ therefore relies on the treatment of α. On the other hand, various ways of dealing with α reflect various researchers' prior beliefs, which also correspond to the "informational content of their assumptions".

From a Bayesian point of view, this is fulfilled by putting a distribution on α, as a prior $\pi(\alpha)$ supported on $[0, 1]$ (possibly also depending on ϕ). The traditional exogeneity assumption such as missing-at-random, in this case, corresponds to a point mass prior on $\alpha = \phi_2 = P(Y = 1|M = 1)$. The more concentrating is the prior, the stronger are the assumptions we impose on the missing mechanism. Such a prior distribution can also come from a previous study based on a different dataset that contain information about α where only summarizing statistics are available instead of the complete data set. Above all, when no informative knowledge about α is available, a uniform prior on $[0, 1]$ is imposed for α, which reduces to Manski's bounds approach.

Given the imposed distribution $\pi(\alpha)$ that reflects researchers' assumptions or beliefs about the missing mechanism, we can deduce a conditional prior for θ through (4.1) given $\phi = (\phi_1, \phi_2)$. As a result, putting a prior on the partially identified parameter can be viewed as a way of incorporating researchers' assumptions on the missing mechanisms. This varies from the traditional exogeneity approach to the most robust bounds approach, which also bridges point identification and partial identification.

4.2 Posterior Consistency for θ

The shape of the posterior of a partially identified parameter still relies upon its prior distribution asymptotically, which distinguishes from the asymptotic posterior behavior in the classical point identified case. On the other hand, the support of the prior distribution of θ is revised after data are observed and eventually converges towards the true identified set asymptotically. The latter corresponds to the frequentist consistency of the posterior distribution for partially identified parameters. Posterior consistency is one of the benchmarks of a Bayesian procedure under consideration, which ensures that with a sufficiently

large amount of data, it is nearly possible to discover the true identified set.

We assume there is a true value of ϕ, denoted by ϕ_0, which induces a true identified set $\Theta(\phi_0)$, and a true F, denoted by F_0. Our goal is to achieve the frequentist *posterior consistency* for the partially identified parameter: for any $\epsilon > 0$ there is $\tau \in (0, 1]$ such that

$$P(\theta \in \Theta(\phi_0)^\epsilon | D_n) \to^p 1, \text{ and } \quad P(\theta \in \Theta(\phi_0)^{-\epsilon} | D_n) \to^p (1 - \tau).$$

Here $\Theta(\phi)^\epsilon$ and $\Theta(\phi)^{-\epsilon}$ are the ϵ-envelope and ϵ-contraction of $\Theta(\phi)$, respectively:

$$\Theta(\phi)^\epsilon = \{\theta \in \Theta : d(\theta, \Theta(\phi)) \leq \epsilon\}, \quad \Theta(\phi)^{-\epsilon} = \{\theta \in \Theta(\phi) : d(\theta, \Theta \backslash \Theta(\phi)) \geq \epsilon\}, \tag{4.2}$$

with $\Theta \backslash \Theta(\phi) = \{\theta \in \Theta; \theta \notin \Theta(\phi)\}$ and $d(\theta, \Theta(\phi)) = \inf_{x \in \Theta(\phi)} \|\theta - x\|$. Note that this result still carries over when θ is point identified, in which case $\Theta(\phi)^\epsilon$ is an ϵ-ball around θ, $\Theta(\phi)^{-\epsilon}$ is empty, and $\tau = 1$.

The likelihood function is endowed with a prior through either the nonparametric prior $\pi(F)$ as described in Section 3.2 or the semi-parametric prior $\pi(\phi)$ as described in Appendix A. We assume that the priors $\pi(F)$ and $\pi(\phi)$ specified for F and ϕ are such that the corresponding posterior distribution of $p(\phi|D_n)$ is consistent.

Assumption 4.1. *At least one of the following holds:*

(i) The measurable function $\phi(F) : \mathcal{F} \to \Phi$ is continuous. The prior $\pi(F)$ is such that the posterior $p(F|D_n)$ satisfies:

$$\int_{\mathcal{F}} m(F) p(F|D_n) dF \to^p \int_{\mathcal{F}} m(F) \delta_{F_0}(dF)$$

for any bounded and continuous function $m(\cdot)$ on $\mathcal{F}$ where δ_{F_0} is the Dirac function at the true distribution function F_0;

(ii) The prior $\pi(\phi)$ is such that the posterior $p(\phi|D_n)$ satisfies:

$$\int_{\Phi} m(\phi) p(\phi|D_n) d\phi \to^p \int_{\Phi} m(\phi) \delta_{\phi_0}(d\phi)$$

for any bounded and continuous function $m(\cdot)$ on Φ where δ_{ϕ_0} is the Dirac function at the true ϕ_0.

Assumptions 4.1 *(i)* and *(ii)* refer to the nonparametric and semi-parametric prior scheme respectively, and are verified by many nonparametric and semi-parametric priors. Examples are: Dirichlet process priors, Polya Tree process priors, Gaussian process priors, etc. For instance, when $\pi(F)$ is the Dirichlet process prior, the second part of Assumption 4.1 *(i)* was proved in Ghosh and Ramamoorthi (2003, Theorem 3.2.7) while the condition that $\phi(F)$ is continuous in F is verified in many examples relevant for applications. For instance, in Example 3.1, $\phi(F) = (E(Y_1|F), E(Y_2|F))^T$ and in Example 3.2,

$\phi(F) = (E(ZY_1|F), E(ZX^T|F), E(ZY_2|F))$, which are all bounded linear functionals of F. We refer to Ghosh and Ramamoorthi (2003) for examples and sufficient conditions for this assumption.

Assumption 4.2 (Prior for ϕ)**.** *For any $\epsilon > 0$ there are measurable sets $A_2 \subset A_1 \subset \Phi$ such that $0 < \pi(\phi \in A_i) \leq 1$, $i = 1, 2$ and*
(i) for all $\phi \in A_1$, $\Theta(\phi_0)^\epsilon \cap \Theta(\phi) \neq \emptyset$; for all $\phi \notin A_1$, $\Theta(\phi_0)^\epsilon \cap \Theta(\phi) = \emptyset$,
(ii) for all $\phi \in A_2$, $\Theta(\phi_0)^{-\epsilon} \cap \Theta(\phi) \neq \emptyset$; for all $\phi \notin A_2$, $\Theta(\phi_0)^{-\epsilon} \cap \Theta(\phi) = \emptyset$.

Assumption 4.2 is satisfied as long as the identified set $\Theta(\phi)$ is bounded and the prior of ϕ spreads over a large support of the parameter space. This assumption allows us to prove the posterior consistency without assuming the prior $\pi(\theta|\phi)$ to be a continuous function of ϕ, and therefore priors like $I_{\phi_1<\theta<\phi_2}$ in the interval censoring data example are allowed. Under this assumption the conditional prior probability of the ϵ-envelope of the true identified set can be approximated by a continuous function, that is, there is a sequence of bounded and continuous functions $h_m(\phi)$ such that (see lemma D.1 in the appendix) almost surely in ϕ:

$$\pi(\theta \in \Theta(\phi_0)^\epsilon|\phi) = \lim_{m\to\infty} h_m(\phi).$$

A similar approximation holds for the conditional prior of the ϵ-contraction $\pi(\theta \in \Theta(\phi_0)^{-\epsilon}|\phi)$.

Assumption 4.3 (Prior for θ)**.** *For any $\epsilon > 0$, and $\phi \in \Phi$, $\pi(\theta \in \Theta(\phi)^{-\epsilon}|\phi) < 1$.*

In the special case when θ is point identified ($\Theta(\phi)$ is a singleton), the ϵ-contraction is empty and thus $\pi(\theta \in \Theta(\phi)^{-\epsilon}|\phi) = 0$.

Assumption 4.3 is an assumption on the prior for θ, which means the identified set should be *sharp* with respect to the prior information. Roughly speaking, the support of the prior should not be a proper subset of any ϵ-contraction of the identified set $\Theta(\phi)$. If otherwise the prior information restricts θ to be inside a strict subset of $\Theta(\phi)$ so that Assumption 4.3 is violated, then that prior information should be taken into account in order to shrink $\Theta(\phi)$ to a sharper set. In that case, the posterior will asymptotically concentrate around a set that is smaller than the set identified by the data alone. Remark that assumption 4.3 is not needed for the first part of Theorem 4.1 below.

The following theorem gives the posterior consistency for partially identified parameters.

Theorem 4.1. *Under Assumptions 4.1 and 4.2, for any $\epsilon > 0$,*

$$P(\theta \in \Theta(\phi_0)^\epsilon|D_n) \to^p 1.$$

If Assumption 4.3 is further satisfied, then there is $\tau \in (0, 1]$ such that

$$P(\theta \in \Theta(\phi_0)^{-\epsilon}|D_n) \to^p (1 - \tau).$$

5 Bayesian Inference of Support Function

Our analysis focuses on identified sets which are closed and convex. These sets are completely determined by their support functions, and efficient estimation of support functions may lead to optimality of estimation and inference of the identified set. As a result, much of the new development in the partially identified literature focuses on the support function, e.g., Kaido and Santos (2013), Kaido (2012), Beresteanu and Molinari (2008), Bontemps et al. (2012).

This section develops Bayesian analysis for the support function $S_\phi(\nu)$ of the identified set $\Theta(\phi)$. We consider a more specific partially identified model: the *moment inequality model* which is described in section 5.1 below. Bayesian inference for the support function has two main interests. First, it provides an alternative way to characterize and perform estimation of the identified set $\Theta(\phi)$, which in many cases is relatively easy for computations and simulations. Second, it allows us to construct a two-sided BCS for $\Theta(\phi)$ that is also asymptotically equivalent to a frequentist confidence set. In this section we first develop a local linearization in ϕ of the support function. As the support function itself is "point identified", we prove that its posterior satisfies the Bernstein-von Mises theorem. This result is *per se* of particular interest in the nonparametric Bayesian literature.

5.1 Moment Inequality Model

The *moment inequality model* assumes that θ satisfies k moment restrictions:

$$\Psi(\theta,\phi) \leq 0, \quad \Psi(\theta,\phi) = (\Psi_1(\theta,\phi), ..., \Psi_k(\theta,\phi))^T \tag{5.1}$$

where $\Psi : \Theta \times \Phi \to \mathbb{R}^k$ is a known function of (θ,ϕ). The identified set can be characterized as:

$$\Theta(\phi) = \{\theta \in \Theta : \Psi(\theta,\phi) \leq 0\}. \tag{5.2}$$

Since most of the partially identified models can be characterized as moment inequality models, model (5.1)-(5.2) has received extensive attention in the literature.

We assume each component of $\Psi(\theta,\phi)$ to be a convex function of θ for every $\phi \in \Phi$, as stated in the next assumption.

Assumption 5.1. *$\Psi(\theta,\phi)$ is continuous in (θ,ϕ) and convex in θ for every $\phi \in \Phi$.*

Let us consider the support function $S_\phi(\cdot) : \mathbb{S}^d \to \mathbb{R}$ of the identified set $\Theta(\phi)$. We restrict its domain to the unit sphere $\mathbb{S}^d$ in $\mathbb{R}^d$ since $S_\phi(\nu)$ is positively homogeneous in ν. Under Assumption 5.1 the support function is the optimal value of an ordinary convex program:

$$S_\phi(\nu) = \sup_{\theta\in\Theta}\{\nu^T\theta;\ \Psi(\theta,\phi) \leq 0\}.$$

Therefore, it also admits a Lagrangian representation (see Rockafellar 1970, chapter 28):

$$S_\phi(\nu) = \sup_{\theta\in\Theta}\{\nu^T\theta - \lambda(\nu,\phi)^T\Psi(\theta,\phi)\}, \tag{5.3}$$

where $\lambda(\nu,\phi) : \mathbb{S}^d \times \mathbb{R}^{d_\phi} \to \mathbb{R}^k_+$ is a k-vector of Lagrange multipliers.

We denote by $\Psi_S(\theta,\phi_0)$ the k_S-subvector of $\Psi(\theta,\phi_0)$ containing the constraints that are strictly convex functions of θ and by $\Psi_L(\theta,\phi_0)$ the k_L constraints that are linear in θ. So $k_S + k_L = k$. The corresponding Lagrange multipliers are denoted by $\lambda_S(\nu,\phi_0)$ and $\lambda_L(\nu,\phi_0)$, respectively, for $\nu \in \mathbb{S}^d$. Moreover, define $\Xi(\nu,\phi) = \arg\max_{\theta\in\Theta}\{\nu^T\theta;\ \Psi(\theta,\phi) \le 0\}$ as the *support set* of $\Theta(\phi)$. Then, by definition,

$$\nu^T\theta = S_\phi(\nu), \quad \forall\theta \in \Xi(\nu,\phi).$$

We also denote by $\nabla_\phi\Psi(\theta,\phi)$ the $k \times d_\phi$ matrix of partial derivatives of Ψ with respect to ϕ, and by $\nabla_\theta\Psi_i(\theta,\phi)$ the d-vector of partial derivatives of Ψ_i with respect to θ for each $i \le k$. In addition, let

$$Act(\theta,\phi) \equiv \{i \le k;\ \Psi_i(\theta,\phi) = 0\}$$

be the set of the inequality active constraint indices. For some $\delta > 0$, let $B(\phi_0,\delta) = \{\phi \in \Phi;\ \|\phi - \phi_0\| \le \delta\}$.

We assume the following:

Assumption 5.2. *The true value ϕ_0 is in the interior of Φ, and Θ is convex and compact.*

Assumption 5.3. *There is some $\delta > 0$ such that for all $\phi \in B(\phi_0,\delta)$, we have:*
(i) the matrix $\nabla_\phi\Psi(\theta,\phi)$ exists and is continuous in (θ,ϕ);
(ii) the set $\Theta(\phi)$ is non empty;
(iii) there exists a $\theta \in \Theta$ such that $\Psi(\theta,\phi) < 0$;
(iv) $\Theta(\phi)$ belongs to the interior of Θ;
(v) for every $i \in Act(\theta,\phi_0)$, with $\theta \in \Theta(\phi_0)$, the vector $\nabla_\theta\Psi_i(\theta,\phi)$ exists and is continuous in (θ,ϕ) for every $\phi \in B(\phi_0,\delta)$ and $\theta \in \Theta(\phi)$.

Assumption 5.3 *(iii)* is the Slater's condition which is a sufficient condition for strong duality to hold. It implies Assumption 5.3 *(ii)*. However, we keep both conditions because in order to establish some technical results we only need condition *(ii)* which is weaker.

The next assumption concerns the inequality active constraints. Assumption 5.4 requires that the active inequality constraints gradients $\nabla_\theta\Psi_i(\theta,\phi_0)$ be linearly independent. This assumption guarantees that a θ which solves the optimization problem (5.3) with $\phi = \phi_0$ satisfies the Kuhn-Tucker conditions. Alternative assumptions that are weaker than Assumption 5.4 could be used, but the advantage of Assumption 5.4 is that it is easy to check.

Assumption 5.4. *For any $\theta \in \Theta(\phi_0)$, the gradient vectors $\{\nabla_\theta\Psi_i(\theta,\phi_0)\}_{i\in Act(\theta,\phi_0)}$ are linearly independent*

The following assumption is key for our analysis, and is sufficient for the differentiability of the support function at ϕ_0:

Assumption 5.5. *At least one of the following holds:*

(i) For the ball $B(\phi_0, \delta)$ in Assumption 5.3, for every $(\nu, \phi) \in \mathbb{S}^d \times B(\phi_0, \delta)$, $\Xi(\nu, \phi)$ is a singleton;

(ii) There are linear constraints in $\Psi(\theta, \phi_0)$, which are also separable in θ, that is, $k_L > 0$ and $\Psi_L(\theta, \phi_0) = A_1\theta + A_2(\phi_0)$ for some function $A_2 : \Phi \to \mathbb{R}^{k_L}$ (not necessarily linear) and some $(k_L \times d)$-matrix A_1.

Assumption 5.5 is particularly important for the linearization of the support function that we develop in Section 5.2. In fact, if one of the two parts of Assumption 5.5 holds then the support function is differentiable at ϕ for every $(\nu, \phi) \in \mathbb{S}^d \times B(\phi_0, \delta)$, and we have a closed form for its derivative. This assumption also plays one of the key roles in the study of asymptotic efficiency by Kaido and Santos (2013).

The last set of assumptions will be used to prove the Bernstein-von Mises theorem for $S_\phi(\cdot)$ and allows to strengthen the result of Theorem 5.1 below. The first three assumptions are (local) Lipschitz equi-continuity assumptions.

Assumption 5.6. *For the ball $B(\phi_0, \delta)$ in Assumption 5.3, for some $K > 0$ and $\forall \phi_1, \phi_2 \in B(\phi_0, \delta)$:*

(i) $\sup_{\nu \in \mathbb{S}^d} \|\lambda(\nu, \phi_1) - \lambda(\nu, \phi_2)\| \leq K\|\phi_1 - \phi_2\|$;

(ii) $\sup_{\theta \in \Theta} \|\nabla_\phi \Psi(\theta, \phi_1) - \nabla_\phi \Psi(\theta, \phi_2)\| \leq K\|\phi_1 - \phi_2\|$;

(iii) $\|\nabla_\phi \Psi(\theta_1, \phi_0) - \nabla_\phi \Psi(\theta_2, \phi_0)\| \leq K\|\theta_1 - \theta_2\|$, *for every* $\theta_1, \theta_2 \in \Theta$;

(iv) If $\Xi(\nu, \phi_0)$ is a singleton for any ν in some compact subset $W \subseteq \mathbb{S}^d$, and if the correspondence $(\nu, \phi) \mapsto \Xi(\nu, \phi)$ is upper hemicontinous on $\mathbb{S}^d \times B(\phi_0, \delta)$ then there exists $\varepsilon = O(\delta)$ such that $\Xi(\nu, \phi_1) \subseteq \Xi^\varepsilon(\nu, \phi_0)$.

Here $\|\nabla_\phi \Psi(\theta, \phi)\|$ denotes the Frobenius norm of the matrix. The above conditions are not stringent. In particular, condition *(iv)* is easy to understand when $\Xi(p, \phi)$ is a singleton, that is, when the optimization problem for the support function has a unique solution, for each $\phi \in B(\phi_0, \delta)$. Then $\Xi(\nu, \phi_1)$ and $\Xi(\nu, \phi_0)$ are singletons that are close to each other, and $\Xi(\nu, \phi_0)^\varepsilon$ is a small ball around $\Xi(\nu, \phi_0)$.

We show in the following example that Assumptions 5.1-5.6 are easily satisfied.

Example 5.1 (Interval censored data - *continued*)**.** The setup is the same as in Example 3.1. Assumption 5.2 is verified if Y_1 and Y_2 are two random variables with finite first moments $\phi_{0,1}$ and $\phi_{0,2}$, respectively. Moreover, $\Psi(\theta, \phi) = (\phi_1 - \theta, \theta - \phi_2)^T$, $\phi = (\phi_1, \phi_2)^T$,

$$\nabla_\phi \Psi(\theta, \phi) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

so that Assumptions 5.1, 5.2 and 5.3 *(i)-(ii)* are trivially satisfied. Assumption 5.3 *(iii)* holds for every θ inside (ϕ_1, ϕ_2); Assumption 5.3 *(iv)* is satisfied if ϕ_1 and ϕ_2 are bounded. To see

that Assumptions 5.3 *(v)* and 5.4 are satisfied note that $\forall\theta < \phi_{0,1}$ we have $Act(\theta, \phi_0) = \{1\}$, $\forall\theta > \phi_{0,2}$ we have $Act(\theta, \phi_0) = \{2\}$ while $\forall\theta \in [\phi_{0,1}, \phi_{0,2}]$ we have $Act(\theta, \phi_0) = \emptyset$. Assumption 5.5 *(i)* and *(ii)* are both satisfied since the support set takes the values $\Xi(1, \phi) = \phi_2$ and $\Xi(-1, \phi) = -\phi_1$ and the constraints in $\Psi(\theta, \phi_0)$ are both linear with $A_1 = (-1, 1)^T$ and $A_2(\phi_0) = \nabla_\phi \Psi(\theta, \phi_0)\phi_0$.

Assumptions 5.6 *(ii)-(iii)* are naturally satisfied because $\nabla_\phi \Phi(\theta, \phi)$ does not depend on (θ, ϕ). The Lagrange multiplier is $\lambda(\nu, \phi) = (-\nu I(\nu < 0), \nu I(\nu \geq 0))^T$ so that Assumption 5.6 *(i)* is satisfied since the norm is equal to 0. Finally, the support set $\Xi(\nu, \phi) = \phi_1 I(\nu < 0) + \phi_2 I(\nu \geq 0)$ is a singleton for every $\phi \in B(\phi_0, \delta)$ and $\Xi(\nu, \phi_0)^\varepsilon = \{\theta \in \Theta; \|\theta - \theta_*\| \leq \varepsilon\}$ where $\theta_* = \Xi(\nu, \phi_0) = \phi_{0,1} I(\nu < 0) + \phi_{0,2} I(\nu \geq 0)$. Therefore, $\|\Xi(\nu, \phi) - \theta_*\| \leq \delta$ and Assumption 5.6 *(iv)* holds with $\varepsilon = \delta$. □

5.2 Asymptotic Analysis

The support function of the closed and convex set $\Theta(\phi)$ admits directional derivatives in ϕ, see e.g. Milgrom and Segal (2002). Moreover, if Assumption 5.5 holds for a particular value (ν, ϕ), then $S_\phi(\nu)$ is differentiable at ϕ and its derivative is equal to the left and right directional derivatives. The next theorem exploits this fact and states that the support function can be locally approximated by a linear function of ϕ.

Theorem 5.1. *If Assumptions 5.1-5.5 hold with $\delta = r_n$ for some $r_n = o(1)$, then there is $N \in \mathbb{N}$ such that for every $n \geq N$, there exist: (i) a real function $f(\phi_1, \phi_2)$ defined for every $\phi_1, \phi_2 \in B(\phi_0, r_n)$, (ii) a Lagrange multiplier function $\lambda(\nu, \phi_0) : \mathbb{S}^d \times \mathbb{R}^{d_\phi} \to \mathbb{R}^k_+$, and (iii) a Borel measurable mapping $\theta_*(\nu) : \mathbb{S}^d \to \Theta$ satisfying $\theta_*(\nu) \in \Xi(\nu, \phi_0)$ for all $\nu \in \mathbb{S}^d$, such that for every $\phi_1, \phi_2 \in B(\phi_0, r_n)$:*

$$\sup_{\nu \in \mathbb{S}^d} \left| (S_{\phi_1}(\nu) - S_{\phi_2}(\nu)) - \lambda(\nu, \phi_0)^T \nabla_\phi \Psi(\theta_*(\nu), \phi_0)[\phi_1 - \phi_2] \right| = f(\phi_1, \phi_2)$$

and $\frac{f(\phi_1, \phi_2)}{\|\phi_1 - \phi_2\|} \to 0$ uniformly in $\phi_1, \phi_2 \in B(\phi_0, r_n)$ as $n \to \infty$.

We remark that the functions λ and θ_* do not depend on the specific choice of ϕ_1 and ϕ_2 inside $B(\phi_0, r_n)$, but only on ν and the true value ϕ_0. The expansion can also be viewed as stochastic when ϕ_1, ϕ_2 are interpreted as random variables associated with the posterior distribution $P(\phi|D_n)$. This interpretation is particularly useful to understand Theorems 5.2 and 5.3.

With the approximation given in the theorem we are now ready to state posterior consistency (with concentration rate) and asymptotic normality of the posterior distribution of $S_\phi(\nu)$. The posterior consistency of the support function is also based upon the posterior concentration rate for ϕ. In a semi-parametric Bayesian model where ϕ is point identified, the posterior of ϕ achieves a near-parametric concentration rate under proper prior conditions. Since our goal is to study the posterior of $S_\phi(\nu)$, we state a high-level assumption on the posterior of ϕ as follows, instead of deriving it from more general conditions.

Assumption 5.7. *The marginal posterior of ϕ is such that, for some $C > 0$,*

$$P(\|\phi - \phi_0\| \leq Cn^{-1/2}(\log n)^{1/2}|D_n) \to^p 1.$$

This assumption is a standard result in semi/non-parametric Bayesian literature. If we place a nonparametric prior on F as described in Section 3.2, the notation used in this assumption is a shorthand for

$$P(\|\phi(F) - \phi(F_0)\| \leq Cn^{-1/2}(\log n)^{1/2}|D_n) \to^p 1.$$

When the likelihood function is unknown, a formal derivation of this assumption for a semi-parametric prior of (ϕ, F) will be presented in Appendix B.

The next theorem gives the contraction rate for the posterior of the support function.

Theorem 5.2. *Under Assumption 5.7 and the Assumptions of Theorem 5.1 with $r_n = \sqrt{(\log n)/n}$, for some $C > 0$,*

$$P\left(\sup_{\nu \in \mathbb{S}^d} |S_\phi(\nu) - S_{\phi_0}(\nu)| < C(\log n)^{1/2} n^{-1/2} \middle| D_n\right) \to^p 1. \tag{5.4}$$

Remark 5.1. The above result holds for both nonparametric and semi-parametric prior on (ϕ, F). The concentration rate, as given in the theorem, is nearly parametric: $\sqrt{\frac{\log n}{n}}$ and is the same as the rate in assumption 5.7. Thus, when the posterior for ϕ contracts at the rate $n^{-1/2}$, the same holds for the posterior of the support function. The posterior probability in the theorem is now converging to zero, instead of being smaller than an arbitrarily small constant. This often gives rise to the term $\sqrt{\log n}$, which arises commonly in the posterior concentration rate literature (e.g., Ghosal et al. 2000, Shen and Wasserman 2001). The same rate of convergence in the frequentist perspective has been achieved by Chernozhukov et al. (2007), Beresteanu and Molinari (2008), Kaido and Santos (2013), among others, when estimating the identified set.

We now state a Bernstein-von Mises (BvM) theorem for the support function. This theorem is valid under the assumption that a BvM theorem holds for the posterior distribution of the identified parameter ϕ. We denote by $\|\cdot\|_{TV}$ the total variation distance, that is, for two probability measures P and Q,

$$\|P - Q\|_{TV} \equiv \sup_B |P(B) - Q(B)|$$

where B is an element of the σ-algebra on which P and Q are defined.

Assumption 5.8. *Let $P_{\sqrt{n}(\phi-\phi_0)|D_n}$ denote the posterior distribution of $\sqrt{n}(\phi - \phi_0)$. We assume*

$$\|P_{\sqrt{n}(\phi-\phi_0)|D_n} - \mathcal{N}(\Delta_{n,\phi_0}, I_0^{-1})\|_{TV} \to^p 0$$

where $\mathcal{N}$ denotes the d_ϕ-dimensional normal distribution, $\Delta_{n,\phi_0} \equiv n^{-1/2}\sum_{i=1}^n I_0^{-1} l_{\phi_0}(X_i)$, l_{ϕ_0} is the semi-parametric efficient score function of the model and I_0 denotes the semi-parametric efficient information matrix.

As we focus on the partial identification features, we state the above assumption as a high-level condition instead of proving it. We refer to Bickel and Kleijn (2012) and Rivoirard and Rousseau (2012) for primitive conditions of this assumption in semi-parametric models. Remark that l_{ϕ_0} and I_0 also depend on the true likelihood function. The *semi-parametric efficient score function* and the *semi-parametric efficient information* contribute to the stochastic local asymptotic normality (LAN, Le Cam 1986) expansion of the integrated likelihood, which is necessary in order to get the BvM result in Assumption 5.8. A precise definition of l_{ϕ_0} and I_0 may be found in van der Vaart (2002) (Definition 2.15). In particular, they become the usual score function (first derivative of the likelihood) and Fisher's information matrix when the true likelihood is fully parametric.

Below we denote $P_{\sqrt{n}(S_\phi(\nu)-S_{\phi_0}(\nu))|D_n}$ as the posterior distribution of $\sqrt{n}(S_\phi(\nu) - S_{\phi_0}(\nu))$.

Theorem 5.3. *Let Assumption 5.8 hold. If the assumptions of Theorem 5.2 and Assumption 5.6 hold with $\delta = r_n = \sqrt{(\log n)/n}$, then for any $\nu \in \mathbb{S}^d$,*

$$\|P_{\sqrt{n}(S_\phi(\nu)-S_{\phi_0}(\nu))|D_n} - \mathcal{N}(\bar{\Delta}_{n,\phi_0}(\nu), \bar{I}_0^{-1}(\nu))\|_{TV} \to^p 0,$$

where $\bar{\Delta}_{n,\phi_0}(\nu) = \lambda(\nu,\phi_0)^T \nabla_\phi \Psi(\theta_*(\nu),\phi_0)\Delta_{n,\phi_0}$ and

$$\bar{I}_0^{-1}(\nu) = \lambda(\nu,\phi_0)^T \nabla_\phi \Psi(\theta_*(\nu),\phi_0) I_0^{-1} \nabla_\phi \Psi(\theta_*(\nu),\phi_0)^T \lambda(\nu,\phi_0).$$

The asymptotic mean and covariance matrix can be easily estimated by replacing ϕ_0 by any consistent estimator $\hat{\phi}$. Thus, $\theta_*(\nu)$ will be replaced by an element $\hat{\theta}_*(\nu) \in \Xi(\nu,\hat{\phi})$ and an estimate of $\lambda(\nu,\phi_0)$ will be obtained by numerically solving the ordinary convex program in (5.3) with ϕ_0 replaced by $\hat{\phi}$.

Remark 5.2. The posterior asymptotic variance of the support function $\bar{I}_0^{-1}$ is the same as that of the frequentist estimator obtained by Kaido and Santos (2013, Theorem 3.2). Both are derived based on a linear expansion of the support function. This implies that the Bayesian estimation of the support function is also asymptotically semi-parametric efficient in the frequentist sense. On the other hand, there is also a major difference between our results and theirs because when studying the posterior distributions, we do not have an empirical process as Kaido and Santos (2013) do. This requires us to develop a different strategy to prove the linear expansion given in Theorem 5.1 as well as the asymptotic normalities given in Theorems 5.3 and 5.4 below. This also achieves a more strengthened result because the expansion in Theorem 5.1 is uniformly valid in a neighborhood of ϕ_0.

Remark 5.3. The support function $S_\phi(\cdot)$ is a stochastic process with realizations in $\mathcal{C}(\mathbb{S}^d)$, the space of bounded continuous functions on $\mathbb{S}^d$. Despite of the pointwise convergence in Theorem 5.3 for each fixed ν, however, the posterior distribution of the process $\sqrt{n}(S_\phi(\cdot) -$

$S_{\phi_0}(\cdot))$ does not converge to a Gaussian measure on $\mathcal{C}(\mathbb{S}^d)$ in the total variation distance. Roughly speaking, the convergence in total variation would require the existence of a Gaussian measure $\mathbb{G}(\cdot)$ on $\mathcal{C}(\mathbb{S}^d)$ such that uniformly in all Borel measurable sets B of $\mathcal{C}(\mathbb{S}^d)$,

$$|P_{\sqrt{n}(S_\phi(\cdot)-S_{\phi_0}(\cdot))|D_n}(B) - \mathbb{G}(B)| \to^p 0, \tag{5.5}$$

where $P_{\sqrt{n}(S_\phi(\cdot)-S_{\phi_0}(\cdot))|D_n}$ denotes the posterior distribution of the centered support function. However, in general (5.5) does not hold uniformly in all the Borel sets B. Such a negative result can be made rigorous, and is generally known, see e.g., Freedman (1999) or Leahu (2011). □

On the positive side, a *weak* Bernstein-von Mises theorem holds with respect to the weak topology. More precisely, let $\mathbb{G}$ be a Gaussian measure on $\mathcal{C}(\mathbb{S}^d)$ with mean function $\bar{\Delta}_{n,\phi_0}(\cdot) = \lambda(\cdot,\phi_0)^T \nabla_\phi \Psi(\theta_*(\cdot),\phi_0)\Delta_{n,\phi_0}$ and covariance operator with kernel

$$\bar{I}_0^{-1}(\nu_1,\nu_2) = \lambda(\nu_1,\phi_0)^T \nabla_\phi \Psi(\theta_*(\nu_1),\phi_0) I_0^{-1} \nabla_\phi \Psi(\theta_*(\nu_2),\phi_0)^T \lambda(\nu_2,\phi_0), \qquad \forall \nu_1,\nu_2 \in \mathbb{S}^d.$$

We then have the following theorem. For a set B, denote by ∂B the boundary set of B.

Theorem 5.4. *Let $\mathcal{B}$ be the class of Borel measurable sets in $\mathcal{C}(\mathbb{S}^d)$ such that $\mathbb{G}(\partial B) = 0$. Under the assumptions of Theorem 5.3,*

$$\sup_{B\in\mathcal{B}} \left| P_{\sqrt{n}(S_\phi(\cdot)-S_{\phi_0}(\cdot))|D_n}(B) - \mathbb{G}(B) \right| \to^p 0. \tag{5.6}$$

Let '$\Rightarrow$' denote weak convergence on the class of probability measures on $\mathcal{C}(\mathbb{S}^d)$. Then equivalently,

$$P_{\sqrt{n}(S_\phi(\cdot)-S_{\phi_0}(\cdot))|D_n} \Rightarrow \mathbb{G}(\cdot). \tag{5.7}$$

5.3 Models with moment equalities

Our analysis carries over when the model contains both moment equalities and inequalities if the moment equality functions are affine functions. This case is more general than the previous one. Suppose that the identified set writes as

$$\begin{aligned} \Theta(\phi) = \{\theta \in \Theta; \quad & \Psi_i(\theta,\phi) \le 0,\ i = 1,\dots,k_1 \quad \text{and} \\ & a_i^T\theta + b_i(\phi) = 0,\ i = k_1+1,\dots,k_1+k_2\} \end{aligned} \tag{5.8}$$

where a_i is a $(d\times 1)$-vector and b_i is a continuous real-valued function of ϕ for all i. Let k_1 denote the number of moment inequalities, k_2 denote the number of moment equalities, and $k = k_1 + k_2$. We then define $\Psi(\theta,\phi)$ as the $(k\times 1)$ vector whose first k_1 components are the functions $\Psi_i(\theta,\phi)$, $i = 1,\dots,k_1$ and the last k_2 components are the functions $a_i^T\theta + b_i(\phi)$, $i = k_1+1,\dots,k_1+k_2$.

The set $\Theta(\phi)$ is closed and convex (with an empty interior) since it is the intersection of a closed and convex set with closed hyperplanes. In this case, the support function still has a Lagrangian representation as:

$$S_\phi(\nu) = \sup_{\theta \in \Theta} \{\nu^T \theta - \lambda(\nu, \phi)^T \Psi(\theta, \phi)\},$$

where $\lambda(\nu, \phi) : \mathbb{S}^d \times \mathbb{R}^{d_\phi} \to \mathbb{R}_+^{k_1} \times \mathbb{R}^{k_2}$ is a k-vector of Lagrange multipliers (see Rockafellar 1970, chapter 28). Assumptions 5.1-5.5 remain unchanged except for Assumption 5.3 *(iii)*, which is replaced by:

Assumption 5.3. *(iii) There is some $\delta > 0$ such that for all $\phi \in B(\phi_0, \delta)$ there exists a $\theta \in \Theta$ such that $\Psi_i(\theta, \phi) < 0$, $\forall i = 1, \ldots, k_1$.*

The results of Section 5.2 are still valid with minor modifications in the proofs. We detail these modifications in Appendix E.5.

6 Bayesian Credible Sets

Inferences can be carried out through finite-sample Bayesian credible sets (BCS's). We study two kinds of BCS's: credible sets for θ and credible sets for the identified set $\Theta(\phi)$.

6.1 Credible set for $\Theta(\phi)$

6.1.1 Two-sided BCS

We focus on the case when the identified set is convex and closed, and aim at constructing two-sided credible sets A_1 and A_2 such that

$$P(A_1 \subset \Theta(\phi) \subset A_2 | D_n) \geq 1 - \tau$$

for $\tau \in (0, 1)$, where the probability is taken with respect to the posterior of ϕ. Our construction is based on the support function. To illustrate why support function can help, for a set $\Theta(\phi)$ recall its ϵ-envelope: $\Theta(\phi)^\epsilon = \{\theta \in \Theta : d(\theta, \Theta(\phi)) \leq \epsilon\}$ and its ϵ-contraction: $\Theta(\phi)^{-\epsilon} = \{\theta \in \Theta(\phi) : d(\theta, \Theta \backslash \Theta(\phi)) \geq \epsilon\}$ where $\epsilon \geq 0$. Let $\hat{\phi}$ be a Bayesian estimator for ϕ_0, which can be, e.g., the posterior mean or mode. We have, for any $c_n \geq 0$,

$$P(\Theta(\hat{\phi})^{-c_n} \subset \Theta(\phi) \subset \Theta(\hat{\phi})^{c_n} | D_n) = P(\sup_{\|\nu\|=1} |S_\phi(\nu) - S_{\hat{\phi}}(\nu)| \leq c_n | D_n).$$

Note that the right hand side of the above equation depends on the posterior of the support function. Let q_τ be the $1 - \tau$ quantile of the posterior of

$$J(\phi) = \sqrt{n} \sup_{\|\nu\|=1} |S_\phi(\nu) - S_{\hat{\phi}}(\nu)|$$

so that

$$P\left(J(\phi) \leq q_\tau \middle| D_n\right) = 1 - \tau. \tag{6.1}$$

The posterior of $J(\phi)$ is determined by that of ϕ. Hence q_τ can be simulated efficiently from the MCMC draws from $p(\phi|D_n)$. Immediately, we have the following theorem:

Theorem 6.1. *Suppose for any $\tau \in (0,1)$, q_τ is defined as in (6.1), then for every sampling sequence D_n,*

$$P(\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}} \subset \Theta(\phi) \subset \Theta(\hat{\phi})^{q_\tau/\sqrt{n}} | D_n) = 1 - \tau.$$

In particular, $\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}}$ is allowed to be an empty set.

Remark 6.1. It is straightforward to construct the one-sided BCS for $\Theta(\phi)$ using the described procedure. For example, let $\tilde{q}_\tau$ be such that

$$P(\sqrt{n} \sup_{\|\nu\|=1} (S_\phi(\nu) - S_{\hat{\phi}}(\nu)) \leq \tilde{q}_\tau | D_n) = 1 - \tau.$$

Then, $P(\Theta(\phi) \subset \Theta(\hat{\phi})^{\tilde{q}_\tau/\sqrt{n}} | D_n) = 1 - \tau$ for every sampling sequence D_n.

6.1.2 Frequentist coverage probability of BCS for $\Theta(\phi)$

The constructed two-sided BCS for the identified set has desired frequentist properties, which follows from the Bernstein-von Mises Theorem (see Theorem 5.3) of the support function. The frequentist coverage probability for a general (two-sided) multi-dimensional BCS has been largely unknown in the literature before. The analysis relies on the following assumption, which requires the asymptotic normality and semi-parametric efficiency of the consistent estimator $\hat{\phi}$. Under mild conditions, it holds for many regular estimators such as the posterior mean, mode and the maximum likelihood estimator.

Assumption 6.1. *The consistent estimator $\hat{\phi}$ satisfies*

$$\sqrt{n}(\hat{\phi} - \phi_0) \to^d \mathcal{N}(0, I_0^{-1})$$

where I_0 denotes the semi-parametric efficient information matrix as in Assumption 5.8.

Theorem 6.2. *Consider the moment inequality model in (5.1)-(5.2). If the Assumptions of Theorem 5.3 and Assumption 6.1 hold, then the constructed two-sided Bayesian credible set has asymptotically correct frequentist coverage probability, that is, for any $\tau \in (0,1)$,*

$$P_{D_n}(\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}} \subset \Theta(\phi_0) \subset \Theta(\hat{\phi})^{q_\tau/\sqrt{n}}) \geq 1 - \tau + o_p(1).^2$$

where $P_{D_n}(.)$ denote the probability measure based on the sampling distribution, fixing $(\theta, \phi) = (\theta_0, \phi_0)$.

[2]The result presented here is understood as: There is a random sequence $\Delta(D_n)$ that depends on D_n such that $\Delta(D_n) = o_p(1)$, and for any sampling sequence D_n, we have $P_{D_n}(\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}} \subset \Theta(\phi_0) \subset \Theta(\hat{\phi})^{q_\tau/\sqrt{n}}) \geq 1 - \tau + \Delta(D_n)$. Similar interpretation applies to (6.2).

Note that in Theorem 6.1, the random set is $\Theta(\phi)$, while in Theorem 6.2 the random sets are $\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}}$ and $\Theta(\hat{\phi})^{q_\tau/\sqrt{n}}$. The rationale of this theorem is that, because the identified set itself is "point identified", its prior does not depend on that of θ and is dominated by the data asymptotically.

Remark 6.2. Note that q_τ depends only on the posterior of ϕ. Hence Theorem 6.2 does not rely on the prior of θ, and shows asymptotic robustness to the prior of ϕ. It also holds when $\Theta(\phi)$ becomes a singleton, and in that case the lower-side $\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}}$ is empty. Therefore the point identified case is also nested. We shall discuss the uniformity issue in more detail in Section 9.

Similarly, we can show that the one-sided BCS as constructed in Remark 6.1 above has asymptotically correct coverage probability too. For example, for $\tilde{q}_\tau$ such that $P(\sqrt{n}\sup_{\|\nu\|=1}(S_\phi(\nu) - S_{\hat{\phi}}(\nu)) \leq \tilde{q}_\tau | D_n) = 1-\tau$, then

$$P_{D_n}(\Theta(\phi_0) \subset \Theta(\hat{\phi})^{\tilde{q}_\tau/\sqrt{n}}) \geq 1 - \tau + o_p(1). \tag{6.2}$$

6.2 Credible set for θ

We now construct the Bayesian credible set for θ. A BCS for θ at level $1-\tau$ is a set $\text{BCS}(\tau)$ such that

$$P(\theta \in \text{BCS}(\tau) | D_n) = 1 - \tau$$

for $\tau \in (0,1)$. One of the popular choices of the credible set is the highest-probability-density (HPD) set, which has been widely used in empirical studies and also used in the Bayesian partially identified literature by e.g., Moon and Schorfheide (2012) and Norets and Tang (2012).

The BCS can be compared with the frequentist confidence set (FCS). A frequentist confidence set $\text{FCS}(\tau)$ for θ_0 satisfies

$$\lim_{n\to\infty} \inf_{\phi\in\Phi} \inf_{\theta_0 \in \Theta(\phi)} P_{D_n}(\theta_0 \in \text{FCS}(\tau)) \geq 1 - \tau.$$

There have been various procedures in the literature to construct a $\text{FCS}(\tau)$ that satisfies the above inequality. One of the important FCS's is based on a consistent estimator $\hat{\phi}$ of ϕ_0 such that $\Theta(\hat{\phi}) \subset \text{FCS}(\tau)$. By using a known likelihood function, Moon and Schorfheide (2012) compared the BCS with this type of FCS and showed that the BCS and FCS are asymptotically different. As Theorem 6.3 below shows, such a comparison still carries over under the more robust semi-parametric Bayesian setup. The following assumption is needed.

Assumption 6.2. *(i) The frequentist FCS(τ) is such that, there is $\hat{\phi}$ with $\|\hat{\phi} - \phi_0\| = o_p(1)$ satisfying $\Theta(\hat{\phi}) \subset$ FCS(τ).*
(ii) $\pi(\theta \in \Theta(\phi)|\phi) = 1$ for all $\phi \in \Phi$; $\sup_{(\theta,\phi)\in\Theta\times\Phi} \pi(\theta|\phi) < \infty$.

Many frequentist FCS's satisfy condition *(i)*, see, e.g., Imbens and Manski (2004), Chernozhukov et al. (2007), Rosen (2008), Andrews and Soares (2010), etc. Condition *(ii)* is easy

to verify since $\Theta \times \Phi$ is compact. When for every $\phi \in \Phi$, $\Theta(\phi)$ is not a singleton, examples of $\pi(\theta|\phi)$ satisfying assumption 6.2 *(ii)* include: the uniform prior with density

$$\pi(\theta|\phi) = \mu(\Theta(\phi))^{-1} I_{\theta \in \Theta(\phi)},$$

where $\mu(\cdot)$ denotes the Lebesgue measure; and the truncated normal prior with density

$$\pi(\theta|\phi) = \left[\int_{\Theta(\phi)} h(x; \lambda, \Sigma) dx \right]^{-1} h(\theta; \lambda, \Sigma) I_{\theta \in \Theta(\phi)},$$

where $h(x; \lambda, \Sigma)$ is the density function of a multinormal distribution $\mathcal{N}(\lambda, \Sigma)$.

Theorem 6.3. *Under Assumption 6.2 and the assumptions of Theorem 5.2,* $\forall \tau \in (0, 1)$, *(i)*

$$P(\theta \in FCS(\tau)|D_n) \to^p 1,$$

(ii)

$$P(\theta \in FCS(\tau), \theta \notin BCS(\tau)|D_n) \to^p \tau.$$

Remark 6.3. Theorem 6.3 *(i)* shows that the posterior probability that θ lies inside the frequentist confidence set is arbitrarily close to one, as $n \to \infty$. This indicates that the posterior will asymptotically concentrate inside the FCS. On the other hand, by *(ii)*, there is a non-negligible probability that FCS is strictly larger than BCS. The prior information on θ still plays a non-negligible role in the posterior as the sample size increases.

Our prior condition in Assumption 6.2 *(ii)* implies that Theorem 6.3 only focuses on partial identification. It can be restrictive in the point identified case. Because our prior is such that $\pi(\theta \in \Theta(\phi)|\phi) = 1$ for each ϕ, when $\Theta(\phi)$ is a singleton $\pi(\theta|\phi)$ becomes a Dirac function and $\sup_{\theta,\phi} \pi(\theta|\phi) < \infty$ cannot be expected to hold in this case. On the other hand, Assumption 6.2 *(ii)* does cover many partially identified models of interest, and it is a prior assumption that has been used frequently elsewhere in the literature, e.g., Moon and Schorfheide (2012) and Gustafson (2012).

7 Projection and Subset Inference

One of the important features of the proposed procedure is that it is relatively easy to marginalize onto low-dimensional subspaces, and the computation is fast. Suppose the dimension of θ is relatively large, but we are interested in only one component of θ, say θ_1. Then projections aim at constructing the BCS's for θ_1 and for its identified set $\widetilde{\Theta}(\phi)_1$.

We illustrate this by using the interval regression example (Example 3.2). Suppose the full parameter θ is high-dimensional. Let $W_1 = ZY_1$, $W_2 = ZY_2$ and $V = Zx^T$. Here $\phi_1 = EW_1$, $\phi_2 = EV$ and $\phi_3 = EW_2$. Let $\phi = (\phi_1^T, vec(\phi_2)^T, \phi_3^T)$, and $e = (1, 0, ..., 0)^T$. The identified set for θ_1 can be expressed using the support function $S_\phi(\cdot)$:

$$\widetilde{\Theta}(\phi)_1 = \{\theta_1 : \exists \omega = (\theta_2, ..., \theta_d) \text{ such that } (\theta_1, \omega) \in \Theta(\phi)\} = [-S_\phi(-e), S_\phi(e)],$$

where the exact expression for $S_\phi(\cdot)$ is given in Appendix C.1 in the supplementary material. We place a Dirichlet process prior $\mathcal{D}ir(\nu_0, Q_0)$ on the joint CDF of (W_1, W_2, V). By the stick-breaking representation (see Sethuraman 1994) the deduced posterior distribution of ϕ is the distribution of the following quantity:

$$\phi|D_n = \rho \sum_{i=1}^{n} \beta_i D_{n,i} + (1-\rho) \sum_{j=1}^{\infty} \alpha_j \xi_j \tag{7.1}$$

where $D_{n,i}$ is the ith observation of the vector $(W_1^T, vec(V)^T, W_2^T)$, ρ is drawn from a Beta distribution $\mathcal{B}e(n, \nu_0)$ independently of the other quantities, $(\beta_1, \ldots, \beta_n)$ is drawn from a Dirichlet distribution of parameters $(1, \ldots, 1)$ on the simplex S_{n-1} of dimension $(n-1)$, $\xi_j \sim iid Q_0$ and $\{\alpha_k\}_{k\geq 1}$ are computed as $\alpha_k = v_k \prod_{l=1}^{k}(1-v_l)$ where $\{v_l\}_{l\geq 1}$ are independent drawings from a beta distribution $\mathcal{B}e(1, \nu_0)$ and $\{v_j\}_{j\geq 1}$ are independent of $\{\xi_j\}_{j\geq 1}$. In practice, we can set a truncation K, so the infinite sum in the posterior representation in (7.1) is replaced with a truncated sum $(1-\rho)\sum_{j=1}^{K} \alpha_j \xi_j$. In addition, $(\alpha_1, ...\alpha_K)$ are normalized so that $\sum_{j=1}^{K} \alpha_j = 1$.

We can place a uniform prior for θ, and draw $\{\theta^{(i)}, \phi^{(i)}\}_{i=1}^{B}$ from the posterior $(\theta, \phi)|D_n$. Then $\{\theta_1^{(i)}\}_{i=1}^{B}$ are the draws from the marginal posterior of θ_1. Let $\theta_1^{(\tau/2)}$ and $\theta_1^{(1-\tau/2)}$ be the $\tau/2$th and $(1-\tau/2)$th sample quantiles of $\{\theta_1^{(i)}\}_{i=1}^{B}$. Then $[\theta_1^{(\tau/2)}, \theta_1^{(1-\tau/2)}]$ is the BCS(τ) of θ_1. Moreover, let q_τ be the $(1-\tau)$th quantile of the posterior of

$$J(\phi) = \sqrt{n} \max\{S_\phi(e) - S_{\hat{\phi}}(e), S_\phi(-e) - S_{\hat{\phi}}(-e)\},$$

which can be approximated by the $(1-\tau)$th sample quantile of $\{J(\phi^{(i)})\}_{i=1}^{B}$. We then construct the BCS(τ) for $\widetilde{\Theta}(\phi)_1$ as $[-S_{\hat{\phi}}(-e) - \frac{q_\tau}{\sqrt{n}}, S_{\hat{\phi}}(e) + \frac{q_\tau}{\sqrt{n}}]$.

We present a simple numerical result for illustration, where $\theta_{01} = 1$, but the total dimension is high: $\dim(\theta_0) = 10$. Let $W_1 \sim \mathcal{N}(0, 0.5I)$ and $W_2 \sim \mathcal{N}(5, I)$. Set $\nu_0 = 3$ and the base measure $Q_0 = \mathcal{N}(0, I)$. $B = 100$ posterior draws are sampled. While the finite sample performance is very robust to the choice of the truncation K, we choose K following the guidance of Ishwaran and James (2002), who obtained an approximation error of order $n\exp(-(K-1)/\nu_0)$ for truncations. Hence, in the simulation with $n = 500, \nu_0 = 3$, the choice $K = 50$ gives an error of order 4×10^{-5}. Table 1 summarizes the true identified set $\widetilde{\Theta}(\phi_0)_1$ for θ_1, and the averaged BCS(0.1) for both θ_1 and the projected set $\widetilde{\Theta}(\phi)_1$ over 50 replications. Results based on various choices of (n, B, K) are reported.

When computing the BCS for $\widetilde{\Theta}(\phi)_1$, it is also interesting to compare the computation time with that of the high-dimensional projection based on the criterion function approach as in, e.g. Chernozhukov et al. (2007), Andrews and Soares (2010) etc, because they have the same asymptotic frequentist coverages as ours. For the moment inequalities $\Psi(\theta, \phi) = (\phi_2\theta - \phi_3, \phi_1 - \phi_2\theta) \leq 0$ we employ the criterion function and construct a confidence set FCS

Table 1: 90% Bayesian credible sets marginalized to the subset for θ_1

n	K	B	$\widetilde{\Theta}(\phi_0)_1$	BCS for θ_1	BCS for $\widetilde{\Theta}(\phi)_1$
500	50	100	[0, 1.667]	[0.007, 1.667]	[-0.174, 1.844]
	50	500		[-0.008 1.666]	[-0.174, 1.837]
	100	100		[-0.012 1.662]	[-0.181, 1.832]
	100	500		[-0.000 1.667]	[-0.169, 1.840]
1000	50	100	[0, 1.667]	[0.023, 1.641]	[-0.121, 1.789]
	50	500		[0.011, 1.641]	[-0.126, 1.786]
	100	100		[0.036, 1.636]	[-0.120, 1.781]
	100	500		[0.025, 1.641]	[-0.121, 1.786]

as in Chernozhukov et al. (2007):

$$Q_n(\theta) = \sum_j \max(\Psi_j(\theta, \hat{\phi}), 0)^2 w_j, \quad \text{FCS}(\tau) = \{\theta : \sqrt{n} Q_n(\theta) \leq c_\tau\},$$

with $w_j = 1$ and $\hat{\phi}$ the sample mean estimator of ϕ. The critical value c_τ is obtained via the bootstrap procedure proposed by Bugni (2010), which requires solving a constrained optimization problem. We use the "fmincon" toolbox in Matlab for the numerical optimization[3], and then project $\text{FCS}(\tau)$ onto the subspace for θ_1 to get the marginal confidence set $\text{FCS}_1(\tau)$. The projection is done through the following steps: generate $\{\theta_j^*\}_{j=1}^M$ uniformly from Θ. Let $\theta_{j,1}^*$ be the first component of θ_j^* and

$$L(\tau) = \min\{\theta_{j,1}^* : \theta_j^* \in \text{FCS}(\tau), j = 1, ..., M\}, \quad U(\tau) = \max\{\theta_{j,1}^* : \theta_j^* \in \text{FCS}(\tau), j = 1, ..., M\}. \tag{7.2}$$

Then $[L(\tau), U(\tau)]$ forms a projected frequentist confidence interval for $\widetilde{\Theta}(\phi_0)_1$. In the simulation, we set a small parameter space $\Theta = \otimes_{i=1}^{10}[-2, 2]$ in order to calculate $L(\tau)$ and $U(\tau)$ efficiently.

Table 2 compares the computation times necessary to obtain the projected sets using our proposed BCS and using the criterion function approach. Reported is the averaged time for one computation over 50 replications, using the same simulated model. We see that the proposed BCS projection computes much faster.

[3]We use the Matlab code of Bugni (2010), downloaded from the online supplement of Econometrica. The optimization is solved constrained on an estimated identified set, which involves an additional parameter t_n. We set $t_n = \log(n)$.

Table 2: Computation times (in seconds) for the projected BCS and criterion-function-FCS

		BCS			FCS		
		K			M		
n	B	50	100	500	30	50	100
500	50	0.065	0.073	0.129	9.169	10.214	10.955
	100	0.128	0.142	0.248	18.963	18.893	19.496
	200	0.244	0.273	0.479	37.067	36.599	37.288
1000	50	0.072	0.079	0.136	14.123	14.248	15.837
	100	0.137	0.155	0.259	26	27.029	28.442
	200	0.269	0.295	0.549	54.027	52.661	54.240

Proposed BCS and criterion-function-based-FCS are compared. B is the number of either posterior draws (for BCS) or Bootstrap draws (for FCS) to compute the critical values; K is the truncation number to approximate the Dirichlet process posterior; M is used in (7.2) for the projected FCS. Computations are conducted using a 2.3 GHz Mac with Intel Core i7 CPU.

8 Posterior consistency for $\Theta(\phi)$

The estimation accuracy of the identified set is often measured, in the literature, by the Hausdorff distance. Specifically, for a point a and a set A, let $d(a, A) = \inf_{x\in A} \|a-x\|$, where $\|\cdot\|$ denotes the Euclidean norm. The Hausdorff distance between sets A and B is defined as

$$d_H(A, B) = \max\left\{\sup_{a\in A} d(a, B), \sup_{b\in B} d(b, A)\right\} = \max\left\{\sup_{a\in A}\inf_{b\in B}\|a-b\|, \sup_{b\in B}\inf_{a\in A}\|b-a\|\right\}.$$

This section aims at deriving a rate $r_n = o(1)$ such that for some constant $C > 0$,

$$P(d_H(\Theta(\phi), \Theta(\phi_0)) < Cr_n|D_n) \to^p 1.$$

The above result is based upon two important features: (1) the posterior concentration rate for ϕ which is stated in assumption 5.7, and (2) the continuity of the Hausdorff distance with respect to ϕ.

The continuity of $d_H(\Theta(\phi), \Theta(\phi_0))$ with respect to ϕ in multi-dimensional models is hard to verify. Hence, instead of assuming the continuity directly, we place a less demanding assumption which implicitly implies the continuity but is relatively easier to verify. With this aim, we consider the *moment inequality model* described in equations (5.1) - (5.2) and place the following assumptions.

Assumption 8.1. *The parameter space $\Theta \times \Phi$ is compact.*

Assumption 8.2. $\{\Psi(\theta,\cdot) : \theta \in \Theta\}$ *is Lipschitz equi-continuous on* Φ*, that is, for some* $K > 0$*,* $\forall \phi_1, \phi_2 \in \Phi$*,*

$$\sup_{\theta\in\Theta} \|\Psi(\theta,\phi_1) - \Psi(\theta,\phi_2)\| \leq K\|\phi_1 - \phi_2\|.$$

Given the compactness of Θ, this assumption is satisfied by many interesting examples of moment inequality models.

Assumption 8.3. *There exists a closed neighborhood* $U(\phi_0)$ *of* ϕ_0*, such that for any* $a_n = O(1)$*, and any* $\phi \in U(\phi_0)$*, there exists* $C > 0$ *that might depend on* ϕ*, so that*

$$\inf_{\theta: d(\theta,\Theta(\phi))\geq Ca_n} \max_{i\leq k} \Psi_i(\theta,\phi) > a_n.$$

Intuitively, when θ is bounded away from $\Theta(\phi)$ (up to a rate a_n), at least one of the moment inequalities is violated, which means $\max_{i\leq k} \Psi_i(\theta,\phi) > 0$. This assumption quantifies how much $\max_{i\leq k} \Psi_i(\theta,\phi)$ will depart from zero. This is a sufficient condition for the partial identification condition (4.5) in Chernozhukov, Hong and Tamer (2007). If we define

$$Q(\theta,\phi) = \|\max(\Psi(\theta,\phi),0)\| = \left[\sum_{i=1}^{k} (\max(\Psi_i(\theta,\phi),0))^2\right]^{1/2}$$

then $Q(\theta,\phi) = 0$ if and only if $\theta \in \Theta(\phi)$. The partial identification condition in Chernozhukov et al. (2007, condition (4.5)) assumes that there exists $K > 0$ so that for all θ,

$$Q(\theta,\phi) \geq Kd(\theta,\Theta(\phi)), \tag{8.1}$$

which says that Q should be bounded below by a number proportional to the distance of θ from the identified set if θ is bounded away from the identified set. Assumption 8.3 is a sufficient condition for (8.1).

Example 8.1 (Interval censored data - *continued*)**.** In the interval censoring data example, $\Psi(\theta,\phi) = (\theta - \phi_2, \phi_1 - \theta)^T$ and for any $\phi = (\phi_1,\phi_2)$ and $\tilde{\phi} = (\tilde{\phi}_1,\tilde{\phi}_2)$ we have: $\|\Psi(\theta,\phi) - \Psi(\theta,\tilde{\phi})\| = \|\phi - \tilde{\phi}\|$. This verifies Assumption 8.2. Moreover, for any θ such that $d(\theta,\Theta(\phi)) \geq a_n$, either $\theta \leq \phi_1 - a_n$ or $\theta \geq \phi_2 + a_n$. If $\theta \leq \phi_1 - a_n$, then $\Psi_2(\theta,\phi) = \phi_1 - \theta \geq a_n$; if $\theta \geq \phi_2 + a_n$, then $\Psi_1(\theta,\phi) = \theta - \phi_2 \geq a_n$. This verifies Assumption 8.3. □

The following theorem shows the concentration rate for the posterior of the identified set.

Theorem 8.1. *Under Assumptions 5.7, 8.1-8.3, for some* $C > 0$*,*

$$P(d_H(\Theta(\phi),\Theta(\phi_0)) > C\sqrt{\frac{\log n}{n}}|D_n) \to^p 0. \tag{8.2}$$

Remark 8.1. The convergence in Hausdorff distance can be implied by that of the support function for convex and close sets (e.g., Beresteanu and Molinari 2008). Therefore, (8.2) is

another statement of result (5.4). However, they are obtained under different assumptions and (8.2) is obtained directly from the perspective of the posterior of the identified set.

Remark 8.2. Recently, Kitagawa (2012) obtained the posterior consistency for $\Theta(\phi)$ in the one-dimensional case: $P(d_H(\Theta(\phi), \Theta(\phi_0)) > \epsilon | D_n) \to 0$ for almost every sampling sequence of D_n. This result was obtained for the case where $\Theta(\phi)$ is a connected interval and $d_H(\Theta(\phi), \Theta(\phi_0))$ is assumed to be a continuous map of ϕ. In multi-dimensional cases where $\Theta(\phi)$ is a more general convex set, however, verifying the continuity of $d_H(\Theta(\phi), \Theta(\phi_0))$ is much more technically involved, due to the challenge of computing the Hausdorff distance in multi-dimensional manifolds. In contrast, our Lipschitz equi-continuity condition in Assumption 8.2 and Assumption 8.3 are much easier to verify in specific examples, as they depend on the moment conditions directly.

9 Further Illustrations and Uniformity

9.1 Missing data: coverage probabilities and prior sensitivity

This subsection illustrates the coverages of the proposed BCS in the missing data problem (example 3.3), previously discussed by Manski (2003). Let Y be a binary variable, indicating whether a treatment is successful ($Y = 1$) or not ($Y = 0$). However, Y is observed subject to missing. We write $M = 0$ if Y is missing, and $M = 1$ otherwise. Hence, we observe (M, MY). The parameter of interest is $\theta = P(Y = 1)$. The identified parameters are denoted by

$$\phi_1 = P(M = 1), \quad \phi_2 = P(Y = 1 | M = 1).$$

Let $\phi_0 = (\phi_{10}, \phi_{20})$ be the true value of $\phi = (\phi_1, \phi_2)$. Then, without further assumption on $P(Y = 1|M = 0)$, θ is only partially identified on $\Theta(\phi) = [\phi_1\phi_2, \phi_1\phi_2 + 1 - \phi_1]$. The support function is easy to calculate and is

$$S_\phi(1) = \phi_1\phi_2 + 1 - \phi_1 \quad S_\phi(-1) = -\phi_1\phi_2.$$

Suppose we observe i.i.d. data $\{(M_i, Y_iM_i)\}_{i=1}^n$, and define $\sum_{i=1}^n M_i = n_1$ and $\sum_{i=1}^n Y_iM_i = n_2$. In this example, the true likelihood function $l_n(\phi) \propto \phi_1^{n_1}(1 - \phi_1)^{n-n_1}\phi_2^{n_2}(1 - \phi_2)^{n_1-n_2}$ is known.

We place independent beta priors, Beta(α_1, β_1) and Beta(α_2, β_2), on (ϕ_1, ϕ_2). The uniform distribution is a special case of Beta prior. Then the posterior of (ϕ_1, ϕ_2) is a product of Beta$(\alpha_1 + n_1, \beta_1 + n - n_1)$ and Beta$(\alpha_2 + n_2, \beta_2 + n_1 - n_2)$. If in addition, we have prior information on θ and place a prior $\pi(\theta|\phi)$ supported on $\Theta(\phi)$, then by integrating out ϕ, we immediately obtain the marginal posterior of θ.

We now present the two-sided BCS for $\Theta(\phi)$ obtained by using the support function of $\Theta(\phi)$. The estimator $\hat{\phi}$ is taken as the posterior mode: $\hat{\phi}_1 = (n_1 + \alpha_1 - 1)/(n + \alpha_1 + \beta_1 - 2)$,

and $\hat{\phi}_2 = (n_2 + \alpha_2 - 1)/(n_1 + \alpha_2 + \beta_2 - 2)$. Then

$$J(\phi) = \sqrt{n} \max \left\{ |\phi_1\phi_2 - \phi_1 - \hat{\phi}_1\hat{\phi}_2 + \hat{\phi}_1|, |\phi_1\phi_2 - \hat{\phi}_1\hat{\phi}_2| \right\}.$$

Let q_τ be the $1-\tau$ quantile of the posterior of $J(\phi)$, which can be obtained by simulating from the Beta distributions. The lower and upper $1-\tau$ level BCS's for $\Theta(\phi)$ are $\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}} \subset \Theta(\phi) \subset \Theta(\hat{\phi})^{q_\tau/\sqrt{n}}$ where

$$\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}} = [\hat{\phi}_1\hat{\phi}_2 + q_\tau/\sqrt{n}, \hat{\phi}_1\hat{\phi}_2 + 1 - \hat{\phi}_1 - q_\tau/\sqrt{n}]$$

and

$$\Theta(\hat{\phi})^{q_\tau/\sqrt{n}} = [\hat{\phi}_1\hat{\phi}_2 - q_\tau/\sqrt{n}, \hat{\phi}_1\hat{\phi}_2 + 1 - \hat{\phi}_1 + q_\tau/\sqrt{n}],$$

which are also two-sided asymptotic $1-\tau$ frequentist confidence intervals of the true $\Theta(\phi_0)$.

Here we present a simple simulated example, where the true $\phi_0 = (0.7, 0.5)$. This implies the true identified interval to be $[0.35, 0.65]$ and about thirty percent of the simulated data are "missing". We set $\alpha_1 = \alpha_2, \beta_1 = \beta_2$ in the prior. In addition, $B = 1,000$ posterior draws $\{\phi^i\}_{i=1}^B$ are sampled from the posterior Beta distribution. For each of them, compute $J(\phi^i)$ and set $q_{0.05}$ as the 95% upper quantile of $\{J(\phi^i)\}_{i=1}^B$ to obtain the critical value of the BCS and construct the two-sided BCS for the identified set. Each simulation is repeated for 500 times to calculate the coverage frequency of the true identified interval. Table 3 presents the results. We see that the coverage probability for the two-sided is close to the desired 95% when sample size increases. In addition, the marginal coverages of the lower and upper sets are close to 97.5% when sample size is relatively large.

Moreover, Figure 1 plots the five conjugate prior specifications used in this study: flat prior, reverse J-shaped with a right tail, J-shaped with a left tail, U-shaped, and uni-mode. These priors reflect different types of prior beliefs: the first prior is used if a researcher has no informative prior information, the second one (resp. third one) is used if one strongly believes that the probability of missing is low (resp. high), the fourth prior is used when one thinks that the probability of missing is either very high or very low, and the last prior corresponds to a symmetric prior belief centered at fifty percent. So Table 3 also provides simple sensitivity analysis of prior specification. The results demonstrate robustness of the coverage probabilities to conjugate prior specification.

9.2 Uniformity: from partial identification to point identification

We have been focusing on partially identified models, and the inference results achieved are for a fixed data generating process. It is interesting to see whether they still hold uniformly over a class of data generating processes, including the case when point identification is nearly achieved. This is important because in many cases it is possible that we actually have point identification and, in that event, $\Theta(\phi)$ degenerates to a singleton. For example, in the interval censored model, when $EY_1 = EY_2$, $\theta = EY$ is point identified.

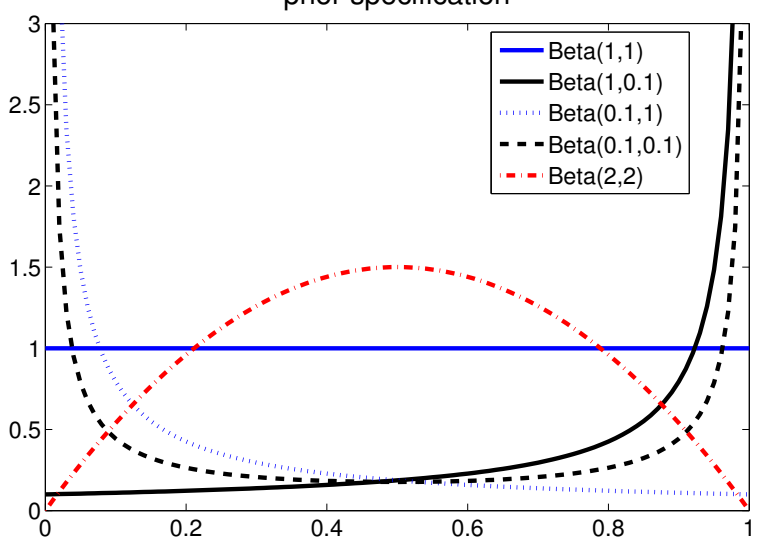
When point identification is indeed achieved, the frequentist coverage probability of the

Table 3: Frequentist coverage probability of BCS and prior sensitivity for missing data

n	α	β	Lower	Upper	Two-sided
50	1	1	0.978	0.944	0.924
	1	0.1	0.964	0.944	0.912
	0.1	1	0.952	0.958	0.916
	0.1	0.1	0.974	0.958	0.938
	2	2	0.958	0.970	0.932
100	1	1	0.982	0.96	0.948
	1	0.1	0.978	0.968	0.950
	0.1	1	0.968	0.968	0.948
	0.1	0.1	0.972	0.972	0.944
	2	2	0.956	0.978	0.944
500	1	1	0.970	0.974	0.950
	1	0.1	0.978	0.978	0.958
	0.1	1	0.974	0.972	0.948
	0.1	0.1	0.972	0.974	0.950
	2	2	0.976	0.978	0.956

Lower, Upper and Two-sided represent the frequencies of the events $\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}} \subset \Theta(\phi_0)$, $\Theta(\phi_0) \subset \Theta(\hat{\phi})^{q_\tau/\sqrt{n}}$, *and* $\Theta(\hat{\phi})^{-q_\tau/\sqrt{n}} \subset \Theta(\phi_0) \subset \Theta(\hat{\phi})^{q_\tau/\sqrt{n}}$ *over 500 replicates. The coverage probability for the two-sided BCS is set to 95%.*

Figure 1: Conjugate priors for sensitivity analysis in the missing data problem



upper-sided BCS $\Theta(\phi) \subset \Theta(\hat{\phi})^{q_\tau/\sqrt{n}}$ and the asymptotic normality for the posterior of the support function still hold because they are generally guaranteed by the semi-parametric Bernstein-von Mises theorem for ϕ when $\Theta(\phi)$ is a singleton (e.g., Rivoirard and Rousseau 2012, Bickel and Kleijn 2011). But the low-side BCS for $\Theta(\phi)$ will be empty with a positive probability. Theorem 6.3, however, does not hold anymore when θ is point identified, as we discussed previously. We further illustrate the uniformity in two examples.

Example 9.1 (Interval censored data - *continued*)**.** We show that the the upper BCS for the identified set has a uniformly correct frequentist asymptotic coverage probability. To simplify our illustration, we assume Y_1 and Y_2 are independent and follow $\mathcal{N}(\phi_{10}, 1)$ and $\mathcal{N}(\phi_{20}, 1)$, respectively. A sequence of different ϕ_0 are considered which includes the case $\phi_{10} - \phi_{20} = o(1)$. When $\phi_{10} = \phi_{20}$, however, we suppose Y_1, Y_2 are sampled independently. Suppose econometricians place independent standard normal priors on ϕ_1 and ϕ_2, then the posteriors are independent, given by $\phi_i | D_n \sim \mathcal{N}(\bar{Y}_i \frac{n}{1+n}, \frac{1}{1+n}), i = 1, 2$, and $\hat{\phi} = \frac{n}{1+n}(\bar{Y}_1, \bar{Y}_2)$ is the posterior mode of the joint distribution. The support function is $S_\phi(1) = \phi_2$, $S_\phi(-1) = -\phi_1$. Let

$$J(\phi) = \sqrt{n} \sup_{\|\nu\|=1} (S_\phi(\nu) - S_{\hat{\phi}}(\nu)) = \sqrt{n} \max \left\{ \phi_2 - \frac{n}{1+n} \bar{Y}_2, \frac{n}{1+n} \bar{Y}_1 - \phi_1 \right\},$$

and let $\tilde{q}_\tau$ be the $1 - \tau$ quantile of the posterior of $J(\phi)$. We now show that the frequentist coverage of BCS(τ)= $[\bar{Y}_1 \frac{n}{1+n} - \frac{\tilde{q}_\tau}{\sqrt{n}}, \bar{Y}_2 \frac{n}{1+n} + \frac{\tilde{q}_\tau}{\sqrt{n}}]$ is valid uniformly for $\phi_0 = (\phi_{10}, \phi_{20}) \in \Phi$, that is,

$$\liminf_{n \to \infty} \inf_{(\phi_{01}, \phi_{02}) \in \Phi} P_{D_n}(\Theta(\phi_0) \subset \text{BCS}(\tau)) = 1 - \tau. \tag{9.1}$$

We can simplify $J(\phi)$ to be $\sqrt{n/(1+n)} \max\{Z_1, Z_2\}$ where $Z_i \sim \mathcal{N}(0, 1)$, $i = 1, 2$ and Z_1 and Z_2 are independent. This implies, for the standard normal's CDF $H(\cdot)$,

$$1 - \tau = P(J(\phi) \leq \tilde{q}_\tau | D_n) = P\left(\max\{Z_1, Z_2\} \leq \tilde{q}_\tau \sqrt{\frac{1+n}{n}} \right) = H\left(\tilde{q}_\tau \sqrt{\frac{1+n}{n}} \right)^2.$$

Hence, $H\left(\tilde{q}_\tau \sqrt{(1+n)/n}\right)^2 = H(\tilde{q}_\tau)^2 + o(1)$ and so $H(\tilde{q}_\tau)^2 \to 1 - \tau$. The event $\{\Theta(\phi_0) \subset \text{BCS}(\tau)\}$ is equivalent to $(\bar{Y}_1 - \phi_{10}) \frac{n}{1+n} \leq \frac{\tilde{q}_\tau}{\sqrt{n}} + \frac{\phi_{10}}{1+n}$ and $(\bar{Y}_2 - \phi_{20}) \frac{n}{1+n} \geq \frac{\phi_{20}}{1+n} - \frac{\tilde{q}_\tau}{\sqrt{n}}$. Hence,

$$\begin{aligned} \inf_{\phi_0 \in \Phi} P_{D_n}(\Theta(\phi_0) \subset \text{BCS}(\tau)) &= \inf_{\phi_0 \in \Phi} H\left((\frac{\tilde{q}_\tau}{\sqrt{n}} + \frac{\phi_{10}}{1+n}) \frac{1+n}{\sqrt{n}} \right) H\left((\frac{\tilde{q}_\tau}{\sqrt{n}} - \frac{\phi_{20}}{1+n}) \frac{1+n}{\sqrt{n}} \right) \\ &= H(\tilde{q}_\tau)^2 + o(1) \to 1 - \tau. \end{aligned}$$

This gives (9.1).

On the other hand, if $\phi_{20} - \phi_{10} = o(1)$, the lower BCS for $\Theta(\phi)$ is empty with a large probability. To see this, for any fixed q, the lower BCS is $A = [\bar{Y}_1 \frac{n}{1+n} + \frac{q}{\sqrt{n}}, \bar{Y}_2 \frac{n}{1+n} - \frac{q}{\sqrt{n}}]$. Let $\Delta_n = \phi_{20} - \phi_{10}$, then $P(A = \emptyset) = P((\bar{Y}_2 - \bar{Y}_1) \frac{n}{1+n} < \frac{2q}{\sqrt{n}}) = H(\sqrt{2}q - \sqrt{\frac{n}{2}} \Delta_n) + o(1)$.

Suppose $\sqrt{n}\Delta_n = o(1)$, then $P(A = \emptyset) \to H(\sqrt{2}q)$. This probability is very large for many reasonable cut-off q. For example, if $q = 1.96$, $H(\sqrt{2}q) = 0.997$.

For a numerical illustration, set $\phi_{10} = 1, \phi_{20} = 1 + \Delta_n$ for a sequence of small Δ_n that decreases to zero, and calculate the frequency that $\Theta(\phi_0) \subset \text{BCS}(0.05)$ and $A = \emptyset$. The model is nearly point identified, and point identification is achieved when $\Delta_n = 0$. Results are reported in Table 4.

Example 9.2 (Missing data example - *continued*)**.** Consider again the missing data example in Section 9.1, where now the true ϕ_{10} is $\phi_{10} = 1 - \Delta_n$ with $\Delta_n \to 0$, that is, the probability of missing is close to zero. So the model is close to point identification. However, suppose we still place priors on ϕ_1 and ϕ_2 and $\Theta(\phi) = [\phi_1\phi_2, \phi_1\phi_2 + 1 - \phi_1]$ as before. Our result shows that

$$P_{D_n}(\Theta(\phi_0) \subset \Theta(\hat{\phi})^{\tilde{q}_\tau/\sqrt{n}}) \to 1 - \tau \tag{9.2}$$

when $\tilde{q}_\tau$ is the $1 - \tau$ quantile of the posterior of

$$\sqrt{n} \max \left\{ \phi_1\phi_2 - \phi_1 - \hat{\phi}_1\hat{\phi}_2 + \hat{\phi}_1, \hat{\phi}_1\hat{\phi}_2 - \phi_1\phi_2 \right\}.$$

It can also be shown that the coverage (9.2) holds uniformly for ϕ_0 inside a compact parameter space. It is also easy to see that, if $\Delta_n = o(n^{-1/2})$, then for any $\tau \in (0, 1)$, the lower BCS(τ) is empty with probability approaching one.

We illustrate the above discussions using a simulated example, where $\phi_{10} = 1 - \Delta_n$ for a sequence of small Δ_n. We use the uniform priors and compute the frequency of the events that $\Theta(\phi_0) \subset \Theta(\hat{\phi})^{\tilde{q}_{0.05}/\sqrt{n}}$ and that the lower BCS is empty. We set $\phi_{20} = 0.5$ so that $\hat{\phi}_2$ has the maximum possible variance. Therefore, our simulation also demonstrates how sensitive the coverage frequency is to the variance of the point identified estimator. The frequency of coverage over 500 replications are summarized in Table 4 below.

We see that the upper BCS with 95% credible level has the coverage probability for the true identified set close to 0.95. Also, the lower BCS is empty almost all the times.

10 Financial Asset Pricing

We develop a detailed application in financial asset pricing model, where the identified set is of direct interest.

10.1 The model

Asset pricing models state that the equilibrium price P_t^i of a financial asset i is equal to

$$P_t^i = E[M_{t+1}P_{t+1}^i|\mathcal{I}_t], \qquad i = 1, \ldots, N$$

where P_{t+1}^i denotes the price of asset i at the period $(t + 1)$, M_{t+1} is the stochastic discount factor (SDF, hereafter) and $\mathcal{I}_t$ denotes the information set at time t. In vectorial form this

Table 4: Frequency of BCS(0.05) coverages for near point identification

			Δ_n			
	n	event	0.1	0.05	0.01	0
Interval censoring	50	Lower BCS= $\emptyset$	0.966	0.94	0.972	0.956
		$\Theta(\phi_0) \subset \Theta(\hat{\phi})^{\tilde{q}/\sqrt{n}}$	0.952	0.964	0.952	0.944
	100	Lower BCS= $\emptyset$	0.964	0.96	0.962	0.962
		$\Theta(\phi_0) \subset \Theta(\hat{\phi})^{\tilde{q}/\sqrt{n}}$	0.962	0.966	0.958	0.95
Missing data	50	Lower BCS= $\emptyset$	0.998	1	1	1
		$\Theta(\phi_0) \subset \Theta(\hat{\phi})^{\tilde{q}/\sqrt{n}}$	0.95	0.952	0.942	0.944
	100	Lower BCS= $\emptyset$	0.99	1	1	1
		$\Theta(\phi_0) \subset \Theta(\hat{\phi})^{\tilde{q}/\sqrt{n}}$	0.952	0.956	0.958	0.952

The frequencies (over 500 replications) that the lower BCS is empty and that the upper BCS covers the true identified set are summarized. The length of the true identified set is Δ_n. The model achieves point identification when $\Delta_n = 0$.

rewrites as

$$\iota = E[M_{t+1}R_{t+1}|\mathcal{I}_t] \tag{10.1}$$

where ι is the N-dimensional vector of ones and R_{t+1} is the N-dimensional vector of gross asset returns at time $(t+1)$: $R_{t+1} = (r_{1,t+1}, \ldots, r_{N,t+1})^T$ with $r_{i,t+1} = P^i_{t+1}/P^i_t$. This model can be interpreted as a model of the SDF M_{t+1} and may be used to detect the SDFs that are compatible with asset return data. Hansen and Jagannathan (1991) have obtained a lower bound on the volatility of SDFs that could be compatible with a given SDF-mean value and a given set of asset return data. Therefore, the set of SDFs M_{t+1} that can price existing assets generally form a proper set.

Let m and Σ denote, respectively, the vector of unconditional mean returns and covariance matrix of returns of the N risky assets, that is, $m = E(R_{t+1})$ and $\Sigma = E(R_{t+1} - m)(R_{t+1} - m)^T$. Denote $\mu = E(M_{t+1})$ and $\sigma^2 = Var(M_{t+1})$, which are partially identified. We assume that m, Σ, μ and σ^2 do not vary with t. Hansen and Jagannathan (1991) showed that given (m, Σ), which are point identified by the observed $\{R_{t+1}\}$, if the SDF M_{t+1} satisfies (10.1), then its variance σ^2 should be no smaller than:

$$\begin{aligned} \sigma^2_\phi(\mu) &= (\iota - \mu m)^T \Sigma^{-1} (\iota - \mu m) \equiv \phi_1 \mu^2 - 2\phi_2 \mu + \phi_3 \\ \text{with } \phi_1 &= m^T \Sigma^{-1} m, \qquad \phi_2 = m^T \Sigma^{-1} \iota, \qquad \phi_3 = \iota^T \Sigma^{-1} \iota. \end{aligned} \tag{10.2}$$

Therefore, an SDF correctly prices an asset only if, for given (m, Σ), its mean μ and variance σ^2 are such that $\sigma^2 \geq \sigma^2_\phi(\mu)$, and in this case the SDF is called *admissible*. Inadmissible SDFs do not satisfy model (10.1).

Define the set of admissible SDF's means and variances:

$$\Theta(\phi) = \left\{(\mu, \sigma^2) \in \Theta;\ \sigma^2_\phi(\mu) - \sigma^2 \leq 0\right\}$$

where $\phi = (\phi_1, \phi_2, \phi_3)^T$ and $\Theta \subset \mathbb{R}_+ \times \mathbb{R}_+$ is a compact set that we can choose based on experiences. Usually, we can fix upper bounds $\bar{\mu} > 0$ and $\bar{\sigma} > 0$ as large as we want and take $\Theta = [0, \bar{\mu}] \times [0, \bar{\sigma}^2]$. In practice, $\bar{\mu}$ and $\bar{\sigma}$ must be chosen sufficiently large such that $\Theta(\phi)$ is non-empty. We also point out that to be consistent with our developed theory, the parameter space is chosen to be compact. Thus, the space for σ^2 includes zero. In practice, one can require $\sigma^2 \geq \epsilon$ for a sufficiently small $\epsilon > 0$. For simplicity, we keep the current parameter space for σ^2, which is also used sometimes in the literature. Making inference on $\Theta(\phi)$ allows to check whether a family of SDF (and then a given utility function) prices a financial asset correctly or not. Frequentist inference for this set is carried out in Chernozhukov, Kocatulum and Menzel (2012).

Using our previous notation, we define $\theta = (\mu, \sigma^2)$ and

$$\Psi(\theta, \phi) = \sigma^2_\phi(\mu) - \sigma^2,$$

which gives a moment inequality model.

10.2 Support function

In this case $\Psi(\theta, \phi)$ is convex in θ. More precisely, $\Psi(\theta, \phi)$ is linear in σ^2 and strictly convex in μ (because Σ is positive definite so $\phi_1 > 0$). Assumptions 5.1- 5.6 are easy to verify except for Assumptions 5.5 and 5.6*(i)* and *(iv)*. However, it can be shown that the support function is differentiable at ϕ_0 without Assumption 5.5 being satisfied. So, our Bayesian analysis on the support function of $\Theta(\phi)$ still goes through. Assumption 5.6 *(i)* and *(iv)* must be checked case by case (that is, for every region of values of ν) since $\lambda(\nu, \phi)$ takes a different expression in each case, see Appendix C.2 in the supplementary material.

We can rewrite the support function to be $S_\phi(\nu) = \Xi(\nu, \phi)^T \nu$, where

$$\begin{aligned}
\Xi(\nu, \phi) &= \arg\max_{\theta \in \Theta} \left\{\nu^T \theta;\ \Psi(\theta, \phi) \leq 0\right\} \\
&= \arg \max_{0 \leq \mu < \bar{\mu},\, 0 < \sigma^2 < \bar{\sigma}^2} \left\{\nu_1 \mu + \nu_2 \sigma^2 - \lambda(\nu, \phi)(\phi_1 \mu^2 - 2\phi_2 \mu + \phi_3 - \sigma^2)\right\}
\end{aligned}$$

where $\nu = (\nu_1, \nu_2)$. The support function and $\Xi(\nu, \phi)$ have explicit expressions, but they are very long and complicated. Thus, we present them in Appendix C.2.

10.3 Dirichlet process prior

Let F denote a probability distribution. The Bayesian model is $R_t|F \sim F$ and $(m, \Sigma) = (m(F), \Sigma(F))$, where

$$m(F) = \int rF(dr), \quad \Sigma(F) = \int rr^T F(dr) - \int rF(dr) \int rF(dr)^T.$$

Let us impose a Dirichlet process prior for F, with parameter v_0 and base probability measure Q_0 on $\mathbb{R}^N$. By Sethuraman (1994)'s decomposition, the Dirichet process prior induces a prior for (m, Σ) as: $m = \sum_{j=1}^{\infty} \alpha_j \xi_j$, and $\Sigma = \sum_{j=1}^{\infty} \alpha_j \xi_j \xi_j^T - \sum_{i=1}^{\infty} \alpha_i \xi_i \sum_{j=1}^{\infty} \alpha_j \xi_j^T$ where ξ_j are independently sampled from Q_0; $\alpha_j = u_j \prod_{l=1}^{j}(1-u_l)$ with $\{u_i\}_{i=1}^n$ drawn from Beta$(1, v_0)$. These priors then induce a prior for ϕ. The posterior distribution for (m, Σ) can be calculated explicitly:

$$\begin{aligned} \Sigma|D_n \quad \sim \quad & (1-\gamma) \sum_{j=1}^{\infty} \alpha_j \xi_j \xi_j^T + \gamma \sum_{t=1}^{n} \beta_t R_t R_t^n \\ & - \left((1-\gamma) \sum_{j=1}^{\infty} \alpha_j \xi_j + \gamma \sum_{t=1}^{n} \beta_t R_t \right) \left((1-\gamma) \sum_{j=1}^{\infty} \alpha_j \xi_j + \gamma \sum_{t=1}^{n} \beta_t R_t \right)^T, \end{aligned}$$

$$m|D_n \sim (1-\gamma) \sum_{j=1}^{\infty} \alpha_j \xi_j + \gamma \sum_{t=1}^{n} \beta_t R_t, \quad \gamma \sim \text{Beta}(n, v_0), \quad \{\beta_j\}_{j=1}^n \sim Dir(1, ..., 1).$$

We can set a truncation $K > 0$, so the infinite sums in the posterior representation are replaced with a truncated sum. We can then simulate the posterior for ϕ based on the distributions of $\Sigma|D_n$, $m|D_n$ and (10.2).

10.4 Simulation

We present a simple simulated example. The returns R_t are generated from a 2-factor model: $R_t = \Lambda f_t + u_t + 2\iota$, where Λ is a $N \times 2$ matrix of factor loadings. The error terms $\{u_{it}\}_{i \le N, t \le n}$ are i.i.d. uniform U$[-2, 2]$. Besides, the components of Λ are standard normal, and the factors are also uniform U$[-2, 2]$. The true $m = ER_t = 2\iota$, $\Sigma = \Lambda\Lambda^T + I_N$.

We set $N = 5, n = 200$. When the posterior is calculated, the DGP's distributions and the factor model structure are treated unknown, and we apply the nonparametric Dirichlet Process prior on the CDF of $R_t - m$, with parameter $v_0 = 3$, and based measure $Q_0 = \mathcal{N}(0, 1)$. We use a uniform prior for (σ^2, μ), and obtain the posterior distributions for $(m, \Sigma, \phi_1, \phi_2, \phi_3, \sigma^2, \mu)$. More concretely, the prior is assumed to be:

$$\pi(\sigma^2, \mu|\phi) = \pi(\sigma^2|\phi, \mu)\pi(\mu); \quad \pi(\sigma^2|\phi, \mu) \sim U[\sigma_\phi^2(\mu), \bar{\sigma}^2], \pi(\mu) \sim U[0, \bar{\mu}],$$

where μ and ϕ are a priori independent. We sample 1,000 draws from the posterior of (ϕ, σ^2, μ). Each time we first draw (m, Σ) from their marginal posterior distributions, based

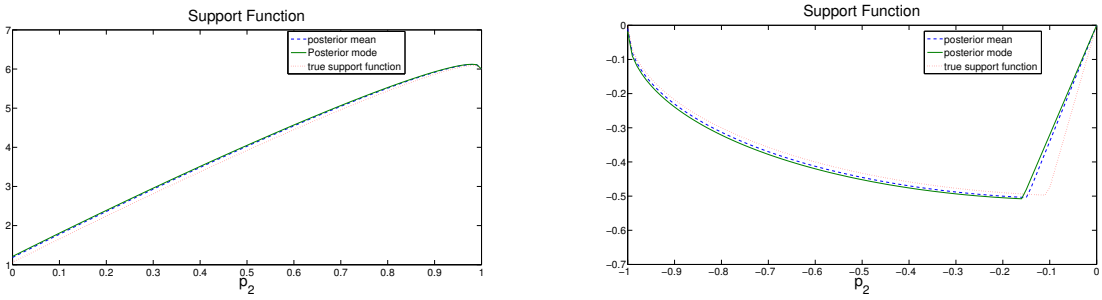
on which obtain the posterior draw of ϕ from (10.2). In addition, draw μ uniformly from $[0, \bar{\mu}]$, and finally σ^2 uniformly from $[\sigma^2_\phi(\mu), \bar{\sigma}^2]$, where $\sigma^2_\phi(\mu)$ is calculated based on the drawn ϕ and μ.

The posterior mean $(\hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3)$ of ϕ is calculated, based on which we calculate a Bayesian estimate of the boundary of the identified set (we set $\bar{\mu} = 1.4$ and $\bar{\sigma}^2 = 6$):

$$\partial\Theta(\hat{\phi}) = \{\mu \in [0, \bar{\mu}], \sigma^2 \in [0, \bar{\sigma}^2] : \sigma^2 = \hat{\phi}_1\mu^2 - 2\hat{\phi}_2\mu + \hat{\phi}_3\},$$

which is helpful to compute the BCS for the identified set. In addition, we estimate the support function $S_\phi(\nu)$ using the posterior mean of ϕ. In Figure 2, we plot the Bayesian estimates of the support function for two cases: $\nu_2 \in [0, 1]$, $\nu_1 = \sqrt{1 - \nu_2^2}$, and $\nu_2 \in [-1, 0]$, $\nu_1 = -\sqrt{1 - \nu_2^2}$.

Figure 2: Posterior estimates of support function. Left panel is for $\nu_2 \in [0, 1]$, $\nu_1 = \sqrt{1 - \nu_2^2}$; right panel is for $\nu_2 \in [-1, 0]$, $\nu_1 = -\sqrt{1 - \nu_2^2}$



Using the support function, we calculate the 95% posterior quantile q_τ for $J(\phi)$, based on which we construct the BCS $\Theta(\hat{\phi})^{q_\tau/\sqrt{n}}$ for the identified set. The boundary of $\Theta(\hat{\phi})^{q_\tau/\sqrt{n}}$ is given by

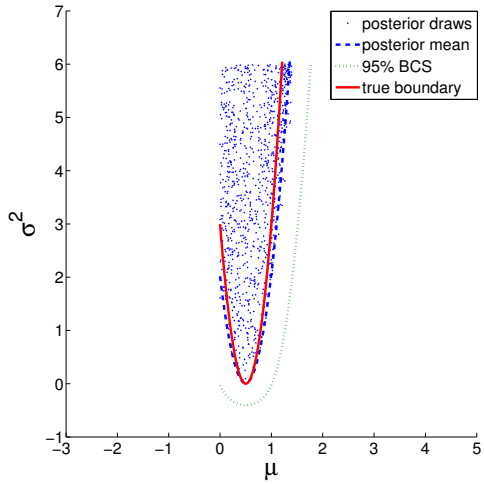
$$\partial\Theta(\hat{\phi})^{q_\tau/\sqrt{n}} = \left\{\mu \in [0, \bar{\mu}], \sigma^2 \in [0, \bar{\sigma}^2] : \inf_z \sqrt{|z - \mu|^2 + |\sigma^2_{\hat{\phi}}(z) - \sigma^2|^2} = q_\tau/\sqrt{n}\right\}.$$

In Figure 3, we plot the posterior draws of (μ, σ^2), $\partial\Theta(\hat{\phi})$, $\partial\Theta(\hat{\phi})^{q_\tau/\sqrt{n}}$ and the boundary of the true identified set. The scatter plot of posterior draws, the estimated boundaries and the BCS show that the true identified set is well estimated.

11 Conclusion

We propose a semi-parametric Bayesian procedure for inference about partially identified models. Bayesian approaches are appealing in many aspects. Classical Bayesian approach in this literature has been assuming a known likelihood function. However, in many applications econometric models only identify a set of moment inequalities, and therefore assuming

Figure 3: 1,000 posterior draws of (μ, σ^2). Solid line is the boundary of the true identified set; dashed line represents the estimated boundary using the posterior mean; dotted line gives the 95% BCS of the identified set. Plots are obtained based on a single set of simulated data. The BCS also covers a part of negative values for σ^2. In practice, we can truncate it to ensure it is always positive.



a known likelihood function suffers from the risk of misspecification, and may result in inconsistent estimations of the identified set. On the other hand, Bayesian approaches that use moment-condition-based likelihoods such as the limited information and exponential tilted empirical likelihood, though guarantee the consistency, lack of probabilistic interpretations, and do not provide an easy way to make inference about the identified set. Our approach, on the contrary, only requires a set of moment conditions but still possesses a pure Bayesian interpretation. Importantly, we can conveniently analyze both the partially identified parameter and its identified set. Moreover, we shed light on many appealing features of our proposed approach such as the computational efficiency for subset inference and projections.

Our analysis primarily focuses on identified sets which are closed and convex. These sets are completely characterized by their support function. By imposing a prior on the support function, we construct its posterior distribution. It is shown that the support function for a very general moment inequality model admits a linear expansion, and its posterior is consistent. The Bernstein-von Mises theorem for the posterior of the support function is proven.

Some researcher may argue that frequentist and Bayesian approaches are asymptotically equivalent because the prior is "washed away" as the sample size increases. So why bother with Bayesian asymptotics? Interestingly, this is no longer the case under partial identification. As was shown by Moon and Schorfheide (2012), when making inference about the partially identified parameter, the BCS can be strictly smaller than the FCS. Therefore, the two fundamental statistical methodologies provide different inferences that are not

necessarily equivalent even asymptotically. This paper completes such a comparison. We also establish the large-sample correspondence of the two approaches for the identified set. It is also illustrated that the results hold in the uniform sense, in particular when point identification is nearly achieved.

Appendix

A Semi-parametric prior

We present in this section a prior scheme for (ϕ, F) which is an alternative to the prior scheme presented in section 3.2. The idea is to reformulate the sampling distribution F in terms of ϕ and a nuisance parameter $\eta \in \mathcal{P}$, where $\mathcal{P}$ is an infinite-dimensional measurable space. Therefore, $F = F_{\phi,\eta} \in \mathcal{F} = \{F_{\phi,\eta}; \phi \in \Phi,\ \eta \in \mathcal{P}\}$. We denote by $l_n(\phi, \eta)$ the model's likelihood function. One of the appealing features of this semi-parametric prior is that it allows us to impose a prior $\pi(\phi)$ directly on the identified parameter ϕ, which is convenient whenever we have good prior information regarding ϕ.

For example, in the interval censored data example (Example 3.1), we can write

$$Y_1 = \phi_1 + u, \quad Y_2 = \phi_2 + v$$

where (u, v) are independent random variables with zero means and have unknown densities (η_1, η_2). Then $\eta = (\eta_1, \eta_2)$, and the likelihood function is

$$l_n(\phi, \eta) = \prod_{i=1}^{n} \eta_1(Y_{1i} - \phi_1)\eta_2(Y_{2i} - \phi_2).$$

We put priors on (ϕ, η). Examples of priors on density functions η_1 and η_2 include mixture of Dirichlet process priors, Gaussian process priors, etc. (see Ghosal et al. (1999) and Amewou-Atisso et al. (2003)). We list some commonly used priors for densities in the examples below.

We place an independent prior as $\pi(\phi, \eta) = \pi(\phi)\pi(\eta)$. Then the joint prior distribution $\pi(\theta, \phi, \eta)$ is naturally decomposed as $\pi(\theta, \phi, \eta) = \pi(\theta|\phi)\pi(\phi)\pi(\eta)$ and the Bayesian experiment is

$$X|\phi, \eta \sim F_{\phi,\eta}, \qquad (\phi, \eta) \sim \pi(\phi)\pi(\eta), \qquad \theta|\phi, \eta \sim \pi(\theta|\phi).$$

The posterior distribution of ϕ has a density function given by

$$p(\phi|D_n) \propto \int_{\mathcal{P}} \pi(\phi, \eta) l_n(\phi, \eta) d\eta. \tag{A.1}$$

Then the marginal posterior of θ is, for any measurable set $B \subset \Theta$:

$$P(\theta \in B|D_n) \propto \int_{\Phi} \pi(\theta \in B|\phi) p(\phi|D_n) d\phi = E[\pi(\theta \in B|\phi)|D_n]$$

where the conditional expectation is taken with respect to the posterior of ϕ. Moreover, the posteriors of $\Theta(\phi)$ and $S_\phi(\cdot)$ are deduced from that of ϕ. Suppose for example we are interested in whether $\Theta(\phi) \cap A$ is an empty set for some $A \subset \Theta$, then the posterior probability $P(\Theta(\phi) \cap A = \emptyset|D_n)$ is relevant.

Example A.1 (Interval regression model - *continued*)**.** Consider Example 3.2, where $\phi = (\phi_1, \phi_2, \phi_3) = (E(ZY_1), E(ZX^T), E(ZY_2))$. Write $ZY_1 = \phi_1 + u_1$, $ZY_2 = \phi_3 + u_3$, and $\text{vec}(ZX^T) = \text{vec}(\phi_2) + u_2$, where u_1, u_2 and u_3 are correlated and their joint unknown probability density function is $\eta(u_1, u_2, u_3)$. The likelihood function is then

$$l_n(\phi, \eta) = \prod_{i=1}^{n} \eta(Z_i Y_{1i} - \phi_1, Z_i Y_{2i} - \phi_3, \text{vec}(Z_i X_i^T) - \text{vec}(\phi_2)).$$

Many nonparametric priors can be used for $\pi(\eta)$ in a "location-model" of this type, where the likelihood takes the form $l_n(\phi, \eta) = \prod_{i=1}^{n} \eta(X_i - \phi)$, including Dirichlet mixture of normals (Ghosal et al. 1999), random Bernstein polynomials (Walker et al. 2007), and finite mixture of normals (Lindsay and Basak 1993).

We provide some examples of prior for the density η.

Example A.2. The finite mixture of normals (e.g., Lindsay and Basak (1993), Ray and Lindsay (2005)) assumes η to take the form

$$\eta(x) = \sum_{i=1}^{k} w_i h(x; \mu_i, \Sigma_i),$$

where $h(x; \mu_i, \Sigma_i)$ is the density of a multivariate normal distribution with mean μ_i and variance Σ_i and $\{w_i\}_{i=1}^k$ are unknown weights such that $\sum_{i=1}^k w_i \mu_i = 0$. Then $\int \eta(x) x dx = \sum_{i=1}^k w_i \int h(x; \mu_i, \Sigma_i) x dx = 0$. We impose prior $\pi(\eta) = \pi(\{\mu_l, \Sigma_l, w_l\}_{l=1}^k)$, then

$$p(\phi|D_n) \propto \int \pi(\phi) \prod_{i=1}^{n} \sum_{j=1}^{k} w_j h(X_i - \phi; \mu_j, \Sigma_j) \pi(\{\mu_l, \Sigma_l, w_l\}_{l=1}^k) dw_j d\mu_j d\Sigma_j.$$

Example A.3. Dirichlet mixture of normals (e.g., Ghosal *et al.* (1999) Ghosal and van der Vaart (2001), Amewou-Atisso, et al. (2003)) assumes

$$\eta(x) = \int h(x - z; 0, \Sigma) dH(z)$$

where $h(x; 0, \Sigma)$ is the density of a multivariate normal distribution with mean zero and variance Σ and H is a probability distribution such that $\int z H(z) dz = 0$. Then $\int x \eta(x) dx = 0$. To place a prior on η, we let H have the Dirichlet process prior distribution $D_\alpha \equiv \mathcal{D}(\nu_0, Q_0)$ where α is a finite positive measure, $\nu_0 = \alpha(\mathcal{X}) \in \mathbb{R}_+$ and $Q_0 = \alpha/\alpha(\mathcal{X})$ is a base probability on $(\mathcal{X}, \mathcal{B}_x)$ such that $Q_0(x) = 0, \forall x \in (\mathcal{X}, \mathcal{B}_x)$. In addition, we place a prior on Σ independent of the prior on H. Then

$$p(\phi|D_n) \propto \int \pi(\phi) \pi(\Sigma) D_\alpha(H) \prod_{i=1}^{n} \int h(X_i - \phi - z; 0, \Sigma) dH(z) d\Sigma dH.$$

Example A.4. Random Bernstein polynomials (e.g., Walker et al. (2007) and Ghosal (2001)) allow to write the density function η as

$$\eta(x) = \sum_{j=1}^{k} [H(j/k) - H((j-1)/k)] \mathcal{B}e(x; j, k-j+1),$$

where $\mathcal{B}e(x; a, b)$ stands for the beta density with parameters $a, b > 0$ and H is a random distribution function with prior distribution assumed to be a Dirichlet process. Moreover, the parameter k is also random with a prior distribution independent of the prior on H. Then $p(\phi|D_n) \propto \int \pi(\phi) \prod_{i=1}^{n} \eta(X_i - \phi)\pi(H)\pi(k)dHdk$. □

Other commonly used priors on density functions are wavelet expansions (Rivoirard and Rousseau (2012)), Polya tree priors (Lavine (1992)), Gaussian process priors (van der Vaart and van Zanten (2008), Castillo (2008)), etc.

B Posterior Concentration for ϕ

This section focuses on the case where the prior for ϕ is specified through the semi-parametric prior described in Appendix A

Much of the literature on posterior concentration rate for Bayesian non-parametrics relies on the notion of *entropy cover number*, which we now define as follows. Recall that for i.i.d. data, the likelihood function can be written as $l_n(\phi, \eta) = \prod_{i=1}^{n} l(X_i; \phi, \eta)$, where $l(x; \phi, \eta)$ denotes the density of the sampling distribution. Let

$$G = \{l(\cdot; \phi, \eta) : \phi \in \Phi, \eta \in \mathcal{P}\}$$

be the family of likelihood functions. We assume $\mathcal{P}$ is a metric space with a norm $\|.\|_\eta$, which then induces a norm $\|.\|_G$ on G such that $\forall l(\cdot; \phi, \eta) \in G$,

$$\|l(\cdot; \phi, \eta)\|_G = \|\phi\| + \|\eta\|_\eta.$$

For instance, in the examples of intervel censoring data and interval regression, $l(x; \phi, \eta) = \eta(x - \phi)$ and $\|\eta\|_\eta = \|\eta\|_1 = \int |\eta(x)|dx$. Then in this case $\|l(., \phi, \eta)\|_G = \|\phi\| + \|\eta\|_1$.

Define the entropy cover number $\mathcal{N}(\rho, G, \|.\|_G)$ to be the minimum number of balls with radius ρ needed to cover G. The importance of the entropy cover number on nonparametric Bayesian asymptotics has been realized for a long time. We refer the audience to Kolmogorov and Tikhomirov (1961) and van der Vaart and Wellner (1996) for good early references.

We first present the assumptions that are sufficient to derive the posterior concentration rate for the point identified ϕ. The first one is placed on the entropy cover number.

Assumption B.1. *Suppose for all n large enough,*

$$\mathcal{N}(n^{-1/2}(\log n)^{1/2}, G, \|.\|_G) \leq n.$$

This condition requires that the "model" G be not too big. Once this condition holds, then for all $r_n \geq n^{-1/2}(\log n)^{1/2}$, $\mathcal{N}(r_n, G, \|.\|_G) \leq \exp(nr_n^2)$. Moreover, it ensures the existence of certain tests as given in Lemma I.2 in the supplementary appendix, and hence it can be replaced by the test condition that are commonly used in the literature of posterior concentration rate, e.g., Jiang (2007), Ghosh and Ramamoorthi (2003). The same condition has been imposed by Ghosal et al. (2000) when considering Hellinger rates, and Bickel and Kleijn (2012) when considering semi-parametric posterior asymptotic normality, among others. When the true density η_0 belongs to the family of location mixtures, this condition was verified by Ghosal et al. (1999, Theorem 3.1).

The next assumption places conditions on the prior for (ϕ, η). For each (ϕ, η), define

$$K_{\phi,\eta} = E\left[\log \frac{l(X;\phi_0,\eta_0)}{l(X;\phi,\eta)}\middle|\phi_0,\eta_0\right] = \int \log\left(\frac{l(x;\phi_0,\eta_0)}{l(x;\phi,\eta)}\right) l(x;\phi_0,\eta_0)dx$$

$$V_{\phi,\eta} = \text{var}\left[\log \frac{l(X;\phi_0,\eta_0)}{l(X;\phi,\eta)}\middle|\phi_0,\eta_0\right] = \int \log^2\left(\frac{l(x;\phi_0,\eta_0)}{l(x;\phi,\eta)}\right) l(x;\phi_0,\eta_0)dx - K_{\phi,\eta}^2.$$

Assumption B.2. *The prior $\pi(\phi, \eta)$ satisfies:*

$$\pi\left(K_{\phi,\eta} \leq \frac{\log n}{n}, \quad V_{\phi,\eta} \leq \frac{\log n}{n}\right) n^M \to \infty$$

for some $M > 2$.

Intuitively, when (ϕ, η) is close to (ϕ_0, η_0), both $K_{\phi,\eta}$ and $V_{\phi,\eta}$ are close to zero. Hence this assumption requires that the prior have sufficient amount of support around the true point identified parameters in terms of the Kullback-Leibler distance. Such a prior condition has also been commonly imposed in the literature on semi-parametric posterior concentration, e.g., Ghosal et al. (1999 (2.10), 2000 condition 2.4), Shen and Wasserman (2001, Theorem 2) and Bickel and Kleijn (2012, (3.13)). Moreover, it has been verified in the literature using the sieve prior (Shen and Wasserman 2001), Dirichlet mixture prior (Ghosal et al. 1999) and Normal mixture prior (Ghosal and van der Vaart 2007).

We are now ready to present the posterior concentration rate for ϕ, whose proof is given in Appendix I of the supplementary appendix.

Theorem B.1. *Suppose the data $X_1, ..., X_n$ are i.i.d. Under Assumptions B.1 and B.2, for some $C > 0$,*

$$P(\|\phi - \phi_0\| \leq Cn^{-1/2}(\log n)^{1/2}|D_n) \to^p 1.$$

All the proofs are given in the supplementary appendix.

References

Amewou-Atisso, M., Ghosal, S. and Ghosh, J. (2003). Posterior consistency for semiparametric regression problems. *Bernoulli*, **9**, 291-312.

Andrews, D. and Guggenberger, P. (2009). Validity of subsampling and plug-in asymptotic inference for parameters defined by moment inequalities. *Econometric Theory*, **25**, 669-709.

Andrews, D. and Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, **81**, 609-666.

Andrews, D. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, **78**, 119-157.

Beresteanu, A., Molchanov, I. and Molinari, F. (2011). Sharp identification regions in models with convex moment predictions. *Econometrica*, **79**, 1785-1821.

Beresteanu, A., Molchanov, I. and Molinari, F. (2012). Partial identification using random set theory. *Journal of Econometrics*, **166**, 17-32.

Beresteanu, A. and Molinari, F. (2008). Asymptotic properties for a class of partially identified models. *Econometrica*, **76**, 763-814.

Bickel, P. and Kleijn, B. (2012). The semiparametric Bernstein-Von Mises theorem. *Annals of Statistics*, **40**, 206-237.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch, Verw. Gebiete*, **65**, 181-237.

Bollinger, C. and Hasselt, M. (2009). A Bayesian analysis of binary misclassification: inference in partially identified models. *Manuscript*. University of Kentucky.

Bontemps, C., Magnac, T. and Maurin, E. (2012). Set identified linear models. *Econometrica*, **80**, 1129-1155.

Bugni, F. (2010). Bootstrap inference in partially identified models defined by moment inequalities: coverage of the identified set. *Econometrica*, **78**, 735-753.

Canay, I. (2010). EL Inference for partially identified models: large deviations optimality and bootstrap validity. *Journal of Econometrics*, **156**, 408-425.

Castillo, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics*, **2**, 1281-1299.

Chandrasekhar, A., Chernozhukov, V., Molinari, F. and Schrimpf, P. (2012). Inference for best linear approximations to set identified functions. *Cemmap Working Papers*, CWP43/12.

Chernozhukov, V., Kocatulum, E. and Menzel, K. (2012). Inference on sets in finance. *Cemmap Working Paper*, CWP46/12.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, **115**, 293-346.

Chernozhukov, V., Hong, H. and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, **75**, 1243-1284.

Chiburis, R. (2009). Approximately most powerful tests for moment inequalities. *Manuscript.* UT Austin.

Choi, I. and Ramamoorthi, R. (2008). Remarks on consistency of posterior distributions. *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, **3**, 170-186

Epstein, L. and Seo, K. (2011). Bayesian inference and non-Bayesian prediction and choice: foundations and an application to entry games with multiple equilibria. *Manuscript*. Boston Univ.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, No.2, 209-230.

Florens, J.P. and Simoni, A. (2011). Bayesian identification and partial identification. *Manuscript*.

Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Annals of Statistics*, **27**, 1119-1141.

Gelfand, A. and Sahu, S. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, **94**, 247-253.

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Annals of Statistics*, **29**, 1264-1280.

Ghosal, S., Ghosh, J. and Ramamoorthi, R. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, **27**, 143-158.

Ghosal, S., Ghosh, J. and van der Vaart, A. (1999). Consistent semiparametric Bayesian inference about a location parameter. *Journal of Statistical Planning and Inference*, **77**, 181-193.

Ghosal, S., Ghosh, J. and van der Vaart, A. (2000) Convergence rates of posterior distributions. *Annals of Statistics*. **28**, 500-531.

Ghosal, S. and van der Vaart, A. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statistics*, **29**, 1233-1263.

Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for noniid obsevations. *Annals of Statistics*, **35**, 192-223.

Ghosh, J.K. and Ramamoorthi, R.V.(2003). *Bayesian Nonparametrics.* Springer-Verlag.

Gustafson, P. (2012). On the behaviour of Bayesian credible intervals in partially identified models. *Electronic Journal of Statistics*, **6**, 2107-2124.

Haile, P. and Tamer, E. (2003). Inference with an incomplete model of English auctions. *Journal of Political Economy*, **111**, 1-51.

Hansen, L. and Jagannathan, R. (1991). Implications of security market data for models of dynamic economies. *Journal of Political Economy*, **99**, 225-262.

Imbens, G. and Manski, C. (2004). Confidence Intervals for partially identified parameters. *Econometrica*, **72**, 1845-1857.

Ishwaran, H. and James, L. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *J. Comp. Graph. Statist.* **11**, 508-532.

Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. *Annals of Statistics*, **35**, 1487-1511.

Kaido, H. and Santos, A. (2013). Asymptotically efficient estimation of models defined by convex moment inequalities, *Econometrica*, forthcoming.

Kaido, H. (2012). A dual approach to inference for partially Identified econometric models. *Manuscript.*

Kim, J. (2002). Limited information likelihood and Bayesian analysis. *Journal of Econometrics*, **107**, 175-193.

Kitagawa, T. (2012). Estimation and inference for set-identified parameters using posterior lower probability. *Manuscript*, UCL.

Kline, B. (2011). The Bayesian and frequentist approaches to testing a one-sided hypothesis about a multivariate mean. *Journal of Statistical Planning and Inference*, **141**, 3131-3141.

Kolmogorov, A. and Tikhomirov, V. (1961). Epsilon-entropy and epsilon-capacity of sets in function spaces. *Amer. Math. Soc. Trans. Ser.* **17**, 277-364.

Lavine, M. (1992). Some aspects of polya tree distributions for statistical modelling. *Annals of Statistics*, **20**, No.3, 1222-1235.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.

Leahu, H. (2011). On the Bernstein-von Mises phenomenon in the Gaussian white noise. *Electronic Journal of Statistics*, **5**, 373-404.

Liao, Y. and Jiang, W. (2010). Bayesian analysis in moment inequality models. *Annals of Statistics*, **38**, 275-316.

Lindsay, B. and Basak, P. (1993). Multivariate normal mixtures: a fast consistent method of moments. *Journal of American Statistical Association*, **88**, 468-476.

Manski, C. and Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica*, **70**, 519-547.

Manski, C. (2003). *Partial identification of probability distributions.* Spinger, New York.

Menzel, K. (2011). Consistent estimation with many moment inequalities. *Manuscript.* NYU.

Milgrom, P. and Segal, I. (2002). Envelope theorems for arbitrary choice sets. *Econometrica*, **70**, 583-601.

Moon, H. R. and Schorfheide, F. (2012). Bayesian and frequentist inference in partially-identified models. *Econometrica*, **80**, 755-782.

Neath, A. and Samaniego, F. (1997). On the efficacy of Bayesian inference for nonidentifiable models. *The American Statistician*, **51**, 325-332.

Norets, A. and Tang, X. (2012). Semiparametric Inference in dynamic binary choice models. *Manuscript.* Princeton Univ.

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

Ray, S. and Lindsay, B. (2005). The topography of multivariate normal mixtures. *Annals of Statistics*, **33**, 2042-2065.

Rivoirard, V. and Rousseau, J. (2012). Bernstein-von Mises theorem for linear functionals of the density. *Annals of Statistics*, **40**, 1489-1523.

Rockafellar, R.T. (1970). *Convex analysis. Princeton University Press.*

Romano, J. and Shaikh, A. (2010). Inference for the identified set in partially identified econometric models. *Econometrica*, **78**, 169-211.

Rosen, A. (2008). Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities. *Journal of Econometrics*, **146**, 107-117.

Schennach, S. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika*, **92**, 31-46.

Sethuraman, J. (1994). A constructive definition of the Dirichlet prior. *Statistica Sinica*, **2**, 639-650.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Annals of Statistics*, **29**, 687-714.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica*, **77**, 1299-1315,

Stoye, J. (2012). New perspectives on statistical decisions under ambiguity. *Annual Review of Economics*, **4**, 257-282.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, **2**, 167-195.

van der Vaart, A.W. (2002). Semiparametric Statistics. In *Ecole d'Eté de St Flour 1999* (eds. P. Bernard), 331-457. Springer Verlag, New York.

van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes.* Springer.

van der Vaart, A.W. and J.H., van Zanten (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Annals of Statistics*, **36**, No.3, 1435-1463.

Walker, S., Lijoi, A. and Prunster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Annals of Statistics*, **35**, 738-746.

Wan, Y. (2011). An integration-based approach to moment inequality models. *Manuscript.* Univ. of Toronto.

Wu, Y. and Ghosal, S. (2008). Posterior consistency for some semi-parametric problems. *Sankhya: the Indian Journal of Statistics*, **70**, 0-46.