Federal Reserve Bank of Philadelphia

Working Papers
Research Department

WP 19-51
December 2019
https://doi.org/10.21799/frbp.wp.2019.51

Bayesian Estimation and Comparison of Conditional Moment Models

Siddhartha Chib
Olin Business School, Washington University in St. Louis

Minchul Shin
Federal Reserve Bank of Philadelphia Research Department

Anna Simoni
CREST, CNRS, Ecole Polytechnique

ISSN: 1962-5361

Bayesian Estimation and Comparison of Conditional Moment Models*

Siddhartha Chib† Minchul Shin‡ Anna Simoni§

December 2019

Abstract

We provide a Bayesian analysis of models in which the unknown distribution of the outcomes is specified up to a set of conditional moment restrictions. This analysis is based on the nonparametric exponentially tilted empirical likelihood (ETEL) function, which is constructed to satisfy a sequence of unconditional moments, obtained from the conditional moments by an increasing (in sample size) vector of approximating functions (such as tensor splines based on the splines of each conditioning variable). The posterior distribution is shown to satisfy the Bernstein-von Mises theorem, subject to a growth rate condition on the number of approximating functions, even under misspecification of the conditional moments. A large-sample theory for comparing different conditional moment models is also developed. The central result is that the marginal likelihood criterion selects the model that is less misspecified, that is, the model that is closer to the unknown true distribution in terms of the Kullback-Leibler divergence. Several examples are provided to illustrate the framework and results.

Keywords: Bayesian inference, Bernstein-von Mises theorem, Conditional moment restrictions, Exponentially tilted empirical likelihood, Marginal likelihood, Misspecification, Posterior consistency.

JEL codes: C11, C14, C13, C52

*__Disclaimer:__ This Philadelphia Fed working paper represents preliminary research that is being circulated for discussion purposes. The views expressed in these papers are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. Any errors or omissions are the responsibility of the authors. Philadelphia Fed working papers are free to download at https://philadelphiafed.org/research-and-data//publications/working-papers.

†Olin Business School, Washington University in St. Louis, Campus Box 1133, 1 Brookings Drive, St. Louis, MO 63130. e-mail: chib@wustl.edu.

‡Federal Reserve Bank of Philadelphia, Ten Independence Mall, Philadelphia, PA 19106. e-mail: visiblehand@gmail.com.

§CREST, CNRS, École Polytechnique, ENSAE, 5, Avenue Henry Le Chatelier, 91120 Palaiseau - France, e-mail: simoni.anna@gmail.com.

1 Introduction

We tackle the problem of prior-posterior inference in models where the only available information about the unknown parameter $\theta \in \Theta \subset \mathbb{R}^p$ is supplied in the form of a set of *conditional* moment restrictions

$$\mathbf{E}^P[\rho(X, \theta)|Z] = 0, \tag{1.1}$$

where $\rho(X, \theta)$ is a d-vector of known functions of a $\mathbb{R}^{d_x}$-valued random vector X and the unknown θ, and P is the unknown conditional distribution of X given a $\mathbb{R}^{d_z}$-valued random vector Z. We suppose that $d \geq p$, letting the number of conditional moments exceed the number of parameters. This is a conditional moment restricted model because the moments constrain the set of possible distributions P. We say that the model is correctly specified if the true data generating process P_* is in the set of distributions constrained to satisfy these moment conditions for some $\theta \in \Theta$, while the model is misspecified if P_* is not in the set of implied distributions for any $\theta \in \Theta$.

A different starting point is that with unconditional moments, say $\mathbf{E}^P[g(X, \theta)] = 0$. Such models have recently attracted interest in the Bayesian community as distributional assumptions are entirely bypassed. Prior-posterior analysis is based on the empirical likelihood, for example, Lazar (2003) and many others, or the exponentially tilted empirical likelihood (ETEL), as in Schennach (2005) and Chib, Shin and Simoni (2018). The latter paper provides the large sample theory under misspecification, and introduces the use of marginal likelihoods for comparing unconditional moment models, including the relevant framework for comparing such models and the large-sample model consistency of the marginal likelihoods.

Although the conditional moments imply that the functions $\rho(X, \theta)$ are uncorrelated with Z, i.e., $\mathbf{E}^P[\rho(X, \theta) \otimes Z] = 0$, where $\otimes$ is the Kronecker product operator, it is inappropriate to use these unconditional restrictions alone as a substitute for the conditional moments. This is because the conditional moments assert even more, that $\rho(X, \theta)$ is uncorrelated with any measurable, bounded function of Z, or if Z is square-integrable, that it is uncorrelated with *any* L^2-measurable function of Z. Thus, in order to assemble the set of unconditional moments that are equivalent to the conditional moments we must consider all such functions, which is only feasible as the sample size n goes to infinity. A general result due to Donald, Imbens and Newey (2003) states that this equivalent set of unconditional moments can be constructed by approximating functions $q^K(Z) = (q_1^K(Z), \ldots, q_K^K(Z))'$, such as tensor product splines obtained from splines of each variable in Z, with the number of such functions, denoted by K, increasing

with n. Then, instead of (1.1), inference based on the expanded unconditional moment conditions

$$\mathbf{E}^P[\rho(X,\theta)\otimes q^K(Z)]=0 \tag{1.2}$$

is valid. This is then how we proceed in this paper.

The transformation into unconditional moments introduces, however, some challenges for the prior-posterior analysis that are different from those addressed in Chib, Shin and Simoni (2018). For one, quantities that are bounded with fixed moment restrictions, now diverge with K. On determining the rate of this divergence (and thus stabilizing the growth), we show that under correct specification of the conditional moments, the posterior distribution of θ satisfies the Bernstein-von Mises (BvM) theorem with asymptotic posterior variance equal to the semiparametric efficiency bound derived in Chamberlain (1987). As a result, Bayesian credible sets are asymptotically valid efficient confidence sets. Conversely, sets based on unconditional versions of the conditional moments with a fixed K, in general, are not optimal.

Second, we consider misspecified conditional moment models, which occur widely in practice, and establish a similar BvM-type phenomenon. Just as in Kleijn and van der Vaart (2012), our theorem establishes that the posterior distribution of the centered and scaled parameter $\sqrt{n}(\theta-\theta_\circ)$, where $\theta_\circ$ is the pseudo-true value, converges to a Normal distribution with a random mean that is bounded in probability. Again, the proof of this result, which requires fixing conditions under which the ETEL function satisfies a stochastic LAN expansion with increasing moments, is substantially more complicated than in the unconditional moment case.

Third, we develop a large sample theory for comparing competing conditional moment models by marginal likelihoods. Despite some similarities, the model comparison framework here is different from the one in our previous work. For one, the comparison of these models does not require that the models are reformulated to have the same number of conditional moments. The models can be compared directly as stated in terms of marginal likelihoods. We show that, under regularity conditions which are different than in the unconditional case, the model picked by the marginal likelihood, in the limit, is the model that is less misspecified. This is also the model that is closest to the true distribution in the Kullback-Leibler divergence.

Conditional moments often supply the only source of information about P. Examples of this include causal inference, as in Rosenbaum and Rubin (1983), where one assumes that the potential outcomes are independent of the treatment variable conditioned on covariates, and in missing at random problems as

considered by Hristache and Patilea (2017), and numerous others. Our analysis makes possible Bayesian inference for a large class of models that appeared to be outside the scope of the Bayesian paradigm. In addition, our results here extend and complete the work of Chib, Shin and Simoni (2018) on unconditional moments. Our treatment also complements the papers on conditional moment models from a Bayesian viewpoint that are based on non-ETEL approaches and reflect other concerns. For instance, Liao and Jiang (2011), Florens and Simoni (2012, 2016), Kato (2013), Chen, Christensen and Tamer (2018) and Liao and Simoni (2019) allow the moment function to contain a non-parametric component that is estimated by a sieve type approximation method, or a Gaussian process prior, permitting the possibility that the non-parametric component is only partially identified, but within a quasi-Bayesian formulation based on a pseudo-likelihood. None of these papers tackles the problem of misspecification, or the problem of model comparisons.

The rest of the paper is organized as follows. In Section 2 we sketch the conditional moment setting more formally and provide a motivating example. In Section 3 we describe the construction of the sequence of unconditional moments by an increasing (in sample size) vector of approximating functions. We then supply results on the large sample behavior of the posterior distribution in both the correct and misspecified moment models. In Section 4 we turn to the problem of model comparisons and determine the large sample behavior of the marginal likelihood. In Section 5 we provide an application of our techniques to a causal inference problem. Section 6 concludes. Technical proofs of the theorems are included in the supplementary appendix.

2 Setting and Motivation

Let $X := (X_1', X_z')'$ be an $\mathbb{R}^{d_x}$-valued random vector and $Z := (Z_1', X_z')'$ be an $\mathbb{R}^{d_z}$-valued random vector. The vectors Z and X have elements in common if the dimension of the subvector X_z is non-zero. Moreover, we denote $W := (X', Z_1')' \in \mathbb{R}^{d_w}$ and its (unknown) joint distribution by P. By abuse of notation we use P also to denote the associated conditional distribution. We suppose that we are given a random sample $W_{1:n} = (W_1, \ldots, W_n)$ of W. Hereafter, we denote by $\mathbf{E}^P[\cdot]$ the expectation with respect to P and by $\mathbf{E}^P[\cdot|\cdot]$ the conditional expectation with respect to the conditional distribution associated with P.

The parameter of interest is $\theta \in \Theta \subset \mathbb{R}^p$, which is related to the conditional distribution P through

the conditional moment restrictions

$$\mathbf{E}^P[\rho(X,\theta)|Z] = 0, \tag{2.1}$$

where $\rho(X,\theta)$ is a d-vector of known functions. Many interesting and important models in statistics fall into this framework.

Example 1 *(Linear model with heteroscedastic error) Suppose that*

$$\mathbf{E}^P[(Y - \theta_0 - \theta_1 X)|Z] = 0, \tag{2.2}$$

where $\rho(X,\theta) = (Y - \theta_0 - \theta_1 X)$, $Z = (1, X)$ *and* $d = 1$. *This conditional moment restriction model is consistent with the data generating process (DGP)* $Y = \theta_0 + \theta_1 X + \varepsilon$, *where* $\varepsilon = h(X)U$, *and* (X, U) *follow some unknown distribution* P, *with* $E(U) = 0$, *and the function* $h(X)$, *the heteroscedasticity function, is unknown. If we specify the restrictions*

$$\mathbf{E}^P[(Y - \theta_0 - \theta_1 X)\,|Z] = 0 \quad \textit{and} \quad \mathbf{E}^P[(Y - \theta_0 - \theta_1 X)^3|Z] = 0, \tag{2.3}$$

where now $\rho(X,\theta)$ *is a* (2×1) *vector of functions, we additonally impose conditional symmetry of* ε.

Of course, the conditional moment model is different from the unconditional moment model. For example, in Example 1, the unconditional moment conditions

$$\mathbf{E}^P[(Y - \theta_0 - \theta_1 X) \otimes (1, X)'] = 0, \tag{2.4}$$

which impose the assumptions that ε has mean zero and is uncorrelated with X, are weaker but, if the conditional model is correct, less informative about θ.

3 Prior-Posterior Analysis

3.1 Expanded Moment Conditions

One way to estimate the conditional moment model is by estimating the conditional expectation directly, as in the frequentist approach of Kitamura, Tripathi and Ahn (2004). This approach does not seem to generalize easily, if at all, to the Bayesian setting. An alternative approach, that we adopt, is based on recognizing that the conditional moments in (2.1) are a functional equation and, therefore, constitute a continuum of unconditional moment conditions. Under certain circumstances, see (Bierens, 1982, Chamberlain, 1987), a countable number of unconditional moment restrictions that are equivalent to the

conditional moment restrictions in (2.1) is guaranteed. This is the basis of the frequentist approaches in Donald and Newey (2001), Ai and Chen (2003) and Carrasco and Florens (2000) where, after transforming the conditional moment restrictions into unconditional moment restrictions, the resulting set of unconditional moments are analyzed under a sieve approach or a Tikhonov regularization approach. Following Donald, Imbens and Newey (2003), the equivalent set of unconditional moments can be obtained through approximating functions.

Let $q^K(Z) = (q_1^K(Z), \ldots, q_K^K(Z))'$, $K > 0$, denote a K-vector of real-valued functions of Z, for instance, splines, truncated power series, or Fourier series. Suppose that these functions satisfy the following condition for the distribution P.

Assumption 3.1 *For all K, $\mathbf{E}^P[q^K(Z)'q^K(Z)]$ is finite, and for any function $a(z) : \mathbb{R}^{d_z} \to \mathbb{R}$ with $\mathbf{E}^P[a(Z)^2] < \infty$ there are $K \times 1$ vectors γ_K such that as $K \to \infty$,*

$$\mathbf{E}^P[(a(Z) - q^K(Z)'\gamma_K)^2] \to 0.$$

Now if $\mathbf{E}^P[\rho(X,\theta)'\rho(X,\theta)] < \infty$, then Donald, Imbens and Newey (2003, Lemma 2.1) established that: (1) if equation (2.1) is satisfied with $\theta = \theta_*$, then $\mathbf{E}^P[\rho(X,\theta_*) \otimes q^K(Z)] = 0$ for all K; (2) if equation (2.1) is not satisfied, then $\mathbf{E}^P[\rho(X,\theta_*) \otimes q^K(Z)] \neq 0$, for all large enough K.

Thus, under the stated assumptions, the conditional moment restrictions are equivalent to the limit of a sequence of unconditional moment restrictions

$$\mathbf{E}^P[g(W,\theta)] = 0, \tag{3.1}$$

where $g(W,\theta) := \rho(X,\theta) \otimes q^K(Z)$, with $K \to \infty$, are the *expanded functions*.

Example 1 (continued) *Let $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ denote the sample data, and $(\tau_1, \ldots, \tau_K)$ the K knots, where the exterior knots τ_1 and τ_K are the minimum and maximum values of $\boldsymbol{x}$, and the interior knots are specified quantile points of $\boldsymbol{x}$. Let $q^K(x) = (q_1(x), \ldots, q_K(x))'$ denote (say) K natural cubic spline basis functions, where $q_j(x)$ is the cubic spline basis function located at τ_j. Let B denote the $(n \times K)$ matrix of these basis functions evaluated at $\boldsymbol{x}$, where the ith row of B is given by $q^K(x_i)'$. Let $(\boldsymbol{y} - \theta_0 - \theta_1\boldsymbol{x})$ and $(\boldsymbol{y} - \theta_0 - \theta_1\boldsymbol{x})^3$ each denote $n \times 1$ vectors where $\boldsymbol{y} = (y_1, \ldots, y_n)$. Then, the expanded functions $g(\boldsymbol{W},\theta) = g(\boldsymbol{x},\theta)$ for the sample observations are the $n \times 2K$ functions*

$$g(\boldsymbol{W},\theta) = [\rho(\boldsymbol{x},\theta) \otimes q^K(x)] = \big[(\boldsymbol{y} - \theta_0 - \theta_1\boldsymbol{x}) \odot B \vdots (\boldsymbol{y} - \theta_0 - \theta_1\boldsymbol{x})^3 \odot B\big], \tag{3.2}$$

where $a \odot B$ denotes the Hadamard product, and $\vdots$ denotes matrix concatenation (column binding).

In our numerical examples we use the natural cubic spline basis of Chib and Greenberg (2010) based on Z to construct $q^K(Z)$, with K fixed at a given value, as in sieve estimation. If Z consists of more than one element, say (Z_1, Z_2, Z_3) where Z_1 and Z_2 are continuous variables and Z_3 is binary, then the basis matrix B is constructed as follows. Let $\boldsymbol{z}_j$ denote the $n \times 1$ sample data on Z_j $(j \leq 3)$. Let $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_1 \odot \boldsymbol{z}_2, \boldsymbol{z}_1 \odot \boldsymbol{z}_3, \boldsymbol{z}_2 \odot \boldsymbol{z}_3)$ denote the $n \times 5$ matrix of the continuous data and interactions of the continuous data and the binary data. Now suppose $(\tau_{j1}, \ldots, \tau_{jK})$ are K knots based on each column of $\boldsymbol{Z}$ and let B_j denote the corresponding $n \times K$ matrix of cubic spline basis functions. Then, the basis matrix B is given by

$$B = \left[B_1 \vdots B_2^* \vdots B_3^* \vdots B_4^* \vdots B_5^* \vdots \boldsymbol{Z}_3 \right],$$

where B_j^* $(j = 2, 3, 4, 5)$ is the $n \times (K-1)$ matrix in which each column of B_j is subtracted from its first and then the first column is dropped, see Chib and Greenberg (2010). Thus, the dimension of this basis matrix is $n \times (5K - 4 + 1)$. To define the expanded functions, let $\rho_l(\boldsymbol{X}, \theta)$ $(l \leq d)$ denote a $n \times 1$ vector of the lth element of $\rho(X, \theta)$ evaluated at the sample data matrix $\boldsymbol{X}$. Then, the expanded functions for the sample observations are obtained by multiplying $\rho_l(\boldsymbol{X}, \theta)$ by the matrix B and concatenating. We use versions of this approach to construct the expanded functions in our examples.

3.2 Posterior distribution

The conditional model (2.1) is semiparametric and is characterized by two parameters: the data distribution P and the structural parameter θ, which is assumed to be finite dimensional. For a given value of K, the prior on (θ, P) is specified as $\pi(\theta)\pi(P|\theta, K)$, where the prior on θ is standard. Our default prior on θ is a product of independent student-t distributions with 2.5 degrees of freedom on each component of θ. The conditional prior on P, given (θ, K), can be viewed as a sieve type prior where the hierarchical parameter K has a degenerate prior with a point mass at a given value. In establishing the asymptotic properties of the posterior distribution, however, we let K grow to infinity with the sample size to ensure that (2.1) and (3.1) are equivalent in the limit. Fixing K at a specific value in a finite-sample analysis only impacts the posterior variance.

Priors on P designed to incorporate overidentifying moment restrictions are those of Schennach (2005), Kitamura and Otsu (2011), Shin (2014) and Florens and Simoni (2019). Our prior $\pi(P|\theta, K)$ follows from Schennach (2005). To construct this prior, we first model the joint data distribution P of W as a mixture of uniform probability densities, a construction which is capable of approximating any

distribution as the number of mixing components increases. Then, a prior is placed on the center of the d_w-dimensional hypercubes such that the corresponding mixture satisfies the moment restrictions for a given (θ, K). The resulting posterior is well-defined for every value of K.

By integrating out P with respect to this prior $\pi(P|\theta, K)$ one gets the integrated likelihood

$$p(W_{1:n}|\theta, K) = \prod_{i=1}^{n} \widehat{p}_i(\theta), \tag{3.3}$$

which is the ETEL function and where $\{\widehat{p}_i(\theta), i = 1, \ldots, n\}$ are the probabilities that minimize the KL divergence between the probabilities $(p_1, \ldots, p_n)$ assigned to each sample observation and the empirical probabilities $(\frac{1}{n}, \ldots, \frac{1}{n})$, subject to the conditions that the probabilities $(p_1, \ldots, p_n)$ sum to one and that the expectation under these probabilities satisfies the given unconditional moment conditions (3.1). That is, $\{\widehat{p}_i(\theta), i = 1, \ldots, n\}$ are the solution of the following problem:

$$\max_{p_1,\ldots,p_n} \sum_{i=1}^{n} [-p_i \log(np_i)] \qquad \text{subject to: } \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i g(w_i, \theta) = 0, \quad p_i \geq 0 \tag{3.4}$$

(see Schennach (2005) for a proof). In practice, we compute the ETEL probabilities from the dual (saddlepoint) representation as

$$\widehat{p}_i(\theta) := \frac{e^{\widehat{\lambda}(\theta)' g(w_i,\theta)}}{\sum_{j=1}^{n} e^{\widehat{\lambda}(\theta)' g(w_j,\theta)}} \quad (i = 1, \ldots, n), \tag{3.5}$$

where $\widehat{\lambda}(\theta) = \arg\min_{\lambda \in \mathbb{R}^{dK}} \frac{1}{n} \sum_{i=1}^{n} e^{\lambda' g(w_i,\theta)}$ is the estimated tilting parameter (see *e.g.* Csiszar (1984)).

By multiplying the ETEL function by the prior density of θ, the posterior distribution now takes the form

$$\pi(\theta|w_{1:n}, K) \propto \pi(\theta) \prod_{i=1}^{n} \frac{e^{\widehat{\lambda}(\theta)' g(w_i,\theta)}}{\sum_{j=1}^{n} e^{\widehat{\lambda}(\theta)' g(w_j,\theta)}}, \tag{3.6}$$

which we summarize by MCMC simulations, for example, the one block tailored Metropolis-Hastings (M-H) algorithm of Chib and Greenberg (1995), or the Tailored Randomized Block Metropolis-Hastings algorithm of Chib and Ramamurthy (2010).

Example 1 (continued) *To illustrate the prior-posterior analysis, we create a set of simulated data* $\{y_i, x_i\}_{i=1}^{n}$ *with covariates* $X \sim \mathcal{U}(-1, 2.5)$*, intercept* $\theta_0 = 1$*, slope* $\theta_1 = 1$*, and* ε_i *is distributed according to* $\varepsilon_i \sim \mathcal{SN}(m(x_i), h(x_i), s(x_i))$*, where* $\mathcal{SN}(m, h, s)$ *is the skew normal distribution with location, scale, and shape parameters given by* (m, h, s)*, each depending on* x_i*. When* s *is zero,* ε_i *is*

normal with mean m and standard deviation h. We set $m(x_i) = -h(x_i)\sqrt{2/\pi}s(x_i)/(\sqrt{1+s(x_i)^2})$ so that $E^P[\varepsilon|X] = 0$.

Suppose that $h(x) = \sqrt{\exp(1+0.7x+0.2x^2)}$ and $s(x) = 1+x^2$. We estimate the model using $E^P[\varepsilon|Z] = 0$, $Z = (1, X)$, without the need to model the heteroscedasticity or the skewness functions. Then, under the default independent student-t prior with mean 0, dispersion 5, and degrees of freedom 2.5, implying a prior standard deviation of $(25\,(2.5)\,/(2.5-2))^{1/2} = 11.18$, the marginal posterior distributions of θ_0 and θ_1 are summarized in panel (a) of Table 1 for two different sample sizes. We note from the dispersion of the posterior distribution, that the posterior distributions of both θ_0 and θ_1 shrink to the true value at the $\sqrt{n}$-rate. In the next section we formally establish this behavior. For comparison, we also compute the posterior distribution of (θ_0, θ_1) under the weaker assumption that ε_i is mean zero and uncorrelated with x_i. The relevant moment restrictions, given as in (2.4), are a subset of the expanded moment conditions. As can be seen from panels (a) and (b) of Table 1, imposing the (correct) conditional moment restrictions leads to about a 25% reduction in the posterior standard deviation of θ_1, for each of the two sample sizes.

Panel (a): $\mathbf{E}^P[\varepsilon\|z] = 0$		Mean	SD	Median	Lower	Upper	Ineff
$n = 500$	θ_0	0.896	0.073	0.895	0.755	1.040	1.107
	θ_1	1.127	0.084	1.126	0.964	1.296	1.117
$n = 2000$	θ_0	0.976	0.034	0.976	0.910	1.042	1.119
	θ_1	1.040	0.041	1.040	0.961	1.121	1.093
Panel (b): $\mathbf{E}^P[\varepsilon] = 0, \mathbf{E}^P[\varepsilon x] = 0$		Mean	SD	Median	Lower	Upper	Ineff
$n = 500$	θ_0	0.854	0.079	0.854	0.704	1.010	1.092
	θ_1	1.198	0.115	1.196	0.980	1.432	1.141
$n = 2000$	θ_0	0.962	0.036	0.962	0.893	1.032	1.092
	θ_1	1.053	0.055	1.053	0.947	1.162	1.101

Table 1: Difference between inferences from conditional (top panel) vs unconditional moments (bottom panel). Data is generated from a regression model with conditional heteroscedasticity and skewness. The true value of θ_0 is 1 and that of θ_1 is 1. Results are based on 20,000 MCMC draws beyond a burn-in of 1000. "Lower" and "Upper" refer to the 0.05 and 0.95 quantiles of the simulated draws, respectively, and "Ineff" to the inefficiency factor.

3.3 Asymptotic properties

Consider now the large sample behavior of the posterior distribution of θ. We let θ_* and P_*, respectively, denote the true value of θ and of the data distribution P. As notation, when the true distribution P_* is

involved, expectations $\mathbf{E}^P[\cdot]$ (resp. $\mathbf{E}^P[\cdot|\cdot]$) are taken with respect to P_* (resp. the conditional distribution associated with P_*). In addition, we denote

$$\rho_\theta(X,\theta) := \frac{\partial \rho(X,\theta)}{\partial \theta'}, \qquad D(Z) := \mathbf{E}^P[\rho_\theta(X,\theta_*)|Z],$$

$$\Sigma(Z) := \mathbf{E}[\rho(X,\theta_*)\rho(X,\theta_*)'|Z], \quad \text{and} \quad \rho_{j\theta\theta}(X,\theta_*) := \partial^2 \rho_j(X,\theta)/\partial\theta\partial\theta'.$$

For a vector a, $\|a\|$ denotes the Euclidean norm. For a matrix A, $\|A\|$ denotes the operator norm (the largest singular value of the matrix). Finally, let $\mathcal{Z} := \text{supp}(Z)$ denote the support of Z and $\ell_{n,\theta}(W_i) := \log \widehat{p}_i(\theta)$.

The first assumption is a normalization for the second moment matrix of the approximating functions which is standard in the literature, see *e.g.* Newey (1997) and Donald et al. (2003).

Assumption 3.2 *For each K there is a constant scalar $\zeta(K)$ such that $\sup_{z\in\mathcal{Z}} \|q^K(z)\| \leq \zeta(K)$, $\mathbf{E}^P[q^K(Z)q^K(Z)']$ has smallest eigenvalue bounded away from zero uniformly in K, and $\sqrt{K} \leq \zeta(K)$.*

The bound $\zeta(K)$ is known explicitly in a number of cases depending on the approximating functions we use. Donald et al. (2003) provide a discussion and explicit formulas for $\zeta(K)$ in the case of splines, power series and Fourier series. We also refer to Newey (1997) for primitive conditions for regression splines and power series.

Assumption 3.3 *The data $W_i := (X_i, Z_i)$, $i = 1,\ldots,n$ are i.i.d. according to P_* and (a) there exists a unique $\theta_* \in \Theta$ that satisfies $\mathbf{E}^P[\rho(X,\theta)|Z] = 0$ for the true P_*; (b) Θ is compact; (c) $\mathbf{E}^P[\sup_{\theta\in\Theta} \|\rho(X,\theta)\|^2|Z]$ is uniformly bounded on $\mathcal{Z}$.*

This assumption is the same as Donald et al. (2003, Assumption 3). Part (d) of this assumption imposes a Lipschitz condition which, together with part (c), allows application of uniform convergence results. The following three assumptions are also the same as the ones required by Donald et al. (2003) to establish asymptotic normality of the Generalized Empirical Likelihood (GEL) estimator.

Assumption 3.4 *(a) $\theta_* \in int(\Theta)$; (b) $\rho(X,\theta)$ is twice continuously differentiable in a neighborhood $\mathcal{U}$ of θ_*, $\mathbf{E}^P[\sup_{\theta\in\mathcal{U}} \|\rho_\theta(X,\theta)\|^2|Z]$ and $\mathbf{E}^P[\|\rho_{j\theta\theta}(X,\theta_*)\|^2|Z]$, $j = 1,\ldots d$, are uniformly bounded on $\mathcal{Z}$; (c) $\mathbf{E}^P[D(X)D(X)']$ is nonsingular.*

Assumption 3.5 *(a) $\Sigma(Z)$ has smallest eigenvalue bounded away from zero; (b) for a neighborhood $\mathcal{U}$ of θ_*, $\mathbf{E}^P[\sup_{\theta\in\mathcal{U}} \|\rho(X,\theta)\|^4|z]$ is uniformly bounded on $\mathcal{Z}$, and for all $\theta \in \mathcal{U}$, $\|\rho(X,\theta) - \rho(X,\theta_*)\| \leq \delta(X)\|\theta - \theta_*\|$ and $\mathbf{E}^P[\delta(X)^2|Z] < \infty$.*

Assumption 3.6 *There is $\gamma > 2$ such that $\mathbf{E}^P[\sup_{\theta\in\Theta} \|\rho(X,\theta)\|^\gamma] < \infty$ and $\zeta(K)^2 K/n^{1-2/\gamma} \to 0$.*

The last assumption is about the prior distribution of θ and is standard in the Bayesian literature on frequentist asymptotic properties of Bayes procedures.

Assumption 3.7 *(a) π is a continuous probability measure that admits a density with respect to the Lebesgue measure; (b) π is positive on a neighborhood of θ_*.*

We are now able to state our first major result in which we establish the asymptotic normality and efficiency of the posterior distribution of the local parameter $h := \sqrt{n}(\theta - \theta_*)$.

Theorem 3.1 (Bernstein-von Mises) *Under Assumptions 3.1-3.7, if $K \to \infty$, $\zeta(K)K^2/\sqrt{n} \to 0$, and if for any $\delta > 0$, $\exists \epsilon > 0$ such that as $n \to \infty$*

$$P\left(\sup_{\|\theta-\theta_*\|>\delta} \frac{1}{n}\sum_{i=1}^n (\ell_{n,\theta}(W_i) - \ell_{n,\theta_*}(W_i)) \leq -\epsilon\right) \to 1, \tag{3.7}$$

then the posterior distribution $\pi(\sqrt{n}(\theta-\theta_*)|W_{1:n})$ converges in total variation towards a random Normal distribution, that is,

$$\sup_B \left|\pi(\sqrt{n}(\theta-\theta_*) \in B|W_{1:n}) - \mathcal{N}_{\Delta_{n,\theta_*},V_{\theta_*}}(B)\right| \xrightarrow{p} 0, \tag{3.8}$$

where $B \subseteq \Theta$ is any Borel set, $\Delta_{n,\theta_*} := -\frac{1}{\sqrt{n}}\sum_{i=1}^n V_{\theta_*} D(Z_i)'\Sigma(Z_i)^{-1}\rho(X_i,\theta_*)$ is bounded in probability and $V_{\theta_*} := \left(\mathbf{E}^P[D(Z)'\Sigma(Z)^{-1}D(Z)]\right)^{-1}$.

We note that the centering Δ_{n,θ_*} of the limiting normal distribution satisfies $\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{d\log \widehat{p}_i(\theta_*)}{d\theta} - V_{\theta_*}^{-1}\Delta_{n,\theta_*} \xrightarrow{p} 0$. We also note that the condition $\zeta(K)K^2/\sqrt{n} \to 0$ in the theorem implies $K/n \to 0$, which is a classical condition in the sieve literature. This condition is required to establish a stochastic Local Asymptotic Normality (LAN) expansion, which is an intermediate step to prove the BvM result, as we explain below. The LAN expansion is not required to establish asymptotic normality of the GEL estimators, which explains why our condition is slightly stronger than the condition $\zeta(K)K/\sqrt{n} \to 0$ required by Donald, Imbens and Newey (2003). On the other hand, our condition is weaker than the condition $\zeta(K)^2K^2/\sqrt{n} \to 0$ required by Donald, Imbens and Newey (2009) to establish the mean square error of the GEL estimators. The asymptotic covariance of the posterior distribution coincides with the semiparametric efficiency bound given in Chamberlain (1987) for conditional moment condition models. This means that, for every $\alpha \in (0,1)$, $(1-\alpha)$-credible regions constructed from the posterior of θ are

$(1-\alpha)$-confidence sets asymptotically. Indeed, they are correctly centered and have correct volume.

The proof of this theorem is given in the supplementary appendix and consists of three steps. In the first step we show consistency of the posterior distribution of θ, namely:

$$\pi\left(\sqrt{n}\|\theta-\theta_*\|>M_n\middle|\, W_{1:n}\right)\xrightarrow{p}0 \tag{3.9}$$

for any $M_n\to\infty$, as $n\to\infty$. To show this, the identification assumption (3.7) is used. In the second step we show that the ETEL function satisfies a stochastic LAN expansion:

$$\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\ell_{n,\theta_*+h/\sqrt{n}}(W_i)-\sum_{i=1}^{n}\ell_{n,\theta_*}(W_i)-h'V_{\theta_*}^{-1}\Delta_{n,\theta_*}-\frac{1}{2}h'V_{\theta_*}^{-1}h\right|=o_p(1), \tag{3.10}$$

where $\mathcal{H}$ denotes a compact subset of $\mathbb{R}^p$ and $V_{\theta_*}^{-1}\Delta_{n,\theta_*}\xrightarrow{d}\mathcal{N}(0,V_{\theta_*}^{-1})$. As the ETEL function is an integrated likelihood, expansion (3.10) is better known as integral LAN in the semiparametric Bayesian literature, see *e.g.* Bickel and Kleijn (2012, Section 4). In the third step of the proof we use arguments, see *e.g.* the proof of Van der Vaart (1998, Theorem 10.1), to show that (3.9) and (3.10) imply asymptotic normality of $\pi(\sqrt{n}(\theta-\theta_*)\in B|W_{1:n})$. While these three steps are classical in proving the Bernstein-von Mises phenomenon, establishing (3.10) raises challenges that are otherwise absent. This is because the ETEL function is a nonstandard likelihood that involves estimated parameters $\|\widehat{\lambda}(\theta_*)\|$ whose dimension is dK, which increases with n. Therefore, we first need to determine the rate of growth of $\|\widehat{\lambda}\|$, of $\|\frac{1}{n}\sum_{i=1}^{n}g(W_i,\theta)\|$ and of the norms of the empirical counterparts of $D(Z)$ and $\Sigma(Z)$. While $\|\widehat{\lambda}(\theta_*)\|$ is expected to converge to zero in the correctly specified case, the rate of convergence is slower than $n^{-1/2}$. In the supplementary appendix we show that $\|\widehat{\lambda}(\theta_*)\|=O_p(\sqrt{K/n})$ under the previous assumptions.

3.4 Misspecified model

We now generalize the preceding BvM result for the important class of misspecified conditional moment models, building on the theory derived in Chib, Shin and Simoni (2018) in connection with misspecified unconditional moment models.

Definition 3.1 (Misspecified model) *We say that the conditional moment conditions model is misspecified if the set of probability measures implied by the moment restrictions does not contain the true data generating process P for any $\theta\in\Theta$, that is, $P\notin\mathcal{P}$ where $\mathcal{P}=\bigcup_{\theta\in\Theta}\widetilde{\mathcal{P}}_\theta$ and $\widetilde{\mathcal{P}}_\theta=\{Q\in\mathbb{M}_{X|Z};\ \mathbf{E}^Q[\rho(X,\theta)|Z]=0\ a.s.\}$ with $\mathbb{M}_{X|Z}$ the set of all conditional probability measures of $X|Z$.*

In essence, if (2.1) is misspecified then there is no $\theta \in \Theta$ such that $\mathbf{E}^P[\rho(X,\theta) \otimes q^K(Z)] = 0$ almost surely for every K large enough. Now, for every $\theta \in \Theta$ define $Q^*(\theta)$ as the minimizer of the Kullback-Leibler divergence of P_* to the model $\mathcal{P}_\theta := \{Q \in \mathbb{M}; \mathbf{E}^Q[g(W,\theta)] = 0\}$, where $\mathbb{M}$ denotes the set of all the probability measures on $\mathbb{R}^{d_w}$. That is, $Q^*(\theta) := \operatorname{arginf}_{Q\in\mathcal{P}_\theta} K(Q||P_*)$, where $K(Q||P_*) := \int \log(dQ/dP_*)dQ$. If we suppose that the dual representation of the Kullback-Leibler minimization problem holds, then the P_*-density of $Q^*(\theta)$ has the closed form: $[dQ^*(\theta)/dP_*](W_i) = \frac{e^{\lambda_\circ' g(W_i,\theta)}}{\mathbf{E}^P[e^{\lambda_\circ' g(W_j,\theta)}]}$, where $\lambda_\circ$ denotes the tilting parameter and is defined in the same way as in the correctly specified case:

$$\lambda_\circ := \lambda_\circ(\theta) := \arg\min_{\lambda\in\mathbb{R}^{dK}} \mathbf{E}^P[e^{\lambda' g(W_i,\theta)}]. \tag{3.11}$$

We also impose a condition to ensure that the probability measures $\mathcal{P} := \bigcup_{\theta\in\theta} \mathcal{P}_\theta$, which are implied by the model, are dominated by the true probability measure P_*. This is required for the validity of the dual theorem. Therefore, following Sueishi (2013, Theorem 3.1), we replace Assumption 3.3 (a) by the following.

Assumption 3.8 *For a fixed $\theta \in \Theta$, there exists $Q \in \mathcal{P}_\theta$ such that Q is mutually absolutely continuous with respect to P, where $\mathcal{P}_\theta := \{Q \in \mathbb{M}; \mathbf{E}^Q[g(W,\theta)] = 0\}$ and $\mathbb{M}$ denotes the set of all the probability measures on $\mathbb{R}^{d_w}$.*

This assumption implies that $\mathcal{P}_\theta$ is non-empty. A similar assumption is also made by Kleijn and van der Vaart (2012) and Chib, Shin and Simoni (2018) to establish the BvM under misspecification. The pseudo-true value of the parameter $\theta \in \Theta$ is denoted by $\theta_\circ$ and is defined as the minimizer of the Kullback-Leibler divergence between the true P_* and $Q^*(\theta)$:

$$\theta_\circ := \operatorname{arginf}_{\theta\in\Theta} K(P_*||Q^*(\theta)), \tag{3.12}$$

where $K(P_*||Q^*(\theta)) := \int \log(dP_*/dQ^*(\theta))dP_*$. Under the preceding absolute continuity assumption, the pseudo-true value $\theta_\circ$ is available as

$$\theta_\circ = \operatorname{argmax}_{\theta\in\Theta} \mathbf{E}^P \log\left(\frac{e^{\lambda_\circ' g(W_i,\theta)}}{\mathbf{E}^P[e^{\lambda_\circ' g(W_j,\theta)}]}\right). \tag{3.13}$$

Note that $\lambda_\circ(\theta_\circ)$, the value of the tilting parameter at the pseudo-true value $\theta_\circ$, is nonzero because the moment conditions do not hold.

Assumption 3.8 implies that $K(Q^*(\theta_\circ)||P_*) < \infty$. We supplement this with the assumption that $K(P_*||Q^*(\theta_\circ)) < \infty$ and that $K(P_*||Q^*(\theta)) < \infty$, $\forall\theta \in \Theta$. Because consistency in misspecified

models is defined with respect to the pseudo-true value $\theta_\circ$, we need to replace Assumption 3.7 *(b)* by the following Assumption 3.9 (b) which, together with Assumption 3.9 *(a)*, requires the prior to put enough mass to balls around $\theta_\circ$.

Assumption 3.9 *(a) π is a continuous probability measure that admits a density with respect to the Lebesgue measure; (b) The prior distribution π is positive on a neighborhood of $\theta_\circ$, where $\theta_\circ$ is as defined in* (3.13)*.*

In the next assumption we denote by $int(\Theta)$ the interior of Θ and by $\mathcal{U}$ a ball centered at $\theta_\circ$ with radius $h/\sqrt{n}$ for some $h \in \mathcal{H}$ and $\mathcal{H}$ a compact subset of $\mathbb{R}^p$.

Assumption 3.10 *The data $W_i := (X_i, Z_i)$, $i = 1, \ldots, n$ are i.i.d. according to P_* and*
(a) The pseudo-true value $\theta_\circ \in int(\Theta)$ is the unique maximizer of

$$\lambda_\circ(\theta)' \mathbf{E}^P[g(W,\theta)] - \log \mathbf{E}^P[\exp\{\lambda_\circ(\theta)' g(W,\theta)\}],$$

where Θ is compact;
(b) $\lambda_\circ(\theta) \in int(\Lambda(\theta))$, where $\Lambda(\theta)$ is a compact set for every $\theta \in \Theta$ and $\lambda_\circ$ is as defined in (3.11)*;*
(c) $\rho(X,\theta)$ is continuous at each $\theta \in \Theta$ with probability one;
(d) $\rho(X,\theta)$ is twice continuously differentiable in the neighborhood $\mathcal{U}$ of $\theta_\circ$, $\mathbf{E}^P[\sup_{\theta\in\mathcal{U}} \|\rho_\theta(x,\theta)\|^4|Z]$ and $\mathbf{E}^P[\sup_{\theta\in\mathcal{U}} e^{\lambda_\circ(\theta_\circ)' g_i(\theta)} \|\rho_{j\theta\theta}(x,\theta)\|^2|Z]$, $j = 1, \ldots d$, are uniformly bounded over $\mathcal{Z}$;
(e) for the neighborhood $\mathcal{U}$ of $\theta_\circ$,

$$\mathbf{E}^P[e^{\lambda_\circ(\theta_\circ)' g(W,\theta_\circ)} \|\rho(X,\theta_\circ)\|^2 \|q^K(Z)\|] = O(K)$$

and for all $\theta \in \mathcal{U}$, $\|\rho(X,\theta) - \rho(X,\theta_\circ)\| \leq \delta(X)\|\theta - \theta_\circ\|$, $\mathbf{E}^P[\delta(X)^2|Z] < \infty$ and

$$\mathbf{E}^P[e^{\lambda_\circ(\theta_\circ)' g(W,\theta_\circ)} \delta(X)^2 \|q^K(Z)\|^2] = O(K),$$

(f) for the neighborhood $\mathcal{U}$ of $\theta_\circ$ and for $\kappa = 1, 2$, $j = 2, 4$ it holds that

$$\mathbf{E}^P[\sup_{\theta\in\mathcal{U}} e^{\kappa\lambda_\circ(\theta_\circ)' g(W,\theta)} \|g(W_i,\theta)\|^j] = O(\zeta(K)^{j-2}K)$$

and $\mathbf{E}^P[\sup_{\theta\in\mathcal{U}} e^{\kappa\lambda_\circ(\theta_\circ)' g(W,\theta)} \|G(W,\theta)\|^j] = O(\zeta(K)^{j-2}K)$, where $\zeta(K)$ is as defined in Assumption 3.2;

(g) $\mathbf{E}[e^{\lambda_\circ(\theta_\circ)'g(W,\theta_\circ)}\rho(X,\theta_\circ)\rho(X,\theta_\circ)'|Z]$ *has smallest eigenvalue bounded away from zero;*

(h) if $\mathcal{H}$ *is a compact subset of* $\mathbb{R}^p$*, it holds*

$$\sup_{h\in\mathcal{H}} \mathbf{E}[g(W_i,\theta_\circ)'] \left(\frac{d\widehat{\lambda}(\theta_\circ)}{d\theta'} - \frac{d\lambda_\circ(\theta_\circ)}{d\theta'} \right) h = O_p(n^{-1/2}),$$

where $\widehat{\lambda}(\theta_\circ)$ *is the solution of* $\mathbf{E}_n[e^{\widehat{\lambda}(\theta_\circ)'g(W_i,\theta_\circ)}g(W_i,\theta_\circ)] = 0$*, and* $\mathbb{E}_n[\cdot] := \frac{1}{n}\sum_{i=1}^n[\cdot]$ *is the empirical mean operator.*

Assumption 3.10 (a) guarantees uniqueness of the pseudo-true value and is a standard assumption in the literature on misspecified models (see *e.g.* White (1982)). Assumption 3.10 (d) is the misspecified counterpart of Assumption 3.4 (a) and 3.5 (b). Remark that the presence of the exponential $e^{\lambda_\circ(\theta_\circ)'g(W,\theta_\circ)}$ inside the expectations in Assumption 3.10 (e)-(g) is due to the fact that in the misspecified case the pseudo-true value of the tilting parameter $\lambda_\circ(\theta_\circ)$ is not equal to zero as it is in the correctly specified case. Assumptions 3.10 (e) and (f) impose an upper bound on the rate at which the norms of K-vector and $(dK \times p)$-matrices are allowed to increase. Assumption 3.10 (g) is the misspecified counterpart of Assumption 3.5 (a). Finally, 3.10 (h) guarantees that one of the terms in the random vector $\Delta_{n,\theta_\circ}$, which is introduced in Theorem 3.2 below, is bounded in probability.

Our next important theorem, the BvM theorem for misspecified models, now follows.

Theorem 3.2 (Bernstein-von Mises (misspecified)) *Let Assumptions 3.1, 3.2, 3.6, 3.8, 3.9 and 3.10 hold. Assume that there exists a constant* $C > 0$ *such that for any sequence* $M_n \to \infty$*,*

$$P\left(\sup_{\|\theta-\theta_\circ\|>M_n/\sqrt{n}} \frac{1}{n}\sum_{i=1}^n (\ell_{n,\theta}(W_i) - \ell_{n,\theta_\circ}(W_i)) \leq -CM_n^2/n \right) \to 1, \tag{3.14}$$

as $n \to \infty$*. If* $K \to \infty$*,* $\zeta(K)K^2\sqrt{K/n} \to 0$*, then the posteriors converge in total variation towards a Normal distribution, that is,*

$$\sup_B \left| \pi(\sqrt{n}(\theta-\theta_\circ) \in B|W_{1:n}) - \mathcal{N}_{\Delta_{n,\theta_\circ},\mathcal{A}_{\theta_\circ}^{-1}}(B) \right| \xrightarrow{p} 0, \tag{3.15}$$

where $B \subseteq \Theta$ *is any Borel set,* $\Delta_{n,\theta_\circ}$ *is a random vector bounded in probability and* $\mathcal{A}_{\theta_\circ}^{-1}$ *is a nonsingular matrix.*

The expressions for $\mathcal{A}_{\theta_\circ}$ is given in (E.30) in the supplementary appendix. Just as in Kleijn and van der Vaart (2012), this theorem establishes that the posterior distribution of the centered and scaled parameter $\sqrt{n}(\theta-\theta_\circ)$ converges to a Normal distribution with a random mean that is bounded in probability. Its

proof is based on the same three steps as the proof of Theorem 3.1 in the correctly specified case with θ_* replaced by the pseudo-true value $\theta_\circ$. There are however important differences in proving that the ETEL function satisfies a stochastic LAN expansion in the misspecified case. First of all the limit of $\widehat{\lambda}(\theta_\circ)$ is $\lambda_\circ(\theta_\circ)$, which is different from zero. Therefore, several terms that were equal to zero in the LAN expansion for the correctly specified case are non-zero in the misspecified case and we have to deal with their limit in distribution. Second, the quantity $\frac{1}{\sqrt{n}}\sum_{i=1}^n g(W_i,\theta_\circ)$ is no longer centered on zero, which leads to an additional bias term. Part of the behavior of this term is controlled by Assumption 3.10 (h).

Furthermore, our proof makes use of a stochastic LAN expansion of the ETEL function, which we prove (under the assumptions of the theorem) takes the form

$$\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \ell_{n,\theta_1}(W_i) - \sum_{i=1}^n \ell_{n,\theta_\circ}(W_i) - h'\mathcal{A}_{\theta_\circ}\Delta_{n,\theta_0} - \frac{1}{2}h'\mathcal{A}_{\theta_\circ}h\right| = o_p(1),$$

where Δ_{n,θ_0} and $\mathcal{A}_{\theta_\circ}$ are as in the statement of Theorem 3.2.

4 Model Comparisons

We now turn our attention to the problem of comparing competing conditional moment models. We suppose that the models in the model space are misspecified, which is arguably the most pervasive case in practice. We are concerned with establishing the large sample optimality of the formal Bayesian rule of picking the model with the largest value of the marginal likelihood.

Let M_ℓ denote the ℓth model in model space. Each model is characterized by a parameter θ^ℓ and an extended moment function $g^\ell(W,\theta^\ell)$. For each model M_ℓ, we impose a prior distribution for θ^ℓ, and obtain the posterior distribution based on (3.6). Let $m(W_{1:n}|M_\ell)$ denote the marginal likelihood of model M_ℓ , which we calculate from the marginal likelihood identity of Chib (1995):

$$\log m(W_{1:n}|M_\ell) = \log \pi(\tilde{\theta}^\ell|M_\ell) + \log p(W_{1:n}|\tilde{\theta}^\ell, M_\ell) - \log \pi(\tilde{\theta}^\ell|W_{1:n}, M_\ell), \tag{4.1}$$

and the method of Chib and Jeliazkov (2001). In this expression, $\tilde{\theta}^\ell$ is any point in the support of the posterior (such as the posterior mean).

Remark 4.1 *Comparison of conditional moment condition models differs in one important aspect from the framework for comparing unconditional moment condition models that was established in Chib, Shin and Simoni (2018), where it is shown that to make the unconditional moment condition models comparable it is necessary to linearly transform the moment functions so that all the transformed moments*

are included in each model. This linear transformation consists of adding an extra parameter different from zero to the components of the vector $g(\theta, W)$ that correspond to the restrictions not included in a specific model. When comparing conditional moment models, however, this transformation is not necessary because the convex hulls associated with different expanded models have the same dimension asymptotically.

4.1 Model selection consistency

Let us suppose that our collection of models among which we want to make a selection contains J models. At least $J-1$ of these models are misspecified and one can be either misspecified or correctly specified. Moreover, suppose that the best model M_ℓ is selected by the size of the marginal likelihoods. Then, in Theorem 4.1 we show that this criterion in the limit picks the model M_ℓ with the smallest KL divergence between P and the corresponding $Q^*(\theta^\ell)$, where $Q^*(\theta^\ell)$ is such that $K(Q^*(\theta^\ell)||P) = \inf_{Q\in\mathcal{P}_{\theta^\ell}} K(Q||P)$ and $\mathcal{P}_{\theta^\ell}$ is defined in Section 3.4.

Theorem 4.1 *Let the assumptions of Theorem 3.2 hold. Let us consider the comparison of $J < \infty$ models M_ℓ, $\ell = 1, \ldots, J$, such that $J-1$ of these models each has at least one misspecified moment condition, that is, M_ℓ does not satisfy Assumption 3.3 (a), $\forall \ell \neq j$, and model M_j can be either correctly specified or contain some misspecified moment condition. Then,*

$$\lim_{n\to\infty} P\left(\log m(W_{1:n}; M_j) > \max_{\ell\neq j} \log m(W_{1:n}; M_\ell)\right) = 1$$

if and only if $K(P||Q^*(\theta^j_\circ)) < \min_{\ell\neq j} K(P||Q^*(\theta^\ell_\circ))$, where $K(P||Q) := \int \log(dP/dQ)dP$.

Note that if one model in the contending set of models is correctly specified, then this model will have zero KL divergence and, therefore, according to Theorem 4.1, that model will have the largest marginal likelihood and will be selected by our procedure.

To understand the ramifications of the preceding result, suppose that we are interested in comparing models with the same moment conditions but different conditioning variables:

$$\text{Model 1: } \mathbf{E}^P[\rho(X,\theta)|Z_1] = 0, \qquad \text{Model 2: } \mathbf{E}^P[\rho(X,\theta)|Z_2] = 0, \tag{4.2}$$

where Z_1 and Z_2 may have some elements in common, in particular Z_2 might be a subvector of Z_1 (or vice versa). A situation of this type, where we are unsure about the validity of instrumental variables, is the following.

Example 2 *(Comparing IV models) Consider the following model with three instruments* (Z_1, Z_2, Z_3)*:*

$$Y = \theta_0 + \theta_1 X + e_1,$$
$$X = f(Z_1, Z_2, Z_3) + e_2,$$
$$Z_1 \sim U[0,1], \qquad Z_2 \sim U[0,1], \quad \text{and} \quad Z_3 \sim \mathcal{B}(0.4),$$

where $(e_1, e_2)'$ *are non-Gaussian and correlated, which makes* X *in the outcome model correlated with the error* e_1*. We let the true value of* $\theta = (\theta_0, \theta_1)$ *be* $(1,1)$*. Moreover, suppose that the* Z_j*'s are relevant instruments, that is,* $cov(X, Z_j) \neq 0$ *for* $j \leq 3$*, and*

$$f(Z_1, Z_2, Z_3) = 6\left(\sqrt{0.3}Z_1 + \sqrt{0.7}Z_2\right)^3 (1 - \sqrt{0.3}Z_1 - \sqrt{0.7}Z_2)Z_3 + Z_1 Z_2 (1 - Z_3). \tag{4.3}$$

We consider a situation in which some instruments are valid and some are not, and we are interested in selecting valid instruments from a set of instruments. To this end, we generate (e_1, e_2, Z_1) *from a Gaussian copula whose covariance matrix is* $\Sigma = [1, 0.7, 0.7; 0.7, 1, 0; 0.7, 0, 1]$ *such that the marginal distribution of* e_1 *is the skewed mixture of two normal distributions* $0.5\mathcal{N}(0.5, 0.5^2) + 0.5\mathcal{N}(-0.5, 1.118^2)$ *and the marginal distribution of* e_2 *is* $\mathcal{N}(0,1)$*. Under this setup,* Z_1 *is now an invalid instrument. We consider the following three models*

$$\mathcal{M}_1 : \mathbf{E}^P[(Y - \theta_0 - \theta_1 X)|Z_1, Z_2, Z_3] = 0, \tag{4.4}$$
$$\mathcal{M}_2 : \mathbf{E}^P[(Y - \theta_0 - \theta_1 X)|Z_1, Z_3] = 0, \tag{4.5}$$
$$\mathcal{M}_3 : \mathbf{E}^P[(Y - \theta_0 - \theta_1 X)|Z_2, Z_3] = 0. \tag{4.6}$$

Because Z_1 *is an invalid instrument, both* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *are wrong.*

In $\mathcal{M}_1$*, our basis matrix* B *is made from the variables* $(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_1 \odot \boldsymbol{z}_2, \boldsymbol{z}_1 \odot \boldsymbol{z}_3, \boldsymbol{z}_2 \odot \boldsymbol{z}_3)$*, each using five knots, concatenated with the vector* $\boldsymbol{z}_3$*. This matrix* B *has 22 columns, which equals the number of expanded moment conditions. The prior for* θ_0 *and* θ_1 *is the product of student-*t *distributions with mean zero, dispersion 5, and degrees of freedom equal to 2.5. Estimation and calculation of the marginal likelihood for* $\mathcal{M}_2$ *and* $\mathcal{M}_3$ *are special cases of* $\mathcal{M}_1$*.*

Table 2 calculates the marginal likelihoods of all the three models for two simulated samples. Note that the model with the valid instruments ($\mathcal{M}_3$*) is correctly specified and it has the highest marginal likelihood, in conformity with our theory.*

Table 2: Model comparison: IV regression example

		$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
$n = 500$	Marginal Likelihood	-3160.65	-3130.36	-3118.76
		(0.032)	(0.123)	(0.004)
$n = 2,000$	Marginal Likelihood	-15350.08	-15262.06	-15217.79
		(0.188)	(0.370)	(0.001)

Note: The posterior summaries are based on 20,000 MCMC draws beyond a burn-in of 1000. Numerical standard errors are in parenthesis.

5 Application: Moment-Based Causal Inference

An important application of our methods is to problems that arise in causal inference. For specificity, we consider here the estimation of causal parameters in the sharp regression-discontinuity (RD) design. Another example, the average treatment effect (ATE) estimation under a conditional independence assumption, is deferred to the supplementary appendix.

RD-ATE in a Sharp design. Suppose that the data arise from the following data generating mechanism,

$$Y = (1 - X)g_0(Z) + Xg_1(Z) + \varepsilon,$$

where $X = 1\{Z \geq \tau\}$ and $E^P[\varepsilon|Z] = 0$. We define the RD-ATE as

$$\text{RD-ATE} = g_1(\tau) - g_0(\tau),$$

where $g_0(\tau)$ is the left limit of $g_0(Z)$ and $g_1(\tau)$ is a right limit of $g_1(Z)$.

For illustrative purposes, suppose that

$$g_0(z) = 0.5 + Z \quad \text{and} \quad g_1(Z) = 0.8 + 2Z,$$

where $Z = 2(Z^* - 1)$ and $Z^* \sim 2\mathcal{B}(2, 4)$, ε is independently drawn from $\mathcal{SN} \sim (m(Z), h(Z), s(Z))$ with $m(Z) = -h(Z)\sqrt{2/\pi}s(Z)/(\sqrt{1 + s(Z)^2})$, $h(Z) = 0.7(2 - Z^2)$, and $s(Z) = 3 + Z^2$. Under this set up, the true value of RD-ATE at the break-point ($\tau = 0$) is 0.3. We estimate the RD-ATE with three different sample sizes, $n = 500, 2000, 8000$.

Our prior-posterior analysis is based on the conditional mean independence assumption $E^P[\varepsilon|Z] = 0$, without any further assumptions about ε. We estimate $g_0(Z)$ and $g_1(Z)$ separately for data on either side of τ using the conditional moment restrictions, $E^P[Y - \theta_{j0} - \theta_{j1}Z|Z] = 0$, where $j = 0, 1$. We use 5

knots to convert the conditional expectation into the expanded moment conditions when $n = 500, 2000$, and 10 knots when $n = 8000$. The prior of $(\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11})$ is an independent student-t prior with mean 0, dispersion 5, and degrees of freedom 2.5.

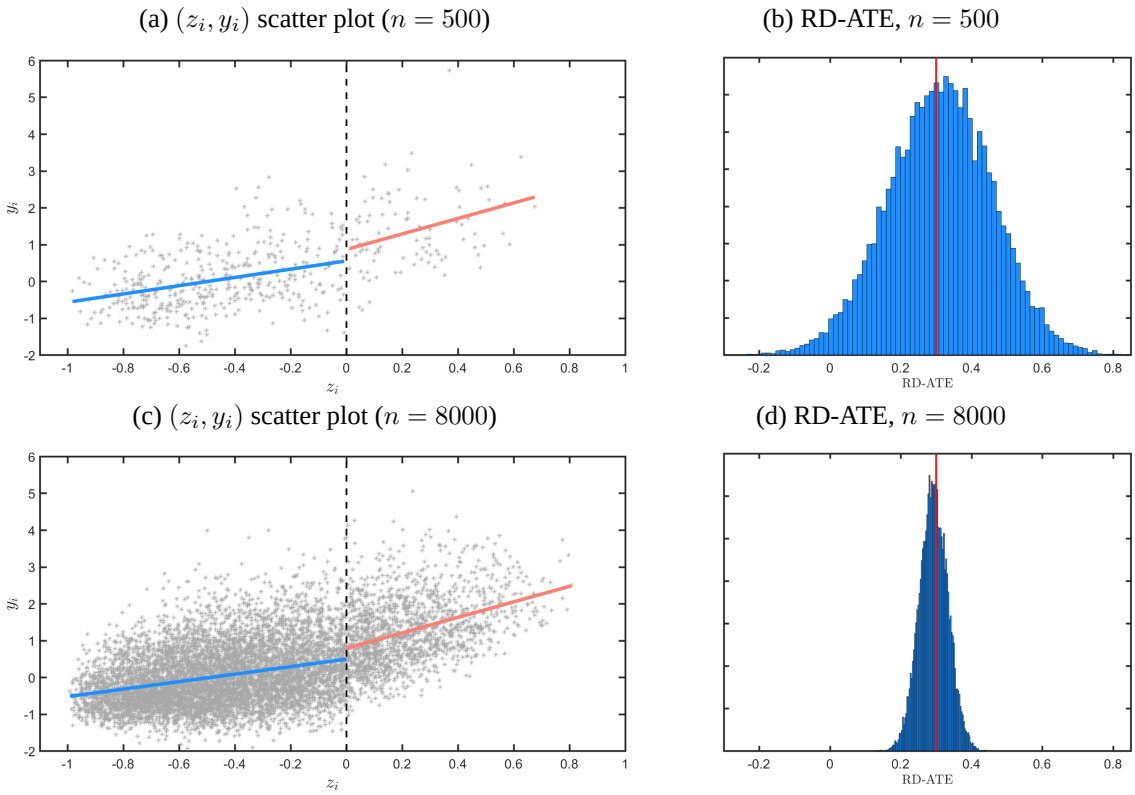

Figure 1: In the left panels, grey dots represent realizations of (z_i, y_i). Blue and red lines are $g_0(z_i)$ and $g_1(z_i)$ evaluated at the posterior mean ($n = 500, 8000$). Right panels have the posterior distributions of the RD-ATE. Results are based on 20,000 MCMC draws beyond a burn-in of 1000.

The results from this analysis are reported in Figure 1 and Table 3. The left panels of the figure have a scatter plot of the data and the estimated regression functions at the posterior mean of the parameters. The right panels of the figure have the histogram approximation to the posterior distribution of the RD-ATE. One can see that the posterior distribution puts high mass around the true RD-ATE value of 0.3, and that the posterior distribution shrinks around this value with n.

Table 3: Posterior summaries for RD-ATE

	Mean	SD	Median	Lower	Upper	Ineff
$n = 500$	0.311	0.147	0.314	0.016	0.594	1.137
$n = 2000$	0.324	0.088	0.324	0.153	0.496	1.093
$n = 8000$	0.293	0.040	0.293	0.214	0.373	1.073

6 Conclusion

In this paper we have developed a Bayesian framework for analyzing an important and broad class of semiparametric models in which the distribution of the outcomes is defined only up to a set of conditional moments, some of which may be misspecified. We have derived BvM theorems for the behavior of the posterior distribution under both correct and incorrect specification of the conditional moments, and developed the theory for comparing different conditional moment models through a comparison of model marginal likelihoods.

Our theory and examples, taken together, show that our framework makes possible the formal Bayesian analysis of a new, large class of problems that were hitherto difficult, or not possible, to tackle from the Bayesian viewpoint.

Supplementary Material

Technical proofs of all the results developed in the paper are in the supplementary appendix.

References

Ai, C. and Chen, X. (2003), 'Efficient estimation of models with conditional moment restrictions containing unknown functions', *Econometrica* **71**(6), 1795–1843.

Bickel, P. J. and Kleijn, B. J. K. (2012), 'The semiparametric Bernstein-von Mises theorem', *Annals of Statistics* **40**(1), 206–237.

Bierens, H. J. (1982), 'Consistent model specification tests', *Journal of Econometrics* **20**(1), 105–134.

Carrasco, M. and Florens, J.-P. (2000), 'Generalization of GMM to a continuum of moment conditions', *Econometric Theory* **16**(6), 797–834.

Chamberlain, G. (1987), 'Asymptotic efficiency in estimation with conditional moment restrictions', *Journal of Econometrics* **34**(3), 305–334.

Chen, X., Christensen, T. and Tamer, E. T. (2018), 'Monte Carlo confidence sets for identified sets', *Econometrica* **86**(6), 1965–2018.

Chib, S. (1995), 'Marginal likelihood from the Gibbs output', *Journal of the American Statistical Association* **90**(432), 1313–1321.

Chib, S. and Greenberg, E. (1995), 'Understanding the Metropolis-Hastings algorithm', *The American Statistician* **49**(4), 327–335.

Chib, S. and Greenberg, E. (2010), 'Additive cubic spline regression with Dirichlet process mixture errors', *Journal of Econometrics* **156**(2), 322–336.

Chib, S. and Jeliazkov, I. (2001), 'Marginal likelihood from the Metropolis-Hastings output', *Journal of the American Statistical Association* **96**(453), 270–281.

Chib, S. and Ramamurthy, S. (2010), 'Tailored Randomized Block MCMC methods with application to DSGE models', *Journal of Econometrics* **155**(1), 19–38.

Chib, S., Shin, M. and Simoni, A. (2018), 'Bayesian estimation and comparison of moment condition models', *Journal of the American Statistical Association* **113**(524), 1656–1668.

Csiszar, I. (1984), 'Sanov property, generalized i-projection and a conditional limit theorem', *Annals of Probability* **12**(3), 768–793.

Donald, S. G., Imbens, G. W. and Newey, W. K. (2003), 'Empirical likelihood estimation and consistent tests with conditional moment restrictions', *Journal of Econometrics* **117**(1), 55 – 93.

Donald, S. G., Imbens, G. W. and Newey, W. K. (2009), 'Choosing instrumental variables in conditional moment restriction models', *Journal of Econometrics* **152**(1), 28 – 36.

Donald, S. G. and Newey, W. K. (2001), 'Choosing the number of instruments', *Econometrica* **69**(5), 1161–1191.

Florens, J.-P. and Simoni, A. (2012), 'Nonparametric estimation of an instrumental regression: A quasi-Bayesian approach based on regularized posterior', *Journal of Econometrics* **170**(2), 458 – 475.

Florens, J.-P. and Simoni, A. (2016), 'Regularizing priors for linear inverse problems', *Econometric Theory* **32**(1), 71–121.

Florens, J.-P. and Simoni, A. (2019), 'Gaussian processes and Bayesian moment estimation', *Journal of Business & Economic Statistics* . published online October 25, 2019.

Hristache, M. and Patilea, V. (2017), 'Conditional moment models with data missing at random', *Biometrika* **104**(3), 735–742.

Kato, K. (2013), 'Quasi-Bayesian analysis of nonparametric instrumental variables models', *Annals of Statistics* **41**(5), 2359–2390.

Kitamura, Y. and Otsu, T. (2011), Bayesian analysis of moment condition models using nonparametric

priors, Technical report, Yale University.

Kitamura, Y., Tripathi, G. and Ahn, H. (2004), 'Empirical likelihood-based inference in conditional moment restriction models', *Econometrica* **72**(6), 1667–1714.

Kleijn, B. and van der Vaart, A. (2012), 'The Bernstein-von-Mises theorem under misspecification', *Electronic Journal of Statistics* **6**, 354–381.

Lazar, N. A. (2003), 'Baysian empirical likelihood', *Biometrika* **90**(2), 319–326.

Liao, Y. and Jiang, W. (2011), 'Posterior consistency of nonparametric conditional moment restricted models', *Annals of Statistics* **39**(6), pp. 3003–3031.

Liao, Y. and Simoni, A. (2019), 'Bayesian inference for partially identified smooth convex models', *Journal of Econometrics* **211**(2), 338 – 360.

Newey, W. K. (1997), 'Convergence rates and asymptotic normality for series estimators', *Journal of Econometrics* **79**(1), 147 – 168.

Rosenbaum, P. R. and Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), 41–55.

Schennach, S. M. (2005), 'Bayesian exponentially tilted empirical likelihood', *Biometrika* **92**(1), 31–46.

Shin, M. (2014), Bayesian GMM, Technical report, University of Pennsylvania.

Sueishi, N. (2013), 'Identification problem of the exponential tilting estimator under misspecification', *Economics Letters* **118**(3), 509 – 511.

Van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics.

White, H. (1982), 'Maximum likelihood estimation of misspecified models', *Econometrica* **50**(1), 1–25.