Estimation and Comparison of Conditional Moment Models

Siddhartha Chib*
Minchul Shin†
Anna Simoni‡

April 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. The prior-posterior analysis is made possible by taking advantage of the nonparametric exponentially tilted empirical likelihood function, 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). We show that subject to a growth rate condition on the number of approximating functions, the posterior distribution satisfies the Bernstein-von Mises theorem, even when the set of conditional moments contain misspecified moment conditions. Large-sample theory for comparing different conditional moment models shows 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. Examples to illustrate the framework and results are provided.

Keywords: Conditional moment restrictions, Bayesian inference, Exponentially tilted empirical likelihood, Marginal likelihood, Posterior consistency.

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

†Department of Economics, University of Illinois, 214 David Kinley Hall, 1407 W. Gregory Dr., Urbana, IL 61801, e-mail: mincshin@illinois.edu.

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

1 Introduction

In this paper we extend Bayesian analysis under the empirical likelihood (for example, see Lazar (2003), Schennach (2005), Fang and Mukerjee (2006), Chang and Mukerjee (2008), Mengersen, Pudlo and Robert (2013), Vexler, Tao and Hutson (2014), Chib, Shin and Simoni (2018)), which deal with unconditional moment restrictions, to the class of problems that are specified by a set of *conditional* moment restrictions of the type

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

where $\rho(X,\theta)$ is a d-vector of known functions of a $\mathbb{R}^{d_x}$-valued random vector X and an unknown parameter vector θ, and P is the unknown conditional distribution of X given a $\mathbb{R}^{d_z}$-valued random vector Z. The parameter $\theta \in \Theta \subset \mathbb{R}^p$ is the parameter of interest. Conditional moment conditions, by virtue of being more informative than unconditional moment restrictions, lead to sharper posterior inferences for parameters and model comparisons. Interestingly, conditional moment restrictions arise in various settings, for example, in causal inference, as in the framework of Rosenbaum and Rubin (1983) where one assumes that the potential outcomes are independent of the treatment variable, conditioned on covariates and, for example, in missing at random problems as considered by Hristache and Patilea (2017). Inference by conditional moment restrictions avoids the full probability modeling of the outcomes that is required by the classical Bayesian machinery, whether by parametric or non-parametric Dirichlet process methods or its variants, and delivers Bayesian prior-posterior analyses under core, minimally supportable assumptions, yet retaining the semi-parametric viewpoint.

Our analysis of conditional moment models relies on the nonparametric exponentially tilted empirical likelihood (ETEL) family (Schennach, 2005) because of its better behavior under misspecification (Schennach, 2007), a central feature of the problems considered in this paper. Our development differs from the Bayesian analysis of related problems developed in Liao and Jiang (2011) (see also Florens and Simoni (2012, 2016) and Kato (2013)). Our analysis covers a broad range of models that considerably enlarge the category of problems that can be subjected to a formal Bayesian analysis, importantly, without the necessity of auxiliary modeling and prior assumptions that would arise in classical Bayesian treatments. As in the frequentist setting, for example, Donald, Imbens and Newey (2003) (see Kitamura, Tripathi and Ahn (2004) for a different approach), our procedure requires that the conditional moment conditions are first transformed into unconditional moment conditions. This transformation is made through approximating functions $q^K(Z) = (q_1^K(Z), \dots, q_K^K(Z))'$, such as tensor product splines from splines of each variable in Z, with the number of such functions, denoted by K, increasing with the sample size n at a certain rate. Thus, instead of (1.1), inference is based on the expanded unconditional moment conditions

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

where $\otimes$ is the Kronecker product operator. As $K \to \infty$, (1.1) and (1.2) are equivalent under mild assumptions. Given this equivalence, the posterior distribution, for each sample size, is based on the ETEL function that is constructed from these expanded moment conditions. This ETEL function can be interpreted in the current problem as a nonparametric likelihood that is consistent with the expanded moment restrictions.

In our theoretical analysis, we study the behavior of the sequence of posterior distributions as K increases with the sample size. The parameter K plays the role of a regularization parameter and our asymptotic theory supplies the rate at which K must increase to ensure that the asymptotic posterior variance achieves the semiparametric efficiency bound derived in Chamberlain (1987). Specifically, we prove that, under regularity conditions, as $K \to \infty$ with the sample size n at a certain rate the posterior distribution of θ satisfies the Bernstein von Mises (BvM) theorem with asymptotic posterior variance equal to the semiparametric efficiency bound. As a result, Bayesian credible sets are asymptotically valid efficient confidence sets. Because the number of unconditional moment conditions must increase with sample size, the details of the theory are different from those in Chib, Shin and Simoni (2018). This is primarily because quantities that are bounded with fixed moment restrictions, now diverge with K, and the rate of this divergence has to be determined to stabilize the growth.

We also provide generalizations of our theory for conditional moment models that are misspecified in the sense that the set of probability measures implied by the moment restrictions does not contain the true data generating process P for every $\theta \in \Theta$. For such models, which can be considered to be the norm in practical settings, we also establish a BvM-type phenomenon. We also develop the theory for comparing different conditional moment models with the aim of finding the model (in the set of contending models) that is closest in the Kullback-Leibler (KL) divergence to the true unknown distribution. The theory is based on the marginal likelihood of each competing model, which we estimate by the method of Chib (1995). Unlike Chib, Shin and Simoni (2018), it is not necessary to reformulate models to have the same number of conditional moment restrictions, as long as each contending model contains at least one misspecified condition. Because of this simplification, the different conditional models can be compared directly in terms of marginal likelihoods. Under regularity conditions, we establish the model-consistency of the sequence of models picked according to the largest value of the marginal likelihood. The theory shows that in the limit we select the model that is less misspecified, that is, the model that contains the smaller number of misspecified moment restrictions. This is also the model that is closet to the true distribution in the KL divergence.

The rest of the paper is organized as follows. Section 2 describes the conditional moment model with motivating examples. In Section 3 we discuss the construction of the sequence of unconditional moments, obtained from the conditional moments by an increasing (in sample size) vector of approximating functions. We then consider the prior-posterior analysis and the asymptotic behavior of the posterior distribution. Section 4 is concerned with the theory of model comparisons based on the large sample behavior of the marginal likelihood. Along with various running examples to illustrate our framework and results, Section 5 discusses issues that arise in higher dimensional problems and Section 6 provides applications of our techniques in two causal inference problems. Proofs of the theorems are included in the appendix, and in a 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 nonzero. 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 taken with respect to P and by $\mathbf{E}^P[\cdot|\cdot]$ the conditional expectation taken 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 following conditional moment restrictions

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

for all $z \in \text{supp}(Z) \triangleq \mathcal{Z}$, 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) Consider the following data generating process (DGP)*

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i, \quad \varepsilon_i = s(x_i)u_i, \tag{2.2}$$

where $(x_i, u_i)'$, $i \leq n$, are independently drawn from some distribution P. Suppose the researcher does not know the form of $s(x)$, the heteroskedasticity function, or the distribution of u_i. Then, the (conditional) moment restrictions implied by the model without knowing the form of the heteroskedasticity is

$$\mathbf{E}^P[(y_i - \theta_0 - \theta_1 x_i)|x_i] = 0 \tag{2.3}$$

where Z in the conditioning set is equal to the scalar x_i, $\rho(X, \theta) = (y_i - \theta_0 - \theta_1 x_i)$, and $d = 1$. Now suppose we want to impose conditional symmetry of ε_i. Then, our model is defined by the following two conditional moment conditions

$$\begin{aligned} \mathbf{E}^P[(y_i - \theta_0 - \theta_1 x_i)|x_i] = 0 \\ \mathbf{E}^P[(y_i - \theta_0 - \theta_1 x_i)^3|x_i] = 0 \end{aligned} \tag{2.4}$$

In this case $\rho(X, \theta)$ is (2×1) vector of functions, $X_{1,i} = y_i$, $X_{z,i} = x_i$ and the conditioning variable Z is the scalar x_i.

It is worth noting that the conditional moment model is different from the unconditional moment model. For example, one could start the Bayesian analysis in Example 1 based on weaker assumptions that ε_i is mean zero and uncorrelated with x_i. In such case, relevant unconditional moment conditions can be written as

$$\mathbf{E}^P[(y_i - \theta_0 - \theta_1 x_i) \otimes (1, x_i)'] = 0, \tag{2.5}$$

More generally, one could always transform a conditional moment restriction model into a set of unconditional moment conditions $\mathbf{E}^P[\rho(X, \theta) \otimes (1, Z')'] = 0$, but this is less informative than the conditional moment model.

3 Prior-Posterior analysis

3.1 Expanded Moment Conditions

Under certain circumstances (Bierens, 1982, Chamberlain, 1987), conditional moment restrictions are equivalent to a countable number of unconditional moment restrictions. The equivalent set of unconditional moments are obtained through approximating functions, similarly as in Donald, Imbens and

Newey (2003). 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 $\mathbf{E}^P[\rho(x,\theta_*) \otimes q^K(z)] \neq 0$, $\forall K$ large enough.

Thus, the prior-posterior analysis can be based on the expanded moment functions

$$g(W,\theta) := \rho(X,\theta) \otimes q^K(Z) \tag{3.1}$$

given the equivalence between the conditional moment restrictions and the limit of a sequence of unconditional moment restrictions.

Example 1 (continued) *Let $(\tau_1, \ldots, \tau_K)$ denote K knots, with the exterior knots τ_1 and τ_K taken to be the min and maximum values of the sample data $\boldsymbol{x} = (x_1, \ldots, x_n)$, and the interior knots taken to be 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 moment conditions for the n sample observations are the $n \times 2K$ conditions*

$$\mathbf{E}^P[\rho(\boldsymbol{x},\theta) \otimes q^K(\boldsymbol{z})] = \mathbf{E}^P\left[(\boldsymbol{y} - \theta_0 - \theta_1\boldsymbol{x}) \odot B \,\vdots\, (\boldsymbol{y} - \theta_0 - \theta_1\boldsymbol{x})^3 \odot B\right] = 0 \tag{3.2}$$

where $a \odot B$ is the operation in which the vector a is multiplied element by element into each column of the matrix B, $\vdots$ denotes matrix concatenation (column binding) and 0 is the matrix of zeros.

In our numerical examples, we use the natural cubic spline basis of Chib and Greenberg (2010) based on z_i to construct $q^K(z_i)$. In the case with one conditioning variable, we set $K \approx n^{1/3-1/24}$ (so that we have $K < n^{1/3}$) as suggested by the asymptotic analysis below. For example, when

$n = 500$, we set $K = 6$ and when $n = 2000$ we set $K = 9$. 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 $Z = (\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_1 \odot \boldsymbol{z}_2, \boldsymbol{z}_1 \odot \boldsymbol{z}_3, \boldsymbol{z}_2 \odot \boldsymbol{z}_3)$ 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 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 moment conditions, 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 $\boldsymbol{x}$. Then the expanded moment conditions for the sample observations are obtained by multiplying $\rho_l(\boldsymbol{x}, \theta)$ into the matrix B, and concatenating, as

$$\mathbf{E}^P[\rho(\boldsymbol{x}, \theta) \otimes q^K(\boldsymbol{z})] = \mathbf{E}^P \left[G(\boldsymbol{w}, \theta) \right] = 0$$

where

$$G(\boldsymbol{w}, \theta) = \left[\rho_1(\boldsymbol{x}, \theta) \odot B \vdots \rho_2(\boldsymbol{x}, \theta) \odot B \vdots \cdots \vdots \rho_d(\boldsymbol{x}, \theta) \odot B \right].$$

We use versions of this approach to construct the expanded moment conditions in our examples.

3.2 Prior-posterior analysis

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. Our prior on P is related to the one proposed by Schennach (2005) for the unconditional moment condition models, extended here to the case where the number of moment restrictions on P is not fixed, but increases with the sample size. Schennach (2005)'s nonparametric prior is a mixture of uniform probability densities, which is capable of approximating any distribution as the number of mixing components increases. Our modified prior on P, which is discussed in the online appendix, restricts P to satisfy the expanded moment restrictions $\mathbf{E}^P[g(W,\theta)] = 0$, given (θ, K).

As in the case of the unconditional moments problem, the posterior distribution of θ, after marginalization over the nonparametric prior on P, has the form

$$\pi(\theta|w_{1:n}, K) \propto \pi(\theta)p(w_{1:n}|\theta, K) \tag{3.3}$$

where

$$p(w_{1:n}|\theta, K) = \prod_{i=1}^{n} \widehat{p}_i(\theta) \tag{3.4}$$

is the Exponential Tilting (ET) empirical likelihood (ETEL) for a given K, and $\{\widehat{p}_i(\theta), i = 1, \ldots, n\}$ 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 satisfy the given unconditional moment conditions (3.1):

$$\max_{p_1,\ldots,p_n} \sum_{i=1}^{n} [-p_i \log(np_i)] \qquad \text{subject to: } \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i g(w_i, \theta) = 0, \quad p_i \geq 0. \tag{3.5}$$

These probabilities are computed conveniently from the dual (saddlepoint) representation as

$$\widehat{p}_i(\theta) := \frac{e^{\widehat{\lambda}(\theta)' g(w_i,\theta)}}{\sum_{j=1}^{n} e^{\widehat{\lambda}(\theta)' g(w_j,\theta)}} \quad (i = 1, \ldots, n) \tag{3.6}$$

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. Therefore, on multiplying the ETEL function by the prior density of θ, the posterior distribution takes the form

$$\pi(\theta|w_{1:n}, K) \propto \pi(\theta) \prod_{i=1}^{n} \frac{e^{\widehat{\lambda}(\theta)' g(w_i,\theta)}}{\sum_{j=1}^{n} e^{\widehat{\lambda}(\theta)' g(w_j,\theta)}}. \tag{3.7}$$

Efficient simulation of θ from this posterior distribution is possible, in small dimension problems, with the one block tailored Metropolis-Hastings (M-H) algorithm of Chib and Greenberg (1995), and, in larger dimension problems, by the Tailored Randomized Block MH algorithm of Chib and Ramamurthy (2010).

Example 1 (continued) *To illustrate the prior-posterior analysis, we create a set of simulated data* $\{y_i, x_i\}_{i=1}^{n}$ *from the regression model in 2.2 with covariates* $x_i \sim \mathcal{U}(-1, 2.5)$*, intercept* $\theta_0 = 1$*, slope* $\theta_1 = 1$*, and* ε_i *is distributed according to*

$$\varepsilon_i \sim \mathcal{SN}(m(x_i), s(x_i), w(x_i)) \tag{3.8}$$

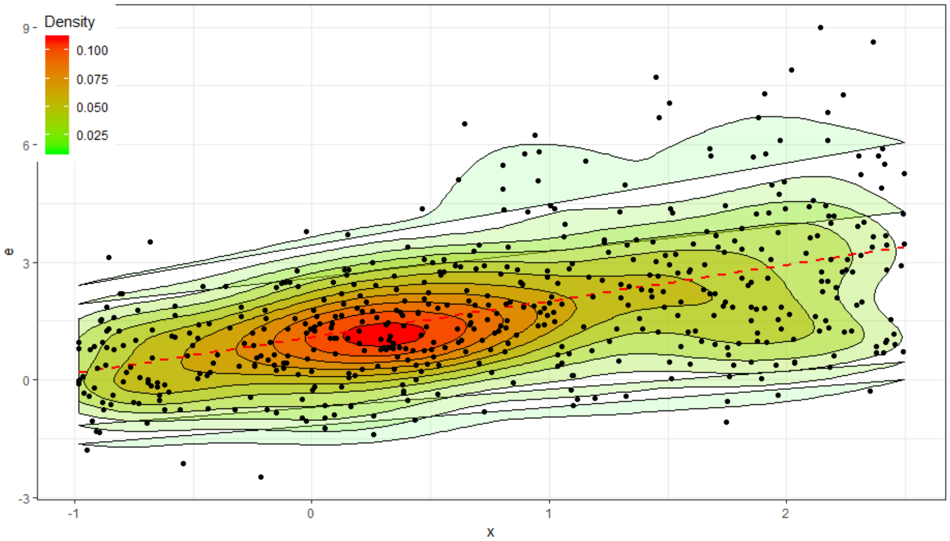


Figure 1: Scatter plot of (x_i, ε_i). Red dashed line is a regression line. Black dots represent realizations of (x_i, ε_i). Contour lines based on the joint density function of (x_i, ε_i) are presented in the figure.

where $\mathcal{SN}(m, s, w)$ is the skew normal distribution with location, scale, and shape parameter given by (m, s, w), each depending on x_i. When w is zero, ε_i is a normal distribution with mean m and standard deviation s. We set $m(x_i) = -s(x_i)\sqrt{2/\pi}w(x_i)/(\sqrt{1 + w(x_i)^2})$ so that $E^P[\varepsilon_i|x_i] = 0$.

As an illustration we generate a set of data $\{y_i, x_i\}_{i=1}^n$ with $s(x_i) = \sqrt{\exp(1 + 0.7x_i + 0.2x_i^2)}$ and $w(x_i) = 1 + x_i^2$. Under this setup ε_i is conditionally heteroscedastic and asymmetric. A sample of 500 realizations of (x_i, ε_i) are presented in Figure 1.

Note that under this model, $E^P[\varepsilon|x] = 0$, and inferences about (θ_0, θ_1) require just this core restriction, without the need to model the heteroskedasticity or the skewness functions. We create the expanded moment conditions in (3.2), with $\rho(X, \theta) = (y - \theta_0 - \theta_1 x)$. 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 distribution of θ_0 and θ_1 are summarized in the panel (a) of Table 1 for two different sample sizes. We note from the dispersion of the posterior distribution, that the posterior distribution 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 (2.5), 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): $\mathbf{E}^P[\varepsilon\|x] = 0$							
		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 vs unconditional moments. 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. Inference in the top panel is based on the single conditional moment restriction; inference in the bottom panel is based on two unconditional moment restrictions. Results are based on 20,000 MCMC draws beyond a burn-in of 1000. The M-H acceptance rate is around 95% in both cases. "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

In this section, we study asymptotic properties of the posterior distribution of θ from a frequentist point of view. This means that we admit the existence of a true value θ_* of the parameter of interest θ and a true value P_* of the data distribution P. When we are using the true distribution P_*, $\mathbf{E}^P[\cdot]$ (resp. $\mathbf{E}^P[\cdot|\cdot]$) has to be understood as the expectation (resp. conditional expectation) 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, $\ell_{n,\theta}(w_i) := \log \widehat{p}_i(\theta)$.

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 bounded.*

This assumption is the same as Donald, Imbens and Newey (2003, Assumption 3). Part (d) of this assumption imposes a Lipschitz condition which, together with part (c), allows to apply uniform convergence results. The following three assumptions are also the same as the ones required by Donald, Imbens and Newey (2003) to establish asymptotic normality of the Generalized Empirical Likelihood 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 bounded; (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 bounded, 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 $\zeta(K)^2K/n^{1-2/\gamma} \to 0$.*

The last assumption is about the prior distribution of θ and is standard in Bayesian literature establishing 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_*.*

Let $\pi(\sqrt{n}(\theta - \theta_*)|w_{1:n})$ denote the posterior distribution of the local parameter $h := \sqrt{n}(\theta - \theta_*)$. We are now able to state our first major result in which we establish the asymptotic normality and efficiency of the posterior distribution of θ.

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 $n \to \infty$*

$$P\left(\sup_{\|\theta - \theta_*\| > \delta} \frac{1}{n} \sum_{i=1}^{n} \left(\ell_{n,\theta}(w_i) - \ell_{n,\theta_*}(w_i) \right) \leq -\epsilon \right) \to 1, \tag{3.9}$$

then the posterior distribution of the local parameter h converges in total variation towards a random Normal distribution, that is,

$$\sup_B \left| \pi(\sqrt{n}(\theta - \theta_*) \in B | w_{1:n}) - \mathcal{N}_{\Delta_{n,\theta_*}, V_{\theta_*}}(B) \right| \xrightarrow{p} 0 \tag{3.10}$$

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. On the other hand, it is slightly stronger than the condition $\zeta(K)K/\sqrt{n} \to 0$ required by Donald, Imbens and Newey (2003) to establish asymptotic normality of the Generalized Empirical Likelihood 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 Appendix and consists of three steps. In the first step we show consistency of the posterior distribution of θ, namely:

$$\pi\left(\sqrt{n}\|\theta - \theta_*\| > M_n \middle| w_{1:n} \right) \xrightarrow{p} 0 \tag{3.11}$$

for any $M_n \to \infty$, as $n \to \infty$. To show this, the identification assumption (3.9) is used. In the second step we show that the ETEL function satisfies a stochastic Local Asymptotic Normality (LAN) expansion:

$$\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\ell_{n,\theta_*+h/\sqrt{n}}(w_i)-\sum_{i=1}^{n}\ell_{n,\theta_*}(w_i)-h'V_{\theta_*}^{-1}\Delta_{n,\theta_*}-\frac{1}{2}h'V_{\theta_*}^{-1}h\right|=o_p(1) \tag{3.12}$$

where $\mathcal{H}$ denotes a compact subset of $\mathbb{R}^p$ and $V_{\theta_*}^{-1}\Delta_{n,\theta_*}\overset{d}{\to}\mathcal{N}(0,V_{\theta_*}^{-1})$. In the third step of the proof we use standard arguments, see *e.g.* the proof of Van der Vaart (1998, Theorem 10.1), to show that (3.11) and (3.12) imply asymptotic normality of $\pi(\sqrt{n}(\theta-\theta_*)\in B|w_{1:n})$. While these three steps are classical in proving Bernstein-von Mises phenomenon, here the main difficulty consists in showing (3.12) because the ETEL function is a nonstandard likelihood function involving estimated parameters whose dimension increases with K. Therefore, we first need to determine the rate of $\|\widehat{\lambda}\|$, $\|\frac{1}{n}\sum_{i=1}^{n}g(w_i,\theta)\|$ and of the norms of the empirical counterparts of $D(z)$, $\Sigma(z)$. For instance, the tilting parameter $\widehat{\lambda}(\theta)$ has dimension dK, where K increases with n. Therefore, 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 Appendix we show that $\|\widehat{\lambda}(\theta_*)\|=O_p(\sqrt{K/n})$ under the previous assumptions.

3.4 Misspecified model

We now turn our attention to conditional moment condition models that are misspecified. By misspecified conditional model we mean the following.

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 every $\theta\in\Theta$, 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:

$$\lambda_\circ:=\lambda_\circ(\theta):=\arg\min_{\lambda\in\mathbb{R}^{dK}}\mathbf{E}^P[e^{\lambda' g(w_i,\theta)}]. \tag{3.13}$$

However, under misspecification the dual theorem is not guaranteed to hold. In fact, when the model is misspecified, the probability measures in $\mathcal{P} := \bigcup_{\theta \in \theta} \mathcal{P}_\theta$, which are implied by the model, might not have a common support with the true P_*, see Sueishi (2013) for a discussion on this point. Following Sueishi (2013, Theorem 3.1), in order to guarantee validity of the dual theorem we introduce the following assumption. This assumption replaces Assumption 3.3 (a) in misspecified models.

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 et al. (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 $Q^*(\theta)$:

$$\theta_\circ := \text{arginf}_{\theta \in \Theta} K(P_* || Q^*(\theta)) \tag{3.14}$$

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

$$\theta_\circ = \text{argmax}_{\theta \in \Theta} \mathbf{E}^P \log \left(\frac{e^{\lambda_\circ' g(w_i,\theta)}}{\mathbf{E}^P[e^{\lambda_\circ' g(w_j,\theta)}]} \right). \tag{3.15}$$

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 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* (3.15).

In the next assumption we denote by $int(\Theta)$ the interior of Θ and by $\mathcal{U}$ a ball centred 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* (3.13)*;*

(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 bounded;*

(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) let $\mathcal{H}$ *be 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.

We are now in a position to state our next important theorem, the Bernstein - von Mises theorem for misspecified models.

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 $M_n \to \infty$,*

$$P\left(\sup_{\|\theta-\theta_\circ\|>M_n/\sqrt{n}} \frac{1}{n}\sum_{i=1}^{n}\left(\ell_{n,\theta}(w_i) - \ell_{n,\theta_\circ}(w_i)\right) \leq -CM_n^2/n\right) \to 1, \tag{3.16}$$

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,

$$\sup_B \left|\pi(\sqrt{n}(\theta - \theta_\circ) \in B|w_{1:n}) - \mathcal{N}_{\Delta_{n,\theta_\circ}, \mathcal{A}_{\theta_\circ}^{-1}}(B)\right| \xrightarrow{p} 0 \tag{3.17}$$

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 (B.20) in the 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 buy 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 Comparison: Misspecified models

The comparison of different moment models, with the aim of selecting the best model, is central in practice (for example, Vexler, Deng and Wilding (2013), Vexler, Yu and Lazar (2017), Chib, Shin and Simoni (2018)). In this section we consider this question for conditional moment models. Our idea is to screen these different models in terms of the marginal likelihood of each competing model, selecting the model with the largest marginal likelihood.

We show that such a marginal likelihood selection criterion selects the model that is the closest to the true distribution P_* in terms of the Kullback-Leibler (KL) divergence. Thus, if all the models are misspecified the marginal likelihood selection procedure isolates the less misspecified one. Because the KL divergence is zero if and only if the true distribution belongs to the model considered, our procedure can determine, up to statistical error, which model is correctly specified if one knows that there is a correctly specified model.

Comparison of conditional moment condition models differs in one important respect from the framework for comparing unconditional moment condition models that was established in Chib et al. (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.

Let M_ℓ denote the ℓ-th model in the comparison set of models. 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 (3.7). Then, we select the model with the largest marginal likelihood, denoted by $m(w_{1:n}|M_\ell)$, which we calculate by the method of Chib (1995) as extended to Metropolis-Hastings samplers in Chib and Jeliazkov (2001). This method makes computation of the marginal likelihood simple and is a key feature of our procedure. The main advantage of the Chib (1995) method is that it is calculable from the same inputs and outputs that are used in the MCMC sampling of the posterior distribution. The starting point of this method is the following identity of the log-marginal likelihood introduced in Chib (1995):

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

where $\tilde{\theta}^\ell$ is any point in the support of the posterior (such as the posterior mean) and the dependence of the terms on the right hand side on the model M_ℓ has been made explicit. The first two terms on the right-hand side of this decomposition are available directly whereas the third term can be estimated from the output of the MCMC simulation of the posterior distribution.

4.1 Model selection consistency results

In this section we establish the consistency of our marginal likelihood based selection procedure. This result is stated in Theorem 4.1 below. It states that if we compare J misspecified models, then the marginal likelihood based selection procedure selects the model with the smallest KL divergence $K(P||Q^*(\theta^\ell))$ between P and $Q^*(\theta^\ell)$, 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. Under Assumption 3.8, $dQ^*(\theta^\ell)/dP = e^{\lambda_\circ(\theta)' A(X,\theta)} / \mathbf{E}^P\left[e^{\lambda_\circ(\theta)' A(X,\theta)}\right]$ by the dual theorem, as defined in Section 3.4. Because the I-projection $Q^*(\theta^\ell)$ on $\mathcal{P}_{\theta^\ell}$ is unique (Csiszar (1975)), which $Q^*(\theta^\ell)$ is closer to P (in terms of $K(P||Q^*(\theta^\ell))$) depends only on the "amount of misspecification" contained in each model $\mathcal{P}_{\theta^\ell}$. We then have the following theorem.

Theorem 4.1 *Let the assumptions of Theorem 3.2 hold. Let us consider the comparison of $J < \infty$ models M_j, $j = 1, \ldots, J$ that each has at least one misspecified moment, that is, M_j does not satisfy Assumption 3.3 (a), $\forall j$. 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 larger marginal likelihood and will be selected by our procedure.

4.2 Remarks and examples

We now explore and explain the various ramifications of the preceding result, which covers both nested and non-nested models. Focusing on the more general non-nested situation, the first setting is one where we have misspecified models that involve different moment conditions and different conditioning variables:

$$\text{Model 1: } \mathbf{E}^P[\rho_1(X_1,\theta_1)|Z_1]=0, \qquad \text{Model 2: } \mathbf{E}^P[\rho_2(X_2,\theta_2)|Z_2]=0, \tag{4.2}$$

where X_1 and X_2 (resp. θ_1 and θ_2) either have all or some or none elements in common and Z_1 and Z_2 may have some elements in common.

Example 2 *(Variable selection, comparing misspecified and non-nested models). Suppose that we compare the following two models*

$$\mathcal{M}_1 : E[(y_i-\theta_0-\theta_2 x_{2,i})|x_{2,i}]=0$$
$$\mathcal{M}_2 : E[(y_i-\theta_0-\theta_3 x_{3,i})|x_{3,i}]=0$$

where the two covariates $x_{2,i}$ and $x_{3,i}$ are competing against each other. We generate data from the following mechanism

$$y_i = 1 + x_{2,i}^2 + x_{3,i}^3 + \varepsilon_i, \quad E[\varepsilon_i|x_{1,i},x_{2,i},x_{3,i}]=0.$$

with x_2 entering as a square and x_3 as a cube. Because relevant covariates in both $\mathcal{M}_1$ and $\mathcal{M}_2$ enter linearly, both models are misspecified, but $\mathcal{M}_2$ is more misspecified. We generate $x_i=[x_{1,i},x_{2,i},x_{3,i}]'$ from a correlated Gaussian copula with the same uniform marginals, $\mathcal{U}[-1,2.5]$. Diagonal elements of the covariance matrix for this Gaussian copula are all 1 and off-diagonal elements are cov(x1,x2) = cov(x1,x3) = 0.5, and cov(x2,x3) = 0. The error ε_i is identically and independently drawn from the skew normal distribution with the following covariate-dependent parameters,

$$m(x_i) = -s(x_i)\sqrt{2/\pi}\frac{w(x_i)}{\sqrt{1+w(x_i)^2}}$$
$$s(x_i) = \sqrt{\exp(1+0.2x_{1,i}^2)}\ ,\ w(x_i)=1+x_{1,i}^2.$$

In Table 2 we present posterior summaries based on both $\mathcal{M}_1$ and $\mathcal{M}_2$, and the corresponding marginal likelihood of both models. $\mathcal{M}_1$ has lower marginal likelihood, therefore, it is less misspecified than $\mathcal{M}_2$ according to our theory.

Table 2: Variable selection, $n = 500$

Comparing misspecified and non-nested models

$\mathcal{M}_1$: ML=-3134.75						
	Mean	SD	Median	Lower	Upper	Ineff
θ_0	4.60	0.20	4.59	4.28	4.95	1.35
θ_2	1.57	0.13	1.57	1.35	1.79	1.26
$\mathcal{M}_2$: ML=-3444.08						
	Mean	SD	Median	Lower	Upper	Ineff
θ_0	3.50	0.07	3.50	3.38	3.61	1.36
θ_3	3.22	0.06	3.22	3.12	3.31	1.12

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

The second situation that we consider is one in which we have different moment conditions but the same conditioning variables:

$$\text{Model 1: } \mathbf{E}^P[\rho_1(X_1, \theta_1)|Z] = 0, \qquad \text{Model 2: } \mathbf{E}^P[\rho_2(X_2, \theta_2)|Z] = 0, \tag{4.3}$$

where X_1 and X_2 (resp. θ_1 and θ_2) either have all or some or none elements in common. An example of this is the case where we are unsure about which covariate we have to include in a regression model, as discussed in the following example.

Example 3 *(Variable selection, selecting a correct model). Consider a linear regression model with two explanatory variables*

$$y_i = \theta_0 + \theta_1 x_{1,i} + \theta_2 x_{2,i} + \varepsilon_i, \quad E[\varepsilon_i | x_{1,i}, x_{2,i}] = 0$$

where $\theta = [\theta_0, \theta_1, \theta_2]'$. Suppose that we are interested in whether $x_{2,i}$ should be included or not. To this end we can compare the following two models

$$\begin{aligned} \mathcal{M}_1 :& E[(y_i - \theta_0 - \theta_1 x_{1,i})|x_{1,i}, x_{2,i}] = 0 \\ \mathcal{M}_2 :& E[(y_i - \theta_0 - \theta_1 x_{1,i} - \theta_2 x_{2,i})|x_{1,i}, x_{2,i}] = 0. \end{aligned}$$

Note that $\mathcal{M}_1$ implies that $E[(y_i - \theta_0 - \beta_1 x_{1,i})|x_{1,i}] = 0$ meaning that the $x_{2,i}$ does not influence the conditional mean of y_i conditional on $x_{1,i}$. Applying our preceding theorem to this problem, when the true θ_2 is zero, then the marginal likelihood is higher for $\mathcal{M}_1$. When the true θ_2 is not zero, then the marginal likelihood is higher for $\mathcal{M}_2$. If both models are wrong, the marginal likelihood selects the model with lower KL divergence.

To see these results in action, we generate sample data on $x_i = [x_{1,i}, x_{2,i}]'$ from a correlated Gaussian copula with the same uniform $\mathcal{U}[-1, 2.5]$ marginal distribution. Now, let us consider two cases. In Case 1, we assume that $\theta = (\theta_0, \theta_1, \theta_2) = [1, 1, 0]'$, and

$$\varepsilon_i \sim \mathcal{SN}(m(x_i), s(x_i), w(x_i))$$

where $\mathcal{SN}(m, s, w)$ is the skew normal distribution with location, scale, and shape parameters specified as follows:

$$\begin{aligned} m(x_i) &= -s(x_i)\sqrt{2/\pi}\frac{w(x_i)}{\sqrt{1 + w(x_i)^2}} \\ s(x_i) &= \sqrt{\exp(1 + 0.5x_{1,i} + 0.1x_{1,i}^2)}, \qquad w(x_i) = 1 + x_{1,i}^2. \end{aligned}$$

In Case 1, $\mathcal{M}_1$ is expected to have higher marginal likelihood. From Table 3 we see that this is what occurs. Table 3 also presents posterior summaries and the corresponding marginal likelihood of models $\mathcal{M}_1$ and $\mathcal{M}_2$.

Table 3: Variable selection (Case 1), $n = 500$

Case 1						
$\mathcal{M}_1$: ML=-3122.68						
	Mean	SD	Median	Lower	Upper	Ineff
θ_0	1.148	0.063	1.149	1.027	1.271	1.108
θ_1	1.063	0.064	1.063	0.937	1.187	1.073
$\mathcal{M}_2$: ML=-3126.73						
	Mean	SD	Median	Lower	Upper	Ineff
θ_0	1.152	0.067	1.152	1.018	1.282	1.167
θ_1	1.073	0.098	1.074	0.883	1.265	1.151
θ_2	-0.013	0.096	-0.014	-0.200	0.172	1.108

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

In Case 2, we assume that $\theta = (1, 1, 0.5)'$, and

$$\begin{aligned} m(x_i) &= -s(x_i)\sqrt{2/\pi}\frac{w(x_i)}{\sqrt{1 + w(x_i)^2}} \\ s(x_i) &= \sqrt{\exp(1 + 0.5x_{1,i} + 0.1x_{1,i}^2 + 0.3x_{2,i})}\ ,\ w(x_i) = 1 + x_{1,i}^2 + 0.5x_{2,i} \end{aligned}$$

In this case, $\mathcal{M}_2$ is expected to have the higher marginal likelihood, precisely what we see in Table 4.

The third situation is the one where we have the same moment conditions but different conditioning variables:

$$\text{Model 1: } \mathbf{E}^P[\rho(X, \theta)|Z_1] = 0, \qquad \text{Model 2: } \mathbf{E}^P[\rho(X, \theta)|Z_2] = 0, \tag{4.4}$$

Table 4: Variable selection (Case 2), $n = 500$

Case 2						
$\mathcal{M}_1$: ML=-3133.14						
	Mean	SD	Median	Lower	Upper	Ineff
θ_0	1.281	0.059	1.281	1.167	1.396	1.111
θ_1	1.455	0.071	1.454	1.316	1.592	1.055
$\mathcal{M}_2$: ML=-3125.88						
	Mean	SD	Median	Lower	Upper	Ineff
θ_0	1.180	0.066	1.180	1.049	1.308	1.159
θ_1	1.112	0.102	1.113	0.912	1.313	1.115
θ_2	0.459	0.098	0.458	0.267	0.651	1.116

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

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). An example of this is when we are unsure about the validity of instrumental variables in an instrumental regression model, as discussed in the following example.

Example 4 *(Comparing IV models) Consider the following model with three instruments* (z_1, z_2, z_3)*:*

$$\begin{aligned} y_i &= \theta_0 + \theta_1 x_i + e_{1,i} \\ x_i &= f(z_{1,i}, z_{2,i}, z_{3,i}) + e_{2,i} \\ z_{1,i} &\sim U[0,1] \text{ and } z_{2,i} \sim U[0,1] \text{ and } z_3 \sim \mathcal{B}(0.4) \end{aligned}$$

where $(e_{1,i}, e_{2,i})'$ *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* z_j*'s are relevant instruments, that is, cov*$(x_i, z_{j,i}) \neq 0$ *for* $j \leq 3$*, and*

$$f(z_{1,i}, z_{2,i}, z_{3,i}) = 6(\sqrt{0.3}z_{1,i} + \sqrt{0.7}z_{2,i})^3(1 - \sqrt{0.3}z_{1,i} - \sqrt{0.7}z_{2,i})z_{3,i} + z_{1i}z_{2,i}(1 - z_{3,i}). \quad (4.5)$$

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 $(e_{1,i}, e_{2,i}, z_{1,i})$ *from a Gaussian copula whose covariance matrix is*

$$\Sigma = \begin{bmatrix} 1 & 0.7 & 0.7 \\ 0.7 & 1 & 0 \\ 0.7 & 0 & 1 \end{bmatrix}$$

such that the marginal distribution of $e_{1,i}$ *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* $e_{2,i}$ *is* $\mathcal{N}(0, 1)$*. Under this setup,* z_1 *is now an* *invalid instrument. We consider the following three models*

$$\mathcal{M}_1 : \mathbf{E}^P[(y_i - \theta_0 - \theta_1 x_i)|z_{1,i}, z_{2,i}, z_{3,i}] = 0 \tag{4.6}$$

$$\mathcal{M}_2 : \mathbf{E}^P[(y_i - \theta_0 - \theta_1 x_i)|z_{1,i}, z_{3,i}] = 0 \tag{4.7}$$

$$\mathcal{M}_3 : \mathbf{E}^P[(y_i - \theta_0 - \theta_1 x_i)|z_{2,i}, z_{3,i}] = 0. \tag{4.8}$$

Because $z_{1,i}$ is invalid instrument, $\mathcal{M}_1$ and $\mathcal{M}_2$ are wrong.

For purpose of inference about θ for $\mathcal{M}_1$, 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 case of $\mathcal{M}_1$.

Table 5 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 5: 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 Empirical Issues in Higher Dimensions

In this section, we discuss two problems that arise in higher dimensions. One is the question of variable selection with a large set of covariates and the other is posterior simulation when the dimension of θ is large.

5.1 Model search

Suppose we have a large set of covariates, but only a small subset of regressors is assumed to be active. That is, the model is sparse. In this situation, one can conduct a "sparsity-endowed model search" by which we mean the comparison of all models that contain at most L covariates. For example, when we have 10 potential covariates and $L = 5$, the sparsity-endowed model search would involve the marginal likelihood based comparison of $\sum_{l=1}^{5} \binom{10}{l} = 2,560$ models. We illustrate this search with the IV regression example.

Example 4 (continued) *(Selecting exogenous regressors in IV regression). Consider the IV model with additional exogenous regressors* ($w = [w_1, w_2, ..., w_{10}]'$)*:*

$$y_i = \theta_0 + \theta_1 x_i + w_i'\gamma + e_{1,i}$$

where $w_{k,i} \sim_{i.i.d.} \mathcal{N}(0,1)$*,* $E[e_{1,i}w_i] = 0$*, and under the true* P*,* $\gamma_1 = \gamma_2 = \gamma_3 = 1$ *and all other* γ_k*'s are zero. The other components of the model are the same as before in the previous Example 2. Then* $\mathcal{M}_3$ *with these additional exogenous regressors can be represented as*

$$\mathbf{E}^P[(y_i - \theta_0 - \theta_1 x_i - w_i'\gamma)|z_{2,i}, z_{3,i}] = 0 \quad \textit{and} \quad \mathbf{E}^P[(y_i - \theta_0 - \theta_1 x_i - w_i'\gamma)w_i] = 0.$$

To illustrate our model selection strategy we set the maximum number of possible covariates at $L = 5$*. For each model we compute the marginal likelihood based on 20,000 MCMC draws following a burn-in of 1,000 MCMC cycles. We set a prior for* γ_k*'s as independent Student-t distribution with mean zero, dispersion 5, and degrees of freedom equal to 2.5. Table 6 presents the model with the highest marginal likelihood among the models with the same number of covariates. As can be seen, our marginal likelihood based model selection strategy is able to locate the true model, that is, to select the correct number of active covariates as well as the correct identity of the active covariates.*

5.2 TaRB-MH

The examples in the paper thus far have intentionally limited the dimension of θ in order to focus on the theoretical aspects of our method. In this section, we present an example where the size of θ is considerably larger. Instead of simulating the posterior distribution with the one-block M-H algorithm, which now, with large size of θ, tends to becomes less efficient (in the simulation sense), we employ the Tailored Randomized Block MH algorithm of Chib and Ramamurthy (2010) to sample the posterior

	Group 1	Group 2	Group 3	Group 4	Group 5
# of models	10	90	360	840	1260
Best model	$\{w_9\}$	$\{w_9, w_{10}\}$	$\{w_1, w_2, w_3\}$	$\{w_1, w_2, w_3, w_9\}$	$\{w_6, w_7, w_8, w_9, w_{10}\}$
Marginal Likelihood	-3142.21	-3141.65	**-3137.57**	-3141.02	-3144.59
	(0.021)	(0.034)	(0.010)	(0.016)	(0.051)

Table 6: Model comparison ($n = 500$): IV regression example with additional covariates. Group l is the set of models with l number of covariates. # of models indicates the total number of models based on l covariates. There are 2,560 models in total. Best model presents the combination of covariates that are selected by marginal likelihood. The summaries are based on 20,000 MCMC draws beyond a burn-in of 1,000. Numerical standard errors are in parenthesis.

distribution. The TaRB-MH algorithm has proved useful in several other highly non-linear settings. As we now show by way of an example, it is similarly useful in sampling the posterior distribution of the conditional moment model. To conserve space, we suppress the details of how this algorithm works since these follow closely the implementation given in the preceding source paper.

Example 4 (continued) *(IV regression with additional exogenous regressors). We generate 1,000 observations from the previous IV regression model with additional exogenous regressors*

$$y_i = \theta_0 + \theta_1 x_i + w_i'\gamma + e_{1,i}$$

where both $[w_{1,i}, w_{2,i}, w_{3,i}, w_{4,i}, w_{5,i}]'$ and $[w_{6,i}, w_{7,i}, w_{8,i}, w_{9,i}, w_{10,i}]'$ are identically and independently drawn from $\mathcal{N}(0, \Sigma(\rho))$ where $\Sigma(\rho)$ is 5×5 matrix. Diagonal elements in $\Sigma(\rho)$ are set to one and off-diagonal elements are set to ρ. We set $\rho = 0.9$ and $\gamma_k = 1$ for all k. Other elements of the DGP are unchanged. Table 7 presents the posterior summaries of the posterior distribution based on the same conditional moment conditions, based on the one-block and TaRB-MH samplers. It is evident that the TaRB-MH sampler dominates the one-block MH sampler in terms of simulation efficiency as measured by the inefficiency factor, the ratio of the numerical variance of the mean to the variance of the mean assuming independent draws: an inefficiency factor close to 1 indicates that the MCMC draws, although serially correlated, are essentially independent.

6 Moment-based Causal Inference

In this section we illustrate the application of our techniques in the estimation of causal parameters in two problems, the average treatment effect (ATE) estimation under a conditional independence as-

	One-block-MH			TaRB-MH		
	Mean	SD	Ineff	Mean	SD	Ineff
θ_0	1.00	0.04	60.30	1.00	0.04	2.55
θ_1	0.95	0.14	128.40	0.95	0.13	4.11
γ_1	1.00	0.03	31.06	0.99	0.03	4.68
γ_2	0.98	0.10	21.13	0.98	0.09	3.04
γ_3	0.85	0.10	11.40	0.85	0.10	2.02
γ_4	1.19	0.11	16.16	1.19	0.10	2.31
γ_5	1.04	0.10	23.77	1.04	0.10	1.59
γ_6	1.02	0.03	33.43	1.02	0.03	2.57
γ_7	0.96	0.10	74.61	0.97	0.09	1.88
γ_8	1.01	0.10	71.14	1.01	0.09	1.98
γ_9	0.92	0.10	26.78	0.92	0.10	3.13
γ_{10}	1.00	0.10	42.85	1.00	0.10	2.33

Table 7: Posterior summary of IV regression example with additional covariates ($n = 1000$). The true value of all parameters (θ's and γ's) are set to one. The summaries are based on 50,000 MCMC draws beyond a burn-in of 10,000 for the one-block-MH sampler and 5,000 draws beyond a burn-in of 1,000 for the TaRB-MH. The M-H acceptance rate is around 52% for the one-block-MH and 93% for TaRB-MH. The average size of blocks in each iteration for the TaRB-MH is around 6.5. "Ineff" is the inefficiency factor.

sumption, and the regression-discontinuity (RD) ATE estimation under a sharp-design.

6.1 Average treatment effect (ATE) estimation

A standard problem in causal inference with a binary treatment $x_i \in \{0, 1\}$, for control and treated, respectively, and covariates $z_i : d \times 1$ assumes that the two potential outcomes y_{i0} and y_{i1} for n randomly chosen subjects satisfy the conditional independence assumption (Rosenbaum and Rubin, 1983)

$$(y_{i0}, y_{i1}) \perp x_i | z_i.$$

If we let $\mathbf{E}^P(y_{i1}|z_i) - \mathbf{E}^P\,(y_{i0}|z_i)$ denote the ATE conditional on z_i for the ith subject, then the goal of the analysis is to calculate the ATE given by

$$\text{ATE} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{E}^P(y_{i1}|z_i) - \mathbf{E}^P\,(y_{i0}|z_i)\right).$$

We show that the conditional moment technique developed above is ideally suited for calculating the posterior distribution of this quantity, under minimal assumptions. We just need to make assumptions about the conditional expectations $\mathbf{E}^P(y_{ij}|z_i)$ $(j=0,1)$ without specifying (or restricting) the conditional distributions of $y_{ij}|z_i$ in any further way. For illustration, suppose that

$$\mathbf{E}^P(y_{ij}|z_i) = z_i'\beta_j \ , \ j=0,1.$$

Also suppose that there are n_0 control subjects, and that the data are organized such that the observations $i \leq n_0$ are the data on the controls, and the observations $i > n_0$ are the data on the treated. Then, the latter conditional expectations imply that estimation of β_0 can be based on the conditional moment conditions

$$\mathbf{E}^P((y_{i0} - z_i'\beta_0)|z_i) = 0 \ \ (i \leq n_0)$$

since y_{i0} is observed for such subjects, and that independently, estimation of β_1 can be based on the conditional moment conditions

$$\mathbf{E}^P((y_{i1} - z_i'\beta_1)|z_i) = 0 \ \ (i > n_0)$$

since y_{i1} is observed for these subjects. Now, suppose that our prior-posterior analysis is applied to these sets of moment conditions to produce the MCMC samples

$$\{\beta_0^g\}_{g=1}^M \sim \pi(\beta_0|\left\{y_{i0}, z_i\right\}_{i=1}^{n_0}) \ \text{ and } \ \{\beta_1^g\}_{g=1}^M \ \sim \pi(\beta_1|\left\{y_{i1}, z_i\right\}_{i>n_0}).$$

Then, the Bayes posterior sample of the ATE is given by the sequence of values

$$\text{ATE}^{(g)} = \frac{1}{n}\sum_{i=1}^n \left(z_i'\beta_1^{(g)} - z_i'\beta_0^{(g)}\right), \qquad g = 1, 2, \ldots, M.$$

As an illustration of this approach, consider n = 500, 1000 and 2000 observations generated from the following DGP. First, suppose that z_1 and z_2 are generated from a Gaussian copula whose covariance matrix has 1 on the diagonal, and 0.7 on the off-diagonal, such that the marginal distribution of z_1 is uniform on $(0,1)$ and the marginal distribution of z_2 is standard normal. Next, conditional on (z_1, z_2), suppose that x is generated as independent Bernoulli

$$x \sim \mathcal{B}(p)$$

where the propensity score, the probability p of being treated, is given by

$$p = \Phi\left(0.5(\sqrt{0.3}z_{1,i} + \sqrt{0.7}z_{2,i})^3(1 - \sqrt{0.3}z_{1,i} - \sqrt{0.7}z_{2,i})\right)$$

and $\Phi(\cdot)$ is the cdf of the standard normal. Finally, suppose that the potential outcomes for each individual in the sample are given by

$$\begin{aligned} y_0 &= 10 + z_1 + 1.5z_2 + \varepsilon_0, \\ y_1 &= 10 + 1.5z_1 - z_2 + \varepsilon_1 \end{aligned} \tag{6.1}$$

where the conditional distribution of ε_j is skewed normal with conditional variance and conditional skewness depending on $z = (z_1, z_2)$. In particular,

$$\varepsilon_j \sim \mathcal{SN}(m_j(z), s_j(z), w_j(z)) \tag{6.2}$$

where

$$s_0(z) = \exp\left(0.5\left(1 + .5z_1 + 1z_1^2 + .3z_2\right)\right), \qquad w_0\left(z\right) = 1 + z_1^2 + .5z_2$$

and

$$s_1(z) = \exp\left(0.5\left(1 + z_1 + .2z_1^2 + .3z_2\right)\right), \qquad w_1\left(z\right) = 1 + z_1^2 + z_2$$

and $m_j(z)$ is fixed based on these functions to ensure that $E(\varepsilon_j|z) = 0$. The observed data is $y = xy_1 + (1-x)y_0$.

There are approximately 42 percent treated subjects that emerge from this design. Also note that, because of the extreme nonlinearity of the propensity score function, standard propensity score matching does not perform well with data generated from this design. In addition, any method that is based on direct modeling of the outcome distributions that is not robust to covariate dependent heteroskedasticity, or to covariate dependent skewness, would also not perform well.

Our results in Table 8, which are based on 5 knots for the $n = 500$ case (implying 13 expanded moment conditions created from z_1, z_2, and $z_1 z_2$) and 7 knots for the larger sample sizes (implying 19 expanded moment conditions), show that the ATE is well inferred in this problem.

	True	Mean	SD	Median	Lower	Upper	Ineff
$n = 500$	0.19	0.12	0.05	0.12	-0.00	0.19	1.27
$n = 1000$	0.21	0.15	0.04	0.15	0.07	0.20	1.09
$n = 2000$	0.23	0.25	0.03	0.26	0.21	0.30	1.19

Table 8: Posterior summary for ATE estimation with three data sets. True ATE for each sample size is indicated by True. The summaries are based on 10,000 MCMC draws beyond a burn-in of 1000. The M-H acceptance rate is around 90% in the estimation of the control and treated models.

6.2 RD ATE in a Sharp design

Consider now the Bayesian analysis of the RD ATE effect under a sharp regression discontinuity design where we suppose that the data arise from the following data generating mechanism,

$$y_i = (1 - x_i)g_0(z_i) + x_i g_1(z_i) + \varepsilon_i,$$

where $x_i = 1\{z_i \geq \tau\}$ and $E^P[\varepsilon_i|z_i] = 0$. We define the regression discontinuity average treatment effect (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 $g_1(z)$.

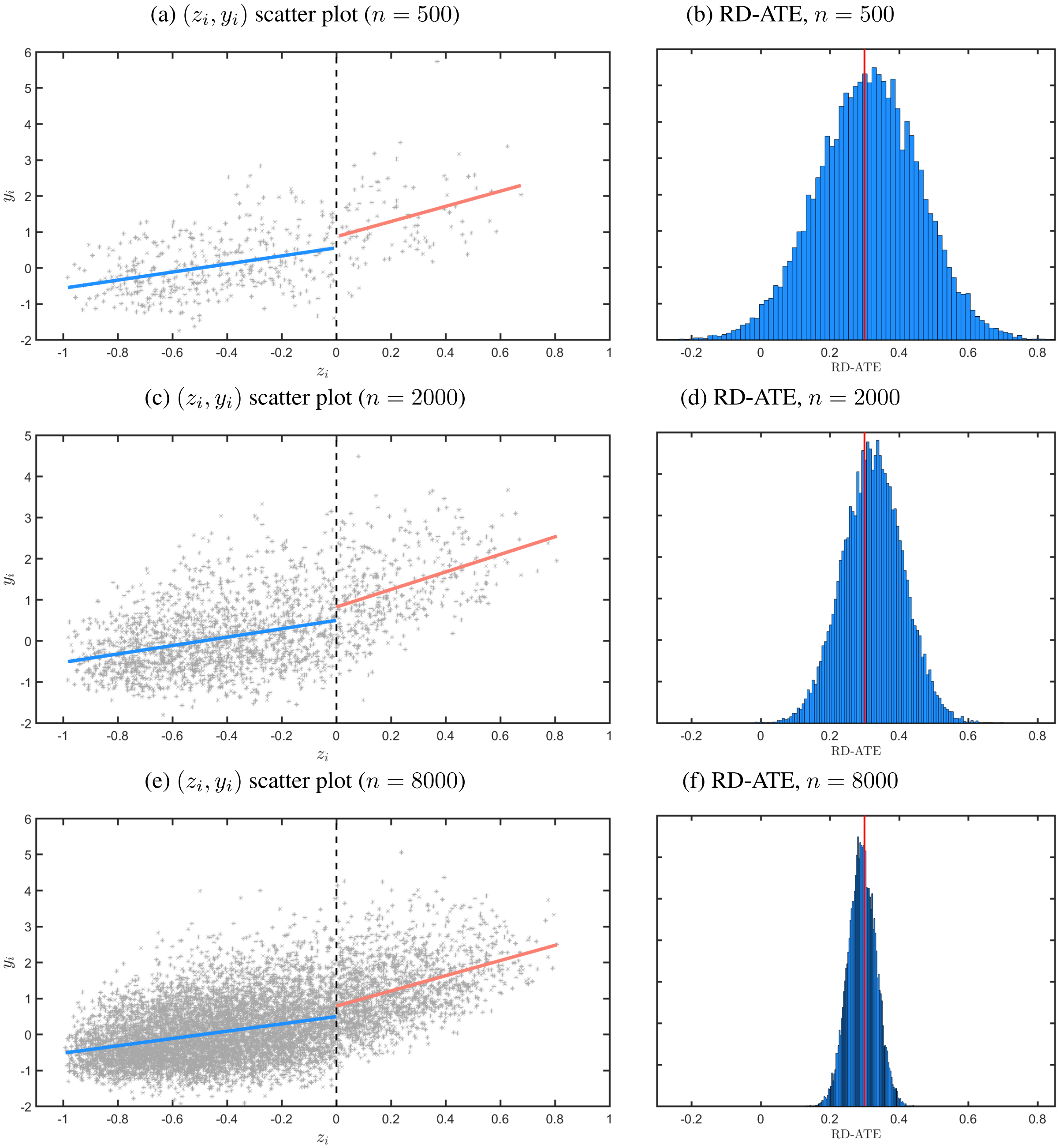
For illustrative purposes, suppose that

$$g_0(z_i) = 0.5 + z_i$$
$$g_1(z_i) = 0.8 + 2z_i$$

where $z_i = 2(z_i^* - 1)$ and $z_i^* \sim 2\mathcal{B}(2, 4)$. ε_i is independently drawn from $\mathcal{SN} \sim (m(z_i), s(z_i), w(z_i))$ with $m(x_i) = -s(z_i)\sqrt{2/\pi}w(z_i)/(\sqrt{1 + w(z_i)^2})$, $s(z_i) = 0.7(2 - z_i^2)$, and $w(z_i) = 3 + z_i^2$. 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 $E^P[\varepsilon_i|z_i] = 0$, without any further assumptions about ε_i. We estimate $g_0(z_i)$ and $g_1(z_i)$ separately for data on either side of τ using the conditional moment restrictions, $E^P[y_i - \theta_{j0} - \theta_{j1}z_i|z_i] = 0$, 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.

The results from this analysis are reported in Figure 2 and Table 9. The left panels of the figure has 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.



(a) (z_i, y_i) scatter plot ($n = 500$)

(b) RD-ATE, $n = 500$

(c) (z_i, y_i) scatter plot ($n = 2000$)

(d) RD-ATE, $n = 2000$

(e) (z_i, y_i) scatter plot ($n = 8000$)

(f) RD-ATE, $n = 8000$

Figure 2: In 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 ($n = 500, 2000, 8000$). Results are based on 20,000 MCMC draws beyond a burn-in of 1000. The M-H acceptance rate is around 95% in both cases. In right panels, the histogram of posterior draws for RD-ATE is presented for $n = 500, 2000, 8000$. In this example, RD-ATE is defined as $g_1(0) - g_0(0)$. The true value of RD-ATE is 0.3.

Table 9: 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

7 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 Bernstein von Mieses 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 examples show that the framework we have developed is both practical and useful.

Appendix

In this Appendix we only provide the main technical results that are new and that we need in order to prove the theorems in the paper. These results are specific to the particular setting with increasing dimension that we are considering. The complete proofs of all the theorems and technical results can be found in the Supplementary Appendix to this paper.

A Notation

For each positive integer K let $q^K(z) := (q_1(z), \dots, q_K(z))'$ be a K-vector of approximating functions. For every $\varepsilon > 0$ and a constant $C > 0$, let

$$\Theta(C, \varepsilon) := \{\|\theta - \theta_*\| \leq C\varepsilon\}.$$

We denote: $W := (X', Z')'$, $w_{1:n} := (w_1, \dots, w_n)$ the n-sample of i.i.d. observations of W, $g(W, \theta) := \rho(X, \theta) \otimes q^K(Z)$ the expanded moment functions and $g_i(\theta) = g(w_i, \theta)$.

Denote $p(w_{1:n}|\theta) := \prod_{i=1}^n \widehat{p}_i(\theta)$,

$$\ell_{n,\theta}(w_i) := \log \widehat{p}_i(\theta) = \log \frac{e^{\widehat{\lambda}(\theta)'g(w_i,\theta)}}{\sum_{j=1}^n e^{\widehat{\lambda}(\theta)'g(w_j,\theta)}}$$

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. Moreover,

$$\tau(\widehat{\lambda},\theta,w) := e^{\widehat{\lambda}(\theta)'g(w,\theta)} \quad \text{and} \quad \tau_n(\widehat{\lambda},\theta) := \frac{1}{n}\sum_{i=1}^n \tau(\widehat{\lambda},\theta,w_i).$$

We use the notation $\mathbb{E}_n[\cdot] := \frac{1}{n}\sum_{i=1}^n[\cdot]$ for the empirical mean and $\mathbf{E}[\cdot]$ for the population mean with respect to the true distribution P_*. For a matrix A, we denote by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ the minimum and maximum eigenvalue of A, respectively.

In addition we denote,

$$\widehat{g}(\theta) := \mathbb{E}_n[g_i(\theta)], \qquad \rho_\theta(X,\theta) := \frac{\partial \rho(X,\theta)}{\partial \theta'}, \qquad \widehat{G}(\theta) := \mathbb{E}_n[G(w_i,\theta)]$$

with $G(w,\theta) := \rho_\theta(x,\theta) \otimes q^K(z)$ a $dK \times p$ matrix,

$$\check{G}(\theta) := \mathbb{E}_n[\tau(\widehat{\lambda},\theta,w_i)G(w_i,\theta)], \qquad \widehat{\Omega}(\theta) := \mathbb{E}_n[g(w_i,\theta)g(w_i,\theta)']$$

a $dK \times dK$ matrix and

$$\check{\Omega}(\theta) := \mathbb{E}_n[\tau(\widehat{\lambda},\theta,w_i)g(w_i,\theta)g(w_i,\theta)'].$$

Their population counterparts in the correctly specified model are $G_* := \mathbf{E}[\rho_\theta(X,\theta_*) \otimes q^K(z)]$ and $\Omega_* := \mathbf{E}[g(w_i,\theta_*)g(w_i,\theta_*)']$, respectively. In addition, $\Sigma(z) := \mathbf{E}[\rho(X,\theta_*)\rho(X,\theta_*)'|z]$, $D(z) := \mathbf{E}[\rho_\theta(X,\theta_*)|z]$, $V_{\theta_*}^{-1} := \mathbf{E}^P[D(Z)'\Sigma(Z)^{-1}D(Z)]$, and $\rho_{j\theta\theta}(x,\theta_*) := \partial^2\rho_j(x,\theta)/\partial\theta\partial\theta'$.

Finally, let CS, M, and MVT refer to the Cauchy-Schwartz, Markov, and Mean Value Theorem, respectively.

B Proofs for Section 3.3

B.1 Proof of Theorem 3.1

The main steps of this proof proceed as in the proof of Van der Vaart (1998, Theorem 10.1) while the proofs of the technical results we need all along this proof are new. For this reason, here we only provide the main technical results that are new. The detailed proof of Theorem 3.1 can be found in the Supplementary Appendix to this paper.

The first technical result that we need to establish a Bernstein - von Mises theoerm is the stochastic local asymptotic normality (LAN) expansion which is given in Lemma B.1 below. The second result that we need is consistency of the posterior distribution, namely $P(\pi(\sqrt{n}\|\theta - \theta_*\| > M_n|w_{1:n}) > 0) \to 0$ for any $M_n \to \infty$, as $n \to \infty$, which is established in Theorem B.1.

We start with stating posterior consistency. The proof of this theorem is similar to the proof of Theorem C.2 in Chib et al. (2018) and is given in the Supplementary Appendix for completeness.

Theorem B.1 (Posterior Consistency) *Let the Assumptions of Lemma B.1 and Assumption 3.7 hold. Moreover, assume that there exists a constant $C > 0$ such that for any sequence $M_n \to \infty$,*

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

as $n \to \infty$. Then,

$$\pi\left(\sqrt{n}\|\theta - \theta_*\| > M_n\middle| w_{1:n}\right) \xrightarrow{p} 0 \tag{B.2}$$

for any $M_n \to \infty$, as $n \to \infty$.

Lemma B.1 (Stochastic LAN) *Let Assumptions 3.1, 3.2, 3.3, 3.5 and 3.6 be satisfied and assume $\zeta(K)K^2/\sqrt{n} \to 0$. Let $\mathcal{H}$ denote a compact subset of $\mathbb{R}^p$. Then,*

$$\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\ell_{n,\theta_*+h/\sqrt{n}}(w_i) - \sum_{i=1}^{n}\ell_{n,\theta_*}(w_i) - h'V_{\theta_*}^{-1}\Delta_{n,\theta_*} - \frac{1}{2}h'V_{\theta_*}^{-1}h\right| = o_p(1) \tag{B.3}$$

where $V_{\theta_*}^{-1}\Delta_{n,\theta_*} \xrightarrow{d} \mathcal{N}(0, V_{\theta_*}^{-1})$ and $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d\ell_{n,\theta_*}(w_i)}{d\theta} - V_{\theta_*}^{-1}\Delta_{n,\theta_*} \xrightarrow{p} 0$.

Proof. Define $\tau_i(\lambda,\theta) := \frac{e^{\lambda' g_i(\theta)}}{\mathbf{E}_n[e^{\lambda' g_j(\theta)}]}$,

$$\check{\Omega}(\theta,\lambda) := \mathbb{E}_n[\tau_i(\lambda,\theta)g_i(\theta)g_i(\theta)'] \qquad \text{and} \qquad \check{G}(\theta,\lambda) := \mathbb{E}_n[\tau_i(\lambda,\theta)G(w_i,\theta)].$$

By the MVT expansion of $\sum_{i=1}^{n}\ell_{n,\theta}(w_i)$ around $\widehat{\lambda}(\theta) = 0$, there exists a $\widetilde{\lambda}_\theta$ lying on the line between $\widehat{\lambda}(\theta)$ and zero such that:

$$\begin{aligned}\sum_{i=1}^{n}\ell_{n,\theta}(w_i) = \sum_{i=1}^{n}\log\tau_i(\widehat{\lambda},\theta) - n\log(n) = -n\log(n) + n\widehat{g}(\theta)'\widehat{\lambda}(\theta) - n\widehat{g}(\theta)'\widehat{\lambda}(\theta)\\ - \frac{1}{2}n\widehat{\lambda}(\theta)'\check{\Omega}(\theta,\widetilde{\lambda}_\theta)\widehat{\lambda}(\theta) + \frac{1}{2}n\left|\mathbf{E}_n[\tau_i(\widetilde{\lambda}_\theta,\theta)g_i(\theta)']\widehat{\lambda}(\theta)\right|^2.\end{aligned} \tag{B.4}$$

By expanding the first order condition for $\widehat{\lambda}(\theta)$ around $\widehat{\lambda}(\theta) = 0$, there exists a $\bar{\lambda}_\theta$ lying on the line between $\widehat{\lambda}(\theta)$ and zero such that : $\widehat{g}(\theta) + \check{\Omega}(\theta,\bar{\lambda}_\theta)\widehat{\lambda}(\theta) = 0$ which gives $\widehat{\lambda}(\theta) = \check{\Omega}(\theta,\bar{\lambda}_\theta)^{-1}\widehat{g}(\theta)$. By replacing this in (B.4) we obtain:

$$\sum_{i=1}^{n} \ell_{n,\theta}(w_i) = -n\log(n) - \frac{1}{2} n\widehat{g}(\theta)'\check{\Omega}(\theta,\bar{\lambda}_\theta)^{-1}\check{\Omega}(\theta,\widetilde{\lambda}_\theta)\check{\Omega}(\theta,\bar{\lambda}_\theta)^{-1}\widehat{g}(\theta)$$
$$+ \frac{1}{2} n \left|\mathbf{E}_n[\tau_i(\widetilde{\lambda}_\theta,\theta)g_i(\theta)']\check{\Omega}(\theta,\bar{\lambda}_\theta)^{-1}\widehat{g}(\theta)\right|^2. \quad \text{(B.5)}$$

Hence, by replacing in $\sum_{i=1}^n \ell_{n,\theta_*+h_n}(w_i)$ the following MVT expansion $\widehat{g}(\theta_* + h/\sqrt{n}) = \widehat{g}(\theta_*) + \widehat{G}(\widetilde{\theta})h/\sqrt{n}$, for $\widetilde{\theta}$ lying between $\theta_* + h/\sqrt{n}$ and θ_*, and by denoting $\theta_1 := \theta_* + h_n$ $h_n := h/\sqrt{n}$ we get

$$\sum_{i=1}^{n} \ell_{n,\theta_*+h_n}(w_i) - \sum_{i=1}^{n} \ell_{n,\theta_*}(w_i)$$
$$= \frac{1}{2} n\widehat{g}(\theta_*)' \left[\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1}\check{\Omega}(\theta_*,\widetilde{\lambda}_{\theta_*})\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1} - \check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\right]\widehat{g}(\theta_*)$$
$$- \sqrt{n}\widehat{g}(\theta_*)'\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\widehat{G}(\widetilde{\theta})h$$
$$- \frac{1}{2} h'\widehat{G}(\widetilde{\theta})'\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\widehat{G}(\widetilde{\theta})h$$
$$+ \frac{1}{2} n \left|\mathbf{E}_n[\tau_i(\widetilde{\lambda}_{\theta_1},\theta_1)g_i(\theta_1)']\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\widehat{g}(\theta_1)\right|^2 - \frac{1}{2} n \left|\mathbf{E}_n[\tau_i(\widetilde{\lambda}_{\theta_*},\theta_*)g_i(\theta_*)']\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1}\widehat{g}(\theta_*)\right|^2. \quad \text{(B.6)}$$

By using the equality $A^{-1}BA^{-1} - C^{-1}DC^{-1} = A^{-1}(B-D)A^{-1} + (A^{-1}-C^{-1})DA^{-1} + C^{-1}D(A^{-1} - C^{-1})$ for matrices A, B, C, D we can write

$$\sum_{i=1}^{n} \ell_{n,\theta_*+h_n}(w_i) - \sum_{i=1}^{n} \ell_{n,\theta_*}(w_i)$$
$$= \frac{1}{2} n\widehat{g}(\theta_*)'\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1}\left[\check{\Omega}(\theta_*,\widetilde{\lambda}_{\theta_*}) - \check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})\right]\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1}\widehat{g}(\theta_*)$$
$$+ \frac{1}{2} n\widehat{g}(\theta_*)'\left[\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1} - \check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\right]\check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1}\widehat{g}(\theta_*)$$
$$+ \frac{1}{2} n\widehat{g}(\theta_*)'\left[\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1} - \check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\right]\check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\widehat{g}(\theta_*)$$
$$- \sqrt{n}\widehat{g}(\theta_*)'\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\widehat{G}(\widetilde{\theta})h$$
$$- \frac{1}{2} h'\widehat{G}(\widetilde{\theta})'\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\widehat{G}(\widetilde{\theta})h$$
$$+ \frac{1}{2} n \left|\mathbf{E}_n[\tau_i(\widetilde{\lambda}_{\theta_1},\theta_1)g_i(\theta_1)']\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\widehat{g}(\theta_1)\right|^2 - \frac{1}{2} n \left|\mathbf{E}_n[\tau_i(\widetilde{\lambda}_{\theta_*},\theta_*)g_i(\theta_*)']\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1}\widehat{g}(\theta_*)\right|^2. \quad \text{(B.7)}$$

Let us analyse the first three terms in (B.7). Since $\bar{\lambda}_{\theta_*}, \bar{\lambda}_{\theta_1}, \widetilde{\lambda}_{\theta_1} \in \Lambda_n$, where Λ_n is as defined in Lemma G.2, in the following we can use the results in Lemma G.4 to get $\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1} \leq C\widehat{\Omega}(\theta_*)^{-1}$, $\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1} \leq C\widehat{\Omega}(\theta_1)^{-1}$ and $\check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})^{-1} \leq C\widehat{\Omega}(\theta_1)^{-1}$ with probability approaching 1 for any $1 < C < \infty$. We start from the first term:

$$\sup_{h\in\mathcal{H}} R_{n,1}(h) := \frac{1}{2}\sup_{h\in\mathcal{H}}\left| n\widehat{g}(\theta_*)'\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1}\left[\check{\Omega}(\theta_*,\widetilde{\lambda}_{\theta_*}) - \check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})\right]\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1}\widehat{g}(\theta_*)\right|$$

$$\leq \frac{1}{2}\|\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1}\sqrt{n}\widehat{g}(\theta_*)\|^2 \sup_{h\in\mathcal{H}}\|\check{\Omega}(\theta_*,\widetilde{\lambda}_{\theta_*}) - \check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})\|$$

$$\leq \left(\min_{1\leq i\leq n}\tau_i(\bar{\lambda}_{\theta_*},\theta_*)\right)^{-2}\|\widehat{\Omega}(\theta_*)^{-1}\sqrt{n}\widehat{g}(\theta_*)\|^2 O_p\left(\zeta(K)K/\sqrt{n}\right) = O_p\left(\zeta(K)K^2/\sqrt{n}\right)$$

by using the first result in Lemma G.7 and because $\|\widehat{\Omega}(\theta_*)^{-1}\sqrt{n}\widehat{g}(\theta_*)\| = \|\Omega_*^{-1}\sqrt{n}\widehat{g}(\theta_*)\|$ with probability approaching 1 by Donald et al. (2003, Lemma A.6) and $\|\Omega_*^{-1}\sqrt{n}\widehat{g}(\theta_*)\| = O_p(\sqrt{K})$ by M. For the second term we use the identity $(A^{-1} - B^{-1}) = A^{-1}(B-A)B^{-1}$ for two matrices A, B, and again the first result in Lemma G.7:

$$\sup_{h\in\mathcal{H}} R_{n,2}(h) := \frac{1}{2}\sup_{h\in\mathcal{H}}\Big| n\widehat{g}(\theta_*)'\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1}\left[\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1}) - \check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1}\right]\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}$$

$$\times\,\check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1}\widehat{g}(\theta_*)\Big|$$

$$\leq \frac{1}{2}\|\check{\Omega}(\theta_*,\bar{\lambda}_{\theta_*})^{-1}\sqrt{n}\widehat{g}(\theta_*)\|^2 O_p\left(\zeta(K)K/\sqrt{n}\right) = O_p\left(\zeta(K)K^2/\sqrt{n}\right).$$

The third term can be treated in a similar way and gives the same rate.

Next, we analyze the last two terms in (B.7). We use again Lemma G.4. Therefore, because $\|\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\widehat{g}(\theta_*)\| = \|\Omega_*^{-1}\sqrt{n}\widehat{g}(\theta_*)\|$ with probability approaching 1 by Donald et al. (2003, Lemma A.6) and $\|\Omega_*^{-1}\sqrt{n}\widehat{g}(\theta_*)\| = O_p(\sqrt{K})$ by M.

$$\sup_{h\in\mathcal{H}}\frac{1}{2}n\left|\mathbf{E}_n[\tau_i(\widetilde{\lambda}_{\theta_1},\theta_1)g_i(\theta_1)']\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\widehat{g}(\theta_1)\right|^2$$

$$\leq \frac{1}{2}C\mathbf{E}_n[\tau_i(\widetilde{\lambda}_{\theta_1},\theta_1)g_i(\theta_1)']^2\|\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\widehat{g}(\theta_1)\|^2 = O_p(K^2/n) \quad \text{(B.8)}$$

where we have used the MVT expansion $\widehat{g}(\theta_1) = \widehat{g}(\theta_*) + \widehat{G}(\widetilde{\theta})h/\sqrt{n}$ for a $\widetilde{\theta}$ lying between θ_1 and θ_* and the result of Lemma G.5. We conclude that

$$\sup_{h\in\mathcal{H}}\Big|\sum_{i=1}^{n}\ell_{n,\theta_*+h_n}(w_i) - \sum_{i=1}^{n}\ell_{n,\theta_*}(w_i) - \sqrt{n}\widehat{g}(\theta_*)'\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\widehat{G}(\widetilde{\theta})h$$

$$-\frac{1}{2}h'\widehat{G}(\widetilde{\theta})'\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\check{\Omega}(\theta_1,\widetilde{\lambda}_{\theta_1})\check{\Omega}(\theta_1,\bar{\lambda}_{\theta_1})^{-1}\widehat{G}(\widetilde{\theta})h\Big|$$

$$= O_p\left(\zeta(K)K^2/\sqrt{n}\right) + O_p(K/\sqrt{n}) = O_p\left(\zeta(K)K^2/\sqrt{n}\right). \quad \text{(B.9)}$$

Under the assumptions of the theorem the term $O_p\left(\zeta(K)K^2/\sqrt{n}\right)$ converges to zero.Moreover, by Lemma G.7 and $\zeta(K)K/\sqrt{n} \to 0$:

$$
\begin{aligned}
&- \sqrt{n}\widehat{g}(\theta_*)'\check{\Omega}(\theta_1, \bar{\lambda}_{\theta_1})^{-1}\check{\Omega}(\theta_1, \widetilde{\lambda}_{\theta_1})\check{\Omega}(\theta_1, \bar{\lambda}_{\theta_1})^{-1}\widehat{G}(\widetilde{\theta})h \\
&\qquad\qquad = -\sqrt{n}\widehat{g}(\theta_*)'\widehat{\Omega}(\theta_*)^{-1}\widehat{G}(\theta_*)h + o_p(1) \\
&\qquad\qquad\qquad = -\frac{h'}{\sqrt{n}}\sum_{i=1}^{n} D(z_i)'\Sigma(z_i)^{-1}\rho(x_i, \theta_*)' + o_p(1)
\end{aligned}
$$

where the $o_p(1)$ term is uniform in $h \in \mathcal{H}$ and where to get the second equality we have used arguments similar to the ones in Donald et al. (2003, Proof of Theorem 5.6). By the Lindberg-Levy central limit theorem, $\Delta_{n,\theta_*} := -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_{\theta_*}D(z_i)'\Sigma(z_i)^{-1}\rho(x_i, \theta_*) \xrightarrow{d} \mathcal{N}(0, V_{\theta_*})$. Similarly as in Donald et al. (2003, Proof of Theorem 5.6) it is possible to show that (by using compactness of $\mathcal{H}$)

$$
h'\widehat{G}(\widetilde{\theta})'\check{\Omega}(\theta_1, \bar{\lambda}_{\theta_1})^{-1}\check{\Omega}(\theta_1, \widetilde{\lambda}_{\theta_1})\check{\Omega}(\theta_1, \bar{\lambda}_{\theta_1})^{-1}\widehat{G}(\widetilde{\theta})h = h'V_{\theta_*}h + o_p(1)
$$

where the $o_p(1)$ term is uniform in $h \in \mathcal{H}$. Finally, remark that

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d\ell_{n,\theta_*}(w_i)}{d\theta} - V_{\theta_*}^{-1}\Delta_{n,\theta_*} \xrightarrow{p} 0.
$$

This establishes the result of the Lemma. □

B.2 Proof of Theorem 3.2

For the misspecified case we use the following notation: $G_i(\theta) := G(w_i, \theta)$, $G_\circ := \mathbb{E}[G_i(\theta_\circ)]$, $\check{G}_\circ := \mathbb{E}[\tau_i(\lambda_\circ(\theta_\circ), \theta_\circ)G_i(\theta_\circ)]$, $\Omega_\circ := \mathbb{E}[\tau(\lambda_\circ(\theta_\circ), \theta_\circ, W_i)g_i(\theta_\circ)g_i(\theta_\circ)']$, $\tau_i(\lambda, \theta) := \frac{e^{\lambda' g_i(\theta)}}{\mathbf{E}_n[e^{\lambda' g_j(\theta)}]}$, $\check{G}(\lambda, \theta) := \mathbb{E}_n[\tau_i(\lambda, \theta)G_i(\theta)]$, $\check{\Omega}(\theta, \lambda) := \mathbb{E}_n[\tau(\lambda, \theta, W_i)g_i(\theta)g_i(\theta)']$. We also use standard notation in empirical process theory: $\mathbb{P}_n := \mathbb{E}_n[\delta_{x_i}]$ where δ_x is the Dirac measure at x, and $\mathbb{G}_n g := \sqrt{n}(\mathbb{P}_n f - \mathbb{E}^P f)$ for every function f.

The proof of Theorem 3.2 proceeds as the proof of Theorem 3.1 and so we omit it. In the Supplementary Appendix we explain the differences between the proofs of Theorem 3.2 and of Theorem 3.1. As for Theorem 3.1, even for the Bernstein - von Mises theorem in the misspecified case we need to establish a stochastic local asymptotic normality (LAN) expansion, which is given in Lemma B.2 below, and consistency of the posterior distribution, namely $P(\pi(\sqrt{n}\|\theta - \theta_\circ\| > M_n|w_{1:n}) > 0) \to 0$ for any $M_n \to \infty$, as $n \to \infty$, which is given in Theorem B.2.

We start with stating posterior consistency. The proof of this theorem is similar to the proof of Theorem C.2 in Chib et al. (2018) and to the proof of Lemma B.1 which is given in the Supplementary Appendix for completeness.

Theorem B.2 (Posterior Consistency) *Let the Assumptions of Lemma B.2 and Assumption 3.9 hold. Moreover, assume that there exists a constant $C > 0$ such that for any sequence $M_n \to \infty$,*

$$P\left(\sup_{\|\theta-\theta_\circ\|>M_n/\sqrt{n}} \frac{1}{n}\sum_{i=1}^{n}\left(\ell_{n,\theta}(w_i) - \ell_{n,\theta_\circ}(w_i)\right) \leq -CM_n^2/n\right) \to 1, \tag{B.10}$$

as $n \to \infty$. Then,

$$\pi\left(\sqrt{n}\|\theta-\theta_\circ\| > M_n \middle| w_{1:n}\right) \xrightarrow{p} 0 \tag{B.11}$$

for any $M_n \to \infty$, as $n \to \infty$.

Lemma B.2 (Stochastic LAN) *Let Assumptions 3.1, 3.2, 3.6 and 3.10 hold and assume $\zeta(K)K^2\sqrt{K/n} \to 0$. Let $\mathcal{H}$ denote a compact subset of $\mathbb{R}^p$ and $\theta_1 := \theta_\circ + h/\sqrt{n}$ with $h \in \mathcal{H}$. Then,*

$$\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\ell_{n,\theta_1}(w_i) - \sum_{i=1}^{n}\ell_{n,\theta_\circ}(w_i) - h'\mathcal{A}_{\theta_\circ}\Delta_{n,\theta_0} - \frac{1}{2}h'\mathcal{A}_{\theta_\circ}h\right| = o_p(1) \tag{B.12}$$

where Δ_{n,θ_0} is a random vector bounded in probability and $\mathcal{A}_{\theta_\circ}$ is a nonsingular matrix.

Proof. We have to analyse the difference $\sum_{i=1}^{n}\ell_{n,\theta_1}(w_i) - \sum_{i=1}^{n}\ell_{n,\theta_\circ}(w_i)$. Because W_i are i.i.d. then $\mathbb{E}[g_i(\theta)] = \mathbb{E}[g_j(\theta)]$, and so we can write:

$$\begin{aligned}\sum_{i=1}^{n}\ell_{n,\theta}(w_i) &= \sum_{i=1}^{n}\log\tau_i(\widehat{\lambda},\theta) - n\log(n) = \sum_{i=1}^{n}\log\frac{e^{\widehat{\lambda}'(g_i(\theta)-\mathbb{E}[g_i(\theta)])}}{\mathbf{E}_n[e^{\widehat{\lambda}'(g_j(\theta)-\mathbb{E}[g_j(\theta)])}]} - n\log n\\ &= n\widehat{\lambda}(\theta)'\mathbb{E}_n(g_i(\theta)-\mathbb{E}[g_i(\theta)]) - n\log\mathbb{E}_n[e^{\widehat{\lambda}(\theta)'(g_j(\theta)-\mathbb{E}[g_j(\theta)])}] - n\log(n).\end{aligned} \tag{B.13}$$

Denote $\overline{g}_i(\theta) := g_i(\theta) - \mathbb{E}[g_i(\theta)]$ and $\overline{G}_i(\theta) := G_i(\theta) - \mathbb{E}[G_i(\theta)]$, so that $\sum_{i=1}^{n}\ell_{n,\theta}(w_i) = n\widehat{\lambda}(\theta)'\mathbb{E}_n[\overline{g}_i(\theta)] - n\log\mathbb{E}_n[e^{\widehat{\lambda}(\theta)'\overline{g}_i(\theta)}] - n\log(n)$. By the MVT there exists a $t \in [0,1]$ such that $\widetilde{\theta} := \theta_\circ + th/\sqrt{n}$ satisfies

$$\sum_{i=1}^{n}\ell_{n,\theta_1}(w_i) = \sum_{i=1}^{n}\ell_{n,\theta_\circ}(w_i) + \frac{h'}{\sqrt{n}}\sum_{i=1}^{n}\dot{\ell}_{n,\theta_\circ}(w_i) + \frac{1}{2}\frac{h'}{\sqrt{n}}\sum_{i=1}^{n}\ddot{\ell}_{n,\widetilde{\theta}}(w_i)\frac{h}{\sqrt{n}} \tag{B.14}$$

where

$$\begin{aligned}\dot{\ell}_{n,\theta_\circ}(w_i) := \sum_{i=1}^{n}\frac{d\ell_{n,\theta_1}(w_i)}{d\theta}\bigg|_{\theta_1=\theta_\circ} &= \frac{d\widehat{\lambda}(\theta_\circ)'}{d\theta}\overline{g}_i(\theta_\circ) + \overline{G}_i(\theta_\circ)'\widehat{\lambda}(\theta_\circ)\\ &\quad + \frac{d\widehat{\lambda}(\theta_\circ)'}{d\theta}\mathbb{E}[g_i(\theta_\circ)] - \mathbb{E}_n\left[\tau_i(\widehat{\lambda}(\theta_\circ),\theta_\circ)\overline{G}_i(\theta_\circ)'\right]\widehat{\lambda}(\theta_\circ)\\ = \frac{d\widehat{\lambda}(\theta_\circ)'}{d\theta}\overline{g}_i(\theta_\circ) + \overline{G}_i(\theta_\circ)'\lambda_\circ(\theta_\circ) &+ \overline{G}_i(\theta_\circ)'(\widehat{\lambda}(\theta_\circ) - \lambda_\circ(\theta_\circ)) + \left(\frac{d\widehat{\lambda}(\theta_\circ)'}{d\theta} - \frac{d\lambda_\circ(\theta_\circ)'}{d\theta}\right)\mathbb{E}[g_i(\theta_\circ)]\end{aligned}$$

$$- \mathbb{E}_n \left[\Big(\tau_i(\widehat{\lambda}(\theta_\circ), \theta_\circ) - \tau_i(\lambda_\circ, \theta_\circ) \Big) G_i(\theta_\circ)' \right] \widehat{\lambda}(\theta_\circ) - \mathbb{E}_n \left[\tau_i(\lambda_\circ, \theta_\circ) G_i(\theta_\circ)' - \check{G}_\circ' \right] \lambda_\circ(\theta_\circ)$$

$$- \mathbb{E}_n[\tau_i(\lambda_\circ, \theta_\circ) G_i(\theta_\circ)' - G_\circ'](\widehat{\lambda}(\theta_\circ) - \lambda_\circ(\theta_\circ)) \quad \text{(B.15)}$$

with

$$\frac{d\widehat{\lambda}(\theta_\circ)'}{d\theta} = -\mathbf{E}_n \left[\tau_i(\widehat{\lambda}, \theta_\circ) G_i(\theta_\circ)'(I + \widehat{\lambda}(\theta_\circ) g_i(\theta_\circ)') \right] \Big(\mathbb{E}_n[\tau_i(\widehat{\lambda}, \theta_\circ) g_i(\theta_\circ) g_i(\theta_\circ)'] \Big)^{-1},$$

$$\frac{d\lambda_\circ(\theta_\circ)'}{d\theta} = -\mathbf{E} \left[\tau(\lambda_\circ, \theta_\circ, W_i) G_i(\theta_\circ)'(I + \lambda_\circ(\theta_\circ) g_i(\theta_\circ)') \right] \Omega_\circ^{-1} \quad \text{(B.16)}$$

and where we have used the first order condition of the pseudo true value $\theta_\circ$, that is:

$$\frac{d\lambda_\circ(\theta_\circ)'}{d\theta} \mathbb{E}[g_i(\theta_\circ)] + G_\circ' \lambda_\circ(\theta_\circ) - \frac{d\lambda_\circ(\theta_\circ)'}{d\theta} \mathbb{E}[\tau_i(\lambda_\circ, \theta_\circ) g_i(\theta_\circ)] - \mathbb{E}[\tau_i(\lambda_\circ, \theta_\circ) G_i(\theta_\circ)'] \lambda_\circ(\theta_\circ) = 0$$

and $\mathbb{E}[\tau_i(\lambda_\circ, \theta_\circ) g_i(\theta_\circ)] = 0$ because it is the first order condition for $\lambda_\circ$. Moreover,

$$\ddot{\ell}_{n,\widetilde{\theta}}(w_i) := \sum_{i=1}^n \frac{d^2 \ell_{n,\theta_1}(w_i)}{d\theta d\theta'} \bigg|_{\theta_1 = \widetilde{\theta}} = \sum_{j=1}^{dK} \frac{d^2 \widehat{\lambda}(\widetilde{\theta})'}{d\theta d\theta'} g_{i,j}(\widetilde{\theta}) + \frac{d\widehat{\lambda}(\widetilde{\theta})'}{d\theta} G_i(\widetilde{\theta}) + \overline{G}_i(\widetilde{\theta})' \frac{d\widehat{\lambda}(\widetilde{\theta})'}{d\theta}$$

$$+ \sum_{j=1}^{dK} \frac{d^2 \overline{g}_{i,j}(\widetilde{\theta})}{d\theta d\theta'} \widehat{\lambda}_j(\widetilde{\theta}) + \mathbb{E}_n \left[\overline{G}_i(\widetilde{\theta})' \widehat{\lambda}(\widetilde{\theta}) \frac{d\tau_i(\widehat{\lambda}(\widetilde{\theta}), \widetilde{\theta})}{d\widehat{\lambda}'} \frac{d\widehat{\lambda}(\widetilde{\theta})}{d\theta'} \right] + \mathbb{E}_n \left[\frac{\tau_i(\widehat{\lambda}(\widetilde{\theta}), \widetilde{\theta})}{d\theta} \widehat{\lambda}(\widetilde{\theta})' \overline{G}_i(\widetilde{\theta}) \right]$$

$$+ \mathbb{E}_n \left[\tau_i(\widehat{\lambda}(\widetilde{\theta}), \widetilde{\theta}) \sum_{j=1}^{dK} \frac{d^2 \overline{g}_{i,j}(\widetilde{\theta})}{d\theta d\theta'} \right] \widehat{\lambda}_j(\widetilde{\theta}) + \mathbb{E}_n \left[\tau_i(\widehat{\lambda}(\widetilde{\theta}), \widetilde{\theta}) \overline{G}_i(\widetilde{\theta})' \right] \frac{d\widehat{\lambda}(\widetilde{\theta})}{d\theta'}. \quad \text{(B.17)}$$

We start with analyzing term (B.15). First, remark that by Lemma G.12 it holds:

$$h' \Big(\frac{d\widehat{\lambda}(\theta_\circ)'}{d\theta} - \frac{d\lambda_\circ(\theta_\circ)'}{d\theta} \Big) \sqrt{n} \mathbb{E}_n[\overline{g}_i(\theta_\circ)] = o_p(1)$$

uniformly in $h \in \mathcal{H}$. Next, we analyse the third term in (B.15). By a MVT expansion of the first order condition for $\widehat{\lambda}(\theta_\circ)$ there exists $\tau \in [0,1]$ such that $\widehat{\lambda}_\tau := \tau(\widehat{\lambda}(\theta_\circ) - \lambda_\circ(\theta_\circ)) + \lambda_\circ(\theta_\circ)$ satisfies $\mathbb{E}_n[e^{\widehat{\lambda}(\theta_\circ)' g_i(\theta_\circ)} g_i(\theta_\circ)] = 0 = \mathbb{E}_n[e^{\lambda_\circ(\theta_\circ)' g_i(\theta_\circ)} g_i(\theta_\circ)] + \check{\Omega}(\theta_\circ, \widehat{\lambda}_\tau)(\widehat{\lambda}(\theta_\circ) - \lambda_\circ(\theta_\circ))$ which implies:

$$(\widehat{\lambda}(\theta_\circ) - \lambda_\circ(\theta_\circ)) = -\check{\Omega}(\theta_\circ, \widehat{\lambda}_\tau)^{-1} \mathbb{E}_n[e^{\lambda_\circ(\theta_\circ)' g_i(\theta_\circ)} g_i(\theta_\circ)]. \quad \text{(B.18)}$$

Therefore,

$$h' \frac{1}{\sqrt{n}} \sum_{i=1}^n \overline{G}_i(\theta_\circ)'(\widehat{\lambda}(\theta_\circ) - \lambda_\circ(\theta_\circ)) = -\sqrt{n} \mathbb{E}_n[e^{\lambda_\circ(\theta_\circ)' g_i(\theta_\circ)} g_i(\theta_\circ)'] \check{\Omega}(\theta_\circ, \widehat{\lambda}_\tau)^{-1} \mathbb{E}_n[\overline{G}_i(\theta_\circ)] h$$

$$= -\mathbb{G}_n[e^{\lambda_\circ(\theta_\circ)' g_i(\theta_\circ)} g_i(\theta_\circ)'] \check{\Omega}(\theta_\circ, \widehat{\lambda}_\tau)^{-1} \mathbb{E}_n[\overline{G}_i(\theta_\circ)] h = O_p(K/\sqrt{n}).$$

Here, to get the term $O_p(K/\sqrt{n})$ we have used the inequality

$$\sup_{h\in\mathcal{H}} \mathbb{E}\left|\mathbb{G}_n[e^{\lambda_\circ(\theta_\circ)'g_i(\theta_\circ)}g_i(\theta_\circ)']\check{\Omega}(\theta_\circ,\widehat{\lambda}_\tau)^{-1}\mathbb{E}_n[\overline{G}_i(\theta_\circ)]h\right| \\ \leq C^{-2}\sup_{h\in\mathcal{H}}\sqrt{\mathbb{E}\left\|\mathbb{G}_n[e^{\lambda_\circ(\theta_\circ)'g_i(\theta_\circ)}g_i(\theta_\circ)']\right\|^2}\sqrt{\mathbb{E}\left\|\mathbb{E}_n[\overline{G}_i(\theta_\circ)]h\right\|^2} = O_p(\sqrt{K}\sqrt{K/n})$$

for which we have used Lemma G.9 and Assumption 3.10 (d) and (f). To control the fourth term in (B.15) we use Assumption 3.10 (h). We now control the fifth term in (B.15). For this, we use again (B.18) and a MVT expansion of $\tau_i(\widehat{\lambda}(\theta_\circ),\theta_\circ)$ around $\lambda_\circ(\theta_\circ)$:

$$\sqrt{n}\widehat{\lambda}(\theta_\circ)'\mathbb{E}_n[\left(\tau_i(\widehat{\lambda}(\theta_\circ),\theta_\circ)-\tau_i(\lambda_\circ(\theta_\circ),\theta_\circ)\right)G_i(\theta_\circ)]h \\ = \sqrt{n}(\widehat{\lambda}(\theta_\circ)-\lambda_\circ(\theta_\circ))'\mathbb{E}_n\left[\frac{\partial\tau_i(\lambda_t,\theta_\circ)}{\partial\lambda}\widehat{\lambda}(\theta_\circ)'G_i(\theta_\circ)\right]h \\ = -\sqrt{n}\mathbb{E}_n[e^{\lambda_\circ(\theta_\circ)'g_i(\theta_\circ)}g_i(\theta_\circ)']\check{\Omega}(\theta_\circ,\widehat{\lambda}_\tau)^{-1}\mathbb{E}_n\left[\frac{\partial\tau_i(\lambda_t,\theta_\circ)}{\partial\lambda}\widehat{\lambda}(\theta_\circ)'G_i(\theta_\circ)\right]h \\ = -\mathbb{G}_n[e^{\lambda_\circ(\theta_\circ)'g_i(\theta_\circ)}g_i(\theta_\circ)']\check{\Omega}(\theta_\circ,\widehat{\lambda}_\tau)^{-1}\mathbb{E}_n\left[\frac{\partial\tau_i(\lambda_t,\theta_\circ)}{\partial\lambda}\lambda_\circ(\theta_\circ)'G_i(\theta_\circ)\right]h + o_p(1) \quad \text{(B.19)}$$

where $\lambda_t = t(\widehat{\lambda}(\theta_\circ)-\lambda_\circ(\theta_\circ))+\lambda_\circ(\theta_\circ)$ for some $t\in[0,1]$. To control the last term in (B.15) we use again (B.18) to get

$$-h'\sqrt{n}\mathbb{E}_n[\tau_i(\lambda_\circ,\theta_\circ)G_i(\theta_\circ)'-G_\circ'](\widehat{\lambda}(\theta_\circ)-\lambda_\circ(\theta_\circ)) \\ = h'\mathbb{E}_n[\tau_i(\lambda_\circ,\theta_\circ)G_i(\theta_\circ)'-G_\circ']\check{\Omega}(\theta_\circ,\widehat{\lambda}_\tau)^{-1}\mathbb{G}_n[e^{\lambda_\circ(\theta_\circ)'g_i(\theta_\circ)}g_i(\theta_\circ)].$$

By putting together these arguments we get:

$$\frac{h}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}_{n,\theta_\circ}(w_i) = h'\mathbb{G}_n(\dot{\mathcal{L}}_{n,\theta_\circ}(w_i)) + \sqrt{n}h'\left(\frac{d\widehat{\lambda}(\theta_\circ)'}{d\theta}-\frac{d\lambda_\circ(\theta_\circ)'}{d\theta}\right)\mathbb{E}[g_i(\theta_\circ)] \\ + \mathbb{G}_n[e^{\lambda_\circ(\theta_\circ)'g_i(\theta_\circ)}g_i(\theta_\circ)']\check{\Omega}(\theta_\circ,\widehat{\lambda}_\tau)^{-1}\mathbb{E}_n\left[\frac{\partial\tau_i(\lambda_t,\theta_\circ)}{\partial\lambda}\lambda_\circ(\theta_\circ)'G_i(\theta_\circ)\right]h \\ - h'\mathbb{G}_n(\tau_i(\lambda_\circ(\theta_\circ),\theta_\circ)G_i(\theta_\circ)')\lambda_\circ(\theta_\circ) \\ + h'\mathbb{E}_n[\tau_i(\lambda_\circ,\theta_\circ)G_i(\theta_\circ)'-G_\circ']\check{\Omega}(\theta_\circ,\widehat{\lambda}_\tau)^{-1}\mathbb{G}_n[e^{\lambda_\circ(\theta_\circ)'g_i(\theta_\circ)}g_i(\theta_\circ)] + o_p(1) =: h'\mathcal{A}_{\theta_\circ}\Delta_{n,\theta_0} + o_p(1) \\ \text{(B.20)}$$

where the $o_p(1)$ is uniform in $h\in\mathcal{H}$, and $\dot{\mathcal{L}}_{n,\theta_\circ}(w_i) := \frac{d}{d\theta}\mathcal{L}_{n,\theta}(w_i)\Big|_{\theta=\theta_\circ}$ with

$$\mathcal{L}_{n,\theta_\circ}(w_i) := \log(dQ^*(\theta_\circ)/dP_*)(w_i) = \log\frac{e^{\lambda_\circ(\theta_\circ)'g_i(\theta_\circ)}}{\mathbb{E}^P[e^{\lambda_\circ(\theta_\circ)'g_i(\theta_\circ)}]}.$$

Moreover, as shown above $\Delta_{n,\theta_0} = O_p(1)$ and $\mathcal{A}_{\theta_\circ}$ is defined below.

We now analyse the limit of (B.17). For this, we use Lemma G.12, the fact that

$$(\widehat{\lambda}(\theta_\circ)-\lambda_\circ(\theta_\circ)) = -\check{\Omega}(\theta_\circ,\widehat{\lambda}_\tau)^{-1}\mathbb{E}_n[e^{\lambda_\circ(\theta_\circ)'g_i(\theta_\circ)}g_i(\theta_\circ)]$$

as shown above, and the fact that $\widehat{\lambda}(\widetilde{\theta})' - \widehat{\lambda}(\theta_\circ)' = (\widetilde{\theta} - \theta_0)\frac{d\widehat{\lambda}(\widetilde{\theta}_2)'}{d\theta}$ for $\widetilde{\theta}_2 = \theta_\circ + th/\sqrt{n}$ and some $t \in [0,1]$, to get

$$
\begin{aligned}
h'\frac{1}{n}\sum_{i=1}^{n}\ddot{\ell}_{n,\widetilde{\theta}}(w_i)h &= h'\sum_{j=1}^{dK}\frac{d^2\lambda_\circ(\theta_\circ)'}{d\theta d\theta'}\mathbb{E}[g_{i,j}(\theta_\circ)]h + h'\frac{d\lambda_\circ(\theta_\circ)'}{d\theta}\mathbb{E}[G_i(\theta_\circ)]h \\
&+ h'\mathbb{E}\left[G_i(\theta_\circ)'\lambda_\circ(\theta_\circ)\frac{d\tau_i(\lambda_\circ(\theta_\circ),\theta_\circ)}{d\lambda_\circ'}\frac{d\lambda_\circ(\theta_\circ)}{d\theta'}\right]h + h'\mathbb{E}\left[\frac{d\tau_i(\lambda_\circ(\theta_\circ),\theta_\circ)}{d\theta}\lambda_\circ(\theta_\circ)'G_i(\theta_\circ)\right]h \\
+h'\mathbb{E}&\left[\tau_i(\lambda_\circ(\theta_\circ),\theta_\circ)\sum_{j=1}^{dK}\frac{d^2 g_{i,j}(\theta_\circ)}{d\theta d\theta'}\right]\lambda_{\circ,j}(\theta_\circ) + h'\mathbb{E}\left[\tau_i(\lambda_\circ(\theta_\circ),\theta_\circ)G_i(\theta_\circ)'\right]\frac{d\lambda_\circ(\theta_\circ)}{d\theta'}h + o_p(1) =: h'\mathcal{A}_{\theta_\circ}h
\end{aligned}
\tag{B.21}
$$

where the $o_p(1)$ is uniform in $h \in \mathcal{H}$. By replacing (B.20) and (B.21) in (B.14) we get the result of the Lemma. □

C Proof of Theorem 4.1

We can write $\log p(w_{1:n}|\theta^\ell; M_\ell) = -n\log n + n\log\widehat{L}(\theta^\ell)$ where

$$
\widehat{L}(\theta^\ell) := \exp\{\widehat{\lambda}(\theta^\ell)'\widehat{g}_i(\theta^\ell)\}\left[\frac{1}{n}\sum_{i=1}^{n}\exp\{\widehat{\lambda}(\theta^\ell)'g_i(w_i,\theta^\ell)\}\right]^{-1}
$$

and $L(\theta^\ell) = \exp\{\lambda_\circ(\theta^\ell)'\mathbb{E}^P[g(w,\theta^\ell)]\}\left(\mathbb{E}^P\left[\exp\{\lambda_\circ(\theta^\ell)'g(w,\theta^\ell)\}\right]\right)^{-1}$. Then, we have:

$$
\begin{aligned}
P\left(\log m(w_{1:n}; M_j) > \max_{\ell\neq j}\log m(w_{1:n}; M_\ell)\right) &= P\Big(n\log\widehat{L}(\theta_\circ^j) + \log\pi(\theta_\circ^j|M_j) - \log\pi(\theta_\circ^j|w_{1:n}, M_j) \\
&\qquad > \max_{\ell\neq j}[n\log\widehat{L}(\theta_\circ^\ell) + \log\pi(\theta_\circ^\ell|M_\ell) - \log\pi(\theta_\circ^\ell|w_{1:n}, M_\ell)]\Big) \\
&= P\Big(n\log L(\theta_\circ^j) + n\log\frac{\widehat{L}(\theta_\circ^j)}{L(\theta_\circ^j)} + \mathcal{B}_j > \max_{\ell\neq j}\left[n\log L(\theta_\circ^\ell) + \mathcal{B}_\ell + n\log\frac{\widehat{L}(\theta_\circ^\ell)}{L(\theta_\circ^\ell)}\right]\Big)
\end{aligned}
\tag{C.1}
$$

where $\forall\ell$, $\mathcal{B}_\ell := \log\pi(\theta_\circ^\ell|M_\ell) - \log\pi(\theta_\circ^\ell|x_{1:n}, M_\ell)$ and $\mathcal{B}_\ell = O_p(1)$ under the assumptions of Theorem 3.2. By definition of $dQ^*(\theta)$ in Section 3.4 we have that: $\log L(\theta_\circ^\ell) = \mathbb{E}^P[\log dQ^*(\theta_\circ^\ell)/dP] = -\mathbb{E}^P[\log dP/dQ^*(\theta_\circ^\ell)] = -K(P||Q^*(\theta_\circ^\ell))$. Remark that $\mathbb{E}^P[\log(dP/dQ^*(\theta_\circ^2))] > \mathbb{E}^P[\log(dP/dQ^*(\theta_\circ^1))]$ means that the KL divergence between P and $Q^*(\theta_\circ^\ell)$, is smaller for model M_1 than for model M_2, where $Q^*(\theta_\circ^\ell)$ minimizes the KL divergence between $Q \in \mathcal{P}_{\theta_\circ^\ell}$ and P for $\ell \in \{1,2\}$ (notice the inversion of the two probabilities).

First, suppose that $\min_{\ell\neq j}\mathbb{E}^P\left[\log\left(dP/dQ^*(\theta_\circ^\ell)\right)\right] > \mathbb{E}^P\left[\log\left(dP/dQ^*(\theta_\circ^j)\right)\right]$. By (C.1):

$$P\left(\log m(w_{1:n};M_j) > \max_{\ell\neq j}\log m(w_{1:n};M_\ell)\right) \geq$$

$$P\left(\log\frac{\widehat{L}(\theta_\circ^j)}{L(\theta_\circ^j)} - \max_{\ell\neq j}\log\frac{\widehat{L}(\theta_\circ^\ell)}{L(\theta_\circ^\ell)} + \frac{1}{n}(\mathcal{B}_j - \max_{\ell\neq j}\mathcal{B}_\ell) > \underbrace{\max_{\ell\neq j}\log L(\theta_\circ^\ell) - \log L(\theta_\circ^j)}_{=:\mathcal{I}_n}\right). \quad \text{(C.2)}$$

This probability converges to 1 because $\mathcal{I}_n = K(P||Q^*(\theta_\circ^j)) - \min_{\ell\neq j} K(P||Q^*(\theta_\circ^\ell)) < 0$ by assumption, and $\left[\log\widehat{L}(\theta^\ell) - \log L(\theta^\ell)\right] \xrightarrow{p} 0$, for every $\theta^\ell \in \Theta^\ell$ and every $\ell \in \{1,2\}$ by Lemma G.10 and by $K/\sqrt{n} \to 0$.

To prove the second direction of the statement, suppose that

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

By (C.1) it holds, $\forall\ell\neq j$

$$P\left(\log m(w_{1:n};M_j) > \max_{\ell\neq j}\log m(w_{1:n};M_\ell)\right) \leq$$

$$P\Big(\log\frac{\widehat{L}(\theta_\circ^j)}{L(\theta_\circ^j)} - \log\frac{\widehat{L}(\theta_\circ^\ell)}{L(\theta_\circ^\ell)} + \frac{1}{n}(\mathcal{B}_j - \mathcal{B}_\ell) > \log\frac{L(\theta_\circ^\ell)}{L(\theta_\circ^j)}\Big). \quad \text{(C.3)}$$

Convergence to 1 of the left hand side implies convergence to 1 of the right hand side which is possible only if $\log L(\theta_\circ^\ell) - \log L(\theta_\circ^j) < 0$. Since this is true for every model ℓ, then this implies that $K(P||Q^*(\theta_\circ^j)) < \min_{\ell\neq j} K(P||Q^*(\theta_\circ^\ell))$ which concludes the proof. □

References

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

Chamberlain, G. (1987), 'Asymptotic Efficiency in Estimation with Conditional Moment Restrictions', *Journal of Econometrics* **34**(3), 305–334.

Chang, I. H. and Mukerjee, R. (2008), 'Bayesian and Frequentist Confidence Intervals Arising from Empirical-type Likelihoods', *Biometrika* **95**(1), 139–147.

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

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

Chib, S. and Greenberg, E. (2010), 'Additive Cubic Spline Regression with Dirichlet Process Mixture Errors', *Journal of Econometrics* **156**, 322–336.

Chib, S. and Jeliazkov, I. (2001), 'Marginal Likelihood from the Metropolis-Hastings Output', *Journal of the American Statistical Association* **96**, 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**, 1656–1668.

Csiszar, I. (1975), 'I-Divergence Geometry of Probability Distributions and Minimization Problems', *Annals of Probability* **3**(1), 146–158.

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.

Fang, K.-T. and Mukerjee, R. (2006), 'Empirical-Type Likelihoods Allowing Posterior Credible Sets with Frequentist Validity: Higher-Order Asymptotics', *Biometrika* **93**(3), 723–733.

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.

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., 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 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', *The Annals of Statistics* **39**(6), pp. 3003–3031.

Mengersen, K. L., Pudlo, P. and Robert, C. P. (2013), 'Bayesian computation via empirical likelihood', *Proceedings of the National Academy of Sciences of the United States of America* **110**(4), 1321–1326.

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.

Schennach, S. M. (2007), 'Point Estimation with Exponentially Tilted Empirical Likelihood', *Annals of Statistics* **35**(2), 634–672.

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.

Vexler, A., Deng, W. and Wilding, G. E. (2013), 'Nonparametric Bayes Factors Based on Empirical Likelihood Ratios', *Journal of Statistical Planning and Inference* **143**(3), 611–620.

Vexler, A., Tao, G. and Hutson, A. D. (2014), 'Posterior expectation based on empirical likelihoods', *Biometrika* **101**(3), 711–718.

Vexler, A., Yu, J. and Lazar, N. (2017), 'Bayesian empirical likelihood methods for quantile comparisons', *Journal of the Korean Statistical Society* **46**(4), 518–538.

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