Bayesian Estimation and Comparison of Conditional Moment Models

Siddhartha Chib∗ Minchul Shin† Anna Simoni‡

December 2019, October 2020

Abstract

We consider the Bayesian analysis of models in which the unknown distribution of the outcomes is specified up to a set of conditional moment restrictions. The nonparametric exponentially tilted empirical likelihood (ETEL) function is constructed to satisfy a sequence of unconditional moments based on an increasing (in sample size) vector of approximating functions (such as tensor splines based on the splines of each conditioning variable). For any given sample size, the number of such expanded moments is based on the predictive likelihood for held-out data. The posterior distribution is shown to satisfy the Bernstein-von Mises theorem, subject to a growth rate condition on the number of approximating functions, even under misspecification of the conditional moments. A large-sample theory for comparing different conditional moment models is developed. The central result is that the marginal likelihood criterion selects the model that is less misspecified. We also introduce sparsity-based model search for high-dimensional conditioning variables, and provide efficient MCMC computations for high-dimensional parameters. Along with clarifying examples, the value of the techniques is illustrated with real-data applications to risk-factor determination in finance, and causal inference under conditional ignorability.

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

1 Introduction

We tackle the problem of prior-posterior inference when the only available information about the unknown parameter \( \theta \in \Theta \subset \mathbb{R}^p \) is supplied by a set of conditional moment (CM) restrictions

\[
E^P[\rho(X, \theta)|Z] = 0,
\]

where \( \rho(X, \theta) \) is a \( d \)-vector of known functions of a \( \mathbb{R}^{d_x} \)-valued random vector \( X \) and the unknown \( \theta \), and \( P \) is the unknown conditional distribution of \( X \) given a \( \mathbb{R}^{d_z} \)-valued random vector \( Z \). Such models

∗Olin Business School, Washington University in St. Louis, Campus Box 1133, 1 Brookings Drive, St. Louis, MO 63130. e-mail: chib@wustl.edu.
†Federal Reserve Bank of Philadelphia, Ten Independence Mall, Philadelphia, PA 19106. e-mail: visiblehand@gmail.com.
‡CREST, CNRS, École Polytechnique, ENSAE, 5, Avenue Henry Le Chatelier, 91120 Palaiseau - France, e-mail: simoni.anna@gmail.com.
are important because many standard models in statistics can be recast in terms of CM restrictions. These models also arise naturally in causal inference, missing data problems, and in models derived from theory in economics and finance. Because the CM conditions constrain the set of possible distributions $P$, we say that the model is correctly specified if the true data generating process $P_\ast$ is in the set of distributions constrained to satisfy these moment conditions for some $\theta \in \Theta$, while the model is misspecified if $P_\ast$ is not in the set of implied distributions for any $\theta \in \Theta$.

A different starting point is when one is given the unconditional moments, say $E^P[g(X, \theta)] = 0$. As in the conditional model, distributional assumptions are entirely bypassed, but prior-posterior analysis can be based on the empirical likelihood, for example, [Lazar (2003)](Lazar2003) and many others, or the exponentially tilted empirical likelihood (ETEL), as in [Schennach (2005)](Schennach2005) and Chib, Shin and Simoni (2018). The latter paper, in particular, provides the large sample behavior of the posterior distribution under misspecification, and a framework, and large-sample model consistency results, for comparing such models by marginal likelihoods.

There are several important reasons for developing a similar set of results for CM models. For one, conditional moments often supply the only source of information about $P$. Examples of this include causal inference, as in [Rosenbaum and Rubin (1983)](RosenbaumRubin1983), where the potential outcomes are assumed independent of the treatment variable conditioned on covariates, and in missing at random problems as considered by [Hristache and Patilea (2017)](HristachePatilea2017), and numerous others. Second, while the conditional moments imply that $\rho(X, \theta)$ is uncorrelated with $Z$, i.e., $E^P[\rho(X, \theta) \otimes Z] = 0$, where $\otimes$ is the Kronecker product operator, the conditional moments assert even more. Specifically, that $\rho(X, \theta)$ is uncorrelated with any measurable, bounded function of $Z$, or if $Z$ is square-integrable, that it is uncorrelated with any $L^2$-measurable function of $Z$. Thus, there is an efficiency loss if this information is ignored.

One can eliminate this efficiency loss by assembling the set of equivalent unconditional moments. This is only feasible as the sample size $n$ goes to infinity. A general result due to [Donald, Imbens and Newey (2003)](DonaldImbensNewey2003) states that this equivalent set of unconditional moments can be constructed by approximating functions $q^K(Z) := (q^K_1(Z), \ldots, q^K_K(Z))'$, such as tensor product splines obtained from splines of each variable in $Z$, with the number of such functions, denoted by $K$, increasing with $n$. Then, instead of (1.1), inference based on the expanded unconditional moment conditions

$$E^P[\rho(X, \theta) \otimes q^K(Z)] = 0$$

(1.2)

is valid. This is then how we proceed in this paper.
Despite the transformation into unconditional moments, the large-sample behavior of the posterior distribution is different from that in Chib, Shin and Simoni (2018). Quantities that are bounded with fixed moment restrictions, now diverge with $K$. This growth can be stabilized, however, if $K$ grows slowly, by a factor that we derive to be essentially $n^{1/6}$ in the correctly specified case, where $n$ is the sample size. Then, under this rate condition and correct specification of the conditional moments, we show that the posterior distribution of $\theta$ satisfies the Bernstein-von Mises (BvM) theorem with asymptotic posterior variance equal to the semiparametric efficiency bound derived in Chamberlain (1987).

Extending this BvM-type phenomenon to the more realistic misspecified case is obviously important. Just as in Kleijn and van der Vaart (2012), we show that the posterior distribution of the centered and scaled parameter $\sqrt{n}(\theta - \theta^*)$, where $\theta^*$ is the pseudo-true value, converges to a Normal distribution with a random mean that is bounded in probability. Interestingly, the rate result shows that $K$ must increase more slowly than in the correctly specified case. The intuition for this is revealing. When the conditional moment conditions are misspecified, the derived unconditional moments magnify the misspecification. Thus, one should limit the growth of the expanded moments with $n$.

While the rate conditions speak to the growth of $K$ with $n$, we also show that for any given $n$, the posterior distribution is robust to the specific choice of $K$, provided one avoids relatively small or relatively large values. One avoids relatively small values for the efficiency reason mentioned above. There are two reasons for avoiding relatively large values. One is because the effective rank of the basis matrix typically becomes less than the column rank. Data-compression methods based on the principal components of the basis matrix can be employed to mitigate this problem. Another is that, as $K$ increases for given $n$, the set of $\theta$ values for which 0 is in the interior of the convex hull of the expanded moments shrinks. In other words, the effective support of the prior (hence, of the posterior) shrinks with $K$, which tends to increase the Bayesian bias. If these two extremes are avoided, the results are robust to choice of $K$. Further fine-tuning of $K$ in the reasonable range is possible. We suggest examining the predictive likelihood (for different values of $K$) for held-out data.

We also obtain an interesting result on the large-sample behavior of marginal likelihoods (used for the comparison of conditional moment models). The central result is that the marginal likelihood criterion selects the model that is less misspecified. Notwithstanding some similarities with the framework in Chib, Shin and Simoni (2018), these marginal likelihoods can be calculated for the CM models as stated, without any reformulation to equalize the number of conditional moments.
Our approach is shown to extend conveniently to the case of a high-dimensional $Z$ and that of a high-dimensional parameter. For the former, one can again employ data-compression methods. In addition, under the assumption that the model depends on a few conditioning variables, a procedure called sparsity-based model search can be used to infer the relevant conditioning variables. On the other hand, in the high-dimensional parameter case, the computational challenges can be overcome by the TaRB-MH algorithm of Chib and Ramamurthy (2010). These techniques are illustrated in real-data applications of risk-factor determination in finance, and causal inference under conditional ignorability.

We note that the framework here departs from other Bayesian treatments of conditional moments, for instance, Liao and Jiang (2011), Florens and Simoni (2012, 2016) and Kato (2013), which consider the nonparametric instrumental variables setting, and Chen, Christensen and Tamer (2018) and Liao and Simoni (2019), which consider models with partially identified parameters in the moment function. Generally, inference is these papers is not fully Bayesian and focuses on different topics. Moreover, these papers ignore the problems of misspecification and model comparisons.

The rest of the paper is organized as follows. Section 2 has the sketch of the conditional moment setting. Section 3 gives the transformation of the conditional moments into unconditional moments by an increasing (in sample size) vector of approximating functions. Results on the large sample behavior of the posterior distribution in both the correct and misspecified moment models are included. Section 4 develops the theory for comparing models and the large sample behavior of the marginal likelihood. In Section 5 two extensions are considered. Section 6 is concerned with real data applications to finance and causal inference. Section 7 concludes. Abbreviated technical proofs of the theorems are in the appendix. Full proofs are in the online supplementary appendix.

### 2 Setting and Motivation

Let $X := (X'_1, X'_z)'$ be an $\mathbb{R}^{d_x}$-valued random vector and $Z := (Z'_1, X'_z)'$ be an $\mathbb{R}^{d_z}$-valued random vector. The vectors $Z$ and $X$ have elements in common if the dimension of the subvector $X_z$ is non-zero. Moreover, we denote $W := (X', Z'_z)' \in \mathbb{R}^{d_w}$ 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 $E^P[\cdot]$ the expectation with respect to $P$ and by $E^P[\cdot|\cdot]$ the conditional expectation with respect to the conditional distribution associated with $P$. 
The parameter of interest is \( \theta \in \Theta \subset \mathbb{R}^p \), which is related to the conditional distribution \( P \) through the conditional moment restrictions

\[
E^P[\rho(X, \theta) | Z] = 0,
\]

(2.1)

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

**Example 1** (Linear model with heteroscedasticity of unknown form) Suppose that

\[
E^P[(Y - \theta_0 - \theta_1 X) | Z] = 0,
\]

(2.2)

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

\[
E^P[(Y - \theta_0 - \theta_1 X) \otimes (1, X)' | Z] = 0 \quad \text{and} \quad E^P[(Y - \theta_0 - \theta_1 X)^3 | Z] = 0,
\]

(2.3)

where now \( \rho(X, \theta) \) is a \((2 \times 1)\) vector of functions, we additionally impose conditional symmetry of \( \varepsilon \).

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

\[
E^P[(Y - \theta_0 - \theta_1 X) \otimes (1, X)'] = 0,
\]

(2.4)

which assert that: (i) \( \varepsilon \) has mean zero and (ii) \( \varepsilon \) is uncorrelated with \( X \), are weaker but, if the CM model is correct, less informative about \( \theta \).

### 3 Prior-Posterior Analysis

#### 3.1 Expanded Moment Conditions

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

5
1987), a countable number of unconditional moment restrictions that are equivalent to the CM restrictions in (2.1) is guaranteed. This is the basis of the frequentist approaches in Donald and Newey (2001), Ai and Chen (2003) and Carrasco and Florens (2000) where, after transforming the CM restrictions into unconditional moment restrictions, the resulting set of unconditional moments are analyzed either under a sieve approach based on truncation or a Tikhonov regularization approach. Following Donald, Imbens and Newey (2003), the equivalent set of unconditional moments can be obtained through approximating functions.

Let \( q^K(Z) := (q^K_1(Z), \ldots, q^K_K(Z))^\prime, K > 0, \) 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 \( K \), \( E^P[q^K(Z)^\prime q^K(Z)] \) is finite, and for any function \( a(z) : \mathbb{R}^d \to \mathbb{R} \) with \( E^P[a(Z)^2] < \infty \) there are \( K \times 1 \) vectors \( \gamma_K \) such that as \( K \to \infty \),

\[
E^P[(a(Z) - q^K(Z)^\prime \gamma_K)^2] \to 0.
\]

Now, let \( \theta_* \) be the value of \( \theta \) that satisfies (2.1) for the true \( P \). If \( E^P[\rho(X, \theta_*)^\prime \rho(X, \theta)] < \infty \), then Donald, Imbens and Newey (2003) Lemma 2.1 established that: (1) if equation (2.1) is satisfied with \( \theta = \theta_* \), then \( E^P[\rho(X, \theta_*) \otimes q^K(Z)] = 0 \) for all \( K \); (2) if equation (2.1) is not satisfied by \( \theta = \theta_* \), then \( E^P[\rho(X, \theta_*) \otimes q^K(Z)] \neq 0 \), for all large enough \( K \).

Henceforth, we let \( g(W, \theta) := \rho(X, \theta) \otimes q^K(Z) \) denote the expanded functions and refer to

\[
E^P[g(W, \theta)] = 0,
\] as the expanded moments. Under the stated assumptions, the expanded moments are equivalent to the CM restrictions (2.1), as \( K \to \infty \).

In our numerical examples, we construct \( q^K(Z) \) using the natural cubic spline basis of Chib and Greenberg (2010), with \( K \) fixed at a given value, as in sieve estimation. If \( Z \) consists of more than one element, say \( (Z_1, Z_2, Z_3) \) where \( Z_1 \) and \( Z_2 \) are continuous variables and \( Z_3 \) is binary, then the basis matrix \( B \) is constructed as follows. Let \( z_j \) denote the \( n \times 1 \) sample data on \( Z_j \) \( (j \leq 3) \). Let \( Z = (z_1, z_2, z_1 \odot z_2, z_1 \odot z_3, z_2 \odot z_3) \) denote the \( n \times 5 \) matrix of the continuous data and interactions of the continuous data and the binary data. Now suppose \( (\tau_{j1}, \ldots, \tau_{jK}) \), for \( j = 1, \ldots, 5 \) 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, $B$ is given by

$$B = \begin{bmatrix} B_1 \colon B_2^* \colon B_3^* \colon B_4^* \colon B_5^* \colon Z_3 \end{bmatrix},$$

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 $B$ matrix is $n \times K^*$, where $K^* = (5K - 4 + 1)$. If $K^*$ is large, in relation to $n$, data-compression methods can be employed. Specifically, let $R$ denote the $K^* \times K^*$ orthogonal matrix of eigenvectors from the singular value decomposition of $B$, and let $e$ denote the corresponding $K^* \times 1$ vector of eigenvalues. Then, after employing the rotation $BR$, the columns of $BR$ corresponding to small values of $e$ are dropped, and the resulting column-reduced $BR$ matrix is taken as the basis matrix. We refer to this as the rotated column reduced basis matrix. To define the expanded functions, let $\rho_l(X, \theta)$ ($l \leq d$) denote a $n \times 1$ vector of the $l$th element of $\rho(X, \theta)$ evaluated at the sample data matrix $X$. Then, the expanded functions for the sample observations are obtained by multiplying $\rho_l(X, \theta)$ by the matrix $B$ (or by the rotated column reduced $B$) and concatenating. We use versions of this approach in our examples.

Example 1 (continued) Let $x := (x_1, \ldots, x_n) \in \mathbb{R}^n$ denote the sample data, and $(\tau_1, \ldots, \tau_K)$ the $K$ knots, where the exterior knots $\tau_1$ and $\tau_K$ are the minimum and maximum values of $x$, and the interior knots are specified quantile points of $x$. 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 $\tau_j$. Let $B$ denote the $(n \times K)$ matrix of these basis functions evaluated at $x$, where the $i$th row of $B$ is given by $q^K(x_i)'$. Let $(y - \theta_0\ell - \theta_1x)$ and $(y - \theta_0\ell - \theta_1x)^3$ each denote $n \times 1$ vectors where $y := (y_1, \ldots, y_n)$ and $\ell$ is the $n$-vector with all components equal to one. Then, the expanded functions $g(w, \theta) = g(x, \theta)$ evaluated at the sample observations are the $n \times 2K$ functions

$$g(w, \theta) = [\rho(x, \theta)' \odot q^K(x)'] = [(y - \theta_0\ell - \theta_1x) \odot B : (y - \theta_0\ell - \theta_1x)^3 \odot B], \quad (3.2)$$

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

3.2 Posterior distribution

The CM model (2.1) is semiparametric and is characterized by two parameters: the data distribution $P$ and the structural parameter $\theta$, which is assumed to be finite dimensional. For a given value of $K$, the prior on $(\theta, P)$ is specified as $\pi(\theta)\pi(P|\theta, K)$, where the prior on $\theta$ is standard. Our default prior on $\theta$
is a product of independent student-\(t\) distributions with 2.5 degrees of freedom on each component of \(\theta\).

We discuss the choice of \(K\) below. In establishing the asymptotic properties of the posterior distribution, however, \(K\) must grow to infinity with the sample size to ensure that (2.1) and (3.1) are equivalent in the limit.

Priors on \(P\) designed to incorporate overidentifying moment restrictions are those of Schennach (2005), Kitamura and Otsu (2011), Shin (2014) and Florens and Simoni (2019). Our prior \(\pi(P|\theta, K)\) follows from Schennach (2005). To construct this prior, we first model the joint data distribution \(P\) of \(W\) as a mixture of uniform probability densities, a construction which is capable of approximating any distribution as the number of mixing components increases. Then, a prior is placed on the center of the \(d_w\)-dimensional hypercubes such that the corresponding mixture satisfies the moment restrictions for a given \((\theta, K)\). The resulting posterior is well-defined for every value of \(K\).

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

\[
 p(W_{1:n}|\theta, K) = \prod_{i=1}^{n} \hat{p}_i(\theta),
\]

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

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

(see Schennach (2005) for a proof). In practice, we compute the ETEL probabilities from the dual (saddlepoint) representation (see e.g. Csiszar (1984)) as

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

where \(\hat{\lambda}(\theta) = \arg \min_{\lambda \in \mathbb{R}^d K} \frac{1}{n} \sum_{i=1}^{n} e^{\lambda g(w_i, \theta)}\) is the estimated tilting parameter.

Let \(\text{Co}(\theta) := \{ p_i g(w_i, \theta), p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \} \) be the convex hull of \(\{g(w_i, \theta)\}_{i=1}^{n}\) for a given \(\theta\) and \(\text{Co}(\theta)\) denote its interior. Let \(H_{n,K} := \{ \theta \in \Theta; 0 \in \text{Co}(\theta) \}\) denote the set of \(\theta\) values for which 0 is in the interior of the convex hull of \(\{g(w_i, \theta)\}_{i=1}^{n}\) (that is, for which the empirical moment conditions hold). Then, the posterior distribution is the truncated distribution given by

\[
 \pi(\theta|w_{1:n}, K) \propto \pi(\theta) \prod_{i=1}^{n} \frac{e^{\hat{\lambda}(\theta)g(w_i, \theta)}}{\sum_{j=1}^{n} e^{\hat{\lambda}(\theta)g(w_j, \theta)}} 1\{\theta \in H_{n,K}\}. \tag{3.6}
\]
Below we provide the large-sample behavior of this posterior distribution under both correct and mis-specified moment conditions.

Note, however, that while the asymptotic theory below provides the large-sample guarantees, the posterior distribution, for any given $n$, is summarized by MCMC simulations. One needs to exercise care in developing an efficient MCMC procedure. Our recommendation, based on extensive experimentation, is to use the one block tailored Metropolis-Hastings (M-H) algorithm of Chib and Greenberg [1995] in low-dimensional problems, and the Tailored Randomized Block Metropolis-Hastings algorithm of Chib and Ramamurthy [2010] otherwise. Each sampling scheme is illustrated below.

Example 1 (continued) To illustrate the role of $K$ in the prior-posterior analysis, we create a set of simulated data $\{y_i, x_i\}_{i=1}^n$, $n = 250$, with covariates $X \sim U(-1, 2.5)$, intercept $\theta_0 = 1$, slope $\theta_1 = 1$, and $\varepsilon$ is distributed according to $\varepsilon \sim SN(m(x_i), h(x_i), s(x_i))$, where $SN(m, h, s)$ is the skew normal distribution with location, scale, and shape parameters given by $(m, h, s)$, each depending on $x_i$. When $s$ is zero, $\varepsilon$ is normal with mean $m$ and standard deviation $h$. We set $m(x_i) = -h(x_i)\sqrt{2/\pi s(x_i)}/(\sqrt{1 + s(x_i)^2})$ so that $E_P[\varepsilon|X] = 0$.

Suppose that $h(x) = \sqrt{\exp(1 + 0.7x + 0.2x^2)}$ and $s(x) = 1 + x^2$. We estimate the model using $E_P[\varepsilon|Z] = 0$, $Z = (1, X)$, without the need to model the heteroscedasticity or the skewness functions. The prior is the default independent student-$t$ distribution with location 0, dispersion 5, and degrees of freedom 2.5, truncated to $H_{n,K}$. The posterior is computed for $K$ given by 2, $2n^{1/6}$, 9 and 20 (the value $2n^{1/6}$ is suggested by our theory below). The results are shown in Table 1. Importantly, when $K$ is close to the value suggested by our theory, the specific choice of $K$ does not meaningfully alter the results. However, when $K = 20$, quite different from the recommended value, the Bayesian bias is larger (especially so for $\theta_2$) and the posterior standard deviation is smaller. This is due to the effect of the prior, in the following way. As $K$ increases for a fixed $n$, the volume of $H_{n,K}$ decreases, equivalently, the support of the prior distribution shrinks. As a consequence, one gets a larger Bayesian bias and a smaller posterior standard deviation.

### 3.3 Asymptotic properties

Consider now the large sample behavior of the posterior distribution of $\theta$. We let $\theta_*$ and $P_*$, respectively, denote the true value of $\theta$ and of the data distribution $P$. As notation, when the true distribution $P_*$ is involved, expectations $E_P[\cdot]$ (resp. $E_P[\cdot|\cdot]$) are taken with respect to $P_*$ (resp. the conditional distribution...
<table>
<thead>
<tr>
<th>$K$</th>
<th>$\theta_1$</th>
<th>SD</th>
<th>Median</th>
<th>Lower</th>
<th>Upper</th>
<th>Ineff</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.07</td>
<td>0.099</td>
<td>1.07</td>
<td>0.88</td>
<td>1.27</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>1.01</td>
<td>0.14</td>
<td>1.01</td>
<td>0.74</td>
<td>1.30</td>
<td>1.10</td>
</tr>
<tr>
<td>5</td>
<td>1.07</td>
<td>0.10</td>
<td>1.07</td>
<td>0.88</td>
<td>1.26</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>1.03</td>
<td>0.12</td>
<td>1.03</td>
<td>0.80</td>
<td>1.25</td>
<td>1.18</td>
</tr>
<tr>
<td>9</td>
<td>1.07</td>
<td>0.10</td>
<td>1.06</td>
<td>0.88</td>
<td>1.26</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td>1.02</td>
<td>0.11</td>
<td>1.02</td>
<td>0.80</td>
<td>1.25</td>
<td>1.11</td>
</tr>
<tr>
<td>20</td>
<td>0.99</td>
<td>0.07</td>
<td>0.99</td>
<td>0.86</td>
<td>1.13</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>1.12</td>
<td>0.10</td>
<td>1.12</td>
<td>0.94</td>
<td>1.31</td>
<td>1.10</td>
</tr>
</tbody>
</table>

Table 1: Simulated data example with $n = 250$: Posterior summary for $K$ given by 2, 2n\(^{1/6}\) and 9. Results based on 20,000 MCMC draws beyond a burn-in of 1000. “Lower” and “Upper” refer to the 0.05 and 0.95 quantiles of the simulated draws, respectively, and “Ineff” to the inefficiency factor.

For a vector $a$, $\|a\|$ denotes the Euclidean norm. For a matrix $A$, $\|A\|$ denotes the operator norm (the largest singular value of the matrix). Finally, let $Z := \text{supp}(Z)$ denote the support of $Z$.

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 $\sup_{z \in Z} \|q^K(z)\| \leq \zeta(K)$, $E_P[q^K(Z)q^K(Z)' \mid Z]$ 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 $W_i := (X_i, Z_{1i})$, $i = 1, \ldots, n$ are i.i.d. according to $P_*$ and (a) there exists a unique $\theta_* \in \Theta$ that satisfies $E_P[\rho(X, \theta)] = 0$ for the true $P_*$; (b) $\Theta$ is compact; (c) $E_P[\sup_{\theta \in \Theta} \|\rho(X, \theta)\|^2 \mid Z]$ is bounded on $Z$. 
This assumption is the same as Donald et al. (2003, Assumption 3). The following three assumptions are also the same as the ones required by Donald et al. (2003) to establish asymptotic normality of the Generalized Empirical Likelihood (GEL) estimator.

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

(c) \( E^P[D(X)D(X)'] \) is nonsingular.

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

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

Part (b) of Assumption 3.5 imposes a Lipschitz condition which allows application of uniform convergence results. The last assumption is about the prior distribution of \( \theta \) and is standard in the Bayesian literature on frequentist asymptotic properties of Bayes procedures.

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

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

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

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

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

\[
\sup_B \left| \pi(\sqrt{n}(\theta - \theta_*) \in B|W_{1:n}, K) - N_{\Delta_{n,\theta_*}, V_{\theta_*}}(B) \right| \to 0, \tag{3.8}
\]

where \( B \subseteq \Theta \) is any Borel set, \( \Delta_{n,\theta_*} := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_{\theta_*} D(Z_i)'\Sigma(Z_i)^{-1} \rho(X_i, \theta_*) \) is bounded in probability and \( V_{\theta_*} := \left( E^P[D(Z)'\Sigma(Z)^{-1}D(Z)] \right)^{-1} \).
We note that the centering $\Delta_{n,\theta^*}$ of the limiting normal distribution satisfies $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d \log \hat{p}_i(\theta^*)}{d\theta} - V_{\theta^*}^{-1} \Delta_{n,\theta^*} \xrightarrow{p} 0$. We also note that the condition $\zeta(K) K^2 / \sqrt{n} \rightarrow 0$ in the theorem implies $K/n \rightarrow 0$, which is a classical condition in the sieve literature. This condition is required to establish a stochastic Local Asymptotic Normality (LAN) expansion, which is an intermediate step to prove the BvM result, as we explain below. The LAN expansion is not required to establish asymptotic normality of the GEL estimators, which explains why our condition is slightly stronger than the condition $\zeta(K) K^2 / \sqrt{n} \rightarrow 0$ required by Donald, Imbens and Newey (2003). On the other hand, our condition is weaker than the condition $\zeta(K) K^2 / \sqrt{n} \rightarrow 0$ required by Donald, Imbens and Newey (2009) to establish the mean square error of the GEL estimators. The asymptotic covariance of the posterior distribution coincides with the semiparametric efficiency bound given in Chamberlain (1987) for conditional moment condition models. This means that, for every $\alpha \in (0, 1)$, $(1 - \alpha)$-credible regions constructed from the posterior of $\theta$ are $(1 - \alpha)$-confidence sets asymptotically. Indeed, they are correctly centered and have correct volume.

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

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

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

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

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

We now generalize the preceding BvM result for the important class of misspecified conditional moment models.

**Definition 3.1 (Misspecified model)** We say that the conditional moment conditions model $E^P[\rho(X, \theta)|Z] = 0$ is misspecified if the set of probability measures implied by the moment restrictions does not contain the true data generating process $P^*$ for any $\theta \in \Theta$, that is, $P_* \notin \mathcal{P}$ where $\mathcal{P} := \bigcup_{\theta \in \Theta} \tilde{P}_\theta$ and $\tilde{P}_\theta = \{ Q \in M_{X|Z}; E^Q[\rho(X, \theta)|Z] = 0 \text{ a.s.} \}$ with $M_{X|Z}$ 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 $E^P[\rho(X, \theta) \otimes q^K(Z)] = 0$ 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 $P_\theta := \{ Q \in M; E^Q[g(W, \theta)] = 0 \}$, where $M$ denotes the set of all the probability measures on $\mathbb{R}^d_{W}$. That is, $Q^*(\theta) := \arg\inf_{Q \in P_\theta} \mathbb{K}(Q||P_*):= \int \log(dQ/dP_*)dQ$. 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: $[dQ^*(\theta)/dP_*](W_i) = e^{\lambda^* g(W_i, \theta)} E^P[e^{\lambda^* g(W_j, \theta)}]$, where $\lambda^*$ denotes the tilting parameter and is defined in the same way as in the correctly specified case:

$$\lambda^* := \lambda^*(\theta) := \arg\min_{\lambda \in \mathbb{R}^d} \mathbb{E}^P[e^{\lambda g(W_i, \theta)}]. \tag{3.11}$$

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

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

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

$$\theta^* := \arg\inf_{\theta \in \Theta} \mathbb{K}(P_*||Q^*(\theta)), \tag{3.12}$$
where $\mathbb{K}(P_\theta|Q^*(\theta)) := \int \log(dP_\theta/dQ^*(\theta))dP_\theta$. Under the preceding absolute continuity assumption, the pseudo-true value $\theta_0$ is available as

$$
\theta_0 = \arg\max_{\theta \in \Theta} \mathbb{E}^P \log \left( \frac{e^{\lambda_0 g(W, \theta)}}{\mathbb{E}^P[e^{\lambda_0 g(W, \theta)}]} \right).
$$

(3.13)

Note that $\lambda_0(\theta_0)$, the value of the tilting parameter at the pseudo-true value $\theta_0$, is nonzero because the moment conditions do not hold.

Assumption 3.8 implies that $\mathbb{K}(Q^*(\theta_0)|P_\theta) < \infty$. We supplement this with the assumption that $\mathbb{K}(P_\theta|Q^*(\theta)) < \infty$, $\forall \theta \in \Theta$ (so that $\mathbb{K}(P_\theta|Q^*(\theta_0)) < \infty$). Because consistency in misspecified models is defined with respect to the pseudo-true value $\theta_0$, we need to replace Assumption 3.7(b) by the following Assumption 3.9(b) which, together with Assumption 3.9(a), requires the prior to put enough mass to balls around $\theta_0$.

**Assumption 3.9**

(a) $\pi$ is a continuous probability measure that admits a density with respect to the Lebesgue measure; (b) The prior distribution $\pi$ is positive on a neighborhood of $\theta_0$, where $\theta_0$ is as defined in (3.13).

In the next assumption we denote by $\text{int}(\Theta)$ the interior of $\Theta$ and by $U$ a ball centered at $\theta_0$ with radius $h/\sqrt{n}$ for some $h \in \mathcal{H}$ and $\mathcal{H}$ a compact subset of $\mathbb{R}^p$.

**Assumption 3.10**

(a) The data $W_i := (X_i, Z_i)$, $i = 1, \ldots, n$ are i.i.d. according to $P_\theta$ and
(b) The pseudo-true value $\theta_0 \in \text{int}(\Theta)$ is the unique maximizer of

$$
\lambda_c(\theta)^T \mathbb{E}^P[g(W, \theta)] - \log \mathbb{E}^P[\exp\{\lambda_c(\theta)^T g(W, \theta)\}],
$$

where $\Theta$ is compact;
(c) $\rho(X, \theta)$ is continuous at each $\theta \in \Theta$ with probability one;
(d) $\rho(X, \theta)$ is twice continuously differentiable in a neighborhood $U$ of $\theta_0$ and for $\kappa = 0, 1, 2$

$$
\mathbb{E}^P[\sup_{\theta \in U} e^{\lambda_c(\theta)^T g(W, \theta)}||\rho_j\theta_0(x, \theta)||^2||q^\kappa(Z)||^2] = O(K), j = 1, \ldots, d;
$$
(e) for a neighborhood $U$ of $\theta_0$ and for $\kappa = 0, 1, 2, j = 0, 2, 4$ it holds that

$$
\mathbb{E}^P \left[ \sup_{\theta \in U} e^{\lambda_c(\theta)^T g(W, \theta)}||\rho(X, \theta)||^2||q^\kappa(Z)||^2 \right] = O(\zeta(K)^{\max(j-2,0)K})
$$

is bounded, where $\zeta(K)$ is as defined in Assumption 3.2;
(f) for a neighborhood $U$ of $\theta_0$ and for $\kappa = 0, 1, 2, j = 2, 4$ it holds that

$$
\mathbb{E}^P \left[ \sup_{\theta \in U} e^{\lambda_c(\theta)^T g(W, \theta)}||\rho_\theta(X, \theta)||^2||q^\kappa(Z)||^2 \right] = O(\zeta(K)^{j-2K}),
$$
where $\zeta(K)$ is as defined in Assumption 3.2.

(g) the matrix $E_P \left[ e^{\lambda_o(\theta_o)^\prime} g(W, \theta_o) \rho(X, \theta_o) \rho(X, \theta_o)^\prime | Z \right]$ has smallest eigenvalue bounded away from zero.

Assumption 3.10 (b) 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 (b). Assumption 3.10 (f) is the misspecified counterpart of Assumption 3.5 (a). Remark that the presence of the exponential $e^{\lambda_o(\theta_o)^\prime} g(W, \theta_o)$ inside the expectations in Assumption 3.10 (d)-(g) is due to the fact that in the misspecified case the pseudo-true value of the tilting parameter $\lambda_o(\theta_o)$ is not equal to zero as it is in the correctly specified case.

Our next important theorem, the BvM theorem for misspecified models, now follows. Denote $\tilde{D}_o(Z) := E^{Q^*(\theta_o)} [\rho(X, \theta_o) | Z]$ and $\tilde{\Sigma}_o(Z) := E^{Q^*(\theta_o)} [\rho(X, \theta_o) \rho(X, \theta_o)^\prime | Z]$ where here and in the following we use the sub/super index $Q^*(\theta_o)$ to denote an expectation, a variance or covariance taken with respect to the probability $Q^*(\theta_o)$.

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

$$P_* \left( \sup_{\|\theta - \theta_o\| > M_n / \sqrt{n}} \frac{1}{n} \sum_{i=1}^{n} (\ell_n,\theta(W_i) - \ell_n,\theta_o(W_i)) \leq - CM^2 / n \right) \to 1,$$

as $n \to \infty$. Let $H$ denote a compact subset of $\mathbb{R}^p$ and $\theta_h := \theta_o + h / \sqrt{n}$ with $h \in H$. Suppose that $\sup_{h \in H} \|\lambda_o(\theta_h)\|^2 \zeta(K) K^2 \sqrt{K / n} \to 0$. Then, the posteriors converge in total variation towards a Normal distribution, that is,

$$\sup_{B} \left| \pi(\sqrt{n}(\theta - \theta_o)) \in B | W_{1:n}, K \right| \to 0,$$

where $B \subseteq \Theta$ is any Borel set, $\Delta_{n,\theta_o}$ is a random vector bounded in probability and $V_{\theta_o}$ is a positive definite matrix. If in addition Assumption 3.11 holds and $K \to \infty$ then $\tilde{V}_{\theta_o}^{-1} = E_P^P [ \tilde{D}_o(Z) (\tilde{\Sigma}_o(Z))^{-1} D_o(Z) ] + E^{Q^*(\theta_o)} \left[ G_i(\theta)^\prime \lambda_o(\theta_o) g_i(\theta_o)^\prime \right] (\Omega^\prime(\theta_o))^{-1} G(\theta_o) - \frac{d\lambda_o(\theta_o)^\prime}{d\theta} \left( E_P^P [G_i(\theta_o)] - E^{Q^*(\theta_o)} [G_i(\theta_o)] \right) - \frac{d^2\lambda_o(\theta_o)^\prime}{d\theta d\theta} \left( E_P^P [g_i(W_i, \theta_o)] \right) - \sum_{j=1}^{dK} \frac{d^2\lambda_o(\theta_o)^\prime}{d\theta d\theta} \left( E^{Q^*(\theta_o)} \left[ \frac{d^2\lambda_j(\theta_o)}{d\theta d\theta} \right] \right) \lambda_o(\theta_o) + \text{Var}_{Q^*(\theta_o)} [G_i(\theta_o)^\prime \lambda_o(\theta_o)] + \frac{d\lambda_o(\theta_o)^\prime}{d\theta} \text{Cov}_{Q^*(\theta_o)} (g_i(\theta_o), G_i(\theta_o)^\prime \lambda_o(\theta_o))$. 

15
Just as in Kleijn and van der Vaart (2012), this theorem establishes that the posterior distribution of the centered and scaled parameter $\sqrt{n} (\theta - \theta_0)$ converges to a Normal distribution with a random mean that is bounded in probability. The rate restriction $\sup_{h \in H} \| \lambda_0 (\theta_h) \|^2 \zeta (K) K^2 \sqrt{K/n} \rightarrow 0$ is much stronger than in the correctly specified case (see Theorem 3.1). The fact that $K$ should be smaller in this case is also intuitive. When the conditional moment conditions are misspecified, limiting the number of implied unconditional moments serves to limit the magnification of the misspecification.

The strategy of the proof of this result is generally similar to the proof of Theorem 3.1 with $\theta_*$ replaced by the pseudo-true value $\theta_0$. However, proving that the ETEL function satisfies a stochastic LAN expansion is more complex, for the following reasons. First, the limit of $\hat{\lambda} (\theta_0)$ is $\lambda_0 (\theta_0)$, which is not zero. Therefore, several terms that were equal to zero in the LAN expansion under correct specification, are non-zero in the misspecified case. The limit in distribution of these terms has to be derived. This explains our stronger assumptions with respect to the correctly specified case. Second, the quantity $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(W_i, \theta_0)$ is no longer centered on zero, which leads to an additional bias term.

Example 1 (continued) Consider now the impact of $K$ in misspecified models. For simplicity, we fix $\theta_0 = 1$ and impose the conditions $E_P [\varepsilon | Z] = 0$ and $E_P [\varepsilon^3 | Z] = 0$, where the second moment condition is incorrect. The pseudo-true value of $\theta_1$ based on 5 million observations is found to be 1.224. For comparison, we omit the second condition and label that as the correctly specified model. The adverse impact of increasing $K$ (for a fixed $n$) on the Bayesian bias and posterior sd are reported in Table 2. As in the case of the correctly specified models, relatively small values of $K$ (around $K = 5$ for $n = 250$) lead to the best results and, in addition, the value of the Bayesian bias increase more quickly when $K$ increases beyond the recommended value. Finally, the posterior sd also declines more sharply as $K$ increases. As pointed out before, these effects are due to the reduction in the support of the prior, now magnified by moment misspecification.

4 Model Comparisons

In practice, we can be unsure about elements of the conditional moment model. For instance, we can be faced with a large number of variables in $Z$, but only some of which are relevant. In such cases, any specific model may be considered to be misspecified, and the goal is to find the best model given the data.

Let $M_\ell$ denote the $\ell$th model in the model space. Each model is characterized by a parameter $\theta^\ell$ and
Table 2: Bayesian bias and posterior sd for different values of $K$ under correct and incorrect conditional moments.

<table>
<thead>
<tr>
<th>$K$</th>
<th>Correctly specified model</th>
<th>Misspecified model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[Bias]</td>
<td>SD</td>
</tr>
<tr>
<td>2</td>
<td>0.046</td>
<td>0.140</td>
</tr>
<tr>
<td>5</td>
<td>0.033</td>
<td>0.112</td>
</tr>
<tr>
<td>9</td>
<td>0.037</td>
<td>0.112</td>
</tr>
<tr>
<td>12</td>
<td>0.084</td>
<td>0.106</td>
</tr>
<tr>
<td>20</td>
<td>0.124</td>
<td>0.094</td>
</tr>
</tbody>
</table>

an extended set of moment functions given by $g^\ell(W, \theta^\ell)$. In addition, each model $M_\ell$ is described by a prior distribution for $\theta^\ell$. The posterior distribution is obtained based on (3.6). The aim is to compare these models by marginal likelihoods, denoted by $m(W_{1:n}|M_\ell, K)$. These are each calculated by the marginal likelihood identity of Chib (1995) (where we explicit the dependence on $M_\ell$ in the notation):

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

(4.1)

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

**Remark 4.1** Comparison of CM condition models differs in one important aspect from the framework for comparing unconditional moment condition models that was established in Chib, Shin and Simoni (2018), where it is shown that to make the unconditional moment condition models comparable it is necessary to linearly transform the moment functions so that all the transformed moments are included in each model. This linear transformation consists of adding an extra parameter different from zero to the components of the vector $g(\theta, W)$ that correspond to the restrictions not included in a specific model. When comparing conditional moment models, however, this transformation is not necessary because the convex hulls associated with different expanded models have the same dimension asymptotically.

### 4.1 Model selection consistency

Suppose that there are $J$ contending models. Suppose also that at least $J - 1$ of these models are misspecified and the remaining one can be either misspecified or correctly specified. Moreover, suppose that a model $M_\ell$ is selected by the size of the marginal likelihoods. Then, in Theorem 4.1 we show that this criterion in the limit picks the model $M_\ell$ with the smallest Kullback-Leibler divergence between $P_\ast$ and the corresponding $Q^\ast(\theta^\ell)$, where $Q^\ast(\theta^\ell) = \arg\inf_{Q \in \mathcal{P}_{\theta^\ell}} K(Q || P_\ast)$ and $\mathcal{P}_{\theta^\ell}$ is defined in Section 3.4.
Theorem 4.1 Let the assumptions of Theorem 3.2 hold. Let us consider the comparison of $J < \infty$ models $M_\ell, \ell = 1, \ldots, J$, such that $J - 1$ of these models each has at least one misspecified moment condition and model $M_j$ can be either correctly specified or contain some misspecified moment condition, that is, $M_\ell$ does not satisfy Assumption 3.3(a), $\forall \ell \neq j$. Then,

$$\lim_{n \to \infty} P_* \left( \log m(W_{1:n}|M_j, K) > \max_{\ell \neq j} \log m(W_{1:n}|M_\ell, K) \right) = 1$$

if and only if $\mathbb{K}(P_*||Q^*(\theta_j)) < \min_{\ell \neq j} \mathbb{K}(P_*||Q^*(\theta_\ell))$, where $\mathbb{K}(P||Q) := \int \log(dP/dQ) dP$.

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

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

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

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

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

$$Y = \theta_0 + \theta_1 X + e_1,$$

$$X = f(Z_1, Z_2, Z_3) + e_2,$$

$$Z_1 \sim U[0, 1], \hspace{1cm} Z_2 \sim U[0, 1], \hspace{1cm} \text{and} \hspace{1cm} Z_3 \sim B(0.4),$$

where $(e_1, e_2)'$ are non-Gaussian and correlated. Thus, $X$ in the outcome model is correlated with the error $e_1$. Let true $\theta = (\theta_0, \theta_1)$ equal $(1, 1)$. Moreover, suppose that the $Z_j$’s are relevant instruments, that is, $\text{cov}(X, Z_j) \neq 0$ for $j \leq 3$, and

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

In practice, some instruments can be valid and some not, and the goal is to select the valid instruments. To this end, we generate $(e_1, e_2, Z_1)$ from a Gaussian copula whose covariance matrix is $\Sigma = [1, 0.7, 0.7; 0.7, 1, 0; 0.7, 0, 1]$ such that the marginal distribution of $e_1$ is the skewed mixture of
two normal distributions \(0.5 \mathcal{N}(0.5, 0.5^2) + 0.5 \mathcal{N}(-0.5, 1.118^2)\) and the marginal distribution of \(e_2\) is \(\mathcal{N}(0, 1)\). Under this setup, \(Z_1\) is an invalid instrument. Consider the following three models

\[
\mathcal{M}_1 : E_P[(Y - \theta_0 - \theta_1 X)|Z_1, Z_2, Z_3] = 0, \\
\mathcal{M}_2 : E_P[(Y - \theta_0 - \theta_1 X)|Z_1, Z_3] = 0, \\
\mathcal{M}_3 : E_P[(Y - \theta_0 - \theta_1 X)|Z_2, Z_3] = 0.
\]

Because \(Z_1\) is an invalid instrument, models \(\mathcal{M}_1\) and \(\mathcal{M}_2\) are misspecified.

In \(\mathcal{M}_1\), the basis matrix \(B\) is made from the variables \((z_1, z_2, z_1 \odot z_2, z_1 \odot z_3, z_2 \odot z_3)\), each using \(K\) knots, concatenated with the vector \(z_3\). In \(\mathcal{M}_2\), \(B\) is made from the variables \((z_1, z_1 \odot z_3)\), each using \(K\) knots, concatenated with the vector \(z_3\). In \(\mathcal{M}_3\), \(B\) is made from the variables \((z_2, z_2 \odot z_3)\), each using \(K\) knots, concatenated with the vector \(z_3\). The number of columns in the \(B\) matrix is \(5(K - 1) + 2\) for \(\mathcal{M}_1\), and \(2(K - 1) + 2\) for \(\mathcal{M}_2\) and \(\mathcal{M}_3\). The prior for \(\theta_0\) and \(\theta_1\) is the product of student-\(t\) distributions with mean zero, dispersion 5, and degrees of freedom equal to 2.5. A repeated sampling experiment is conducted. The marginal likelihood of each model is calculated in 200 repeated samples. Table 3 reports the model selection frequency for \((n = 100, K = 4)\), \((n = 250, K = 5)\), and \((n = 1000, K = 6)\), where \(K\) is based on \(K = 2n^{1/6}\). Note that the model with the valid instruments, i.e., \(\mathcal{M}_3\), is selected more frequently as the number of observation gets larger, in conformity with the theory.

<table>
<thead>
<tr>
<th></th>
<th>(\mathcal{M}_1)</th>
<th>(\mathcal{M}_2)</th>
<th>(\mathcal{M}_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n = 100)</td>
<td>0%</td>
<td>48%</td>
<td>52%</td>
</tr>
<tr>
<td>(n = 250)</td>
<td>0%</td>
<td>36%</td>
<td>64%</td>
</tr>
<tr>
<td>(n = 1000)</td>
<td>0%</td>
<td>2%</td>
<td>98%</td>
</tr>
</tbody>
</table>

Table 3: Model comparison: IV regression example. Each entry in the table presents the model selection frequency in 100 repetitions; \((n = 100, K = 4)\), \((n = 250, K = 5)\), and \((n = 1000, K = 6)\), where \(K\) is based on \(K = 2n^{1/6}\). Each result from 10,000 MCMC draws beyond a burn-in of 1000.

5 Additional topics

5.1 High dimensional \(Z\)

We now consider the case where \(Z\) lies in a high-dimensional space. If all the elements of \(Z\) are relevant, then the situation can be challenging, but there is an interesting sub-case that is worth discussing. Suppose that the conditional expectation depends only on a few elements of \(Z\) or, in other words, most
of the elements of $Z$ are redundant. In this case, one can find the relevant elements of $Z$ by estimating and comparing models that condition on different subsets of $Z$, where the cardinality of these subsets is say 2 or 3. The relevant elements of $Z$ correspond to the model with the largest marginal likelihood. We refer to this procedure as sparsity-based model selection. The next example provides an illustration.

**Example 3 (Sparsity-based model selection)** Recall our Example 2, but assume that one has nine additional potential $Z$’s

$$Z_j = \frac{9}{10} Z_1 + \frac{1}{10} \eta_j, \quad \eta_j \sim \text{Unif}([0, 1])$$

for $j = 4, 5, \ldots, 12$. Recall, $Z_1$ is an invalid instrument. Therefore, $Z_j$’s for $j = 4, \ldots, 12$ are also invalid. Suppose that $Z_3$ affects the conditional expectation, but that one is unsure about the remaining elements of $Z$. Suppose one believes that at most three elements of $Z$ affect the conditional expectation (the sparsity assumption). In this situation one can compute marginal likelihoods of the following 66 models:

$$E^P[(Y - \theta_0 - \theta_1 X)|Z_j, Z_3] = 0, \quad j \in \{1, 2, 4, 5, \ldots, 12\}, \quad (5.1)$$

and

$$E^P[(Y - \theta_0 - \theta_1 X)|Z_j, Z_k, Z_3] = 0, \quad j, k \in \{1, 2, 4, 5, \ldots, 12\} \text{ and } k \neq j, \quad (5.2)$$

with the correct model given by

$$E^P[(Y - \theta_0 - \theta_1 X)|Z_2, Z_3] = 0 \quad (5.3)$$

Sample data (size $n = 250$) is generated from the design in Example 2. Estimation and marginal likelihood computations are based on expanded moments from $K = 3$ basis functions for each conditioning $Z_j$. A summary of the marginal likelihood results appears in Figure 1, sorted by the size of the marginal likelihood. The top ranked model is the true model. As shown in the right panel of the same figure, which reports the posterior model probabilities (under a uniform prior on model space), the support for the true model is decisive.

### 5.2 High dimensional $\theta$: TaRB-MH

It is also important to consider the estimation of conditional moment models that contain a high-dimensional $\theta$. While the regularizing role of the Bayesian prior is important, MCMC sampling of the posterior simulation becomes more complicated. From our experience, the one-block M-H algorithm
<table>
<thead>
<tr>
<th>Ranking</th>
<th>Model</th>
<th>log(ML)</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(2,3)</td>
<td>-1388.73</td>
<td>0.912</td>
</tr>
<tr>
<td>2</td>
<td>(4,3)</td>
<td>-1392.48</td>
<td>0.021</td>
</tr>
<tr>
<td>3</td>
<td>(5,3)</td>
<td>-1393.12</td>
<td>0.011</td>
</tr>
<tr>
<td>4</td>
<td>(7,3)</td>
<td>-1393.14</td>
<td>0.011</td>
</tr>
<tr>
<td>5</td>
<td>(5,9,3)</td>
<td>-1393.78</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Figure 1: Left figure presents log marginal likelihood for each of 66 models in the model space. Arrows point to the best model \((Z_2, Z_3)\), top 20 model \((Z_6, Z_{12}, Z_3)\), and the worst model \((Z_1, Z_4, Z_3)\). Right table presents log marginal likelihood and posterior model probability for top 5 models. Posterior model probabilities are computed with a uniform prior on model space.

that we have used above tends to be inefficient. A straightforward alternative sampling scheme is offered by the Tailored Randomized Block MH algorithm of [Chib and Ramamurthy (2010)](#). This algorithm, which has proved useful in several similarly complex settings, trades more computations for gains in simulation efficiency.

**Example 3 (continued) (IV regression with additional exogenous regressors).** Consider the previous IV regression model, but now with 18 additional exogenous regressors \(w\)

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

where \(w_i = [w_i^{(1)'}, w_i^{(2)'}, w_i^{(3)'}, \ldots]'\), and each group \(w_i^{(j)'}, j \leq 3\), are identically and independently drawn from \(N_6(0, \Sigma(\rho))\), where \(\Sigma(\rho)\) is a 6 \(\times\) 6 matrix in correlation form with each off-diagonal element set equal to 0.97. In addition, \(\gamma\) is a vector of ones. In total, there are 20 unknown parameters. Other elements of the DGP are unchanged. Suppose one has 1500 observations from this DGP, and we estimate \([\theta_0, \theta_1, \gamma]'\) from \(E^P[(Y - \theta_0 - \theta_1 X)|Z_2, Z_3, W] = 0\) with the expanded moment conditions similar to \(M_3\) in Example 2. The basis function matrix is formed with \(Z_2, Z_3,\) and \(Z_2Z_3\), concatenated with columns in \(W\). We set \(K = 6\), following our recommendation, which leads to 36 expanded moment conditions. The training sample prior is based on the first 10% of the sample, and estimation on the remaining 90%. The prior is a product of independent student-\(t\) distributions with 5 degrees of freedom,
centered on the two-stage least squares (2SLS) estimate, and scale equal to two times the 2SLS standard error.

<table>
<thead>
<tr>
<th></th>
<th>TaRB-MH</th>
<th>One-block-MH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>1.00</td>
<td>0.03</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.94</td>
<td>0.09</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>0.99</td>
<td>0.03</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>1.06</td>
<td>0.15</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>1.11</td>
<td>0.16</td>
</tr>
<tr>
<td>$\gamma_4$</td>
<td>1.03</td>
<td>0.15</td>
</tr>
<tr>
<td>$\gamma_5$</td>
<td>1.11</td>
<td>0.16</td>
</tr>
<tr>
<td>$\gamma_6$</td>
<td>0.79</td>
<td>0.15</td>
</tr>
<tr>
<td>$\gamma_7$</td>
<td>0.95</td>
<td>0.03</td>
</tr>
<tr>
<td>$\gamma_8$</td>
<td>1.21</td>
<td>0.16</td>
</tr>
<tr>
<td>$\gamma_9$</td>
<td>1.13</td>
<td>0.15</td>
</tr>
<tr>
<td>$\gamma_{10}$</td>
<td>1.28</td>
<td>0.15</td>
</tr>
<tr>
<td>$\gamma_{11}$</td>
<td>0.80</td>
<td>0.15</td>
</tr>
<tr>
<td>$\gamma_{12}$</td>
<td>0.84</td>
<td>0.16</td>
</tr>
<tr>
<td>$\gamma_{13}$</td>
<td>1.04</td>
<td>0.03</td>
</tr>
<tr>
<td>$\gamma_{14}$</td>
<td>0.90</td>
<td>0.17</td>
</tr>
<tr>
<td>$\gamma_{15}$</td>
<td>0.90</td>
<td>0.16</td>
</tr>
<tr>
<td>$\gamma_{16}$</td>
<td>0.92</td>
<td>0.15</td>
</tr>
<tr>
<td>$\gamma_{17}$</td>
<td>0.97</td>
<td>0.15</td>
</tr>
<tr>
<td>$\gamma_{18}$</td>
<td>0.99</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Table 4: Posterior summary of IV regression example with additional covariates ($n = 1500$). The true value of all parameters ($\theta$’s and $\gamma$’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 3,000 draws beyond a burn-in of 1,000 for the TaRB-MH. The M-H acceptance rate is around 41% for the one-block-MH and 86% for TaRB-MH. “Ineff” is the inefficiency factor.

The results appear in Table 4. In implementing the TaRB-MH sampling scheme, the probability of starting a new block is set to 0.3, so that the each block, within each MCMC iteration, contains 6 parameters on average. For comparison, results from the single-block sampling scheme (on the same conditional moments) are also included. It is evident that the two MCMC samplers produce identical posterior moments, but 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 are essentially independent. Therefore, armed with the TaRB-MH sampler, computational
efficiency is retained, even in higher-dimensional \( \theta \) problems.

6 Applications

6.1 Asset pricing

A key question in finance concerns the makeup of the pricing kernel, or the stochastic discount factor (SDF). Factors in the SDF are the risk factors that explain the cross-section of expected equity returns and, for this reason, establishing the identity of these risk factors has been a long-standing quest in finance.

Following notation from Chib and Zeng (2020), write the SDF \( M_t \) at time (month) \( t \) as

\[
M_t = 1 - b'(x_t - \mu_x) \tag{6.1}
\]

where \( x_t \) is a \((k_x \times 1)\) vector of risk factors (empirically these are the excess returns on portfolios of stocks), and \( b \) is the unknown risk-factor premia and \( \mu_x = E_P(x_t) \). The parameters \((b, \mu_x)\) are unknown. Suppose that there are other factors (excess returns on other portfolios) that are collected in a \((k_w \times 1)\) vector \( w_t \). Let \( f_t := (x_t', w_t')' \) be a \((k_f \times 1)\)-vector, where \( k_f = k_x + k_w \). If \( x_t \) are risk factors, then finance theory dictates that the restriction \( E_P(M_t f_t) = 0 \) holds. Given a sample of observations \( \{f_t\}_{t=1}^n \), one can estimate \((b, \mu_x)\) based on the following moment conditions

\[
E_P[(1 - b'(x_t - \mu_x))f_t] = 0, \quad E_P[x_t - \mu_x|f_{t-1}] = 0,
\]

where the second conditional moment restriction identifies \( \mu_x \).

As an example, consider the data at [http://apps.olin.wustl.edu/faculty/chib/rpackages/czfactor/czfactor.pdf](http://apps.olin.wustl.edu/faculty/chib/rpackages/czfactor/czfactor.pdf) on monthly excess returns (Jan 1974 – Dec 2018) on \( k_f = 12 \) potential risk-factors. Thus, in this situation, there are 12 conditioning variables, an illustration of a modestly high-dimensional \( Z \). Let \( x_t \) be the excess return on the market portfolio (denoted Mkt in the data).

Now construct the expanded moment conditions as

\[
E_P \left[ (x_t - \mu_x) \otimes [q^K(f_1,t-1), q^K(f_2,t-1), \ldots, q^K(f_{12},t-1)] \right] = 0, \tag{6.2}
\]

where \( q^K(f_{1,t-1}) \) consist of \( K = 3 \) basis functions, and \( \tilde{q}^K(f_{j,t-1}) \) (\( j \geq 2 \)) each consist of 2 basis functions derived from \( q^K(f_{j,t-1}) \) by subtracting the second and third columns from the first and then
dropping the first. Along with these 25 expanded moment conditions, the pricing conditions supply an additional twelve, for a total of 37 moment conditions.

For the prior, one can employ the training sample approach. From the first 80 observations (the training sample) the hyperparameters of the independent student-t distribution of $(b, \mu_x)$ with 2.5 degrees of freedom are set as follows. The center of the prior density is set to the Generalized Method of Moments (GMM) estimate, and the scale to two times the GMM standard error. This black-box prior is particularly convenient if the analysis has to be repeated for different possible variables in the SDF. The remaining 459 observations are used to construct a joint posterior distribution of $b$ and $\mu_x$.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b$</td>
<td>2.981</td>
<td>0.730</td>
<td>2.955</td>
<td>1.819</td>
<td>4.211</td>
</tr>
<tr>
<td>$\mu_x$</td>
<td>0.006</td>
<td>0.001</td>
<td>0.006</td>
<td>0.004</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Table 5: Asset pricing data: Summary of the posterior distribution based on 50,000 MCMC draws after 1,000 burn-in.

The posterior summary of $(b, \mu_x)$ from 50,000 MCMC draws is given in Table 5 and the prior and posterior density of $b$ are presented in Figure 2. This summary, specifically the lower and upper limits of the marginal posterior of $b$ confirm that $b$ is non-zero, and hence that the Mkt variable is a risk-factor.

![Figure 2: Asset pricing data: Prior and posterior density of $b$. Prior is Student-t density based on the first 80 observation. Posterior density is based on the remaining 459 observations. 50,000 MCMC draws after 1,000 burn-in.](image)
6.2 ATE under conditional ignorability

For another important application of the methods in this paper, consider the problem of estimating the average treatment effect (ATE) under the assumption of conditional ignorability. In the frequentist literature, the ATE is commonly estimated by propensity score methods and, on the Bayesian side, from models of the potential outcomes. These models generally have a nonparametric mean function, but parametric noise. By adopting the conditional moment perspective, however, one can evade the burden of distributional assumptions.

Data is from the 1997 Child Development Supplement to the Panel Study of Income Dynamics (Guo and Fraser 2015, Section 5.8.2), where the ATE is calculated by the propensity score. The research question is the effect of childhood welfare dependency on academic achievement. The latter, the dependent variable \( y \), is measured by the child’s score on the “letter-word identification” section of the Woodcock-Johnson Revised Tests of Achievement. The treatment variable \( x \) equals one if the child received AFDC (Aid to Families with Dependent Children) benefits at any time from birth to 1997 (the survey year) and equals zero if the child never received benefits during that period. It is assumed that the potential outcomes are independent of \( x \), conditioned on \( z_1, z_2, \ldots, z_6 \) (the assumption of conditional ignorability), where

- \( z_1 \): mratio97, the ratio of family income to the poverty line in 1997
- \( z_2 \): pcged97, the caregiver’s years of schooling
- \( z_3 \): pcg_adc, the number of years in which the caregiver received AFDC in her childhood
- \( z_4 \): age97, the child’s age in 1997
- \( z_5 \): race, one for African-American children and zero for other
- \( z_6 \): male, one if the child is male and zero if female.

Two observations from the sample are dropped. These have values of mratio97 larger than 9 standard deviation from the mean of mratio97. Apart from mratio97 and age97, the other variables are categorical. There are \( n_0 = 727 \) control subjects and \( n_1 = 274 \) treated subjects. The ATE is expected to be negative, reflecting the hypothesis that welfare dependency has an adverse effect on academic achievement.
To answer the research question, suppose that the potential outcomes for the controls satisfy the conditional moments

$$E^{P}((y_{i0} - \beta_{00} - h_{01}(z_1) - \beta_{02}z_2 - \beta_{03}z_3 - h_{04}(z_4) - \beta_{05}z_5 - \beta_{06}z_6)|z_i) = 0$$

and those for the treated satisfy the conditional moments

$$E^{P}((y_{i1} - \beta_{10} - h_{11}(z_1) - \beta_{12}z_2 - \beta_{13}z_3 - h_{14}(z_4) - \beta_{15}z_5 - \beta_{16}z_6)|z_i) = 0$$

where \(\{h_{01}, h_{04}, h_{11}, h_{14}\}\) are four non-parametric functions. These are each modeled by natural cubic splines with 5 knots. Thus, the parameters \(\theta_j\) of the \(j\)th potential outcome model consist of \((\beta_{j0}, \beta_{j2}, \beta_{j3}, \beta_{j5}, \beta_{j6})\) plus the eight spline coefficients. Special cases of this model, mentioned below, are considered and evaluated by marginal likelihoods. For example, models in which the \(h\) functions are linear are of interest.

The expanded moments are constructed as follows. The basis matrix has cubic spline basis functions for \((z_1, z_4)\), each with 5 knots, concatenated with \((z_2, z_3, z_5, z_6)\) (as is) because the latter variables are all essentially categorical. In total, this produces 13 expanded unconditional moments for the estimation of the \(y_0\) and \(y_1\) models. The prior distribution on the parameters is a product of student-t distributions with 2.5 degrees of freedom with mean of the intercept equal to the mean of the first 50 observations, the mean of the slopes equal to 0, and dispersion equal to 5.

Four models are estimated and evaluated. In the baseline model, the \(h\) functions are linear. In the second model, only the effect of \(z_1\) is nonparametric. In the third model, only the effect of \(z_4\) is assumed to be nonparametric and, finally, in the fourth model, both \(z_1\) and \(z_4\) are nonparametric. The results given in Table 6 show that the model best supported by these data is the third.

<table>
<thead>
<tr>
<th></th>
<th>Non-treated</th>
<th>Treated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>-4823.76</td>
<td>-1555.553</td>
</tr>
<tr>
<td>(z_1) nonparametric</td>
<td>-4829.23</td>
<td>-1556.25</td>
</tr>
<tr>
<td>(z_4) nonparametric</td>
<td><strong>-4813.85</strong></td>
<td><strong>-1555.77</strong></td>
</tr>
<tr>
<td>(z_1) and (z_4) nonparametric</td>
<td>-4818.98</td>
<td>-1556.57</td>
</tr>
</tbody>
</table>

Table 6: Academic achievement data: Marginal likelihoods of 4 competing causal models, based on 20,000 MCMC draws beyond a burn-in 1000.
Consider now posterior inference on the ATE. By definition, the sample version of the ATE is

\[
\text{ATE} = \frac{1}{n} \sum_{i=1}^{n} \left( E^P(y_{i1}|z_i, \theta_1) - E^P(y_{i0}|z_i, \theta_0) \right),
\]

where, in the model selected by the preceding comparison,

\[
E^P(y_{ij}|z_i, \theta_j) = \beta_{j0} + \beta_{j1}z_1 + \beta_{j2}z_2 + \beta_{j3}z_3 + h_{j4}(z_4) + \beta_{j5}z_5 + \beta_{j6}z_6.
\]

Clearly, if we evaluate the latter expression at each posterior draw of \((\theta_0, \theta_1)\), we produce a sample of the ATE from its posterior distribution. We summarize this sample in Table 7 and Figure 3. One can see

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Propensity Score Matching</td>
<td>-5.682</td>
<td>1.976</td>
<td>-9.496</td>
<td>-1.502</td>
<td></td>
</tr>
<tr>
<td>Bayesian ATE</td>
<td>-5.251</td>
<td>1.387</td>
<td>-5.261</td>
<td>-7.982</td>
<td>-2.511</td>
</tr>
</tbody>
</table>

Table 7: Academic achievement data: Summary of the posterior distribution of the ATE from model in which the effect of \(z_4\) is non-parametric.

that the ATE posterior point estimate is similar in size to the propensity score estimate, but the posterior standard deviation is smaller, leading to a less dispersed interval estimate. As a takeaway, it is striking that the Bayesian analysis of this important problem can be prosecuted under such minimal assumptions.

Figure 3: Academic achievement data: Posterior density of the ATE from model in which the effect of \(z_4\) is non-parametric. Red dashed line is at -5.251, the posterior mean. Posterior density based on 20,000 posterior draws after 1,000 burn-in.
7 Conclusion

In this paper we have developed perhaps the first Bayesian framework for analyzing an important and broad class of semiparametric models in which the distribution of the outcomes is defined only up to a set of conditional moments, some of which may be misspecified. We have derived BvM theorems for the behavior of the posterior distribution under both correct and incorrect specification of the conditional moments, and developed the theory for comparing different conditional moment models through a comparison of model marginal likelihoods. In addition, we have discussed settings with a high-dimensional Z and θ, the former addressed by a sparsity-based model search procedure, and the latter by the TaRB-MH MCMC algorithm for efficient posterior sampling.

Our theory and various examples, taken together, show that the developments in this paper make possible, for the first time, the formal (and practical) Bayesian analysis of a new, large class of problems that were hitherto difficult, or not possible, to tackle from the Bayesian viewpoint. This enlargement of the scope of Bayesian inference should prove useful for a variety of new applications.

We conclude by noting that a R-package that implements the framework of the paper is available from the authors. Within an user-friendly interface, the package includes both the single block tailored MCMC and the TaRB-MH algorithms for prior-posterior analysis and marginal likelihood estimation. Thus, the methods developed in this paper can be implemented readily, absent any set up costs.

A Appendix: Proofs

For ready access, this appendix provides proofs of some of the results in the paper. Complete details, supporting lemmas, and the proof of the misspecified case, are in the Supplementary Material (SM henceforth).

Let \( g_i(\theta) = g(W_i, \theta), \rho_i(\theta) \triangleq \rho(X_i, \theta), q_i^K \triangleq q^K(Z_i). \) Also let \( p(W_{1:n}|\theta, K) \triangleq \prod_{i=1}^{n} \tilde{p}_i(\theta), \) \( \ell_{n,\theta}(W_i) \triangleq \log \tilde{p}_i(\theta) = \log \frac{e^{\lambda(\theta)^j g(W_i, \theta)}}{\sum_{j=1}^{n} e^{\lambda(\theta)^j g(W_j, \theta)}}, \) where \( \lambda(\theta) \triangleq \arg \min_{\lambda \in \mathbb{R}^{dK}} \frac{1}{n} \sum_{i=1}^{n} e^{\lambda g(W_i, \theta)} \) is the estimated tilting parameter.

As notation, let \( \mathbb{E}_n[\cdot] \triangleq \frac{1}{n} \sum_{i=1}^{n} [\cdot] \) denote the empirical mean, and \( \mathbb{E}_P[\cdot] \) the population mean with respect to the true distribution \( P_\ast. \) Moreover, for a function \( \lambda(\theta) \) of \( \theta, \) define \( \tau_\ast(\lambda, \theta) \triangleq \frac{e^{\lambda(\theta)^j g_i(\theta)}}{\mathbb{E}_n[e^{\lambda(\theta)^j g_i(\theta)}]}, \) \( \tau_\ast(\lambda, \theta) \triangleq e^{\lambda(\theta)^j g_i(\theta)}. \) For a matrix \( A, \) let \( \lambda_{\min}(A) \) and \( \lambda_{\max}(A) \) denote the minimum and maximum eigenvalues of \( A, \) respectively, and \( \|A\| \) the operator norm of \( A. \) Let \( C \) be a generic positive constant.

Further, let \( \tilde{g}(\theta) \triangleq \mathbb{E}_n[g_i(\theta)], \rho_\theta(X, \theta) \triangleq \frac{\partial \rho(X, \theta)}{\partial \theta}, \tilde{G}(\theta) \triangleq \mathbb{E}_n[G(W_i, \theta)] \) with \( G(W, \theta) \triangleq \rho_\theta(X, \theta) \otimes q^K(Z) \) a \( dK \times p \) matrix, \( G_i(\theta) \triangleq G(W_i, \theta), \tilde{G}(\lambda, \theta) \triangleq \mathbb{E}_n[\tau_\ast(\lambda, \theta) G(W_i, \theta)] \) and
\[ \bar{G}^\omega(\lambda, \theta) \triangleq \mathbb{E}_n[\tau^*_i(\lambda, \theta)G(W_i, \theta)]. \]

We also use the notation: \( \hat{\Omega}(\theta) \triangleq \mathbb{E}_n[g(W_i, \theta)g(W_i, \theta)'] \) a \( dK \times dK \) matrix, \( \hat{\pi}(\lambda, \theta) \triangleq \mathbb{E}_n[\tau_i(\lambda, \theta)g_i(\theta)g_i(\theta)'], \quad \bar{\Omega}(\lambda, \theta) \triangleq \Omega(\lambda, \theta) - \mathbb{E}_n[\tau_i(\lambda, \theta)g_i(\theta)]\mathbb{E}_n[\tau_i(\lambda, \theta)g_i(\theta)'], \quad \underline{\Omega}(\lambda, \theta) \triangleq \mathbb{E}_n[\tau^*_i(\lambda, \theta)g_i(\theta)g_i(\theta)'], \) and \( \Omega^*(\lambda, \theta) \triangleq \mathbb{E}_n[\tau^*_i(\lambda, \theta)g_i(\theta)]\mathbb{E}_n[\tau^*_i(\lambda, \theta)g_i(\theta)'] \).

Their population counterparts in the correctly specified model are \( G_* \triangleq \mathbb{E}^P[G(W_i, \theta_*)] \) and \( \Omega_* \triangleq \mathbb{E}^P[g(W_i, \theta_*)g(W_i, \theta_*)'] \), respectively. In addition, \( \Sigma(Z) \triangleq \mathbb{E}^P[\rho(X, \theta_*)\rho(X, \theta_*)'|Z] \), \( D(Z) \triangleq \mathbb{E}^P[\rho(x, \theta_*)|Z] \), \( V_{\theta_*}^{-1} \triangleq \mathbb{E}^P[D(Z)'\Sigma(Z)^{-1}D(Z)] \) and \( \rho_{\theta*}(x, \theta_*) \triangleq \partial^2 \rho_j(x, \theta)/\partial \theta \partial \theta' \). Finally, let CS, M, T, J refer to the Cauchy-Schwartz, Markov, triangular, Jensen’ inequalities, respectively, and MVT the Mean Value Theorem.

### A.1 Proofs for Section 3.3

#### A.1.1 Proof of Theorem 3.1

The main steps of this proof proceed as in the proof of [Van der Vaart (1998, Theorem 10.1)](Van der Vaart (1998, Theorem 10.1)) and so we detail them in the Supplementary Material (SM). On the other hand the technical results that we need all along this proof are new and we detail them here. The first result establishes that the posterior of \( \theta \) concentrates and puts all its mass on \( \Theta_n \triangleq \{\|\theta - \theta_*\| \leq M_n/\sqrt{n}\} \) as \( n \to \infty \).

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

\[
\mathbb{P}_* \left( \sup_{\|\theta - \theta_*\| > M_n/\sqrt{n}} \frac{1}{n} \sum_{i=1}^n (\ell_{n, \theta}(W_i) - \ell_{n, \theta_*}(W_i)) \leq -CM_n^2/n \right) \to 1, \quad \text{(A.1)}
\]

as \( n, K \to \infty \). Then, for any \( M_n \to \infty \),

\[
\pi \left( \sqrt{n}\|\theta - \theta_*\| > M_n \big| W_{1:n}, K \right) \to 0 \quad \text{(A.2)}
\]

as \( n, K \to \infty \).

**Proof.** See the SM.

The second technical result that we need in order to prove Theorem 3.1 is the stochastic LAN of the ETEL function.

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

\[
\sup_{h \in \mathcal{H}} \sum_{i=1}^n \ell_{n, \theta_* + h/\sqrt{n}}(W_i) - \sum_{i=1}^n \ell_{n, \theta_*}(W_i) - h'V_{\theta_*}^{-1} \Delta_{n, \theta_*} + \frac{1}{2} h'V_{\theta_*}^{-1} h = o_p(1), \quad \text{(A.3)}
\]

29
where \( o_p(1) \) goes to zero as \( n, K \to \infty \), \( V_{\theta_*}^{-1} \Delta_{n, \theta_*} \overset{d}{\to} \mathcal{N}(0, V_{\theta_*}^{-1}) \) and \( \frac{1}{\sqrt{n}} \sum i=1^n \frac{d\ell_n, \theta_*(W_i)}{d\theta} - V_{\theta_*}^{-1} \Delta_{n, \theta_*} \overset{p}{\to} 0. \)

**Proof.** The main steps in the proof are as follows. Full details can be found in the SM. By a MVT

\[
\sum_{i=1}^n \ell_{n, \theta}(W_i) = \sum_{i=1}^n \log \tau_i(\hat{\lambda}, \theta) - n \log(n) = -n \log(n) + n\tilde{g}(\theta)\hat{\lambda}(\theta) - n\tilde{g}(\theta)\hat{\lambda}(\theta) - \frac{1}{2}n\hat{\lambda}(\theta)'\overline{\Omega}(\hat{\lambda}, \theta)\hat{\lambda}(\theta). \tag{A.4}
\]

Next, denote \( \theta_1 \equiv \theta_* + h_n, h_n \equiv h/\sqrt{n} \) and consider a first order MVT expansion of \( \hat{\lambda}(\theta_1) \) around the value \( h = 0 \) for which there exists a \( \theta_* \), lying between \( \theta_1 \) and \( \theta_* \) such that: \( \hat{\lambda}(\theta_1) = \hat{\lambda}(\theta_*) + \frac{\hat{\lambda}(\theta_*)'}{d\theta'}h_n. \)

By using this expansion and (A.4) we get:

\[
\sum_{i=1}^n \ell_{n, \theta_1}(W_i) - \sum_{i=1}^n \ell_{n, \theta_*}(W_i) = -\frac{1}{2}n\hat{\lambda}(\theta_*)'\left(\overline{\Omega}(\hat{\lambda}, \theta_1) - \overline{\Omega}(\hat{\lambda}, \theta_*)\right)\hat{\lambda}(\theta_*) - n\hat{\lambda}(\theta_*)'\overline{\Omega}(\hat{\lambda}, \theta_1)\frac{d\lambda(\theta_*)}{d\theta'}h_n - \frac{n}{2}h_n'\frac{d\lambda(\theta_*)}{d\theta'}\overline{\Omega}(\hat{\lambda}, \theta_1)\frac{d\lambda(\theta_*)}{d\theta'}h_n. \tag{A.5}
\]

By CS, T, Lemmas G.2 \[ G.3 \] G.5 G.6 G.7 and G.10 in the SM:

\[
\sup_{h \in H} \left| n\hat{\lambda}(\theta_*)'\left(\overline{\Omega}(\hat{\lambda}, \theta_1) - \overline{\Omega}(\hat{\lambda}, \theta_*)\right)\hat{\lambda}(\theta_*) \right| \\
\leq n \sup_{h \in H} \left| \hat{\lambda}(\theta_*) \right| \left| \overline{\Omega}(\hat{\lambda}, \theta_1) - \overline{\Omega}(\hat{\lambda}, \theta_*) \right| \left| \hat{\lambda}(\theta_*) \right| \\
+ n \sup_{h \in H} \left( \left| E_n \left[ \tau_i(\hat{\lambda}_1(\theta_1)g_i(\theta_1)', - \hat{\lambda}(\theta_*)'\overline{\Omega}(\hat{\lambda}, \theta_1)\frac{d\lambda(\theta_*)}{d\theta'}h_n \right] \hat{\lambda}(\theta_*) \right|^2 + \left| E_n \left[ \tau_i(\hat{\lambda}_1(\theta_1)g_i(\theta_1)', - \hat{\lambda}(\theta_*)'\overline{\Omega}(\hat{\lambda}, \theta_1)\frac{d\lambda(\theta_*)}{d\theta'}h_n \right] \hat{\lambda}(\theta_*) \right|^2 \right) \\
= O_p \left( K \left( \frac{\zeta(K)K}{\sqrt{n}} + \frac{K}{n} \right) \right). \tag{A.6}
\]

By expanding the first order condition for \( \hat{\lambda}(\theta_*) \) around the value \( \hat{\lambda}(\theta_*) \) and zero such that: \( \tilde{g}(\theta_*) + \tilde{\Omega}(\hat{\lambda}_1(\theta_1), \theta_*)\hat{\lambda}(\theta_*) = 0 \) which gives \( \hat{\lambda}(\theta_*) = -\tilde{\Omega}^{-1}(\hat{\lambda}_1(\theta_1), \theta_*)^{-1} \tilde{g}(\theta_*) \). Moreover, by taking the total derivative of the first order condition for \( \hat{\lambda} \) we get:

\[
\forall \theta \in \Theta \quad \frac{d\hat{\lambda}(\theta)'}{d\theta} = -E_n \left[ e^{\hat{\lambda}(\theta)'}g_i(\theta)G_i(\theta)'(I + \hat{\lambda}(\theta)g_i(\theta)') \tilde{\Omega}^\circ(\hat{\lambda}, \theta)^{-1} \right]. \tag{A.7}
\]

By replacing these results in (A.5) we obtain:
\[
\sum_{i=1}^{n} \ell_{n, \theta_i}(W_i) - \sum_{i=1}^{n} \ell_{n, \theta_*}(W_i) = O_p(\zeta(K)K^2/\sqrt{n})
\]
\[
+ h' \frac{d\hat{\lambda}(\theta)}{d\theta} \hat{\Omega}(\tilde{\lambda}(\theta), \theta_1) \hat{\Omega}^\circ(\tilde{\lambda}_\theta, \theta_*)^{-1} \sqrt{n}\hat{g}(\theta_*) - \frac{1}{2} h' \frac{d\hat{\lambda}(\theta)}{d\theta} \hat{\Omega}(\tilde{\lambda}(\theta), \theta_1) \frac{d\hat{\lambda}(\theta)}{d\theta'} h
\]
\[
= O_p(\zeta(K)K^2/\sqrt{n}) - h' \mathbb{E}_n \left[ \tau_i(\lambda, \theta_*)G_i(\theta_*)' \left( I + \tilde{\lambda}(\theta_*)g_i(\theta_*)' \right) \right] \hat{\Omega}(\tilde{\lambda}(\theta), \theta_1) \hat{\Omega}^\circ(\tilde{\lambda}_\theta, \theta_*)^{-1} \sqrt{n}\hat{g}(\theta_*)
\]
\[
\times \hat{\Omega}^\circ(\tilde{\lambda}_\theta, \theta_*)^{-1} \sqrt{n}\hat{g}(\theta_*) - \frac{1}{2} h' \frac{d\hat{\lambda}(\theta)}{d\theta} \hat{\Omega}(\tilde{\lambda}(\theta), \theta_1) \frac{d\hat{\lambda}(\theta)}{d\theta'} h, \quad (A.8)
\]
where the \(O_p(\zeta(K)K^2/\sqrt{n})\) is uniform in \(h \in \mathcal{H}\). The rates of the second and third terms in (A.8) can be derived as followed.

**Behaviour of the second term in (A.8).** The second term in (A.8) can be written as

\[
h' \mathbb{E}_n \left[ e^{\tilde{\lambda}(\theta_*)'g_i(\theta_*)} G_i(\theta_*)' \left( I + \tilde{\lambda}(\theta_*)g_i(\theta_*)' \right) \right] \hat{\Omega}^\circ(\tilde{\lambda}(\theta), \theta_1) \hat{\Omega}^\circ(\tilde{\lambda}_\theta, \theta_*)^{-1} \sqrt{n}\hat{g}(\theta_*)
\]
\[
= h' \hat{G}_i(\theta_*)' \hat{\Omega}^\circ(\tilde{\lambda}(\theta), \theta_1) \hat{\Omega}^\circ(\tilde{\lambda}_\theta, \theta_*)^{-1} \sqrt{n}\hat{g}(\theta_*)
\]
\[
+ h' \mathbb{E}_n \left[ e^{\tilde{\lambda}(\theta_*)'g_i(\theta_*)} G_i(\theta_*)' \left( I + \tilde{\lambda}(\theta_*)g_i(\theta_*)' \right) \right] \hat{\Omega}^\circ(\tilde{\lambda}(\theta), \theta_1) \hat{\Omega}^\circ(\tilde{\lambda}_\theta, \theta_*)^{-1} \sqrt{n}\hat{g}(\theta_*)
\]
\[
= B_1 + B_2 \quad (A.9)
\]

Let \(\tilde{G}_* \triangleq \sum_{i=1}^{n} D(Z_i) \otimes q_i^{K}/n\) and \(\tilde{\Omega}_* \triangleq \sum_{i=1}^{n} \Sigma(Z_i) \otimes q_i^{K}(q_i^{K})'/n\), then

\[
B_1 = h' \left( \hat{G}_i(\theta_*)' \hat{\Omega}^\circ(\tilde{\lambda}(\theta), \theta_1) \hat{\Omega}^\circ(\tilde{\lambda}_\theta, \theta_*)^{-1} - \tilde{G}_i(\theta_*) \tilde{\Omega}_*^{-1} \right) \sqrt{n}\hat{g}(\theta_*) + h' \tilde{G}_i(\theta_*) \tilde{\Omega}_*^{-1} \sqrt{n}\hat{g}(\theta_*)
\]
and \(\tilde{G}_i(\theta_*) \tilde{\Omega}_*^{-1} \sqrt{n}\hat{g}(\theta_*) \xrightarrow{d} \mathcal{N}(0, \Omega_{\theta_*}^{-1})\) by the Lindberg-Levy central limit theorem as in Donald et al. (2003) Proof of Theorem 5.4). Next, the first term in \(B_1\) is upper bounded by

\[
\left| h' \left( \hat{G}_i(\theta_*)' \hat{\Omega}^\circ(\tilde{\lambda}(\theta), \theta_1) \hat{\Omega}^\circ(\tilde{\lambda}_\theta, \theta_*)^{-1} - \tilde{G}_i(\theta_*) \tilde{\Omega}_*^{-1} \right) \sqrt{n}\hat{g}(\theta_*) \right|
\]
\[
\leq C \left| h' \left( \hat{G}_i(\theta_*)' \hat{\Omega}^\circ(\tilde{\lambda}(\theta), \theta_1) \hat{\Omega}^\circ(\tilde{\lambda}_\theta, \theta_*)^{-1} - \tilde{G}_i(\theta_*) \tilde{\Omega}_*^{-1} \right) \right|.
\]

Consider the second term on the right hand side of (A.10). Use the equality \(A^{-1}BC^{-1} - A_*^{-1} = A_*^{-1}[(A_* - A)A_*^{-1}B + (B - A_*) + (A_* - C)]C^{-1}\) for square matrices \(A, B, C, A_*\). Denote \(\kappa_n \triangleq \kappa(\theta_1) \triangleq \frac{1}{E_n[|e^{\lambda(\theta_1)g_i(\theta_1)|}]}. \) Then, by T and CS, that term is upper bounded as:

\[
\left| h' \tilde{G}_i \left( \hat{\Omega}^\circ(\tilde{\lambda}(\theta), \theta_1) \hat{\Omega}^\circ(\tilde{\lambda}_\theta, \theta_*)^{-1} - \tilde{\Omega}_* \right) \right| \sqrt{n}\hat{g}(\theta_*)
\]
\[
\leq \|h' \tilde{G}_i \tilde{\Omega}_*^{-1}\| \|\tilde{\Omega}_* - \hat{\Omega}(\tilde{\lambda}(\theta), \theta_1) \hat{\Omega}^\circ(\tilde{\lambda}_\theta, \theta_*)^{-1} + \tilde{\Omega}_* - \hat{\Omega}(\tilde{\lambda}_\theta, \theta_*)\| \|\hat{\Omega}(\tilde{\lambda}_\theta, \theta_*)^{-1} \sqrt{n}\hat{g}(\theta_*)\|.
\]
From Donald et al. (2003, Lemma A.3) with \( \bar{\beta} = \theta_* \), \( a_i(\bar{\beta}) = b_i(\bar{\beta}) = D(Z_i) \) and \( U_i = \Sigma(Z_i) \), and from Donald et al. (2003, Lemma A.6) it follows that \( \sup_{h \in \mathcal{H}} \| h' \hat{G}_i' \hat{\Omega}^{-1}_* \|^2 \leq C \sup_{h \in \mathcal{H}} h' V_{\theta_*}^{-1} h + o_p(1) = O_p(1) \). By Lemmas [G.5 and G.11] in the SM: \( \sup_{h \in \mathcal{H}} \| \hat{\Omega}_* - \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*) \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*)^{-1} \| \leq C O_p(\zeta(K) K/\sqrt{n} + \zeta(K) \sqrt{K/n}) \). By Lemmas G.3 and G.11 in the SM: \( \| [\hat{\Omega}_* - \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*)] - \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*) \| \leq C O_p(\zeta(K) K/\sqrt{n} + \zeta(K) K/\sqrt{n}) \). By using Lemma G.12 in the SM we obtain that \( \sup_{h \in \mathcal{H}} \| [\hat{\Omega}_* - \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*)] - \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*) \| \) in (A.11) is at most of the order \( O_p(\zeta(K) K/\sqrt{n}) \).

Finally, \( \| \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*)^{-1} \sqrt{n} \tilde{g}(\theta_*) \| \leq \| \hat{\Omega}_G^{-1}(\theta_*) \| = O_p(\sqrt{K}) \). Therefore, (A.11) is at most of the order \( O_p(\zeta(K) K/\sqrt{n}) \) uniformly over \( h \in \mathcal{H} \).

Now, consider the first term in (A.10). By the second result of Lemma G.7 in the SM with \( \tilde{\lambda}(\theta) = \tilde{\lambda}(\theta) \) and \( \tilde{\lambda}(\theta) = \tilde{\lambda}(\theta) \), we have \( \sup_{h \in \mathcal{H}} \| \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*)^{-1} \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*)^{-1} \sqrt{n} \hat{g}(\theta_*) \| = O_p(\sqrt{K}) \) since \( \| \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*)^{-1} \sqrt{n} \hat{g}(\theta_*) \| = O_p(\sqrt{K}) \) by Donald et al. (2003, Lemma A.9) under Assumptions 3.2, 3.3, 3.5 and 3.6. The first term \( \| h'(\hat{G}_i(\tilde{\lambda}, \tilde{\theta}_*) - \hat{G}_i') \| \) is upper bounded by \( \sup_{h \in \mathcal{H}} \| h' \| \| G(\tilde{\lambda}, \tilde{\theta}_*) - \hat{G}(\tilde{\theta}_*) \| \) in (A.11) is at most of the order \( O_p(\zeta(K) K/\sqrt{n}) \) by Lemmas G.9 and G.5 in the SM. On combining these results, and eliminating the negligible terms,\n
\[
\sup_{h \in \mathcal{H}} \mathcal{B}_1 = O_p(\zeta(K) K/\sqrt{n}) - h' V_{\theta_*}^{-1} \Delta_n, \theta_*
\]

and \( -\Delta_n, \theta_* = \mathcal{N}(0, V_{\theta_*}) \).

As for the second term in (A.9), by CS:

\[
\mathcal{B}_2 \triangleq h' \mathbb{E}_n \left[ e^{\hat{\lambda}(\theta_*) g_i(\tilde{\theta}_*)} G_i(\tilde{\theta}_*) \hat{\lambda}(\tilde{\theta}_*) g_i(\tilde{\theta}_*)' \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*)^{-1} \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*)^{-1} \sqrt{n} \hat{g}(\theta_*) \right] \leq \left( \mathbb{E}_n \left[ e^{\hat{\lambda}(\theta_*) g_i(\tilde{\theta}_*)} \left| h' G_i(\tilde{\theta}_*) \hat{\lambda}(\tilde{\theta}_*) \right|^2 \right] \right)^{1/2} \times \left( \mathbb{E}_n \left[ e^{\hat{\lambda}(\theta_*) g_i(\tilde{\theta}_*)} \left| g_i(\tilde{\theta}_*)' \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*)^{-1} \hat{\Omega}_G(\tilde{\lambda}, \tilde{\theta}_*)^{-1} \sqrt{n} \hat{g}(\theta_*) \right|^2 \right] \right)^{1/2} =: \mathcal{A}_1 \times \mathcal{A}_2.
\]

Let \( \Lambda_n \) be as defined in Lemma G.2 in the SM. Since \( \sup_{h \in \mathcal{H}} \| \hat{\lambda}(\tilde{\theta}_*) \| \in \Lambda_n \) by Lemma G.5 in the SM, and by CS it holds that:

\[
\sup_{h \in \mathcal{H}} \mathcal{A}_1 \leq e^{\max 1 \leq i \leq n, \lambda \in \Lambda_n \sup_{h \in \mathcal{H}} \| \lambda h' G_i(\tilde{\theta}_*) \|} \sup_{h \in \mathcal{H}} \left( \mathbb{E}_n \left[ \left| h' G_i(\tilde{\theta}_*) \hat{\lambda}(\tilde{\theta}_*) \right|^2 \right] \right)^{1/2} \leq C \sup_{h \in \mathcal{H}} \| \hat{\lambda}(\tilde{\theta}_*) \| \sup_{h \in \mathcal{H}} \left( \mathbb{E}_n \left[ \| G_i(\tilde{\theta}_*) h \|^2 \right] \right)^{1/2} \leq C \sup_{h \in \mathcal{H}} \| \hat{\lambda}(\tilde{\theta}_*) \| O_p \left( \left\{ \mathbb{E}_n \left[ \left( \sup_{h \in \mathcal{H}} \| \rho(X, \tilde{\theta}_*) h \|^2 \right) Z \right] \right\} \left\| q_i K \right\|^2 \right) \right)^{1/2} \]

by Assumption 3.4 (b) and again Lemma G.5 in the SM. Next, by using the very last result of Lemma G.11 in the SM with \( \lambda_1 = 0, \lambda_2 = \lambda \) and \( \theta_{h_1} = \theta_1 \) to get the first inequality below we have:
\[ \sup_{\theta \in \mathcal{H}} \mathcal{A}_2 = \sup_{\theta \in \mathcal{H}} \left( \sqrt{n} g(\theta) \Omega^\top(\theta_\ast, \theta_\ast)^{-1} \Omega(\theta, \theta) \Omega^\top(\theta, \theta \ast)^{-1} \Omega(\theta, \theta \ast)^{-1} \sqrt{n} g(\theta) \right)^{1/2} \]

\[ \leq \sup_{\theta \in \mathcal{H}} \left( \left[ \lambda_{\min}(\Omega^\top(\theta_\ast, \theta_\ast)) \right]^{-1} \lambda_{\max}(\Omega(\theta, \theta)) \left[ \lambda_{\min}(\Omega^\top(\theta, \theta \ast)) \right]^{-1} \lambda_{\max}(\Omega^\top(\theta, \theta \ast)) \right)^{1/2} \times \left[ \lambda_{\min}(\Omega^\top(\theta_\ast, \theta_\ast)) \right]^{-1/2} \sqrt{n} \|g(\theta)\| \leq C \mathcal{O}_p(\sqrt{K}) \]

since by Lemma [G.7] in the SM if \( K \zeta(K)/n \to 0 \): \( 1/C \leq \inf_{\theta \in \mathcal{H}} \lambda_{\max}(\Omega^\top(\overline{\theta}, \overline{\theta}_\ast)) \leq \sup_{\theta \in \mathcal{H}} \lambda_{\max}(\Omega^\top(\overline{\theta}, \overline{\theta}_\ast)) \leq C \) w.p.a. 1 (and similarly for \( \Omega^\top(\overline{\theta}_\ast, \theta_\ast) \) and \( \Omega^\top(\overline{\theta}, \theta_\ast) \)). Therefore, sup_{\theta \in \mathcal{H}} \mathcal{A}_1 \times \mathcal{A}_2 = \mathcal{O}_p(K \sqrt{K/n}) and (A.9) is equal to \( \mathcal{O}_p(\zeta(K) K \sqrt{K/n} + N(0, V^{-1}_\ast)) \).

**Behaviour of the third term in (A.8).** By using the expression for \( \frac{d\tilde{\lambda}(\overline{\theta}, \theta)}{d\theta} \) given in (A.7) we get:

\[
h' \frac{d\tilde{\lambda}(\overline{\theta}, \theta)}{d\theta} \Omega(\overline{\theta}, \theta) \frac{d\tilde{\lambda}(\overline{\theta}, \theta)}{d\theta} =
\]

\[
h' \mathbb{E}_n \left[ e^{\tilde{\lambda}(\overline{\theta}, \theta) g(\theta)} G_i(\overline{\theta}, \theta)^{\top} \bar{\Omega}(\overline{\theta}, \theta)^{-1} \Omega(\bar{\theta}, \theta) \bar{\Omega}(\bar{\theta}, \theta)^{-1} \Omega(\bar{\theta}, \theta)^{-1} e^{\tilde{\lambda}(\overline{\theta}, \theta) g(\theta)} G_i(\overline{\theta}, \theta) \right] h + h' \mathbb{E}_n \left[ e^{\tilde{\lambda}(\overline{\theta}, \theta) g(\theta)} G_i(\overline{\theta}, \theta)^{\top} \bar{\Omega}(\overline{\theta}, \theta)^{-1} \Omega(\bar{\theta}, \theta) \bar{\Omega}(\bar{\theta}, \theta)^{-1} \Omega(\bar{\theta}, \theta)^{-1} e^{\tilde{\lambda}(\overline{\theta}, \theta) g(\theta)} G_i(\overline{\theta}, \theta) \right] h + 2h' \mathbb{E}_n \left[ e^{\tilde{\lambda}(\overline{\theta}, \theta) g(\theta)} G_i(\overline{\theta}, \theta)^{\top} \bar{\Omega}(\overline{\theta}, \theta)^{-1} \Omega(\bar{\theta}, \theta) \bar{\Omega}(\bar{\theta}, \theta)^{-1} \Omega(\bar{\theta}, \theta)^{-1} e^{\tilde{\lambda}(\overline{\theta}, \theta) g(\theta)} G_i(\overline{\theta}, \theta) \right] h = \mathcal{A}_3 + \mathcal{A}_4 + \mathcal{A}_5. \tag{A.12} \]

Let us start with analysing term \( \mathcal{A}_3 \):

\[
\mathcal{A}_3 = h' \bar{G}(\overline{\theta}, \theta) \Omega(\overline{\theta}, \theta)^{-1} \bar{G}(\overline{\theta}, \theta) h + h' \left[ \bar{G}(\overline{\theta}, \theta) - \bar{G}(\overline{\theta}, \theta)^{\top} \right] \Omega(\overline{\theta}, \theta)^{-1} \Omega(\overline{\theta}, \theta)^{-1} \left[ \bar{G}(\overline{\theta}, \theta) - \bar{G}(\overline{\theta}, \theta)^{\top} \right] \Omega(\overline{\theta}, \theta)^{-1} \bar{G}(\overline{\theta}, \theta) h \]

\[
+ h' \bar{G}(\overline{\theta}, \theta)^{\top} \left[ \Omega(\overline{\theta}, \theta)^{-1} \Omega(\overline{\theta}, \theta)^{-1} \right] \Omega(\overline{\theta}, \theta)^{-1} \bar{G}(\overline{\theta}, \theta) h. \tag{A.13} \]

By Lemma [G.13] in the SM with \( h = h_1 = h_2 = \sup_{h \in \mathcal{H}} h' \bar{G}(\overline{\theta}, \theta)^{\top} \bar{G}(\overline{\theta}, \theta) h = \mathcal{O}_p(h'^{1/2} h + o_p(1)) \), where the \( o_p(1) \) term is uniform in \( h \in \mathcal{H} \). To control the last term in (A.15) we use the equality \( A^{-1} B A^{-1} - C^{-1} = C^{-1} [(C - A) A^{-1} B + (B - A) A^{-1}] C^{-1} \) for square matrices \( A, B \) and \( C \). Therefore,

\[
\left| h' \bar{G}(\overline{\theta}, \theta) \right| \left[ \bar{G}(\overline{\theta}, \theta)^{\top} \right] \Omega(\overline{\theta}, \theta)^{-1} \left[ \bar{G}(\overline{\theta}, \theta)^{\top} \right] \Omega(\overline{\theta}, \theta)^{-1} \Omega(\overline{\theta}, \theta)^{-1} \bar{G}(\overline{\theta}, \theta) h \right| \leq C \left| h' \bar{G}(\overline{\theta}, \theta) \right| \left[ \Omega(\overline{\theta}, \theta)^{-1} \Omega(\overline{\theta}, \theta)^{-1} \right] \left[ \Omega(\overline{\theta}, \theta)^{-1} \bar{G}(\overline{\theta}, \theta) h \right| \times \left| \Omega(\overline{\theta}, \theta)^{-1} \right| \left| \Omega(\overline{\theta}, \theta)^{-1} \bar{G}(\overline{\theta}, \theta) h \right| \tag{A.14} \]

First, by Lemma [G.7] in the SM with \( \tilde{\lambda} = \lambda \): \( \sup_{h \in \mathcal{H}} \| \Omega(\overline{\theta}, \theta) \| = \mathcal{O}_p(\zeta(K) K/\sqrt{n}) \) and if \( \zeta(K) K/\sqrt{n} \to 0 \), \( \sup_{h \in \mathcal{H}} \| \Omega(\overline{\theta}, \theta)^{-1} \| \leq C \). To control term \( \| \Omega(\overline{\theta}, \theta)^{-1} \| \) we use Lemma [G.12] in the SM with \( \kappa_n(\theta_1, \tau) = \kappa(\theta_1) \) and \( \tilde{\lambda}_r(\theta) = \lambda_1(\theta_1) \) to get that \( \| \Omega(\overline{\theta}, \theta)^{-1} \| = \mathcal{O}_p(\zeta(K) K/\sqrt{n}) + C \), where we have used the fact that \( \| \Omega(\overline{\theta}, \theta) \| \leq C \) by Donald et al. [2003, Lemma A.6]. Next, we consider the term \( \| \Omega(\overline{\theta}, \theta)^{-1} \bar{G}(\overline{\theta}, \theta) h \| \) in (A.14). By Lemma [G.12] in the SM with \( \kappa_n(\theta_1, \tau) = \kappa(\theta_1) \) and
\( \tilde{\lambda}_r(\theta) = \tilde{\lambda}(\theta_1) \), Lemmas \( \text{G.10} \) and \( \text{G.7} \) in the SM, and by \( \text{Donald et al. (2003, Lemma A.6)} \) we have:

\[
\sup_{h \in H} \| \tilde{\Omega}(\tilde{\lambda}, \theta_1) - \tilde{\Omega}^*(\tilde{\lambda}, \tilde{\theta}_s) \| = O_p(\zeta(K)K/\sqrt{n}).
\]

Term \( \sup_{h \in H} \| \tilde{\Omega}^*(\tilde{\lambda}, \tilde{\theta}_s)^{-1/2} \tilde{G}(\tilde{\theta}_s) h \| \) in (A.14) is upper bounded by

\[
\sup_{h \in H} \| \tilde{\Omega}^*(\tilde{\lambda}, \tilde{\theta}_s)^{-1/2} \tilde{\Omega}(\tilde{\theta}_s)^{-1/2} \tilde{G}(\tilde{\theta}_s) h \| C \sup_{h \in H} h' V_{\tilde{\theta}_s}^{-1} h + o_p(1) \text{ if } \zeta(K)K/\sqrt{n} \to 0,
\]

by Lemma \( \text{G.7} \) in the SM with \( \bar{\lambda} = \tilde{\lambda} \) and \( \text{Donald et al. (2003, Lemma A.6)} \). So, (A.14) is at most of the order \( O_p(\zeta(K)K/\sqrt{n}) \).

Finally, we analyse the second term in (A.13). By (G.8) in Lemma \( \text{G.9} \) in the SM with \( \tilde{\lambda}(\theta_h) = \tilde{\lambda}(\tilde{\theta}_s) \) and \( \theta_h = \tilde{\theta}_s \), by Lemmas \( \text{G.2, G.7 and G.9} \) in the SM with \( \bar{\lambda} = \tilde{\lambda} \), and by using the same argument as in \( \text{Donald et al. (2003, proof of Theorem 5.4)} \), under the assumption \( \zeta(K)K/\sqrt{n} \to 0 \) we conclude that:

\[
\sup_{h \in H} A_3 = \sup_{h \in H} h' V_{\tilde{\theta}_s}^{-1} h + o_p(1) + O_p(\zeta(K)K/\sqrt{n}).
\]

We now analyse term \( A_4 \) in (A.12). By the matrix extension of the CS inequality (see \( \text{Tripathi (1999)} \)) we have:

\[
\sup_{h \in H} A_4 \leq \sup_{h \in H} \left\| \tilde{\Omega}^*(\tilde{\lambda}, \tilde{\theta}_s)^{-1/2} \tilde{\Omega}(\tilde{\lambda}, \theta_1) \tilde{\Omega}^*(\tilde{\lambda}, \tilde{\theta}_s)^{-1/2} \right\| \sup_{h \in H} E_n \left[ e^{\tilde{\lambda}(\tilde{\theta}_s) g_1(\tilde{\theta}_s) h' G_i(\tilde{\theta}_s) \tilde{\lambda}(\tilde{\theta}_s)} \right]^2 \\
\leq C \sup_{h \in H} \left\| \tilde{\lambda}(\tilde{\theta}_s) \right\|^2 \sup_{h \in H} E_n \left[ \| G_i(\tilde{\theta}_s) h \|^2 \right]^2 \\
\leq C \sup_{h \in H} \left\| \tilde{\lambda}(\tilde{\theta}_s) \right\|^2 2O_p \left( E^D \left[ \sup_{h \in H} \| \rho_0(X, \tilde{\theta}_s) h \|^2 \right] Z \right) \| d_1 \|^2 = O_p(K^2/n)
\]

by Assumption \( \text{3.4 (b), Lemma G.5, Lemma G.7} \) with \( \bar{\lambda} = \tilde{\lambda} \), Lemma \( \text{G.12} \) in the SM with \( \bar{\lambda}_r = \tilde{\lambda} \) and \( \kappa_n(\theta_1, \tau) = \kappa(\theta_1) \), and \( \text{Donald et al. (2003, Lemma A.6)} \).

Next, we analyse term \( A_5 \) in (A.12). By CS:

\[
A_5 \leq \left\| \left( E_n \left[ e^{\tilde{\lambda}(\tilde{\theta}_s) g_1(\tilde{\theta}_s) h' G_i(\tilde{\theta}_s) \tilde{\lambda}(\tilde{\theta}_s)} \right]^2 \right] \right\|^{1/2} \left\| \tilde{\Omega}^*(\tilde{\lambda}, \tilde{\theta}_s)^{-1/2} \tilde{\Omega}(\tilde{\lambda}, \theta_1) \tilde{\Omega}^*(\tilde{\lambda}, \tilde{\theta}_s)^{-1/2} \right\| \\
\times \| \tilde{\Omega}^*(\tilde{\lambda}, \tilde{\theta}_s)^{-1/2} E_n \left[ e^{\tilde{\lambda}(\tilde{\theta}_s) g_1(\tilde{\theta}_s) h' G_i(\tilde{\theta}_s) \tilde{\lambda}(\tilde{\theta}_s)} \right] h \|.
\]

By T. Lemma \( \text{G.7} \) in the SM under the assumption \( \zeta(K)K/\sqrt{n} \to 0 \), Lemmas \( \text{G.5 and G.9} \) in the SM with \( \bar{\lambda}(\theta_h) = \tilde{\lambda}(\tilde{\theta}_s) \) and \( \theta_h = \tilde{\theta}_s \) and \( \text{Donald et al. (2003, Lemma A.6)} \), we have: \( \sup_{h \in H} A_5 = O_p(K/\sqrt{n} + K^2/n) \), where we have also used: \( \sup_{h \in H} E_n \left[ e^{\tilde{\lambda}(\tilde{\theta}_s) g_1(\tilde{\theta}_s) h' G_i(\tilde{\theta}_s) \tilde{\lambda}(\tilde{\theta}_s)} \right]^2 = O_p(K^2/n) \) as in term \( A_4 \). We conclude that

\[
\sup_{h \in H} h' \frac{d \tilde{\lambda}(\tilde{\theta}_s)}{d \theta} \tilde{\Omega}(\tilde{\lambda}, \theta_1) \frac{d \tilde{\lambda}(\tilde{\theta}_s)}{d \theta} h = \sup_{h \in H} h' V_{\tilde{\theta}_s}^{-1} h + o_p(1) + O_p(\zeta(K)K/\sqrt{n})
\]

and by plugging this and the rate for the second term of (A.8) in (A.8) we obtain that uniformly in \( h \in H \):

\[
\sum_{i=1}^n \ell_{n, \theta_1}(W_i) - \sum_{i=1}^n \ell_{n, \tilde{\theta}_s}(W_i) = O_p(\zeta(K)K^2/\sqrt{n}) + h' V_{\tilde{\theta}_s}^{-1} \Delta_{n, \theta_1} - \frac{1}{2} h' V_{\tilde{\theta}_s}^{-1} h
\]

where \( \Delta_{n, \theta_1} = -N(0, V_{\theta_1}) \).
A.2 Proof for Section 4

We can write $\log p(W_{1:n}|\theta^\ell; M_\ell, K) = \sum_{i=1}^n \ell_{n,\theta^\ell}(W_i) = -n \log n + n \log \hat{L}(\theta^\ell)$ where $\hat{L}(\theta^\ell) \triangleq \exp \{ \hat{\lambda}(\theta^\ell)'g(\theta^\ell) \} \left[ \frac{1}{n} \sum_{i=1}^n \exp \{ \hat{\lambda}(\theta^\ell)'g_i(W_i, \theta^\ell) \} \right]^{-1}$ and $L(\theta^\ell) = \exp \{ \lambda_0(\theta^\ell) \mathbb{E}^P [g(W, \theta^\ell)] \} \left( \mathbb{E}^P [\exp \{ \lambda_0(\theta^\ell) g(W, \theta^\ell) \}] \right)^{-1}$. Then, by using (4.1) we have:

$$P_\ast \left( \log m(W_{1:n}|M_j, K) > \max_{\ell \neq j} \log m(W_{1:n}|M_\ell, K) \right) = P_\ast \left( n \log L(\theta_0^j) + \log \frac{\hat{L}(\theta_0^j)}{L(\theta_0^j)} + B_j > \max_{\ell \neq j} \left[ n \log L(\theta_0^\ell) + B_\ell + n \log \frac{\hat{L}(\theta_0^\ell)}{L(\theta_0^\ell)} \right] \right) \quad (A.15)$$

where $\forall \ell$, $B_\ell \triangleq \log \pi(\theta_0^\ell|M_\ell) - \log \pi(\theta_0^\ell|W_{1:n}, M_\ell, K)$ and $B_\ell = O_p(1)$ under the assumptions of Theorem 3.2. By definition of $Q^*(\theta)$ in Section 3.4 we have that:

$$\log L(\theta_0^j) = \mathbb{E}^P [\log dQ^*(\theta_0^j)/dP] = -\mathbb{E}^P [\log dP_\ast/dQ^*(\theta_0^j)] = -\mathbb{K}(P_\ast||Q^*(\theta_0^j))$$

Remark that $\mathbb{E}^P [\log (dP_\ast/dQ^*(\theta_0^j))] > \mathbb{E}^P [\log (dP_\ast/dQ^*(\theta_0^j))]$ means that the Kullback-Leibler divergence between $P_\ast$ and $Q^*(\theta_0^j)$ is smaller for model $M_1$ than for model $M_2$, where $Q^*(\theta_0^j)$ minimizes the Kullback-Leibler divergence between $P_\ast$ and $Q^*(\theta_0^j)$, is smaller for model $M_1$ than for model $M_2$, where $Q^*(\theta_0^j)$ minimizes the Kullback-Leibler divergence between $Q \in \mathcal{P}_{\theta_0}$ and $P_\ast$ for $\ell \in \{1, 2\}$ (notice the inversion of the two probabilities in the definition of the Kullback-Leibler divergences).

First, suppose that $\min_{\ell \neq j} \mathbb{E}^P [\log \left( dP_\ast/dQ^*(\theta_0^j) \right)] > \mathbb{E}^P [\log \left( dP_\ast/dQ^*(\theta_0^j) \right)]$. By (A.15):

$$P_\ast \left( \log m(W_{1:n}|M_j, K) > \max_{\ell \neq j} \log m(W_{1:n}|M_\ell, K) \right) \geq P_\ast \left( \log \frac{\hat{L}(\theta_0^j)}{L(\theta_0^j)} - \max_{\ell \neq j} \log \frac{\hat{L}(\theta_0^\ell)}{L(\theta_0^\ell)} + \frac{1}{n} (B_j - \max_{\ell \neq j} B_\ell) > \max_{\ell \neq j} \log L(\theta_0^\ell) - \log L(\theta_0^j) \right) \quad (A.16)$$

This probability converges to 1 because $\mathcal{I}_n = \mathbb{K}(P_\ast||Q^*(\theta_0^j)) - \min_{\ell \neq j} \mathbb{K}(P_\ast||Q^*(\theta_0^j)) < 0$ by assumption, and $\left[ \log \hat{L}(\theta_0^j) - \log L(\theta_0^j) \right] \to 0$, for every $\ell \in \{1, 2\}$ by Lemma G.26 in the SM if $\max \{ \sqrt{K}, \|\lambda_0(\theta_0)\| \} \sqrt{K/n} \to 0$.

To prove the second direction of the statement, suppose that $\lim_{n \to \infty} P_\ast (\log m(W_{1:n}|M_j, K) > \max_{\ell \neq j} \log m(W_{1:n}|M_\ell, K)) = 1$. By (A.15) it holds, $\forall \ell \neq j$

$$P_\ast \left( \log m(W_{1:n}|M_j, K) > \max_{\ell \neq j} \log m(W_{1:n}|M_\ell, K) \right) \leq P_\ast \left( \log \frac{\hat{L}(\theta_0^j)}{L(\theta_0^j)} - \log \frac{\hat{L}(\theta_0^\ell)}{L(\theta_0^\ell)} + \frac{1}{n} (B_j - B_\ell) > \log \frac{L(\theta_0^j)}{L(\theta_0^j)} \right)$$

Convergence to 1 of the left hand side implies convergence to 1 of the right hand side which is possible only if $\log L(\theta_0^j) - \log L(\theta_0^j) < 0$. Since this is true for every model $\ell$, then this implies that $\mathbb{K}(P||Q^*(\theta_0^j)) < \min_{\ell \neq j} \mathbb{K}(P||Q^*(\theta_0^j))$ which concludes the proof.
References


