Centre de recherche en économie et statistique

Série des Documents de Travail

n° 2016-21

Bayesian Empirical Likelihood Estimation and Comparison of Moment Condition Models

S.Chib [1]
M.Shin [2]
A.Simoni [3]

[1] Olin Business School, Washington University. E-mail: chib@wustl.edu

[2] Department of Economics, University of Illinois. E-mail : mincshin@illinois.edu

[3] CREST. E-mail : simoni.anna@gmail.com

Bayesian Empirical Likelihood Estimation and Comparison of Moment Condition Models

Siddhartha Chib* Minchul Shin† Anna Simoni‡

Washington University in St. Louis *University of Illinois* *CNRS and CREST*

This version: June, 2016

Abstract

In this paper we consider the problem of inference in statistical models characterized by moment restrictions by casting the problem within the Exponentially Tilted Empirical Likelihood (ETEL) framework. Because the ETEL function has a well defined probabilistic interpretation and plays the role of a likelihood, a fully Bayesian framework can be developed. We establish a number of powerful results surrounding the Bayesian ETEL framework in such models. One major concern driving our work is the possibility of misspecification. To accommodate this possibility, we show how the moment conditions can be reexpressed in terms of additional nuisance parameters and that, even under misspecification, the Bayesian ETEL posterior distribution satisfies a Bernstein-von Mises result. A second key contribution of the paper is the development of a framework based on marginal likelihoods (MLs) and Bayes factors to compare models defined by different moment conditions. Computation of the MLs is by Chib (1995)'s method. We establish the consistency of the Bayes factors and show that the ML favors the model with the minimum number of parameters and the maximum number of valid moment restrictions. When the models are misspecified, the ML model selection procedure selects the model that is closer to the (unknown) true data generating process in terms of the Kullback-Leibler divergence. The ideas and results in this paper provide a further broadening of the theoretical underpinning and value

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

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

‡CREST, 15, Boulevard Gabriel Péri, 92240 Malakoff, France, e-mail: simoni.anna@gmail.com

of the Bayesian ETEL framework with likely far-reaching practical consequences. The discussion is illuminated through several examples.

Key words: Bayes factor consistency; Bernstein-von Mises theorem; Estimating Equations; Exponentially Titled Empirical Likelihood; Generalized Method of Moments; Kullback-Leibler divergence; Marginal Likelihood; Misspecification; Model comparison; Count regression.

1 Introduction

Over the last few decades, empirical likelihood (EL) based methods have emerged as a powerful analytical and inference tool for semiparametric frequentist inference about parameters θ that are implicit functionals of the unknown data distribution P (see *e.g.* Owen (1988), Qin and Lawless (1994), Kitamura and Stutzer (1997), Owen (2001), Schennach (2007), Chen and Van Keilegom (2009), and references therein). The EL can also be used in a Bayesian framework in place of the data distribution P, as suggested in Lazar (2003). In fact, Grendar and Judge (2009) show that the EL is the mode of the posterior of P under a general prior on P. In another important paper, Schennach (2005) shows that a nonparametric likelihood closely related to EL, called the exponentially tilted empirical likelihood (ETEL), arises after marginalizing over the unknown P when P is modeled by a nonparametric prior that gives preference to distributions having a small support and favors entropy-maximizing distributions. By combining either one of these nonparametric likelihoods with a prior $\pi(\theta)$ on θ, a large class of models, hitherto difficult to analyze from the Bayesian perspective, can be subjected to a full Bayesian semiparametric analysis. For instance, the class of moment condition models, in which the functionals of P are a set of one or more moment restrictions of the type $\mathbf{E}^P[g(X,\theta)] = 0$, where $g(X,\theta)$ is a known vector-valued function of a random vector X and an unknown parameter vector θ, can be analyzed in this way, thus providing a Bayesian counterpoint to frequentist estimating equation or generalized method of moment approaches.

Not surprisingly, there is a growing Bayesian literature based on such an approach. On the application side, for example, quantile moment condition models are discussed in Lancaster and Jun (2010), Kim and Yang (2011), Yang and He (2012), Xi et al. (2016), complex surveys in Rao and Wu (2010), and small area estimation in Chaudhuri and Ghosh (2011), Porter et al. (2015), Chaudhuri et al. (2017). On the theory side, Yang and He (2012) establishes the asymptotic normality of the Bayesian EL posterior distribution of the quantile regression parameter, and Fang and Mukerjee (2006) and Chang and Mukerjee (2008) study higher-order asymptotic and coverage properties of the Bayesian EL/ETEL posterior distribution for the population mean, while Schennach (2005) and Lancaster and Jun (2010) consider the large-sample behavior of the Bayesian ETEL posterior distribution under the assumption that all moment restrictions are valid.

The goal of this paper is to establish a number of powerful results surrounding the Bayesian ETEL framework in moment condition models, complementing and extending the aforementioned papers in important directions. One major goal is the Bayesian analysis of models that are potentially misspecified. For this reason, our analysis is built on the ETEL function which, as shown by Schennach (2007), leads to frequentist estimators of θ that have the same orders of bias and variance (as a function of the sample size) as the EL estimators but, importantly, maintain the root n convergence even under model misspecification. We show that the ETEL framework is an immensely useful organizing framework within which a fully Bayesian treatment of correctly and misspecified moment condition models can be developed. We show that even under misspecification, the Bayesian ETEL posterior distribution has desirable properties, and that it satisfies the Bernstein - von Mises (BvM) theorem.

Another key focus of the paper is the development of a framework based on marginal likelihoods (MLs) and Bayes factors for comparing different moment restricted models and for discarding any misspecified moment restrictions. Essentially, each set of moment restrictions, and the different sets of restrictions on the parameters, define different models. Our proposal

is to compare these various models based on the corresponding ML, and to select the model with the larger ML. It turns out that in order to compare different models, in particular those defined by different sets of moment conditions, it is necessary to linearly transform the moment functions $g(X, \theta)$ so that all the transformed moments are included in each model. This linear transformation simply consists of adding an extra parameter different from zero to the components of the vector $g(X, \theta)$ that correspond to the restrictions not included in a specific model. We compute the ML based on the method of Chib (1995) as extended to Metropolis-Hastings samplers in Chib and Jeliazkov (2001). This method makes exact (up to simulation error) computation of the ML extremely simple and is a key feature of both our numerical and theoretical analysis. Our asymptotic theory shows that the ML-based selection procedure is consistent in the sense that: *(i)* it discards misspecified moment restrictions, *(ii)* it selects the model that contains the maximum number of valid moment restrictions when comparing two correctly specified models, and *(iii)* it selects the model that is the "less misspecified" when comparing two misspecified models. These important Bayes factor consistency results are based on the asymptotic behavior of the ETEL function for both correctly and misspecified models and the validity of the BvM theorem for both correctly and misspecified models. These results on the model comparison problem complement and substantially extend the work of Variyath et al. (2010), which focuses on EL information criteria, and that of Hong and Preston (2012), where Bayes factors are constructed based on the ML obtained from an approximation to the true P, and Vexler et al. (2013) where Bayes factors are constructed from the EL.

The rest of the article is organized as follows. In Section 2 we describe the moment condition model, define the notion of misspecification in this setting, and then discuss the prior-posterior analysis with the ETEL function. We then provide the first pair of major results dealing with the asymptotic behavior of the posterior distribution for both correctly and misspecified models. Section 3 introduces our model selection procedure based on MLs and Bayes factors and the consistency results regarding Bayes factors. Throughout the

paper, for expository purposes, we include numerical examples. The numerical illustrations are continued further in Section 4 where the problems of variable selection and link estimation are illustrated in the setting of a count regression model. Our conclusions are in Section 5 and proofs of our results are collected in the Appendix and in a Supplementary Appendix.

2 Setting

Suppose that X is an $\mathbb{R}^{d_x}$-valued random vector with (unknown) distribution P. Suppose that the operating assumption is that the distribution P satisfies the d unconditional moment restrictions

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

where $\mathbf{E}^P$ denotes the expectation taken with respect to P, $g : \mathbb{R}^{d_x} \times \Theta \mapsto \mathbb{R}^d$ is a vector of known functions with values in $\mathbb{R}^d$, $\theta := (\theta_1, \ldots, \theta_p)' \in \Theta \subset \mathbb{R}^p$ is the parameter vector of interest, and 0 is the $d \times 1$ vector of zeros. We assume that $\mathbf{E}^P[g(X,\theta)]$ is bounded for every $\theta \in \Theta$. We also suppose that we are given a random sample $x_{1:n} := (x_1, \ldots, x_n)$ on X and that $d \geq p$.

When the number of moment restrictions d exceeds the number of parameters p, the parameter θ in such a setting is said to be overidentified (overrestricted). In such a case, there is a possibility that a subset of the moment condition may be invalid in the sense that the true data generating process is not contained in the collection of probability measures that satisfy the moment conditions for all $\theta \in \Theta$. That is, there is no parameter θ in Θ that is consistent with the moment restrictions (2.1) under the true data generating process P. To deal with possibly invalid moment restrictions, we reformulate the moment conditions in terms of an additional nuisance parameter $V \in \mathfrak{V} \subset \mathbb{R}^d$. For example, if the k-th moment condition is not expected to be valid, we subtract $V = (V_1, \ldots, V_d)$ from the moment restrictions where V_k is a free parameter and all other elements of V are zero. To accommodate this situation, we rewrite the above conditions as the following augmented

moment conditions

$$\mathbf{E}^P[g^A(X,\theta,V)] = 0 \tag{2.2}$$

where $g^A(X,\theta,V) := g(X,\theta) - V$. Note that in this formalism, the parameter V indicates which moment restrictions are active where for 'active moment restrictions' we mean the restrictions for which the corresponding components of V is zero. In order to guarantee identification of θ, at most $(d-p)$ elements of V can be different than zero. If all the elements of V are zero, we recover the restrictions in (2.1).

Let $d_v \leq (d-p)$ be the number of non-zero elements of V and let $v \in \mathcal{V} \subset \mathbb{R}^{d_v}$ be the vector that collects all the non-zero components of V. We call v the augmented parameter and θ the parameter of interest. Therefore, the number of active moment restrictions is $d - d_v$. In the following, we write $g^A(X,\theta,v)$ as a shorthand for $g^A(X,\theta,V)$ with v the vector obtained from this V by collecting only its non-zero components.

The central problem of misspecifcation of the moment conditions, mentioned in the preceding paragraph, can now be formally defined in terms of the augmented moment conditions.

Definition 2.1 (Misspecified model)**.** *We say that the augmented moment condition model is misspecified if the set of probability measures implied by the moment restrictions does not contain the true data generating process P for every $(\theta,v) \in \Theta \times \mathcal{V}$, that is, $P \notin \mathcal{P}$ where $\mathcal{P} = \bigcup_{(\theta,v)\in\Theta\times\mathcal{V}} \mathcal{P}_{(\theta,v)}$ and $\mathcal{P}_{(\theta,v)} = \{Q \in \mathbb{M};\, \mathbf{E}^Q[g^A(X,\theta,v)] = 0\}$ with $\mathbb{M}$ the set of all probability measures on $\mathbb{R}^{d_x}$.*

In a nutshell, a set of augmented moment conditions is misspecified if there is no pair (θ,v) in $(\Theta \times \mathcal{V})$ that satisfies $\mathbf{E}^P[g^A(X,\theta,v)] = 0$ where P is the true data generating process. On the other hand, if such a pair of values (θ,v) exists then the set of augmented moment conditions is correctly specified.

Throughout the paper, we use a location parameter model as a running example to understand the various concepts and ideas.

Example (Linear regression model). Suppose that we are interested in estimating the following linear regression model with an intercept and two predictors:

$$y_i = \mu + \beta_1 z_{1,i} + \beta_2 z_{2,i} + e_i \tag{2.3}$$

where $(z_{1,i}, z_{2,i}, e_i)'$ is independently drawn from some distribution P for $i = 1, 2, ..., n$. Under the assumption that $\mathbf{E}[e_i|z_{j,i}] = 0$ for $j = 1, 2$, we can use the following moment restrictions to estimate $\theta := (\mu, \beta_1, \beta_2)$:

$$\mathbf{E}^P[e_i(\theta)] = 0, \quad \mathbf{E}^P[e_i(\theta)z_{1,i}] = 0, \quad \mathbf{E}^P[e_i(\theta)z_{2,i}] = 0, \quad \mathbf{E}^P[(e_i(\theta))^3] = v, \tag{2.4}$$

where $e_i(\theta) := (y_i - \mu - \beta_1 z_{1,i} - \beta_2 z_{2,i})$. The first three moment restrictions are derived from the standard orthogonality condition and identify θ. The last restriction potentially serves as additional information. Hence, by using notation in (2.1) and (2.2), $x_i := (y_i, z_{1,i}, z_{2,i})$, $g(x_i, \theta) = (e_i(\theta), e_i(\theta)z_{1,i}, e_i(\theta)z_{2,i}, e_i(\theta)^3)'$, $g^A(x_i, \theta, V) = g(x_i, \theta) - (0, 0, 0, V_4)'$ and $v = V_4$. If one believes that the underlying distribution of e_i is indeed symmetric, then one could use this information by setting v to zero. Otherwise, it is desirable to treat v as an unknown object. If the distribution of e_i is skewed and v is forced to be zero, then the model becomes misspecified because there is no (μ, β_1, β_2) that is consistent with the above four moment restrictions altogether under P. When the augmented parameter v is treated as a free parameter, the model is correctly specified even under asymmetry.

2.1 Prior-Posterior analysis

Following Schennach (2005), we now discuss the prior-posterior analysis of (θ, v) with the ETEL function. The ETEL function has been shown by Schennach (2005) to have a sound Bayesian interpretation in that it arises by marginalization over a nonparametric prior on P that favors distributions that are close to the empirical distribution function in terms of Kullback-Leibler (KL) divergence. We note that Schennach (2005) did not involve the

augmented parameter v though the framework there is readily adapted to this case.

In particular, suppose that *(i)* $g^A(x,\theta,v)$ is continuous in x for every $(\theta,v) \in \Theta \times \mathcal{V}$ (or has a finite number of step discontinuities) and *(ii)* the interior of the convex hull of $\bigcup_{i=1}^n g^A(x_i,\theta,v)$ contains the origin. Then, adapting the arguments of Schennach (2005), the posterior distribution of (θ,v) after marginalization over P has the form

$$\pi(\theta,v|x_{1:n}) \propto \pi(\theta,v)p(x_{1:n}|\theta,v) \tag{2.5}$$

where $\pi(\theta,v)$ is the prior of (θ,v) and $p(x_{1:n}|\theta,v)$ is the ETEL function defined as

$$p(x_{1:n}|\theta,v) = \prod_{i=1}^n p_i^*(\theta,v) \tag{2.6}$$

and $p_i^*(\theta,v)$ are the probabilities that minimize the KL divergence between the probabilities $(p_1,...,p_n)$ assigned to each sample observation and the empirical probabilities $(\frac{1}{n},...,\frac{1}{n})$, subject to the conditions that the probabilities $(p_1,...,p_n)$ sum to one and that the expectation under these probabilities satisfies the given moment conditions:

$$\begin{aligned} & \max_{p_1,\ldots,p_n} \sum_{i=1}^n [-p_i \log(np_i)] \\ \text{subject to} \quad & \sum_{i=1}^n p_i = 1 \quad \text{and} \quad \sum_{i=1}^n p_i g^A(x_i,\theta,v) = 0. \end{aligned} \tag{2.7}$$

For numerical and theoretical purposes below, the preceding probabilities are computed more conveniently from the dual (saddlepoint) representation as, for $i = 1,\ldots,n$

$$p_i^*(\theta,v) := \frac{e^{\widehat{\lambda}(\theta,v)'g^A(x_i,\theta,v)}}{\sum_{j=1}^n e^{\widehat{\lambda}(\theta,v)'g^A(x_j,\theta,v)}}, \qquad \text{where } \widehat{\lambda}(\theta,v) = \arg\min_{\lambda\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \exp\left(\lambda' g^A(x_i,\theta,v)\right). \tag{2.8}$$

Therefore, the posterior distribution takes the form

$$\pi(\theta, v|x_{1:n}) \propto \pi(\theta, v) \prod_{i=1}^{n} \frac{e^{\widehat{\lambda}(\theta,v)' g^A(x_i,\theta,v)}}{\sum_{j=1}^{n} e^{\widehat{\lambda}(\theta,v)' g^A(x_j,\theta,v)}}, \tag{2.9}$$

which may be called the Bayesian Exponentially Tilted Empirical Likelihood (BETEL) posterior distribution. It can be efficiently simulated by MCMC methods. For example, the one block tailored Metropolis-Hastings algorithm (Chib and Greenberg, 1995) is implemented as follows. Let $q(\theta, v|x_{1:n})$ denote a student-t distribution whose location parameter is the mode of the log BETEL posterior distribution and whose dispersion matrix is the negative inverse Hessian matrix of the log BETEL posterior at the mode. Then, a sample of draws from the BETEL posterior can be obtained by repeating the following steps for $s = 1, ..., S$ starting from some initial value $(\theta^{(0)}, v^{(0)})$:

1. Draw $(\theta^\dagger, v^\dagger)$ from $q(\theta, v|x_{1:n})$ and solve for $p_i^*(\theta^\dagger, v^\dagger)$, $1 \leq i \leq n$, from the EL saddle-point problem (2.8).

2. Calculate the M-H probability of move

$$\alpha\left((\theta^{s-1}, v^{s-1}), (\theta^\dagger, v^\dagger)\middle|\, x_{1:n}\right) = \min\left\{1, \frac{\pi(\theta^\dagger, v^\dagger|x_{1:n})}{\pi(\theta^{s-1}, v^{s-1}|x_{1:n})} \frac{q(\theta^{s-1}, v^{s-1}|x_{1:n})}{q(\theta^\dagger, v^\dagger|x_{1:n})}\right\}.$$

3. Set $(\theta^s, v^s) = (\theta^\dagger, v^\dagger)$ with probability $\alpha((\theta^{s-1}, v^{s-1}), (\theta^\dagger, v^\dagger)|x_{1:n})$. Otherwise, set $(\theta^s, v^s) = (\theta^{s-1}, v^{s-1})$. Go to step 1.

Note that when the dimension of (θ, v) is large, the TaRB-MH algorithm of Chib and Ramamurthy (2010) can be used instead for improved simulation efficiency.

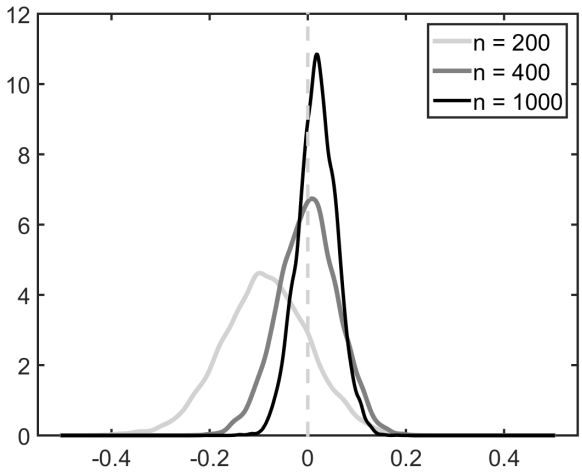
Example (Linear regression model, continued). To illustrate the BETEL posterior distribution, we generate $(y_1, y_2, ..., y_n)$ in (2.3) without predictors (i.e., $\beta_1 = 0$ and $\beta_2 = 0$).

Suppose that the distribution of the e_i is skewed:

$$e_i \sim \begin{cases} N(1, 0.5^2) & \text{with probability } 0.5 \\ N(-1, 1^2) & \text{with probability } 0.5. \end{cases} \tag{2.10}$$

We employ the first and fourth moment restrictions in (2.4) (that is, $g^A(x_i, \theta, v) = (e_i(\theta), e_i(\theta) - v)'$) and compute the BETEL posterior distribution for μ (Figure 1). These two moment restrictions form a correctly specified moment condition model. As the number of observations increases the BETEL posterior distribution shrinks around the true value ($\mu = 0$) and becomes similar to a Gaussian distribution, indicating that the BvM theorem seems to hold for the BETEL posterior distribution. In the next two sections, we explore the behavior of the BETEL posterior distribution under fairly general assumptions and prove that the BETEL posterior distribution shrinks at the $\sqrt{n}$-rate.

Figure 1: BETEL Posterior Distribution for μ



Notes: This figure presents the BETEL posterior distribution of the location parameter μ with $n = 200, 400, 1000$ where n is the number of observations. Prior distribution for μ and v are set to be normal distribution with mean 0 and variance 10. We generate 25,000 posterior draws using the one block tailored Metropolis-Hastings algorithm described in Section 2.1. Our proposal density is set to be a t-distribution with mean as the posterior mode, variance as the 1.5 times negative inverse Hessian of the log-BETEL posterior at the posterior mode, 15 as the degrees of freedom. The rejection probabilities are about 25% for all cases.

Notation. In Sections 2.2 and 2.3 we use the following notations. For ease of exposition, we denote $\psi := (\theta, v)$, $\psi \in \Psi$ with $\Psi := \Theta \times \mathcal{V}$. Moreover, $\|\cdot\|_F$ denotes the Frobenius norm. The notation '$\xrightarrow{p}$' is for convergence in probability with respect to the product measure $P^n = \bigotimes_{i=1}^n P$. The log-likelihood function for one observation is denoted by $l_{n,\psi}$:

$$l_{n,\psi}(x) := \log \frac{e^{\widehat{\lambda}(\psi)' g^A(x,\psi)}}{\sum_{j=1}^n e^{\widehat{\lambda}(\psi)' g^A(x_j,\psi)}} = -\log n + \log \frac{e^{\widehat{\lambda}' g^A(x,\psi)}}{\frac{1}{n}\sum_{j=1}^n \left[e^{\widehat{\lambda}' g^A(x_j,\psi)}\right]}$$

so that the log-ETEL function writes $\log p(x_{1:n}|\psi) = \sum_{i=1}^n l_{n,\psi}(x_i)$. For a set $\mathcal{A} \subset \mathbb{R}^m$, we denote by $int(\mathcal{A})$ its interior relative to $\mathbb{R}^m$. Further notations are introduced as required.

2.2 Asymptotic Properties: correct specification

In this section, we establish that when the model is correctly specified the BETEL posterior distribution has good frequentist asymptotic properties as the sample size n increases. Namely, we show that the BETEL posterior distribution has a Gaussian limiting distribution and that it concentrates on a $n^{-1/2}$-ball centred at the true value of the parameter. These properties have been informally discussed in Schennach (2005) but without specifying the assumptions required. We provide these assumptions, which are standard in the Empirical Likelihood and Bayesian literatures, and in Theorem 2.1 below we provide the asymptotic results. The proof of this theorem is quite standard (see *e.g.* Lehmann and Casella (1998)) and so we postpone it to the Supplementary Appendix.

Let θ_* be the true value of the parameter of interest θ and v_* be the true value of the augmented parameter. So, $\psi_* := (\theta_*, v_*)$. The true value v_* is equal to zero when the non-augmented model (2.1) is correctly specified. Moreover, let $\Delta := \mathbf{E}^P[g^A(X,\psi_*)g^A(X,\psi_*)']$, $\Gamma := \mathbf{E}^P\left[\frac{\partial}{\partial\psi'} g^A(X,\psi_*)\right]$. The first assumption requires that the augmented model is correctly specified in the sense that there is a value of ψ such that (2.2) is satisfied by P, and that this value is unique. A necessary condition for the latter is that $(d-p) \geq d_v \geq 0$.

Assumption 1. *Model (2.2) is such that $\psi_* \in \Psi$ is the unique solution to $\mathbf{E}^P[g^A(X,\psi)] = 0$.*

The next two assumptions include assumptions on the smoothness of the function $g^A(x,\psi)$ and on its moments, and assumptions on the parameter space.

Assumption 2. *(a) X_i, $i = 1,\dots,n$ are i.i.d. random variables that take values in $(\mathcal{X}, \mathfrak{B}_\mathcal{X})$ with probability distribution P, where $\mathcal{X} \subseteq \mathbb{R}^{d_x}$; (b) for every $0 \leq d_v \leq d - p$, $\psi \in \Psi \subset \mathbb{R}^p \times \mathbb{R}^{d_v}$ where Θ and $\mathcal{V}$ are compact and connected and $\Psi := \Theta \times \mathcal{V}$; (c) $g(x,\theta)$ is continuous at each $\theta \in \Theta$ with probability one; (d) $\mathbf{E}^P[\sup_{\psi\in\Psi} \|g^A(X,\psi)\|^\alpha] < \infty$ for some $\alpha > 2$; (e) Δ is nonsingular.*

Assumption 3. *(a) $\psi_* \in int(\Psi)$; (b) $g^A(x,\psi)$ is continuously differentiable in a neighborhood $\mathfrak{U}$ of ψ_* and $\mathbf{E}^P[\sup_{\psi\in\mathfrak{U}} \|\partial g^A(X,\psi)/\partial\psi'\|_F] < \infty$; (c) $rank(\Gamma) = p$.*

Assumption 2 and 3 are the same as the assumptions of Newey and Smith (2004, Theorem 3.2) and Schennach (2007, Theorem 3). The next assumption concerns the prior distribution and is a standard assumption for asymptotic properties of Bayesian procedures. It requires the prior to put enough mass to balls around the true value ψ_* and allows for a $n^{-1/2}$-contraction rate of the posterior distribution.

Assumption 4. *(a) π is a continuous probability measure that admits a density with respect to the Lebesgue measure; (b) π is positive on a neighborhood of ψ_*.*

For a correctly specified moment conditions model, the asymptotic normality of the BETEL posterior is established in the following theorem where we denote by $\pi(\sqrt{n}(\psi - \psi_*)|x_{1:n})$ the posterior distribution of $\sqrt{n}(\psi - \psi_*)$.

Theorem 2.1 (Bernstein - von Mises – correct specification)**.** *Under Assumptions 1 - 4 and if in addition, for any $\delta > 0$, there exists an $\epsilon > 0$ such that, as $n \to \infty$*

$$P\left(\sup_{\|\psi-\psi_*\|>\delta} \frac{1}{n}\sum_{i=1}^{n} (l_{n,\psi}(x_i) - l_{n,\psi_*}(x_i)) \leq -\epsilon\right) \to 1, \tag{2.11}$$

then the posteriors converge in total variation towards a normal distribution, that is,

$$\sup_{B} \left| \pi(\sqrt{n}(\psi - \psi_*) \in B | x_{1:n}) - \mathcal{N}_{0,(\Gamma'\Delta^{-1}\Gamma)^{-1}}(B) \right| \xrightarrow{p} 0 \tag{2.12}$$

where $B \subseteq \Psi$ is any Borel set.

The result of this theorem means that the posterior distribution $\pi(\psi|x_{1:n})$ of ψ is asymptotically normal, centered on the true value ψ_* and with variance $n^{-1}(\Gamma'\Delta^{-1}\Gamma)^{-1}$. The posterior distribution has the same asymptotic variance as the efficient generalized method of moment estimator of Hansen (1982) (see also Chamberlain (1987)). Assumption (2.11) is an identifiability condition which is standard in the literature (see *e.g.* Lehmann and Casella (1998, Assumption 6.B.3)) and which controls the behavior of the log-ETEL function at a distance from ψ_*. Controlling this behavior is important because the posterior involves integration over the whole range of ψ. To understand the meaning of this assumption, remark that asymptotically the log-ETEL function $\psi \mapsto \sum_{i=1}^{n} l_{n,\psi}(x_i)$ is maximized at the true value ψ_* because the model is correctly specified. Hence, Assumption (2.11) means that if the parameter ψ is "far" from the true value ψ_* then the log-ETEL function has also to be small, that is, has to be far from the maximum value $\sum_{i=1}^{n} l_{n,\psi_*}(x_i)$.

2.3 Asymptotic Properties: misspecification

In this section, we consider the case where the model is misspecified in the sense of Definition 2.1 and establish that, even in this case, the BETEL posterior distribution has good frequentist asymptotic properties as the sample size n increases. Namely, we show that the BETEL posterior is asymptotically normal and that it concentrates on a $n^{-1/2}$-ball centred at the pseudo-true value of the parameter. To the best of our knowledge, these properties have not been established yet for misspecified models.

Because in misspecified models there is no value of ψ for which the true data distribution P satisfies the restriction (2.2), we need to define a pseudo-true value for ψ. The latter

is defined as the value of ψ that minimizes the KL divergence $K(P||Q^*(\psi))$ between the true data distribution P and a distribution $Q^*(\psi)$ defined as $Q^*(\psi) := \text{arginf}_{Q\in\mathcal{P}_\psi} K(Q||P)$, where $K(Q||P) := \int \log(dQ/dP)dQ$ and $\mathcal{P}_\psi$ is defined in Definition 2.1. We remark that these two KL divergences are the population counterparts of the KL divergences used for the definition of the ETEL function in (2.6): the empirical counterpart of $K(Q||P)$ is used to construct the $p_i^*(\psi)$ probabilities and the empirical counterpart of $K(P||Q^*(\psi))$ is proportional to the negative log-ETEL function. Roughly speaking, the pseudo-true value is the value of ψ for which the distribution that satisfies the corresponding restrictions (2.2) is the closest to the true P, in the KL sense. By using the dual representation of the KL minimization problem, the P-density $dQ^*(\psi)/dP$ admits a closed-form: $dQ^*(\psi)/dP = e^{\lambda_\circ(\psi)'g^A(X,\psi)}/\mathbf{E}^P\left[e^{\lambda_\circ(\psi)'g^A(X,\psi)}\right]$ where $\lambda_\circ(\psi)$ is the pseudo-true value of the tilting parameter defined as the solution of $\mathbf{E}^P[\exp\{\lambda' g^A(X,\psi)\}g^A(X,\psi)] = 0$ which is unique by the strict convexity of $\mathbf{E}^P[\exp\{\lambda' g^A(X,\psi)\}]$ in λ. Therefore,

$$\begin{aligned} \lambda_\circ(\psi) &:= \arg\min_{\lambda\in\mathbb{R}^d} \mathbf{E}^P\left[e^{\lambda' g^A(X,\psi)}\right], \\ \psi_\circ &:= \arg\max_{\psi\in\Psi} \mathbf{E}^P \log\left[\frac{e^{\lambda_\circ(\psi)'g^A(X,\psi)}}{\mathbf{E}^P\left[e^{\lambda_\circ(\psi)'g^A(X,\psi)}\right]}\right]. \end{aligned} \tag{2.13}$$

However, in a misspecified model, the dual theorem is not guaranteed to hold and so $\psi_\circ$ defined in (2.13) is not necessarily equal to the pseudo-true value defined as the KL-minimizer. In fact, when the model is misspecified, the probability measures in $\mathcal{P} := \bigcup_{\psi\in\Psi}\mathcal{P}_\psi$, which are implied by the model, could not have a common support with the true P, see Sueishi (2013) for a discussion on this point. Following Sueishi (2013, Theorem 3.1), in order to guarantee identification of the pseudo-true value by (2.13) we introduce the following assumption. This assumption replaces Assumption 1 in misspecified models.

Assumption 5. *For a fixed $\psi\in\Psi$, there exists $Q\in\mathcal{P}_\psi$ such that Q is mutually absolutely continuous with respect to P, where $\mathcal{P}_\psi$ is defined in Definition 2.1.*

A similar assumption is also made by Kleijn and van der Vaart (2012) to establish the

BvM under misspecification. Moreover, because consistency in misspecified models is defined with respect to the pseudo-true value $\psi_\circ$, we need to replace Assumption 4 *(b)* by the following assumption which, together with Assumption 4 *(a)*, requires the prior to put enough mass to balls around $\psi_\circ$.

Assumption 6. *The prior distribution π is positive on a neighborhood of $\psi_\circ$.*

In addition to these assumptions, to prove Theorem 2.2 below we also use Assumptions 2 *(a)-(d)* and 3 *(b)* in the previous section. Finally, in order to guarantee $n^{-1/2}$-convergence of $\widehat{\lambda}$ towards $\lambda_\circ$ and $n^{-1/2}$-contraction of the posterior distribution of ψ around $\psi_\circ$, we introduce Assumptions 7 and 8. These assumptions require the pseudo-true values $\lambda_\circ$ and $\psi_\circ$ to be in the interior of a compact parameter space, and the function $g^A(x,\psi)$ to be sufficiently smooth and uniformly bounded as a function of ψ. These assumptions are not new in the literature and are also required by Schennach (2007, Theorem 10) (adapted to account for the augmented model).

Assumption 7. *(a) there exists a function $M(\cdot)$ such that $\mathbf{E}^P[M(X)] < \infty$ and $\|g^A(x,\psi)\| \leq M(x)$ for all $\psi \in \Psi$; (b) $\lambda_\circ(\psi) \in int(\Lambda(\psi))$ where $\Lambda(\psi)$ is a compact set; (c) it holds $\mathbf{E}^P\left[\sup_{\psi\in\Psi,\lambda\in\Lambda(\psi)} e^{\{\lambda' g^A(X,\psi)\}}\right] < \infty$.*

Assumption 8. *(a) the pseudo-true value $\psi_\circ \in int(\Psi)$ is the unique maximizer of*

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

where Ψ is compact; (b) $S_{jl}(x_i,\psi) := \partial^2 g^A(x_i,\psi)/\partial\psi_j\partial\psi_l$ is continuous in ψ for $\psi \in \mathcal{U}_\circ$, where $\mathcal{U}_\circ$ denotes a ball centred at $\psi_\circ$ with radius $n^{-1/2}$; (c) there exists $b(x_i)$ satisfying $\mathbf{E}^P\left[\sup_{\psi\in\mathcal{U}_\circ}\sup_{\lambda\in\Lambda(\psi)}\exp\{\kappa_1\lambda' g^A(X,\psi)\}b(X)^{\kappa_2}\right] < \infty$ for $\kappa_1 = 0,1,2$ and $\kappa_2 = 0,1,2,3,4$ such that $\|g^A(x_i,\psi)\| < b(x_i)$, $\|\partial g^A(x_i,\psi)/\partial\psi'\|_F \leq b(x_i)$ and $\|S_{jl}(x_i,\psi)\| \leq b(x_i)$ for $j,l = 1,\ldots,p$ for any $x_i \in (\mathcal{X},\mathfrak{B}_\mathcal{X})$ and for all $\psi \in \mathcal{U}_\circ$.

A first step to establish the BvM theorem is to prove that the misspecified model satisfies a stochastic Local Asymptotic Normality (LAN) expansion around the pseudo-true value

$\psi_\circ$. Namely, that the log-likelihood ratio $l_{n,\psi} - l_{n,\psi_\circ}$, evaluated at a local parameter around the pseudo-true value, is well approximated by a quadratic form. Such a result is established in Theorem A.1 in the Appendix. The limit of the posterior distribution of $\sqrt{n}(\psi - \psi_\circ)$ is a Gaussian distribution with mean and variance defined in terms of the population counterpart of $l_{n,\psi}(x)$, which we denote by $\mathfrak{L}_{n,\psi}(x) := \log \frac{\exp(\lambda_\circ(\psi)' g^A(x,\psi))}{\mathbf{E}^P[\exp(\lambda_\circ(\psi)' g^A(x,\psi))]} - \log n$ and which involves the pseudo-true value $\lambda_\circ$. With this notation, the variance and mean of the Gaussian limiting distribution are $V_{\psi_\circ}^{-1} := -(\mathbf{E}^P[\ddot{\mathfrak{L}}_{n,\psi_\circ}])^{-1}$ and $\Delta_{n,\psi_\circ} := \frac{1}{\sqrt{n}} \sum_{i=1}^n V_{\psi_\circ}^{-1} \dot{\mathfrak{L}}_{n,\psi_\circ}(x_i)$, respectively, where $\dot{\mathfrak{L}}_{n,\psi_\circ}$ and $\ddot{\mathfrak{L}}_{n,\psi_\circ}$ denote the first and second derivatives of the function $\psi \mapsto \mathfrak{L}_{n,\psi}$ evaluated at $\psi_\circ$.

A second key ingredient for establishing the BvM theorem is the requirement that, as $n \to \infty$, the posterior of ψ concentrates and puts all its mass on $\Psi_n := \{\|\psi - \psi_\circ\| \leq M_n/\sqrt{n}\}$, where M_n is any sequence such that $M_n \to \infty$. We prove this result in Theorem A.2 in the Appendix. Here, we state the BvM theorem. Let $\pi(\sqrt{n}(\psi - \psi_\circ)|x_{1:n})$ denote the posterior distribution of $\sqrt{n}(\psi - \psi_\circ)$.

Theorem 2.2 (Bernstein - von Mises – misspecification)**.** *Assume that the matrix $V_{\psi_\circ}$ is nonsingular and that Assumptions 2 (a)-(d), 3 (b), 4 (a), 6, 5, 7, and 8 hold. If in addition there exists a constant $C > 0$ such that for any sequence $M_n \to \infty$, as $n \to \infty$*

$$P\left(\sup_{\psi \in \Psi_n^c} \frac{1}{n} \sum_{i=1}^n (l_{n,\psi}(x_i) - l_{n,\psi_\circ}(x_i)) \leq -\frac{CM_n^2}{n}\right) \to 1, \tag{2.14}$$

then the posteriors converge in total variation towards a normal distribution, that is,

$$\sup_B \left|\pi(\sqrt{n}(\psi - \psi_\circ) \in B|x_{1:n}) - \mathcal{N}_{\Delta_{n,\psi_\circ}, V_{\psi_\circ}^{-1}}(B)\right| \xrightarrow{p} 0 \tag{2.15}$$

where $B \subseteq \Psi$ is any Borel set.

Condition (2.14) involves the log-likelihood ratio $l_{n,\psi}(x) - l_{n,\psi_\circ}(x)$ and is an identifiability condition, standard in the literature, and with a similar interpretation as condition (2.11).

Theorem 2.2 states that, in misspecified models, the sequence of posterior distributions converges in total variation to a sequence of normal distributions with random mean and fixed covariance matrix $V_{\psi_\circ}^{-1}$. Unlike Theorem 2.1 for correctly specified models, in Theorem 2.2 the centering $\Delta_{n,\psi_\circ}$ of the limiting distribution is in general non-zero since $\lambda_\circ \neq 0$. We stress that the BvM result of Theorem 2.2 for the BETEL posterior distribution does not directly follow from results in Kleijn and van der Vaart (2012) because the ETEL function contains random quantities.

As the next lemma shows, the quantity $\Delta_{n,\psi_\circ}$ relates to the Schennach (2007)'s ETEL frequentist estimator $\widehat{\psi}$ (whose definition is recalled in (A.1) in the Appendix for convenience). Because of this connection, it is possible to write the location of the normal limit distribution in a more familiar form in terms of the semi-parametric efficient frequentist estimator $\widehat{\psi}$.

Lemma 2.1. *Assume that the matrix $V_{\psi_\circ}$ is nonsingular and that Assumptions 2 (a)-(d), 3 (b), 5, 7, and 8 hold. Then, the ETEL estimator $\widehat{\psi}$ satisfies*

$$\sqrt{n}(\widehat{\psi} - \psi_\circ) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_{\psi_\circ}^{-1} \dot{\mathfrak{L}}_{n,\psi_\circ} + o_p(1). \tag{2.16}$$

Therefore, Lemma 2.1 implies that the BvM theorem 2.2 can be reformulated with the sequence $\sqrt{n}(\widehat{\psi} - \psi_\circ)$ as the location for the normal limit distribution, that is,

$$\sup_B \left| \pi(\psi \in B | x_{1:n}) - \mathcal{N}_{\widehat{\psi}, n^{-1} V_{\psi_\circ}^{-1}}(B) \right| \xrightarrow{p} 0. \tag{2.17}$$

Two remarks are in order: (I) the limit distribution of $\sqrt{n}(\widehat{\psi} - \psi_\circ)$ is centred on zero because $\mathbf{E}^P[\dot{\mathfrak{L}}_{n,\psi_\circ}] \xrightarrow{p} 0$ at the rate $n^{-1/2}$; (II) the asymptotic covariance matrix of $\sqrt{n}(\widehat{\psi} - \psi_\circ)$ is $V_{\psi_\circ}^{-1} \mathbf{E}^P[\dot{\mathfrak{L}}_{n,\psi_\circ} \dot{\mathfrak{L}}'_{n,\psi_\circ}] V_{\psi_\circ}^{-1}$ (which is also derived in Schennach (2007, Theorem 10)) and, because of misspecification, it does not coincide with the limiting covariance matrix in the BvM theorem. This consequence of misspecification is also discussed in Kleijn and van der Vaart (2012).

Example (Misspecified model and pseudo-true value). We again consider the model described in (2.3) without predictors (i.e., $\beta_1 = 0$ and $\beta_2 = 0$). Suppose that the distribution of the e_i is skewed as in (2.10). In this example, we consider the following two moment conditions $\mathbf{E}^P[(y_i - \mu)] = 0$ and $\mathbf{E}^P[(y_i - \mu)^3] = 0$. This example is different from the previous one in Figure 1 in that these two moment restrictions form a misspecified model because here the augmented parameter v is forced to be zero. In turn, μ has to satisfy both the moment restrictions, which is impossible under P. Instead, for each μ the ETEL likelihood function is defined by the probability measure $Q^*(\mu)$ that is the closest to the true generating process P in terms of KL divergence among the pairs (Q, μ) that are consistent with the given moment restrictions. In Figure 2 (left panel), we present $\mathbf{E}^P[\log(dQ^*(\mu)/dP)]$. The value that maximizes this function is different from the true value ($\mu = 0$) and it is peaked around -0.28. This value is the pseudo-true value, $\mu_\circ$. In the right panel of Figure 2, we present the BETEL posterior distribution with $n = 200, 400, 1000$. Unlike the correctly specified case in Figure 1, the BETEL posterior distribution shrinks toward the pseudo-true value, in conformity with our theoretical result.

3 Bayesian Model Selection

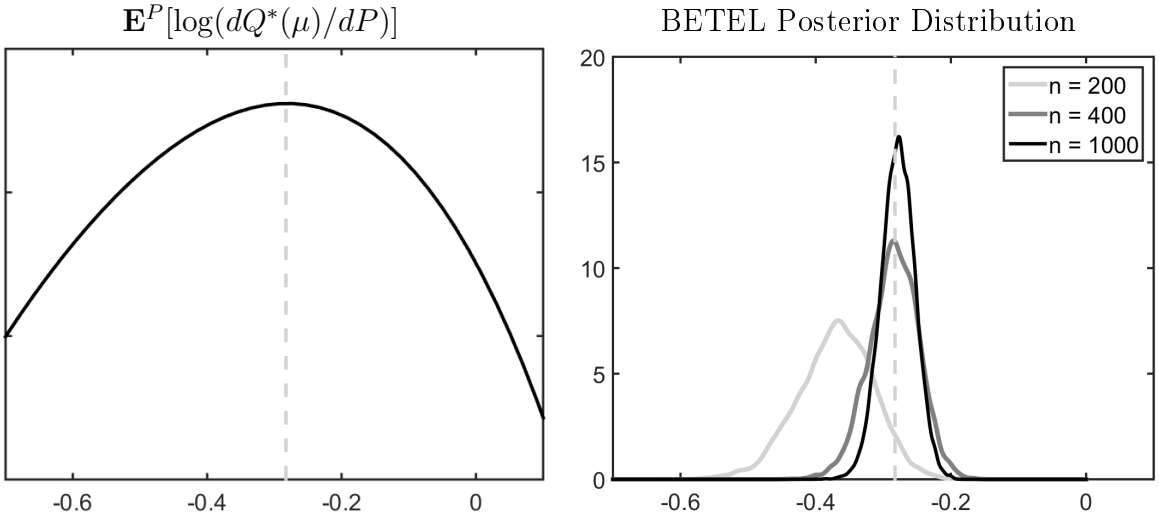
3.1 Basic idea

Now suppose that there are candidate models indexed by ℓ. Suppose that model ℓ is characterized by

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

with $\theta^\ell \in \Theta^\ell \subset \mathbb{R}^{p_\ell}$. Different models involve different parameters of interest θ^ℓ and/or different g^ℓ functions. To make these models all comparable, we need a grand model that nests all the models that we want to compare. The grand model is constructed such that: (1) it includes all the moment restrictions in the models and, (2) if the same moment restriction

Figure 2: BETEL Posterior Distribution under Misspecification



Notes: Left panel presents the $\mathbf{E}^P[\log(dQ^*(\mu)/dP)]$ where $Q^*(\mu)$ is defined as $Q^*(\mu) := \text{arginf}_{Q \in \mathcal{P}_\psi} K(Q||P)$ with $\psi := (\mu, 0)$. For each μ, we approximate this function based on the dual representation in (2.13) using one million simulation draws from P. In the right panel, we present the BETEL posterior distribution of the location parameter μ with $n = 200, 400, 1000$ where n is the number of observations. The prior distribution for μ is set to be a normal distribution with mean 0 and variance 10. Vertical dashed lines indicate the pseudo-true parameter value, $\mu_\circ \approx -0.28$. We generate 25,000 posterior draws using the one block tailored Metropolis-Hastings algorithm described in Section 2.1. Our proposal density is set to be a t-distribution with mean as the posterior mode, variance as the 1.5 times negative inverse Hessian of the log-BETEL posterior at the posterior mode, 15 as the degrees of freedom. The rejection probabilities are about 44% for all cases.

is included in two or more models but involves a different parameter in different models, then the grand model includes the moment restriction that involves the largest parameter. We write the grand model as $\mathbf{E}^P[g^G(X, \theta^G)] = 0$ where g^G has dimension d, then each model can be obtained from this grand model by first subtracting a vector of nuisance parameters V and then by restricting θ^G and V. More precisely, a model can be obtained by setting equal to zero the components of θ^G that are not present in the original model, by letting free the components of V that correspond to the moment restrictions not present in the original model and by setting equal to zero the components of V that correspond to moment restrictions present in the original model. With this formulation, model ℓ, denoted by M_ℓ, is then defined as

$$E^P[g^A(X, \theta^\ell, v^\ell)] = 0, \qquad \theta^\ell \in \Theta^\ell \subset \mathbb{R}^{p_\ell} \tag{3.2}$$

where $g^A(X, \theta^\ell, v^\ell) = g^G(X, \theta^\ell) - V^\ell$ with $V^\ell \in \mathfrak{V} \subset \mathbb{R}^d$ and with $v^\ell \in \mathcal{V}^\ell \subset \mathbb{R}^{d_{v_\ell}}$ being the vector that collects all the non-zero components of V^ℓ. We assume that $0 \leq d_{v_\ell} \leq d - p_\ell$ in order to guarantee identification of θ^ℓ. The parameter v^ℓ is the augmented parameter and θ^ℓ is the parameter of interest for model ℓ. In the following we use the notation $\psi^\ell := (\theta^\ell, v^\ell) \in \Psi^\ell$ with $\Psi^\ell := \Theta \times \mathcal{V}^\ell$.

For expositional simplicity, we suppose in the following of this section that there are two models M_1 and M_2 and denote by $B_{12} := m(x_{1:n}; M_1)/m(x_{1:n}; M_2)$ the Bayes factor for their comparison. If there are more than two models, we can do pairwise comparison. In practice, researchers may want to select one of the two models and they do not know whether the models are misspecified. We base the model selection procedure on Marginal Likelihood (ML) and select the model with the larger ML. The reason why we need a grand model that nests M_1 and M_2 in order to be able to make model selection is that MLs of two different models with different sets of moment restrictions and different parameters may not be comparable. In fact, when we have different sets of moment restrictions, we need to be careful about dealing and interpreting unused moment restrictions. This can be best explained by an example.

Example (Linear regression model, continued). Consider again the linear regression model example from the previous section. Suppose we do not know whether e_i is symmetric or not. In this case, one might tempt to compare the following two candidate models:

$$\begin{aligned} &\text{Model 1} : \mathbf{E}^P[e_i(\theta)] = 0, \quad \mathbf{E}^P[e_i(\theta) z_{1,i}] = 0, \quad \mathbf{E}^P[e_i(\theta) z_{2,i}] = 0. \\ &\text{Model 2} : \mathbf{E}^P[e_i(\theta)] = 0, \quad \mathbf{E}^P[e_i(\theta) z_{1,i}] = 0, \quad \mathbf{E}^P[e_i(\theta) z_{2,i}] = 0, \quad \mathbf{E}^P[(e_i(\theta))^3] = 0. \end{aligned} \tag{3.3}$$

where $\theta = (\mu, \beta_1, \beta_2)$ and $e_i(\theta) = (y_i - \mu - \beta_1 z_{1,i} - \beta_2 z_{2,i})$. It turns out that the MLs from Model 1 and Model 2 are not comparable. This is because Model 1 completely ignores uncertainty coming from the fourth moment restriction while Model 2 puts strong confidence about the fourth moment restriction. Therefore, one has to define the grand model

$g^G(x_i, \theta) := (e_i(\theta), e_i(\theta)z_{1,i}, e_i(\theta)z_{2,i}, e_i(\theta)^3)'$ and its augmented version. With respect to this augmented grand model, Model 1 and Model 2 write as M_1 and M_2, respectively

$$\begin{aligned} M_1 &: \mathbf{E}^P[e_i(\theta)] = 0, \quad \mathbf{E}^P[e_i(\theta)z_{1,i}] = 0, \quad \mathbf{E}^P[e_i(\theta)z_{2,i}] = 0, \quad \mathbf{E}^P[(e_i(\theta))^3] - v = 0 \\ M_2 &: \mathbf{E}^P[e_i(\theta)] = 0, \quad \mathbf{E}^P[e_i(\theta)z_{1,i}] = 0, \quad \mathbf{E}^P[e_i(\theta)z_{2,i}] = 0, \quad \mathbf{E}^P[(e_i(\theta))^3] = 0. \end{aligned} \tag{3.4}$$

It is important to realize how Model 1 in (3.3) and M_1 deal with uncertainty about the fourth moment restriction: Model 1 in (3.3) ignores its uncertainty completely while M_1 models the degree of uncertainty through the augmented parameter v.

In what follows, we show how to construct and compute the ML for a model. Then, in Section 3.3 we formally show that, with probability approaching one as the number of observation increases, the ML-based selection procedure favors the model with the minimum number of parameters of interest and the maximum number of valid moment restrictions. More importantly, we consider the situation where both models are misspecified. In this case, our model selection procedure selects the model that is closer to the true data generating process in terms of KL-divergence.

3.2 Marginal Likelihood (ML)

For each model M_ℓ, we impose a prior distribution for ψ^ℓ on Ψ^ℓ, and obtain the BETEL posterior distribution based on (2.9). Then, we select the model with the largest ML. We compute the ML by the method of Chib (1995) as extended to Metropolis-Hastings samplers in Chib and Jeliazkov (2001). This method makes computation of the ML extremely simple and is a key feature of our procedure. The main advantage of the Chib (1995) method is that it is calculable from the same inputs and outputs that are used in the MCMC sampling of the posterior distribution. The starting point of this method is the following identity of the log-ML introduced in Chib (1995)

$$\log m(x_{1:n}|M_\ell) = \log \pi(\tilde{\psi}^\ell|M_\ell) + \log p(x_{1:n}|\tilde{\psi}^\ell, M_\ell) - \log \pi(\tilde{\psi}^\ell|x_{1:n}, M_\ell), \tag{3.5}$$

where $\tilde{\psi}^\ell$ is any point in the support of the posterior (such as the posterior mean) and the dependence on the model M_ℓ has been made explicit. The first two terms on the right-hand side of this decomposition are available directly whereas the third term can be estimated from the output of the MCMC simulation of the BETEL posterior distribution. For example, in the context of the one block MCMC algorithm given above, from Chib and Jeliazkov (2001), we have that

$$\pi(\tilde{\psi}^\ell|x_{1:n}, M_\ell) = \frac{\mathbf{E}_1\left\{\alpha\left(\psi^\ell, \tilde{\psi}^\ell|x_{1:n}, M_\ell\right) q(\tilde{\psi}^\ell)x_{1:n}, M_\ell)\right\}}{\mathbf{E}_2\left\{\alpha(\tilde{\psi}^\ell, \psi^\ell|x_{1:n}, M_\ell)\right\}}$$

where $\mathbf{E}_1$ is the expectation with respect to $\pi(\psi^\ell|x_{1:n}, M_\ell)$ and $\mathbf{E}_2$ is the expectation with respect to $q(\psi^\ell|x_{1:n}, M_\ell)$. These expectations can be easily approximated by simulations.

3.3 Consistency of the ML-based selection procedure

In this section we establish consistency of our ML-based selection procedure for three cases: the case where the models that we compare contain only valid moment restrictions, the case where one model contains only valid moment restrictions and the other one contains at least one invalid moment restriction, and the case where both the models are misspecified. Our proofs of consistency are based on: (I) the results of the BvM theorems for correctly and misspecified models stated in Sections 2.2 and 2.3, and (II) the asymptotic analysis of the behavior of the ETEL function under correct and misspecification which we develop in the Appendix (see Lemmas B.1 and B.3).

The first theorem states that, if the active moment restrictions are all valid, then the ML selects the model that contains the maximum number of overidentifying conditions, that is, the model with the maximum number of active moment restrictions and the smallest number of parameters of interest. For a model M_ℓ, the dimension of the parameter of interest θ^ℓ to be estimated is p_ℓ while the number of active moment restrictions (included in the model for the estimation of θ^ℓ) is $(d - d_{v_\ell})$.

Consider two generic models M_1 and M_2. Then, $d_{v_2} < d_{v_1}$ means that model M_2 contains

more active restrictions than model M_1, and $p_2 < p_1$ means that model M_1 contains more parameters of interest to be estimated than M_2.

Theorem 3.1. *Let Assumptions 2 – 4 and (2.11) hold, and consider two different models M_1 and M_2 that both satisfy Assumption 1, that is, they are both correctly specified. Then, if $p_2 + d_{v_2} < p_1 + d_{v_1}$:*

$$\lim_{n\to\infty} P\left(\log m(x_{1:n}; M_1) < \log m(x_{1:n}; M_2)\right) = 1.$$

The result of the theorem implies that $B_{12} < 1$ with probability approaching 1.

Example (Model selection when models are correctly specified). As in the previous example, we generate $(y_1, y_2, ..., y_n)$ from the model described in (2.3) without predictors (i.e., $\beta_1 = 0$ and $\beta_2 = 0$). Suppose that e_i is generated from the standard normal distribution and we compare the following two models:

$$\begin{aligned} M_1 &: \mathbf{E}^P[e_i(\theta)] = 0 \quad \text{and} \quad \mathbf{E}^P[(e_i(\theta))^3] = v \\ M_2 &: \mathbf{E}^P[e_i(\theta)] = 0 \quad \text{and} \quad \mathbf{E}^P[(e_i(\theta))^3] = 0. \end{aligned} \tag{3.6}$$

where $\theta = (\mu, 0, 0)$ and $e_i(\theta) = (y_i - \mu)$. Under the standard normal distribution, both models are correctly specified. M_1 has one active moment restriction while M_2 has two active moment restrictions. In Table 1, we report the percentage of times that the ML selects each of the correctly specified model M_1 and M_2 out of 500 trials. Model M_2, the model with the larger number of valid restrictions, is selected 99% times by sample size of $n = 1000$ and 2000.

Next, suppose that some of the models that we consider are misspecified in the sense of Definition 2.1. This means that one or more of the active moment restrictions are invalid, or in other words, that one or more components of V are incorrectly set equal to zero. Indeed, all the models for which the active moment restrictions are valid are not misspecified even

Table 1: Model selection among valid models

Model	M_1	M_2
$n = 250$	3	97
$n = 500$	1.6	98.4
$n = 1000$	1	99
$n = 2000$	1	99

Note: This table presents the frequency (%) of the corresponding model selected by the model selection criteria out of 500 trials. For each case, we compute the ML by the method of Chib (1995) as described in Section 3.2. Other computational details can be found from the note under Figure 1 and 2.

if some invalid moment restrictions are included among the inactive moment restrictions. This is because there always exists a parameter $v \in \mathbb{R}^{d_{v_\ell}}$ that equates the invalid moment restriction. In this case, the true v_* for this model will be different from the zero vector: $v_* \neq 0$ and the true value of the corresponding tilting parameter λ will be zero.

The following theorem establishes that the ML selection criterion does not select models that contain misspecified moment restrictions with probability approaching one. As for Theorem 3.1, the results of the next two theorems are presented for two generic models M_1 and M_2 where M_1 does not use misspecified moments while M_2 does.

Theorem 3.2. *Let Assumptions 2 - 8, (2.11) and (2.14) be satisfied. Let us consider two different models M_1 and M_2 where M_1 satisfies Assumption 1 whereas M_2 does not. Then,*

$$\lim_{n\to\infty} P\left(\log m(x_{1:n}; M_1) > \log m(x_{1:n}; M_2)\right) = 1.$$

The result of the theorem implies that $B_{12} > 1$ with probability approaching 1.

Example (Model selection when one of the models is misspecified). We consider the same setup as in the previous example, but e_i is generated from the following skewed distribution

$$e_i \sim \begin{cases} N(1/2, 0.5^2) & \text{with probability } 0.5 \\ N(-1/2, 1.118^2) & \text{with probability } 0.5. \end{cases} \tag{3.7}$$

Parameters in this mixture distribution are chosen so that e_i has mean 0 and variance 1. We compare two models defined in (3.6). Under the skewed distribution, M_2 becomes a misspecified model because the third moment cannot be zero. M_1 remains correctly specified as the moment restrictions do not restrict the skewness of the underlying distribution. In Table 2, we report the percentage of times that the ML selects each model out of 500 trials. As we can see, the frequency of selecting the correctly specified model over the misspecified model approaches 100% as the number of observation increases.

Table 2: Model selection when one of the models is misspecified

Model	M_1	M_2
$n = 250$	95	5
$n = 500$	99.2	0.8
$n = 1000$	100	0
$n = 2000$	100	0

Note: This table presents the frequency (%) of the corresponding model selected by the ML-model selection criterion out of 500 trials. For each case, we compute the ML by the method of Chib (1995) as described in Section 3.2. Other computational details can be found from the note under Figure 1 and 2.

Finally, we consider the case where all models are wrong in the sense of Definition 2.1. The next theorem establishes that if we compare two misspecified models, then the ML-based selection procedure selects the model with the smallest KL divergence between P and $Q^*(\psi^\ell)$, where $dQ^*(\psi^\ell)/dP = \arg\inf_{Q\in\mathcal{P}_{\psi^\ell}} K(Q||P) = e^{\lambda_\circ(\psi)' g^A(X,\psi)}/\mathbf{E}^P\left[e^{\lambda_\circ(\psi)' g^A(X,\psi)}\right]$ with the second equality holding by the dual theorem, as defined in Section 2.3. Because the projection $Q^*(\psi^\ell)$ on $\mathcal{P}_{\psi^\ell}$ is unique (Csiszar (1975)), which $Q^*(\psi^\ell)$ is closer to P depends only on the "amount of misspecification" contained in each model $\mathcal{P}_{\psi^\ell}$.

Theorem 3.3. *Let Assumptions 2 - 8 and (2.14) be satisfied. Let us consider two different models M_1 and M_2 that both use misspecified moments, that is, neither M_1 nor M_2 satisfy Assumption 1. If $K(P||Q^*(\psi^1)) < K(P||Q^*(\psi^2))$, where $K(P||Q) := \int \log(dP/dQ)dP$, then*

$$\lim_{n\to\infty} P\left(\log m(x_{1:n}; M_1) > \log m(x_{1:n}; M_2)\right) = 1.$$

Remark that the condition $K(P||Q^*(\psi^1)) < K(P||Q^*(\psi^2))$ given in the theorem does not depend on a particular value of ψ^1 and ψ^2. Indeed, the result of the theorem hinges on the fact that ML selects the model with the $Q^*(\psi^\ell)$ the closer to P, that is, the model that contains the "less misspecified" moment restrictions for every value of ψ^ℓ.

Example (Model selection when both models are misspecified). We consider the same setup as in the previous example with e_i being generated from the skewed distribution with mean zero and standard deviation 1. In this example, we compare the following two models:

$$\begin{aligned} M_3 &: \mathbf{E}^P[e_i(\theta)] = 0 \quad \text{and} \quad \mathbf{E}^P[(e_i(\theta))^3] = v \quad \text{and} \quad \mathbf{E}^P[(e_i(\theta))^2 - 2] = 0 \\ M_4 &: \mathbf{E}^P[e_i(\theta)] = 0 \quad \text{and} \quad \mathbf{E}^P[(e_i(\theta))^3] = 0 \quad \text{and} \quad \mathbf{E}^P[(e_i(\theta))^2 - 2] = 0. \end{aligned} \tag{3.8}$$

Thus, in this example, we introduce an additional moment restriction that governs the variance of the distribution. When the underlying distribution has variance 1, both M_3 and M_4 are misspecified due to the new moment restriction: $\mathbf{E}^P[(e_i(\theta))^2 - 2] = 0$. Because we know the true generating data process, we can compute the KL divergence from P to $Q^*(\psi_\circ)$ as well as the pseudo-true. Using the 10,000,000 simulated draws from P, we approximate the $K(P|Q^*(\psi_\circ))$ for each models. It turns out that M_3 is closer to the data generating process in terms of the KL divergence (0.056 for M_3 and 0.073 for M_4). In Table 3, we report the percentage of times that the ML selects each model out of 500 trials. The frequency of selecting M_3 over M_4 is seen to increase toward 100%, in conformity with the stated result.

4 Poisson Regression

The techniques discussed in the previous sections have wide-ranging applications to various statistical settings, such as generalized linear models, and to many different fields of applications, such as biostatistics and economics. In fact, the methods discussed above can be applied to virtually any problem that, in the frequentist setting, would be approached by

Table 3: Model selection when one of the models is misspecified

Model	M_3	M_4
$n = 250$	87.2	12.8
$n = 500$	88.6	11.4
$n = 1000$	92.4	7.6
$n = 2000$	92.2	7.8

Note: This table presents the frequency (%) of the corresponding model selected by the model selection criteria out of 500 trials. For each case, we compute the ML by the method of Chib (1995) as described in Section 3.2. Other computational details can be found from the note under Figure 1 and 2.

generalized method of moments or estimating equation techniques. To illustrate some of the possibilities, we consider two important problems in the context of Poisson regression that hitherto could not have been handled similarly from the Bayesian perspective.

4.1 Variable selection

Consider the poisson regression model

$$\begin{aligned} y_i|\beta, x_i &\sim Poisson(\lambda_i) \\ \log(\lambda_i) &= \beta' x_i. \end{aligned} \tag{4.1}$$

where $\beta = [\beta_1, \beta_2, \beta_3]'$ and $x_i = [x_{i,1}, x_{i,2}, x_{i,3}]'$. In this setting, suppose we wish to learn about β under the moment conditions

$$\begin{aligned} E\left[\left(y_i - \exp(\beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i})\right) x_i\right] &= 0 \\ E\left[\left(\frac{y_i - \exp(\beta' x_i)}{\sqrt{\exp(\beta' x_i)}}\right)^2 - 1\right] &= v. \end{aligned} \tag{4.2}$$

The first type of moment restriction (one for each $x_{j,i}$ for $j = 1, 2, 3$) is derived from the fact that the conditional expectation of y_i is $\exp(\beta' x_i)$ and this identifies β. The second type of restriction is an overidentifying restriction that is related to the Poisson assumption. More specifically, if $v = 0$, that moment condition asserts that the conditional variance of y_i is

equal to the conditional mean. In general, the Poisson assumption can be questioned by supposing that $v \neq 0$.

Suppose that we are interested in excluding the one redundant regressor from the model. To solve this problem, one can create the following two models based on (4.2) with the following restrictions:

$$\begin{aligned} &M_1 : \beta_1 \text{ and } \beta_2 \text{ are free parameters but } \beta_3 = 0 \text{ and } v = 0. \\ &M_2 : \beta_1,\ \beta_2,\ \beta_3 \text{ are free parameters but } v = 0. \end{aligned} \tag{4.3}$$

Note that both models have the same number of active moment restrictions, but they differ in that β_3 is forced to be zero in M_1.

In this subsection, we generate n realizations of $\{y_i, x_i\}$ from the above model with $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 0$. Thus, $x_{i,3}$ is a redundant regressor. Each explanatory variable $x_{i,j}$ is generated *i.i.d.* from normal distributions with mean zero and standard deviation $1/3$. The prior distribution of β_j's is an independent normal distribution with mean 0 and variance 10. We compute the ML's of M_1 and M_2 and select the model with the higher ML. We repeat this exercise 500 times for samples of sizes $n = 250, 500, 1000$. In Table 4, we report the percentage of times that the ML criterion picks M_1 and M_2. As can be seen, M_1 is selected by the ML criterion with frequency approaching one.

Table 4: Variable selection in Poisson regression

Model	M_1	M_2
$n = 250$	97.2	2.8
$n = 500$	98	2
$n = 1000$	99.4	0.6

Note: This table presents the percentage of times each model is selected by the ML criterion in 500 trials. The ML is computed by the method of Chib (1995) as described in Section 3.2. Other computational details can be found from the note under Figure 1 and 2.

4.2 Distributional specification

Another interesting question relates to performance of the ML criterion in differentiating the Poisson model from another other distributions such as the negative Binomial distribution.

Case 1 (DGP is Poisson). As in previous example, suppose that the true DGP corresponds to the Poisson distribution but one considers two models based on (4.2) with the following restrictions:

$$\begin{aligned} &M_3 : v = 0 \\ &M_4 : v \text{ is a free parameter} \end{aligned} \tag{4.4}$$

where β_1, β_2, and β_3 are treated as free parameters. In addition, because the augmented parameter v is free, M_4 allows for the possibility that the underlying distribution has variance different from its mean. Suppose that the prior distribution for v is a normal distribution with mean zero and variance 10. Also suppose that the prior of β in M_3 and M_4 is as in the previous experiment.

In Table 5, we report the percentage of times that the ML criterion selects M_3 and M_4 in 500 trials. It is seen that M_3 is selected more frequently and that this frequency increases with n. This is in conformity with our theory result because under the assumed Poisson DGP, model M_3 involves an additional valid moment restriction. As our theory suggests, the ML criterion selects the model with larger valid restrictions.

Case 2 (DGP is Negative Binomial). Now suppose that y_i is generated from the negative binomial distribution

$$\begin{aligned} &y_i|\beta, x_i \sim NB\left(\frac{p}{1-p}\lambda_i,\ p\right) \\ &\log(\lambda_i) = \beta' x_i. \end{aligned} \tag{4.5}$$

Table 5: Model selection when the DGP is Poisson distribution

Model	M_3	M_4
$n = 250$	97	3
$n = 500$	98.6	1.4
$n = 1000$	99.4	0.6

Note: This table presents the percentage of times each model is selected by the ML criterion in 500 trials. We compute the ML by the method of Chib (1995) as described in Section 3.2. Other computational details can be found from the note under Figure 1 and 2.

where NB denotes the negative binomial distribution and its parameters are chosen so that

$$\begin{aligned} E[y_i|\beta, x_i] &= \lambda_i \\ Var(y_i|\beta, x_i) &= \frac{1}{p}\lambda_i. \end{aligned} \tag{4.6}$$

In this formulation, the last moment restriction in (4.2) is invalid and the assertion that $v = 0$ makes M_3 misspecified as long as $p \neq 1$. For our experiment, we set $p = 1/2$ and compare the performance of the ML criterion in selecting M_3 and M_4.

In Table 6, we report the percentage of times that the ML criterion selects M_3 and M_4 in 500 trials. This time, as can be seen, M_4 is selected more frequently over M_3, the misspecified model, and this frequency increases with n in keeping with our theoretical results.

Table 6: Model selection when the DGP is Negative Binomial

Model	M_3	M_4
$n = 250$	2	98
$n = 500$	0	100
$n = 1000$	0	100

Note: This table presents the percentage of times each model is selected by the ML criterion in 500 trials. we compute the ML by the method of Chib (1995) as described in Section 3.2. Other computational details can be found from the note under Figure 1 and 2.

5 Conclusion

In this paper we have developed a fully Bayesian framework for estimation and model comparisons in statistical models that are defined by moment restrictions. The Bayesian analysis of such models has always been viewed as a challenge because traditional Bayesian semiparametric methods, such as those based on Dirichlet process mixtures and variants thereof, are not suitable for such models. What we have shown in this paper is that the Exponentially Tilted Empirical Likelihood (ETEL) framework is an immensely useful organizing framework within which a fully Bayesian treatment of such models can be developed. We have established a number of new, powerful results surrounding the Bayesian ETEL framework including the treatment of models that are possibly misspecified. We show how the moment conditions can be reexpressed in terms of additional nuisance parameters and that the Bayesian ETEL posterior distribution satisfies a Bernstein-von Mises (BvM) theorem. We have also developed a framework for comparing moment condition models based on marginal likelihoods (MLs) and Bayes factors and provided a suitable large sample theory for Bayes factor consistency. Our results show that the ML favors the model with the minimum number of parameters and the maximum number of valid moment restrictions that are relevant. When the models are misspecified, the ML model selection procedure selects the model that is closer to the (unknown) true data generating process in terms of the Kullback-Leibler divergence. The ideas and results illumined in this paper now provide the means for analyzing a whole array of models from the Bayesian viewpoint. This broadening of the scope of Bayesian techniques to previously intractable problems is likely to have far-reaching practical consequences.

Appendix

A Proofs for Sections 2.2 and 2.3

In this appendix we prove Theorem 2.2 and Lemma 2.1. Theorem 2.1 is proved in the Supplementary Appendix. It is useful to introduce some notation that will be used in this section. The estimator $\widehat{\psi} := (\widehat{\theta}, \widehat{v})$ denotes Schennach (2007)'ETEL estimator of ψ:

$$\widehat{\psi} := \arg\max_{\psi \in \Psi} \frac{1}{n} \sum_{i=1}^{n} \left[\widehat{\lambda}(\psi)' g^A(x_i, \psi) - log \frac{1}{n} \sum_{j=1}^{n} \exp\{\widehat{\lambda}(\psi)' g^A(x_j, \psi)\} \right] \tag{A.1}$$

where $\widehat{\lambda}(\psi) = \arg\min_\lambda \frac{1}{n} \sum_{i=1}^{n} \left[\exp\{\lambda' g^A(x_i, \psi)\}\right]$. The log-likelihood ratio is:

$$l_{n,\psi}(x) - l_{n,\psi_\circ}(x) \quad = \quad \log \frac{e^{\widehat{\lambda}(\psi)' g^A(x,\psi)}}{\frac{1}{n}\sum_{j=1}^{n} \left[e^{\widehat{\lambda}(\psi)' g^A(x_j,\psi)} \right]} - \log \frac{e^{\widehat{\lambda}(\psi_\circ)' g^A(x,\psi_\circ)}}{\frac{1}{n}\sum_{j=1}^{n} \left[e^{\widehat{\lambda}(\psi_\circ)' g^A(x_j,\psi_\circ)} \right]}. \tag{A.2}$$

A.1 Proof of Theorem 2.2.

The main steps of this proof proceed as in the proof of Van der Vaart (2000, Theorem 10.1) and Kleijn and van der Vaart (2012, Theorem 2.1) while the proofs of the technical theorems and lemmas that we use all along this proof are new. Let us consider a reparametrization of the model centred around the pseudo-true value $\psi_\circ$ and define a local parameter $h = \sqrt{n}(\psi - \psi_\circ)$. Denote by π^h and $\pi^h(\cdot|x_{1:n})$ the prior and posterior distribution of h, respectively. Denote by Φ_n the normal distribution $\mathcal{N}_{\Delta_{n,\psi_\circ}, V_{\psi_\circ}^{-1}}$ and by ϕ_n its Lebesgue density. For a compact subset $K \subset \mathbb{R}^p$ such that $\pi^h(h \in K|x_{1:n}) > 0$ define, for any Borel set $B \subseteq \Psi$,

$$\pi_K^h(B|x_{1:n}) := \frac{\pi^h(K \cap B|x_{1:n})}{\pi^h(K|x_{1:n})}$$

and let Φ_n^K be the Φ_n distribution conditional on K. The proof consists of two steps. In the first step we show that the Total Variation (TV) norm of $\pi_K^h(\cdot|x_{1:n}) - \Phi_n^K$ converges to zero in probability. In the second step we show that the TV norm of $\pi^h(\cdot|x_{1:n}) - \Phi_n$ converges to

zero in probability.

Let Assumption 8 *(a)* hold. For every open neighborhood $\mathcal{U} \subset \Psi$ of ψ° and a compact subset $K \subset \mathbb{R}^p$, there exists an N such that for every $n \geq N$:

$$\psi^\circ + K\frac{1}{\sqrt{n}} \subset \mathcal{U}. \tag{A.3}$$

Define the function $f_n : K \times K \to \mathbb{R}$

$$f_n(k_1, k_2) := \left(1 - \frac{\phi_n(k_2)s_n(k_1)\pi^h(k_1)}{\phi_n(k_1)s_n(k_2)\pi^h(k_2)}\right)_+$$

where $(a)_+ = max(a, 0)$, here π^h denotes the Lebesgue density of the prior π^h for h and $s_n(h) = p(x_{1:n}|\psi_\circ + h/\sqrt{n})/p(x_{1:n}|\psi_\circ)$. The function f_n is well defined for n sufficiently large because of (A.3) and Assumption 8 *(a)*. Remark that by (A.3) and since the prior for ψ puts enough mass on $\mathcal{U}$, then π^h puts enough mass on K and as $n \to 0$: $\pi^h(k_1)/\pi^h(k_2) \to 1$. Because of this and by the stochastic LAN expansion (A.8) in Theorem A.1:

$$\begin{aligned}\log\frac{\phi_n(k_2)s_n(k_1)\pi^h(k_1)}{\phi_n(k_1)s_n(k_2)\pi^h(k_2)} = &-\frac{1}{2}(k_2 - \Delta_{n,\psi^\circ})'V_{\psi^\circ}(k_2 - \Delta_{n,\psi^\circ}) + \frac{1}{2}(k_1 - \Delta_{n,\psi^\circ})'V_{\psi_\circ}(k_1 - \Delta_{n,\psi_\circ}) \\ &+ k_1'V_{\psi_\circ}\Delta_{n,\psi_\circ} - \frac{1}{2}k_1'V_{\psi_\circ}k_1 - k_2'V_{\psi_\circ}\Delta_{n,\psi_\circ} + \frac{1}{2}k_2'V_{\psi_\circ}k_2 + o_p(1) = o_p(1). \end{aligned} \tag{A.4}$$

Since, for every n, f_n is continuous in (k_1, k_2) and $K \times K$ is compact, then

$$\sup_{k_1,k_2\in K} f_n(k_1, k_2) \xrightarrow{p} 0, \qquad \text{as } n \to \infty. \tag{A.5}$$

Suppose that the subset K contains a neighborhood of 0 (which guarantees that $\Phi_n(K) > 0$ and then that Φ_n^K is well defined) and let $\Xi_n := \{\pi^h(K|x_{1:n}) > 0\}$. Moreover, for a given $\eta > 0$ define the event

$$\Omega_n := \left\{\sup_{k_1,k_2\in K} f_n(k_1, k_2) \leq \eta\right\}.$$

The TV distance $\|\cdot\|_{TV}$ between two probability measures P and Q, with Lebesgue densities p and q respectively, can be expressed as: $\|P-Q\|_{TV} = 2\int(1-p/q)_+ dQ$. Therefore, by the Jensen inequality and convexity of the functions $(\cdot)_+$,

$$\begin{aligned}\frac{1}{2}\mathbf{E}^P\|\pi_K^h(\cdot|x_{1:n}) - \Phi_n^K\|_{TV}1_{\Omega_n\cap\Xi_n} &= \mathbf{E}\int_K\left(1-\frac{d\Phi_n^K(k_2)}{d\pi_K^h(k_2|x_{1:n})}\right)_+ d\pi_K^h(k_2|x_{1:n})1_{\Omega_n\cap\Xi_n}\\ &\leq \mathbf{E}^P\int_K\int_K f_n(k_1,k_2)d\Phi_n^K(k_1)d\pi_K^h(k_2|x_{1:n})1_{\Omega_n\cap\Xi_n}\\ &\leq \mathbf{E}^P\sup_{k_1,k_2\in K} f_n(k_1,k_2)1_{\Omega_n\cap\Xi_n} \quad \text{(A.6)}\end{aligned}$$

that converges to zero by (A.5). By (A.5) and (A.6), it follows that (by remembering that $\|\cdot\|_{TV}$ is upper bounded by 2)

$$\mathbf{E}^P\|\pi_K^h(\cdot|x_{1:n}) - \Phi_n^K\|_{TV}1_{\Xi_n} \leq \mathbf{E}^P\|\pi_K^h(\cdot|x_{1:n}) - \Phi_n^K\|_{TV}1_{\Omega_n\cap\Xi_n} + 2P(\Omega_n^c\cap\Xi_n) = o(1) \quad \text{(A.7)}$$

In the second step of the proof let K_n be a sequence of balls in the parameter space of h centred at 0 with radii $M_n\to\infty$. For each $n\geq 1$, (A.7) holds for these balls. Moreover, by (A.10) in Theorem A.2: $P(\Xi_n)\to 1$. Therefore, by the triangular inequality, the TV distance is upper bounded by

$$\begin{aligned}\mathbf{E}^P\|\pi^h(\cdot|x_{1:n}) - \Phi_n\|_{TV} &\leq \mathbf{E}^P\|\pi^h(\cdot|x_{1:n}) - \Phi_n\|_{TV}1_{\Xi_n} + \mathbf{E}^P\|\pi^h(\cdot|x_{1:n}) - \Phi_n\|_{TV}1_{\Xi_n^c}\\ &\leq \mathbf{E}^P\|\pi^h(\cdot|x_{1:n}) - \pi_{K_n}^h(\cdot|x_{1:n})\|_{TV} + \mathbf{E}^P\|\pi_{K_n}^h(\cdot|x_{1:n}) - \Phi_n^{K_n}\|_{TV}1_{\Xi_n}\\ &\quad + \mathbf{E}^P\|\Phi_n^{K_n} - \Phi_n\|_{TV} + 2P(\Xi_n^c)\\ &\leq 2\mathbf{E}^P(\pi_{K_n^c}^h(\cdot|x_{1:n})) + \mathbf{E}^P\|\pi_{K_n}^h(\cdot|x_{1:n}) - \Phi_n^{K_n}\|_{TV}1_{\Xi_n} + o(1) \xrightarrow{p} 0\end{aligned}$$

since $\mathbf{E}^P(\pi^h(K_n^c|x_{1:n})) = o(1)$ by (A.10) and $\mathbf{E}^P\|\pi_{K_n}^h(\cdot|x_{1:n}) - \Phi_n^{K_n}\|_{TV}1_{\Xi_n} = o_P(1)$ by (A.7), and where in the third line we have used the fact that: $\mathbf{E}^P\|\pi^h(\cdot|x_{1:n}) - \pi_{K_n}^h(\cdot|x_{1:n})\|_{TV} = 2\mathbf{E}^P(\pi^h(K_n^c|x_{1:n}))$ and $\|\Phi_n^{K_n} - \Phi_n\|_{TV} = \|\Phi_n^{K_n^c}\|_{TV} = o_p(1)$ by Kleijn and van der Vaart (2012,

Lemma 5.2) since Δ_{n,ψ_0} is uniformly tight.

□

The next theorem establishes that the misspecified model satisfies a stochastic Local Asymptotic Normality (LAN) expansion around the pseudo-true value $\psi_\circ$.

Theorem A.1 (Stochastic LAN)**.** *Assume that the matrix $V_{\psi_\circ}$ is nonsingular and that Assumptions 2 (a)-(d), 3 (b), 5, 7, and 8 hold. Then for every compact set $K \subset \mathbb{R}^p$,*

$$\sup_{h\in K}\left|\log\frac{p(x_{1:n}|\psi_\circ + h/\sqrt{n})}{p(x_{1:n}|\psi_\circ)} - h'V_{\psi_\circ}\Delta_{n,\psi_\circ} + \frac{1}{2}h'V_{\psi_\circ}h\right| \xrightarrow{p} 0 \tag{A.8}$$

where $\psi_\circ$ is as defined in (2.13), $V_{\psi_\circ} = -\mathbf{E}^P[\ddot{\mathfrak{L}}_{n,\psi_\circ}]$ and $\Delta_{n,\psi_\circ} = \frac{1}{\sqrt{n}}\sum_{i=1}^n V_{\psi_\circ}^{-1}\dot{\mathfrak{L}}_{n,\psi_\circ}(x_i)$ is bounded in probability.

Proof. See Supplementary Appendix

□

The next theorem establishes that the posterior of ψ concentrates and puts all its mass on Ψ_n as $n \to \infty$.

Theorem A.2 (Posterior Consistency)**.** *Assume that the stochastic LAN expansion (A.8) holds for $\psi_\circ$ defined in (2.13). Moreover, let Assumptions 4 (a), 5 and 6 hold and assume that there exists a constant $C > 0$ such that for any sequence $M_n \to \infty$,*

$$P\left(\sup_{\psi\in\Psi_n^c}\frac{1}{n}\sum_{i=1}^n\left(l_{n,\psi}(x_i) - l_{n,\psi_\circ}(x_i)\right) \le -\frac{CM_n^2}{n}\right) \to 1 \tag{A.9}$$

as $n \to \infty$. Then,

$$\pi\left(\sqrt{n}\|\psi - \psi_\circ\| > M_n \middle| x_{1:n}\right) \xrightarrow{p} 0 \tag{A.10}$$

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

Proof. See Supplementary Appendix

□

A.2 Proof of Lemma 2.1.

By Theorem 10 of Schennach (2007), which is valid under Assumptions 2 *(a)-(c)*, 5, 7 *(c)*, *(e)* and 8: $\sqrt{n}(\widehat{\psi}-\psi_\circ)=O_p(1)$. Denote $\widehat{h}:=\sqrt{n}(\widehat{\psi}-\psi_\circ)$ and $\widetilde{h}:=\Delta_{n,\psi_\circ}$. Because of (A.8), we have:

$$\sum_{i=1}^{n}\left(l_{n,\psi_\circ+\widehat{h}/\sqrt{n}}-l_{n,\psi_\circ}\right)(x_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{h}'\dot{l}_{n,\psi_\circ}(x_i)-\frac{1}{2}\widehat{h}'V_{\psi_\circ}\widehat{h}+o_p(1) \tag{A.11}$$

$$\sum_{i=1}^{n}\left(l_{n,\psi_\circ+\widetilde{h}/\sqrt{n}}-l_{n,\psi_\circ}\right)(x_i) = \frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\widetilde{h}'\dot{l}_{n,\psi_\circ}(x_i)+o_p(1). \tag{A.12}$$

By definition of $\widehat{\psi}$ as the maximizer of $\sum_{i=1}^{n}l_{n,\psi}(x_i)$, the left hand side of (A.11) is not smaller than the left hand side of (A.12). It follows that the same relation holds for the right hand sides of (A.11) and (A.12), and by taking their difference we obtain:

$$-\frac{1}{2}\left(\widehat{h}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{\psi_\circ}^{-1}\dot{l}_{n,\psi_\circ}(x_i)\right)'V_{\psi_\circ}\left(\widehat{h}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{\psi_\circ}^{-1}\dot{l}_{n,\psi_\circ}(x_i)\right)+o_p(1)\geq 0. \tag{A.13}$$

Because $-V_{\psi_\circ}$ is negative definite, $-\frac{1}{2}\left(\widehat{h}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{\psi_\circ}^{-1}\dot{l}_{n,\psi_\circ}(x_i)\right)'V_{\psi_\circ}\left(\widehat{h}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{\psi_\circ}^{-1}\dot{l}_{n,\psi_\circ}(x_i)\right)\leq 0$. This and (A.13) imply that

$$\left\|V_{\psi_\circ}^{-1/2}\left(\widehat{h}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{\psi_\circ}^{-1}\dot{l}_{n,\psi_\circ}(x_i)\right)\right\|\xrightarrow{p}0$$

which in turn implies that

$$\left\|\left(\widehat{h}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{\psi_\circ}^{-1}\dot{l}_{n,\psi_\circ}(x_i)\right)\right\|\xrightarrow{p}0$$

which establishes the result of the lemma.

□

B Proofs for Section 3.3

In this appendix we prove Theorems 3.1 – 3.3. It is useful to introduce some notation that will be used throughout this section. We use the notation $\psi^\ell = (\theta^\ell, v^\ell)$ and the estimator $\widehat{\psi}^\ell := (\widehat{\theta}^\ell, \widehat{v}^\ell)$ denotes Schennach (2007)'ETEL estimator of ψ^ℓ in model M_ℓ:

$$\widehat{\psi}^\ell := \arg\max_{\psi^\ell \in \Psi^\ell} \frac{1}{n}\sum_{i=1}^n \left[\widehat{\lambda}(\psi^\ell)' g^A(x_i, \psi^\ell) - log\frac{1}{n}\sum_{j=1}^n \exp\{\widehat{\lambda}(\psi^\ell)' g^A(x_j, \psi^\ell)\}\right] \tag{B.1}$$

where $\widehat{\lambda}(\psi^\ell) = \arg\min_\lambda \frac{1}{n}\sum_{i=1}^n \left[\exp\{\lambda' g^A(x_i, \psi^\ell)\}\right]$. Denote $\widehat{g}^A(\psi^\ell) := \frac{1}{n}\sum_{i=1}^n g^A(x_i, \psi^\ell)$, $\widehat{g}^A_\ell := \widehat{g}^A(\psi^\ell)$,

$$\widehat{L}(\psi^\ell) := \exp\{\widehat{\lambda}(\psi^\ell)'\widehat{g}^A(\psi^\ell)\}\left[\frac{1}{n}\sum_{i=1}^n \exp\{\widehat{\lambda}(\psi^\ell)' g^A(x_i, \psi^\ell)\}\right]^{-1}$$

and $L(\psi^\ell) = \exp\{\lambda_\circ(\psi^\ell)'\mathbf{E}^P[g^A(x, \psi^\ell)]\}\left(\mathbf{E}^P\left[\exp\{\lambda_\circ(\psi^\ell)' g^A(x, \psi^\ell)\}\right]\right)^{-1}$. Moreover, we use the notation $\Sigma_\ell = \left(\Gamma_\ell' \Delta_\ell^{-1} \Gamma_\ell\right)^{-1}$ where $\Gamma_\ell := \mathbf{E}^P\left[\frac{\partial}{\partial \psi^{\ell\prime}} g^A(X, \psi^\ell_*)\right]$ $\Delta_\ell := \mathbf{E}^P[g^A(X, \psi^\ell_*) g^A(X, \psi^\ell_*)']$. In the proofs, we omit measurability issues which can be dealt with in the usual manner by replacing probabilities with outer probabilities.

B.1 Proof of Theorem 3.1

By Lemmas B.1 and B.2 we obtain

$$P\left(\log m(x_{1:n}; M_1) < \log m(x_{1:n}; M_2)\right) = P\Big(-\frac{n}{2}\widehat{g}_1^{A\prime}\Delta^{-1}\widehat{g}_1^A + \frac{n}{2}\widehat{g}_2^{A\prime}\Delta^{-1}\widehat{g}_2^A + o_p(n^{-1})$$
$$+ \log\frac{\pi(\widehat{\psi}^1)}{\pi(\widehat{\psi}^2)} - \frac{(p_1 + d_{v_1} - p_2 - d_{v_2})}{2}(\log n - \log(2\pi)) + \frac{1}{2}\left(\log|\Sigma_1| - \log|\Sigma_2|\right) < 0\Big). \tag{B.2}$$

Since for $\ell = 1, 2$, $n\widehat{g}_\ell^{A\prime}\Delta^{-1}\widehat{g}_\ell^A \xrightarrow{d} \chi^2_{d-(p_\ell + d_{v_\ell})}$, then $\widehat{g}_\ell^{A\prime}\Delta^{-1}\widehat{g}_\ell^A = O_p(n^{-1})$. Therefore,

$$
\begin{aligned}
P\left(\log m(x_{1:n};M_1) < \log m(x_{1:n};M_2)\right) &\geq P\Big(\frac{n}{2}\widehat{g}_2^{A'}\Delta^{-1}\widehat{g}_2^{A} + o_p(n^{-1}) \\
&< \log n\Big[\frac{(p_1+d_{v_1}-p_2-d_{v_2})}{2} - \frac{(p_1+d_{v_1}-p_2-d_{v_2})}{2\log n}\log(2\pi) \\
&\qquad - \frac{\log[\pi(\widehat{\psi}^1)/\pi(\widehat{\psi}^2)]}{\log n} - \frac{1}{2\log n}\left(\log|\Sigma_1| - \log|\Sigma_2|\right)\Big]\Big) \\
&= P\Big(\frac{n}{2}\widehat{g}_2^{A'}\Delta^{-1}\widehat{g}_2^{A} + o_p(n^{-1}) < \log n\Big[\frac{(p_1+d_{v_1}-p_2-d_{v_2})}{2} + \mathcal{O}_p((\log n)^{-1})\Big]\Big) \quad \text{(B.3)}
\end{aligned}
$$

Because the left hand side of the inequality inside the probability in the last line is $\mathcal{O}_p(1)$ and the right hand side is strictly positive as $n \to \infty$ (since $(p_1 + d_{v_1} > p_2 + d_{v_2})$) and converges to $+\infty$, then the probability converges to 1.

□

B.2 Proof of Theorem 3.2

We can write $\log p(x_{1:n}|\psi^\ell; M_\ell) = -n\log n + n\log\widehat{L}(\psi^\ell)$. By Lemmas B.1 and B.2 we obtain, for every $\psi^1 \in \Psi^1$ and $\psi^2 \in \Psi^2$

$$
\begin{aligned}
P\left(\log m(x_{1:n};M_1) > \log m(x_{1:n};M_2)\right) &= P\Big(-\frac{n}{2}\widehat{g}_1^{A'}\Delta^{-1}\widehat{g}_1 + o_p(n^{-1}) \\
&\qquad - n\log L(\psi^2) - n\left[\log\widehat{L}(\psi^2) - \log L(\psi^2)\right] + \log[\pi(\widehat{\psi}^1)/\pi(\psi^2)] \\
&\qquad - \frac{(p_1+d_{v_1})}{2}(\log n - \log(2\pi)) - \frac{1}{2}\log|\Sigma_1| + \log\pi(\psi^2|x_{1:n};M_2) > 0\Big) \\
&= P\Big(\mathcal{B}n^{-1} + o_p(n^{-2}) - \log L(\psi^2) - \left[\log\widehat{L}(\psi^2) - \log L(\psi^2)\right] > 0\Big). \quad \text{(B.4)}
\end{aligned}
$$

where $\mathcal{B} := -\frac{n}{2}\widehat{g}_1^{A'}\Delta^{-1}\widehat{g}_1^{A} + \log[\pi(\widehat{\psi}^1)/\pi(\psi^2)] - \frac{(p_1+d_{v_1})}{2}(\log n - \log(2\pi)) - \frac{1}{2}\log|\Sigma_1| + \log\pi(\psi^2|x_{1:n};M_2)$. Remark that $\mathcal{B}n^{-1} = o_p(1)$ by Lemma B.1 and because, under the assumptions of Theorem 2.2 and of Lemma 2.1, equation (2.17) holds, that is, $\pi(\psi^2|x_{1:n};M_2)$ is asymptotically equal to a $\mathcal{N}_{\widehat{\psi},n^{-1}V_{\psi_\circ}^{-1}}$. Moreover, $\left[\log\widehat{L}(\psi^2) - \log L(\psi^2)\right] \xrightarrow{p} 0$, $\forall\psi^2 \in \Psi^2$ by Lemma B.3. Therefore,

we conclude that

$$P\left(\log m(x_{1:n};M_1) > \log m(x_{1:n};M_2)\right) = P\Big(o_p(1) - \log L(\psi^2) > 0\Big) \xrightarrow{p} 1$$

since $\log L(\psi^2) = \lambda_\circ(\psi)'\mathbf{E}^P[g^A(x,\psi^2)] - \log \mathbf{E}^P\left[\exp\{\lambda_\circ(\psi)' g^A(x,\psi^2)\}\right] < 0$ for every $\psi^2 \in \Psi^2$ by the Jensen's inequality.

□

B.3 Proof of Theorem 3.3

We can write $\log p(x_{1:n}|\psi^\ell; M_\ell) = -n\log n + n\log \widehat{L}(\psi^\ell)$. Then, we have:

$$\begin{aligned} P\left(\log m(x_{1:n};M_1) > \log m(x_{1:n};M_2)\right) &= P\Big(-n\log n + n\log \widehat{L}(\psi^1_\circ) + n\log n - n\log \widehat{L}_{\ell_2}(\psi^2_\circ) \\ &\quad + \log \pi(\psi^1_\circ|M_1) - \log \pi(\psi^2_\circ|M_2) - \log \pi(\psi^1_\circ|x_{1:n},M_1) + \log \pi(\psi^2_\circ|x_{1:n},M_2) \\ &= P\Big(n\left[\log L(\psi^1_\circ) - \log L(\psi^2_\circ)\right] + n\left[\log \widehat{L}(\psi^1_\circ) - \log L(\psi^1_\circ)\right] \\ &\quad - n\left[\log \widehat{L}(\psi^2_\circ) - \log L(\psi^2_\circ)\right] + \mathcal{B} > 0\Big) \quad \text{(B.5)} \end{aligned}$$

where $\mathcal{B} := \log \pi(\psi^1_\circ|M_1) - \log \pi(\psi^2_\circ|M_2) - \log \pi(\psi^1_\circ|x_{1:n},M_1) + \log \pi(\psi^2_\circ|x_{1:n},M_2)$ and $\mathcal{B} = O_p(1)$ under the assumptions of Theorem 2.2. Moreover, $\left[\log \widehat{L}(\psi^\ell) - \log L(\psi^\ell)\right] \xrightarrow{p} 0$, $\forall \psi^\ell \in \Psi^\ell$ and $\ell \in \{1,2\}$ by Lemma B.3. Therefore,

$$\begin{aligned} P\left(\log m(x_{1:n};M_1) > \log m(x_{1:n};M_2)\right) &= P\Big(\left[\log L(\psi^1_\circ) - \log L(\psi^2_\circ)\right] \\ &\quad + \left[\log \widehat{L}(\psi^1_\circ) - \log L(\psi^1_\circ)\right] - \left[\log \widehat{L}(\psi^2_\circ) - \log L(\psi^2_\circ)\right] + \frac{1}{n}\mathcal{B} > 0\Big) \\ &= P\Big(\left[\log L(\psi^1_\circ) - \log L(\psi^2_\circ)\right] + o_p(1) > 0\Big). \quad \text{(B.6)} \end{aligned}$$

Next, by definition of $dQ^*(\psi)$ in Section 2.3 we have that: $\log L(\psi^\ell_\circ) = \mathbf{E}^P[\log dQ^*(\psi^\ell_\circ)/dP] = -\mathbf{E}^P[\log dP/dQ^*(\psi^\ell_\circ)]$. Therefore, by replacing this in (B.6) we obtain:

$$P\left(\log m(x_{1:n};M_1) > \log m(x_{1:n};M_2)\right)$$
$$= P\Big(\mathbf{E}^P\left[\log\left(dP/dQ^*(\psi_\circ^2)\right)\right] - \mathbf{E}^P\left[\log\left(dP/dQ^*(\psi_\circ^1)\right)\right] + o_p(1) > 0\Big). \quad \text{(B.7)}$$

This probability converges to 1 if $\mathbf{E}^P\left[\log\left(dP/dQ^*(\psi_\circ^2)\right)\right] > \mathbf{E}^P\left[\log\left(dP/dQ^*(\psi_\circ^1)\right)\right]$, that is, if the KL divergence between P and $Q^*(\psi_\circ^\ell)$, is smaller for model M_1 than for model M_2, where $Q^*(\psi_\circ^\ell)$ minimizes the KL divergence between $Q \in \mathcal{P}_{\psi_\circ^\ell}$ and P for $\ell \in \{1,2\}$ (remark the inversion of the two probabilities). This means that the ML-based selection procedures selects the misspecified model that is the closest to the true DGP P, as measured by the KL divergence.

□

B.4 Technical Lemmas

Lemma B.1. *Let Assumptions 1-3 hold for ψ^ℓ. Then,*

$$\log p(x_{1:n}|\widehat{\psi}^\ell;M_\ell) = -n\log n - \frac{n}{2}\widehat{g}_\ell^{A'}\Delta_\ell^{-1}\widehat{g}_\ell^{A} + o_p\left(n^{-1}\right)$$
$$= -n\log n - \frac{\chi^2_{d_\ell - p}}{2} + o_p\left(n^{-1}\right) \quad \text{(B.8)}$$

where $\chi^2_{d-(p_\ell+d_{v_\ell})}$ denotes a chi square distribution with $(d-(p_\ell+d_{v_\ell}))$ degrees of freedom.

Proof. See Supplementary Appendix

□

Lemma B.2. *Let Assumptions 1 - 3 and (2.12) hold for ψ^ℓ. Then,*

$$-\log\pi(\widehat{\psi}^\ell|x_{1:n};M_\ell) = -\frac{(p_\ell+d_{v_\ell})}{2}[\log n - \log(2\pi)] + \frac{1}{2}\log|\Sigma_\ell| + o_p(1).$$

Proof. See Supplementary Appendix

□

Lemma B.3. *Let M_ℓ be a misspecified model (that is, a model that does not satisfy Assumption 1) and let $g^A(x,\psi^\ell)$ and ψ^ℓ be the corresponding moment functions and parameters. Then, under Assumptions 2 (a)-(c), 5 and 7,*

$$\sup_{\psi^\ell\in\Psi^\ell}\left|\log\frac{\exp\{\widehat{\lambda}(\psi^\ell)'\widehat{g}^A(\psi^\ell)\}}{\frac{1}{n}\sum_{i=1}^n\exp\{\widehat{\lambda}(\psi^\ell)'g^A(x_i,\psi^\ell)\}}-\log\frac{\exp\{\lambda_\circ(\psi^\ell)'\mathbf{E}^P[g^A(x,\psi^\ell)]\}}{\mathbf{E}^P\left[\exp\{\lambda_\circ(\psi^\ell)'g^A(x,\psi^\ell)\}\right]}\right|\xrightarrow{p}0.$$

Proof. See Supplementary Appendix

□

References

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

I. H. Chang and R. Mukerjee. Bayesian and frequentist confidence intervals arising from empirical-type likelihoods. *Biometrika*, 95(1):139–147, 2008.

S. Chaudhuri and M. Ghosh. Empirical likelihood for small area estimation. *Biometrika*, 98 (2):473–480, 2011.

S. Chaudhuri, D. Mondal, and T. Yin. Hamiltonian monte carlo sampling in bayesian empirical likelihood computation. *Journal of the Royal Statistical Society, Series B*, forthcoming, 2017.

S. X. Chen and I. Van Keilegom. A review on Empirical Likelihood methods for regression. *TEST*, 18(3):415–447, 2009.

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

S. Chib and E. Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

S. Chib and I. Jeliazkov. Marginal likelihood from the metropolis hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001.

S. Chib and S. Ramamurthy. Tailored randomized block mcmc methods with application to dsge models. *Journal of Econometrics*, 155(1):19–38, 2010.

I. Csiszar. i-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3(1):146–158, 1975.

K.-T. Fang and R. Mukerjee. Empirical-type likelihoods allowing posterior credible sets with frequentist validity: Higher-order asymptotics. *Biometrika*, 93(3):723–733, 2006.

M. Grendar and G. Judge. Asymptotic equivalence of empirical likelihood and bayesian map. *The Annals of Statistics*, 37(5A):2445–2457, 2009.

P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982.

H. Hong and B. Preston. Bayesian averaging, prediction and nonnested model selection. *Journal of Econometrics*, 167(2):358–369, 2012.

M.-O. Kim and Y. Yang. Semiparametric approach to a random effects quantile regression model. *Journal of the American Statistical Association*, 106(496):1405–1417, 2011.

Y. Kitamura and M. Stutzer. An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(4):pp. 861–874, 1997.

B. Kleijn and A. van der Vaart. The bernstein-von-mises theorem under misspecification. *Electron. J. Statist.*, 6:354–381, 2012.

T. Lancaster and S. J. Jun. Bayesian quantile regression methods. *Journal of Applied Econometrics*, 25(2):287–307, 2010.

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

E. L. Lehmann and G. Casella. *Theory of Point Estimation (Springer Texts in Statistics)*. Springer, 2nd edition, 1998.

W. K. Newey and R. J. Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

A. B. Owen. Empirical Likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.

A. B. Owen. *Empirical Likelihood*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 2001.

A. T. Porter, S. H. Holan, and C. K. Wikle. Bayesian semiparametric hierarchical empirical likelihood spatial models. *Journal of Statistical Planning and Inference*, 165:78–90, 2015.

J. Qin and J. Lawless. Empirical Likelihood and general estimating equations. *Ann. Statist.*, 22(1):300–325, 03 1994.

J. Rao and C. Wu. Bayesian pseudo-empirical-likelihood intervals for complex surveys. *Journal of the Royal Statistical Society Series B*, 72(4):533–544, 2010.

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

S. M. Schennach. Point estimation with exponentially tilted empirical likelihood. *Ann. Statist.*, 35(2):634–672, 04 2007.

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

A. W. Van der Vaart. *Asymptotic Statistics*. Lectures on Probability Theory., Ecole d'Ete de Probailites de St. Flour XX, 2000.

A. M. Variyath, J. Chen, and B. Abraham. Empirical likelihood based variable selection. *Journal of Statistical Planning and Inference*, 140(4):971–981, 2010.

A. Vexler, W. Deng, and G. E. Wilding. Nonparametric bayes factors based on empirical likelihood ratios. *Journal of Statistical Planning and Inference*, 143(3):611–620, 2013.

R. Xi, Y. Li, and Y. Hu. Bayesian quantile regression based on the empirical likelihood with spike and slab priors. *Bayesian Analysis*, forthcoming, 2016.

Y. Yang and X. He. Bayesian empirical likelihood for quantile regression. *The Annals of Statistics*, 40(2):1102–1131, 2012.