LARGE SAMPLE JUSTIFICATIONS FOR THE BAYESIAN EMPIRICAL LIKELIHOOD

Sueishi, Naoya

(Citation)
Econometric Theory,40(4):926-956

(Issue Date)
2022-12-05

(Resource Type)
journal article

(Version)
Accepted Manuscript

Large Sample Justifications for the Bayesian Empirical Likelihood

Naoya Sueishi*

Kobe University†

August 25, 2020

Last Revised: September 3, 2022

Abstract

This study investigates the asymptotic properties of the Bayesian empirical likelihood (BEL), which uses the empirical likelihood as an alternative to a parametric likelihood for Bayesian inference. We establish two asymptotic equivalence results based on the Bernstein–von Mises (BvM) theorem by introducing a new formulation of the moment restriction model. First, the limiting posterior distribution of the BEL is the same as that of a parametric Bayesian method that uses the likelihood of a least favorable model of the moment restriction model. Second, the limiting posterior distribution is also the same as that of a semiparametric Bayesian method that places priors on both a finite-dimensional parameter of interest and an infinite-dimensional nuisance parameter. Because parametric and semiparametric Bayesian methods are legitimate Bayesian procedures, the equivalence results provide a large sample justification for the BEL as a Bayesian inference method. Moreover, the BvM theorem provides a frequentist justification for BEL posterior inference.

*I would like to thank Patrik Guggenberger and two anonymous referees for their comments and suggestions. I would also like to thank Mototsugu Shintani, Kohtaro Hitomi, Yoshihiko Nishiyama, and Takahide Yanagi for their comments. This research was supported by JSPS KAKENHI Grant Number 18K01547.

†Graduate School of Economics, 2-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan. Email: sueishi@econ.kobe-u.ac.jp. Phone: +81-78-803-6827.

Keywords: Bayesian empirical likelihood; Bernstein–von Mises theorem; Convolution theorem; Local asymptotic normality

1 Introduction

Specifying a statistical model via a set of moment restrictions of the form $\mathbb{E}_Q[m_\theta(X)] = 0$ is common in econometrics. The model has the finite-dimensional parameter of interest θ and the infinite-dimensional nuisance parameter Q, which is the distribution of the observation X. One advantage of this modeling is that a full specification of the distribution of the observation is not necessary, thereby mitigating the risk of model misspecification. However, it is often difficult to estimate θ by an extremum estimator such as GMM (Hansen 1982) when the objective function has some local maxima or minima. Although a Bayesian method is a useful alternative in such a circumstance (see Fernández-Villaverde 2010), classical Bayesian inference is not feasible because a likelihood function is not specified.

Existing Bayesian procedures for the moment restriction model can be classified into two approaches: the parametric approach and the semiparametric approach. The first approach uses a parametric alternative to the likelihood function (quasi-likelihood function) and places a prior only on the finite-dimensional parameter. The second approach places priors on the finite-dimensional parameter and the infinite-dimensional nuisance parameter. The posterior distribution for the parameter of interest is obtained by integrating out the nuisance parameter.

Although parametric methods are computationally more tractable than semiparametric ones, parametric methods may not have a theoretical base because a quasi-likelihood function is not a genuine likelihood function. Therefore, the main issue of the parametric approach is justification for the use of the quasi-likelihood function. Owen (2001) suggested to use the empirical likelihood (EL) and pointed out its resemblance with the likelihood of a least favorable model of a semiparametric model. Lazar (2003) employed the EL and investigated whether it satisfies the validity condition of Monahan and Boos (1992). Schennach (2005) provided a theoretical basis for the Bayesian exponentially tilted empirical likelihood (BETEL) that uses the ETEL as the quasi-likelihood. Chib et al. (2018) investigated the asymptotic properties of the BETEL posterior. See also Kim (2002) and Ragusa (2007) for different approaches.

The main issue of the semiparametric approach is how to place priors on both finite- and infinite-dimensional parameters without contradiction. If θ and Q are separately drawn

from parametric and nonparametric priors, respectively, then Q may not satisfy the moment restrictions for the realized value of θ; that is, it may be the case that $\mathbb{E}_Q[m_\theta(X)] \neq 0$. Chamberlain and Imbens (2003) extended the Bayesian bootstrap method of Rubin (1981) to the overidentified setting and introduced a Dirichlet prior on Q, which leads to a Dirichlet posterior. The posterior for θ is obtained by solving an augmented set of moment restrictions, where the expectation is taken with respect to the Dirichlet posterior. Thus, they avoided placing a prior on the finite-dimensional parameter of interest. Kitamura and Otsu (2011) proposed the exponential tilting projection method, which first draws Q from a nonparametric prior and then projects it onto the space of probability measures that satisfy the moment restrictions for a given value of θ. Shin (2015) considered a similar approach to Kitamura and Otsu (2011). Bornn et al. (2019) and Florens and Simoni (2021) proposed different procedures. Andrews and Mikusheva (2022) considered a similar issue when the moment restriction model is weakly identified.

This study reconsiders theoretical justifications for the Bayesian EL (BEL) that uses the EL function of Qin and Lawless (1994) as an alternative to a parametric likelihood function. Because of a similarity between the EL and the parametric likelihood, employing the EL as the quasi-likelihood seems to be a natural choice, and there have been some studies on the BEL. The examples include Fang and Mukerjee (2006), Yang and He (2012), Vexler et al. (2014), Chaudhuri et al. (2017), Cheng and Zhao (2019), and Bedoui and Lazar (2020). We justify the BEL by showing its asymptotic equivalence with legitimate Bayesian procedures.

For this purpose, we first introduce a new formulation of the moment restriction model. The model is specified as a set of probability measures that are indexed by the finite-dimensional parameter of interest and the infinite-dimensional nuisance parameter. Then, we establish two versions of the Bernstein–von Mises (BvM) theorem. The first one shows that the posterior of the BEL is asymptotically equivalent to that of an infeasible parametric Bayesian method that uses the likelihood of a least favorable model of the moment restriction model. The second one shows that the posterior of the BEL is asymptotically equivalent to that of a semiparametric Bayesian method that introduces priors on both finite- and infinite-dimensional parameters. Because the parametric and semiparametric Bayesian methods are legitimate Bayesian procedures based on genuine likelihood functions, the equivalence results provide a

large sample justification for the BEL as a Bayesian inference method.

The first equivalence result substantiates the idea of Owen (2001), who suggested that the BEL can be justified by the similarity of the EL with the likelihood of a least favorable model. Our result is one possible way to justify the BEL in line with his idea. The second equivalence result is similar to Schennach (2005), who showed that the posterior of the BETEL can be represented as a certain limit of the posterior of a legitimate Bayesian procedure. Florens and Simoni (2021) and Andrews and Mikusheva (2022) also discussed asymptotic equivalence results between quasi-parametric and semiparametric Bayesian methods.

Our formulation of the moment restriction model is also beneficial in clarifying the relationship between efficient frequentist estimators and Bayesian point estimators. The semiparametric BvM theorem of Bickel and Kleijn (2012) showed that the mean of the limiting marginal posterior distribution is asymptotically equivalent to best regular estimators in general semiparametric models. Although our BvM theorem implies that the mean of the limiting posterior distribution of the BEL is asymptotically equivalent to the EL estimator, it is not immediate that the EL estimator is best regular under the formulation of our model. Thus, we show that the EL estimator is indeed a best regular estimator by establishing the convolution theorem for the moment restriction model based on the local asymptotic normality (LAN) of our model. The result also implies that BEL point estimators are asymptotically efficient in the frequentist sense.

This paper is also closely related to Chernozhukov and Hong (2003). They proposed the Laplace type estimator (LTE), which is a quasi-Bayesian estimator that uses a general statistical function in place of a parametric likelihood function. The BEL can be viewed as a special case of the LTE. However, they only investigated the frequentist properties of the LTE and emphasized that their approach falls outside the Bayesian approach. Our new finding is the Bayesian interpretation of the BEL.

The remainder of the paper proceeds as follows. Section 2 introduces our formulation of the moment restriction model. Section 3 gives an overview of the results of the paper. Section 4 shows the convolution theorem. Sections 5 and 6 show the BvM theorems for the BEL and the semiparametric Bayesian method, respectively. Section 7 concludes. Appendix contains some lemmas and their proofs. Proofs for the main theorems are given in the supplemental

material.

2 Model

Let $\{X_1, \ldots, X_n\}$ be a random sample from a true distribution P, which is defined on a measurable space $(\mathcal{X}, \mathcal{A})$. The true parameter of interest, $\theta_0 \in \Theta \subset \mathbb{R}^p$, is characterized as a unique vector that satisfies the vector of moment restrictions:

$$\mathbb{E}[m_{\theta_0}(X)] = \int m_{\theta_0} dP = 0, \tag{2.1}$$

where $m : \mathcal{X} \times \Theta \to \mathbb{R}^l$ is a known vector-valued function. We assume $l \geq p$ and θ_0 is an inner point of Θ. The moment restriction model is semiparametric in the sense that P is an infinite-dimensional nuisance parameter.

The moment restriction model can also be represented as a set of probability measures (see, e.g., Kitamura and Stutzer 1997, Chen et al. 2007, and Kitamura 2007). Let H be a set of probability measures on $(\mathcal{X}, \mathcal{A})$ that is dominated by a measure ξ. Moreover, let $\mathcal{Q}_\theta = \{\eta \in H : \int m_\theta d\eta = 0\}$, which is a set of probability measures that satisfy the moment restrictions for a given value of θ. Then, $\mathcal{Q} = \cup_{\theta \in \Theta} \mathcal{Q}_\theta$ is a set of probability measures that is compatible with the moment restrictions.

A feature of the moment restriction model is that the distribution of the observation itself is the nuisance parameter. Therefore, the nuisance parameter must depend on the parameter of interest. This feature makes it hard to conduct a semiparametric Bayesian inference. If the finite- and infinite-dimensional parameters, say $\tilde{\theta}$ and $\tilde{\eta}$, are separately drawn from parametric and nonparametric priors, then $\tilde{\eta}$ may not satisfy the moment restriction for given $\tilde{\theta}$.

The feature of the model is also inconvenient in investigating the semiparametric efficiency bound for estimating θ_0. In the analysis of semiparametric models, it is often of interest to compare the efficiency bound that can be achieved when the true nuisance parameter is known and that which can be achieved when the true nuisance parameter is unknown. However, if the true nuisance parameter of the moment restriction model is known, then θ_0 is determined by (2.1). Thus, no estimation problem arises.

To address these issues, we rewrite the model in a way that it is indexed by the finite-dimensional parameter of interest θ and the infinite-dimensional nuisance parameter η. For

given $\theta \in \Theta$ and $\eta \in H$, let $P_{\theta,\eta} \in \mathcal{Q}_\theta$ satisfy

$$-\int \log \frac{dP_{\theta,\eta}}{d\eta} d\eta = \inf_{Q \in \mathcal{Q}_\theta} -\int \log \frac{dQ}{d\eta} d\eta. \tag{2.2}$$

Thus, $P_{\theta,\eta}$ is the projection of η onto $\mathcal{Q}_\theta$ in terms of the Kullback–Leibler (KL) divergence. We define $-\int \log \frac{dQ}{d\eta} d\eta = \infty$ if Q is not absolutely continuous with respect to η. Our moment restriction model is defined as $\mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$. Because the true nuisance parameter η_0 must satisfy $P = P_{\theta_0,\eta_0}$, η_0 coincides with P. We assume that $P_{\theta,\eta} = P$ if and only if $\theta = \theta_0$ and $\eta = \eta_0$.

We give a sufficient condition that guarantees the existence of the solution to (2.2). Many studies have investigated the solution to the minimization problem of the form (2.2). Examples of such studies include Borwein and Lewis (1991), Csiszár (1995), Kitamura (2007), and Komunjer and Ragusa (2016). Let $(\gamma^*_\eta(\theta), \lambda^*_\eta(\theta))$ solve

$$\max_{\gamma \in \mathbb{R}, \lambda \in \mathbb{R}^l} \left[\gamma - \int \phi^*_+(\gamma + \lambda' m_\theta) d\eta \right], \tag{2.3}$$

where

$$\phi^*_+(v) = \begin{cases} -1 - \log(-v) & v < 0 \\ \infty & v \geq 0. \end{cases}$$

The solution to (2.3) exists under the conditions give below. The conditions also ensure that the unique solution to (2.2) exists and is obtained by solving the corresponding dual problem (2.3).

Assumption 2.1

(i) For each $\eta \in H$, there exists $M < \infty$ such that $\sup_{\theta \in \Theta} \|m_\theta\| \leq M$ a.e. with respect to η.

(ii) For each $\theta \in \Theta$ and $\eta \in H$, there exists $Q \in \mathcal{Q}_\theta$ and a positive constant $C < \infty$ such that $\frac{1}{C} \leq \frac{dQ}{d\eta} \leq C$ a.e. with respect to η.

(iii) For each $\eta \in H$, there exists $\epsilon > 0$ such that $\inf_{\theta \in \Theta} -\gamma^*_\eta(\theta) - \lambda^*_\eta(\theta)' m_\theta \geq \epsilon$ a.e. with respect to η.

Conditions (i) and (iii) are similar to condition 1 of Assumption 3 in Chen et al. (2007), who obtained the projection by using a different method from ours. Condition (i) guarantees that $P_{\theta,\eta}$ is an element of $\mathcal{Q}_\theta$ (see Komunjer and Ragusa 2016). In particular, P_{θ_0,η_0} satisfies (2.1). See also Theorem 1 of Schennach (2007) about condition (i). Condition (ii) requires that

$\mathcal{Q}_\theta$ contains a probability measure that is mutually absolutely continuous with respect to η. A related issue is also discussed by Sueishi (2013) and Chib et al. (2018). Condition (ii) implies that the minimization problem is feasible. Conditions (i) and (ii) guarantee that a constraint qualification condition is satisfied; that is, the primal problem (2.2) has the equivalent dual problem (2.3).

Proposition 2.1

Suppose that Assumption 2.1 holds. Then, there exists a unique solution to (2.2) a.e. with respect to η. Moreover, $P_{\theta,\eta}$ satisfies

$$\frac{dP_{\theta,\eta}}{d\eta} = \frac{1}{1+\lambda_\eta(\theta)'m_\theta} \tag{2.4}$$

a.e. with respect to η, where $\lambda_\eta(\theta) = \arg\max_{\lambda\in\mathbb{R}^l} \int \log(1+\lambda' m_\theta) d\eta$.

The proof is given in the Appendix. The existence of $\lambda_\eta(\theta)$ is guaranteed by Assumption 2.1. By construction, $P_{\theta,\eta}$ satisfies $\int m_\theta dP_{\theta,\eta} = 0$ for all θ and η. Thus, it is compatible with the moment restrictions. Under Assumption 2.1, we have $\inf_{\theta\in\Theta}(1+\lambda_\eta(\theta)'m_\theta) > 0$ a.e. with respect to η. Because $P_{\theta,\eta}$ is absolutely continuous with respect to η, the density of $P_{\theta,\eta}$ with respect to ξ exists.

Condition (i) may be restrictive in certain cases. Assuming Θ is bounded, a linear instrumental variable model satisfies the condition if $\mathcal{X}$ is bounded, but not if $\mathcal{X}$ is unbounded. Although some economic variables have bounded support, it is generally not innocuous to assume boundedness of $\mathcal{X}$. Note, however, that the asymptotic properties of the BEL do not depend on the validity of condition (i). When $\mathcal{X}$ is unbounded, Komunjer and Ragusa (2016) showed potential approaches to relax condition (i). For instance, we may modify the tail of the logarithmic function in (2.2) so that it diverges to infinity sufficiently first (see Section 5 of Komunjer and Ragusa 2016). We do not further pursue this approach.

Another way to relax Assumption 2.1 (i) is to replace the role of Q and η in (2.2), that is, we solve

$$\int \log \frac{d\tilde{P}_{\theta,\eta}}{d\eta} d\tilde{P}_{\theta,\eta} = \inf_{Q\in\mathcal{Q}_\theta} \int \log \frac{dQ}{d\eta} dQ.$$

If $\int \exp(\lambda' m_\theta) d\eta < \infty$ for all $\lambda \in \mathbb{R}^l$, then there exists a unique $\tilde{P}_{\theta,\eta} \in \mathcal{Q}_\theta$ that satisfies

$$\frac{d\tilde{P}_{\theta,\eta}}{d\eta} = \frac{\exp(\tilde{\lambda}_\eta(\theta)' m_\theta)}{\int \exp(\tilde{\lambda}_\eta(\theta)' m_\theta) d\eta}, \tag{2.5}$$

where $\tilde{\lambda}_\eta(\theta) = \arg\max_{\lambda \in \mathbb{R}^l} \int \exp(\lambda' m_\theta) d\eta$.

The formulation (2.5) is used by Kitamura and Otsu (2011). If a Dirichlet process is employed for the prior of η in (2.5), then the resulting semiparametric Bayesian procedure is the exponentially tilted Dirichlet process approach of Kitamura and Otsu (2011). Moreover, if $\prod_{i=1}^n d\tilde{P}_{\theta,\mathbb{P}_n}(X_i)$ is used as the alternative to the parametric likelihood function, where $\mathbb{P}_n$ is the empirical distribution of $\{X_1, \ldots, X_n\}$, then the resulting quasi-Bayesian procedure is the BETEL of Schennach (2005) and Chib et al. (2018), although Schennach (2005) and Chib et al. (2018) did not consider $\tilde{P}_{\theta,\eta}$ as their model.

Our formulation of the moment restriction model is useful in the following aspects. First, our model facilitates the comparison between the BEL and the semiparametric Bayesian method because the EL ratio can be written as $\prod_{i=1}^n dP_{\theta,\mathbb{P}_n}/d\mathbb{P}_n(X_i)$. Second, the formulation is convenient in deriving the efficiency bound for the moment restriction model. As will be discussed later, the score function for θ_0 is orthogonal to the score function for η_0. This implies that knowing the true distribution η_0 does not change the efficiency bound for estimating θ_0.

3 Overview

This section gives an informal explanation for the BvM theorem by referring to Chapter 10 of van der Vaart (1998). Moreover, we explain why the BEL is asymptotically equivalent to legitimate Bayesian methods. Rigorous arguments are given in subsequent sections.

First, we consider the parametric Bayesian method. For now, suppose that η_0 is known. Then, we obtain a parametric model $\theta \mapsto P_{\theta,\eta_0}$. Let $p_{\theta,\eta}$ be the density of $P_{\theta,\eta}$ with respect to ξ. Also, let $\pi_\Theta(\theta)$ be a prior density for θ. Then, the posterior density of $\sqrt{n}(\theta - \theta_0)$ is obtained by the Bayes rule:

$$p^0_{\sqrt{n}(\theta-\theta_0)|X_1,\ldots,X_n}(h) = \frac{\prod_{i=1}^n p_{\theta_n(h),\eta_0}(X_i)\pi_\Theta(\theta_n(h))}{\int \prod_{i=1}^n p_{\theta_n(h),\eta_0}(X_i)\pi_\Theta(\theta_n(h))dh}, \tag{3.1}$$

where $\theta_n(h) = \theta_0 + h/\sqrt{n}$ and h is a vector in $\mathbb{R}^p$. If $\pi_\Theta(\theta)$ is continuous at θ_0, $\pi_\Theta(\theta_n(h))$ converges to $\pi_\Theta(\theta_0)$ as $n \to \infty$. Thus, the asymptotic behavior of (3.1) is determined by the likelihood ratio $\prod_{i=1}^n p_{\theta_n(h),\eta_0}/p_{\theta_0,\eta_0}(X_i)$. Suppose that the likelihood ratio has the following LAN expansion:

$$\log \prod_{i=1}^n \frac{p_{\theta_n(h),\eta_0}}{p_{\theta_0,\eta_0}}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h' \dot{\ell}_{\theta_0,\eta_0}(X_i) - \frac{1}{2} h' I_{\theta_0,\eta_0} h + o_P(1), \tag{3.2}$$

where $\dot{\ell}_{\theta_0,\eta_0}$ is the score function for θ_0 and I_{θ_0,η_0} is the Fisher information matrix for known η_0. Then, the likelihood ratio can be approximated by that of a normal distribution:

$$\frac{dN(h, I^{-1}_{\theta_0,\eta_0})}{dN(0, I^{-1}_{\theta_0,\eta_0})}(\Delta_n),$$

where $\Delta_n = n^{-1/2} \sum_{i=1}^n I^{-1}_{\theta_0,\eta_0} \dot{\ell}_{\theta_0,\eta_0}(X_i)$ and $dN(\mu, \Sigma)$ denotes the density of the normal distribution. Hence, the right-hand side of (3.1) can be approximated by

$$\frac{dN(h, I^{-1}_{\theta_0,\eta_0})(\Delta_n)}{\int dN(h, I^{-1}_{\theta_0,\eta_0})(\Delta_n) dh} = dN(\Delta_n, I^{-1}_{\theta_0,\eta_0})(h).$$

This means that the posterior distribution of $\sqrt{n}(\theta - \theta_0)$ is approximated by the normal distribution with mean Δ_n and variance matrix $I^{-1}_{\theta_0,\eta_0}$. Moreover, the center of the posterior distribution is asymptotically equivalent to the best regular estimators for θ_0, which include the maximum likelihood estimator because any best regular estimator $\hat{\theta}_n$ for θ_0 satisfies $\sqrt{n}(\hat{\theta}_n - \theta_0) = \Delta_n + o_P(1)$.

Next, we consider the BEL. The BEL posterior density for $\sqrt{n}(\theta - \theta_0)$ is obtained by replacing η_0 with $\mathbb{P}_n$ in the right-hand side of (3.1). Suppose that we have

$$\log \prod_{i=1}^n \frac{dP_{\theta_n(h),\mathbb{P}_n}}{dP_{\theta_0,\mathbb{P}_n}}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h' \dot{\ell}_{\theta_0,\eta_0}(X_i) - \frac{1}{2} h' I_{\theta_0,\eta_0} h + o_P(1). \tag{3.3}$$

That is, the EL ratio has the same expansion as (3.2). Then the BEL posterior is asymptotically equivalent to (3.1). Condition (3.3) requires that the asymptotic property of the likelihood ratio does not change even if the true nuisance parameter is replaced with its estimator. This is not true for general semiparametric models. A key condition for (3.3) is that the score function for θ_0 is orthogonal to the score function for η_0. The orthogonality of the score functions is discussed in Section 4.

Finally, we consider our semiparametric Bayesian method. Let Π_H be a prior for η defined on H. The marginal posterior density for $\sqrt{n}(\theta - \theta_0)$ is obtained by integrating out the nuisance

parameter. The asymptotic behavior of the marginal posterior is determined by the integrated likelihood ratio:

$$s_n(h) = \int_H \prod_{i=1}^{n} \frac{p_{\theta_n(h),\eta}}{p_{\theta_0,\eta_0}}(X_i) d\Pi_H(\eta). \tag{3.4}$$

Under the conditions given in Section 6, we obtain

$$\log \frac{s_n(h)}{s_n(0)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h' \dot{\ell}_{\theta_0,\eta_0}(X_i) - \frac{1}{2} h' I_{\theta_0,\eta_0} h + o_P(1).$$

This property is called the integral LAN property by Bickel and Kleijn (2012). Again, the score function in the right-hand side is the same as the one of the parametric model that knows the true η_0. The integral LAN property implies the asymptotic equivalence between parametric and semiparametric Bayesian methods.

4 Convolution Theorem

This section derives the efficiency bound by establishing the convolution theorem. There are some studies on the semiparametric efficiency bound for the moment restriction model. The seminal work by Chamberlain (1987) derived the local asymptotic minimax bound. Recently, Dovonon and Atchadé (2020) showed the convolution theorem as well as the minimax theorem based on the results of Begun et al. (1983). See also Severini and Tripathi (2001) for the derivation of the efficiency bound. Although the efficiency bound derived in this section is not new, our result is useful in understanding the connection between Bayesian point estimators and semiparametrically efficient frequentist estimators. In particular, the result is used to show that BEL estimators are asymptotically efficient in the frequentist sense.

As we will see below, our model satisfies two important properties. First, it satisfies the LAN property. Second, the score functions for θ_0 and η_0 are orthogonal to each other. These properties facilitate the derivation of the efficiency bound. Moreover, they are crucial to show that the BEL is asymptotically equivalent to parametric and semiparametric Bayesian methods.

A general convolution theorem states that if a statistical model satisfies the LAN property and a parameter of interest is differentiable, then the lower bound of the asymptotic variance for a class of regular estimators can be derived (see Theorem 3.11.2 of van der Vaart and

Wellner 1996). Thus, we shall show the LAN property of the moment restriction model and the differentiability of the parameter of interest. Besides, we define a class of regular estimators for the moment restriction model.

To show the LAN property, we consider a set of one-dimensional parametric models $t \mapsto P_t$, which are defined on $[0, t_\epsilon)$ for some $t_\epsilon > 0$ and satisfy $P_t \in \mathcal{P}$ and $P_0 = P$. Specifically, we consider parametric models of the form $P_t = P_{\theta_0 + th, \eta_t}$, where $h \in \mathbb{R}^p$ and a map $t \mapsto \eta_t$ is a perturbation of η_0 and coincides with η_0 at $t = 0$. Because we only need to investigate the limiting property of P_t as $t \to 0$, t_ϵ can be arbitrarily close to 0.

We next define the tangent set of the moment restriction model. Let p_t be the density of P_t with respect to ξ. For given h and η_t, suppose that there exists $\dot{\ell}_{\theta_0,\eta_0} : \mathcal{X} \to \mathbb{R}^p$ and $\dot{l} : \mathcal{X} \to \mathbb{R}$ such that

$$\int \left(\frac{\sqrt{p_t} - \sqrt{p}_0}{t} - \frac{1}{2}(h' \dot{\ell}_{\theta_0,\eta_0} + \dot{l}) \sqrt{p}_0 \right)^2 d\xi \to 0 \tag{4.1}$$

as $t \to 0$. Then, we say that the parametric submodel is Hellinger differentiable at $t = 0$ with score function $h' \dot{\ell}_{\theta_0,\eta_0} + \dot{l}$. A different choice of h and η_t yields a different score function. A set of possible score functions is called the tangent set and denoted by $\dot{\mathcal{P}}_P$. Moreover, a set of $\dot{l}$ is called the tangent set for η_0 and denoted by ${}_\eta\dot{\mathcal{P}}_P$.

To specify the tangent set, we find candidates for $\dot{\ell}_{\theta_0,\eta_0}$ and $\dot{l}$. They are typically given by

$$\dot{\ell}_{\theta_0,\eta_0}(x) = \left. \frac{\partial \log p_{\theta,\eta_0}(x)}{\partial \theta} \right|_{\theta=\theta_0} \quad \text{and} \quad \dot{l}(x) = \left. \frac{\partial \log p_{\theta_0,\eta_t}(x)}{\partial t} \right|_{t=0}.$$

Thus, $\dot{\ell}_{\theta_0,\eta_0}$ is the parametric score function for θ_0 when η_0 is known, while $\dot{l}$ is the score function for η_0 when θ_0 is known. However, the definition of the Hellinger differentiability does not necessarily require the existence of the above derivatives.

The derivative of $\log p_{\theta,\eta_0}$ can be obtained by a direct calculation. Let $\lambda(\theta) = \lambda_{\eta_0}(\theta)$. Then $\lambda(\theta)$ satisfies $\lambda(\theta_0) = 0$ because $P_{\theta_0,\eta_0} = \eta_0$. Moreover, because $\int m_\theta dP_{\theta,\eta_0} = 0$ for any $\theta \in \Theta$, the implicit function theorem yields

$$\frac{\partial \lambda(\theta_0)}{\partial \theta'} = \mathbb{E}[m_{\theta_0}(X) m_{\theta_0}(X)']^{-1} \mathbb{E}[\nabla m_{\theta_0}(X)],$$

where $\nabla m_\theta = \partial m_\theta / \partial \theta'$. Thus, we have

$$\left. \frac{\partial \log p_{\theta,\eta_0}(x)}{\partial \theta} \right|_{\theta=\theta_0} = -\mathbb{E}[\nabla m_{\theta_0}(X)]' \mathbb{E}[m_{\theta_0}(X) m_{\theta_0}(X)']^{-1} m_{\theta_0}(x). \tag{4.2}$$

The Fisher information is given by

$$I_{\theta_0,\eta_0} = \mathbb{E}[\nabla m_{\theta_0}(X)]' \mathbb{E}[m_{\theta_0}(X) m_{\theta_0}(X)']^{-1} \mathbb{E}[\nabla m_{\theta_0}(X)]. \tag{4.3}$$

Now we specify the tangent set for η_0. For any semiparametric model, the largest possible tangent set is the set of all zero mean L_2 functions on $\mathcal{X}$. The tangent set is determined by the feature of the semiparametric model. We specify ${}_\eta\dot{\mathcal{P}}_P$ as the following linear space of functions:

$${}_\eta\dot{\mathcal{P}}_P = \left\{ \dot{l} \in L_2(P) : \mathbb{E}[\dot{l}(X)] = 0 \text{ and } \mathbb{E}[\dot{l}(X) m_{\theta_0}(X)] = 0 \right\}. \tag{4.4}$$

The second restriction of (4.4) is implied by the construction of our model. Because P_{θ_0,η_t} satisfies $\int m_{\theta_0} dP_{\theta_0,\eta_t} = 0$ for any $t \in [0, t_\epsilon)$, differentiating both sides of the equation at $t = 0$ yields the second restriction. That is, if our model satisfies (4.1) for some η_t, then $\dot{l}$ must be an element of (4.4) as long as (2.1) is satisfied. Therefore, the set (4.4) is the largest tangent set for our model. Notice that the second restriction implies that all score functions in ${}_\eta\dot{\mathcal{P}}_P$ must be orthogonal to the score function for θ_0. Therefore, the score function (4.2) coincides with the efficient score function of the semiparametric model.

We can find a path $t \mapsto \eta_t$ that yields (4.4) as the tangent set. For instance, we can choose

$$\eta_t = (1 + t\dot{l})P \tag{4.5}$$

for a given $\dot{l} \in_\eta \dot{\mathcal{P}}_P$. Then, we have $\partial \log d\eta_t / \partial t|_{t=0} = \dot{l}$. Moreover, because η_t satisfies $\eta_t \in \mathcal{Q}_{\theta_0}$, we have $P_{\theta_0,\eta_t} = \eta_t$. Therefore, we also have $\partial \log p_{\theta_0,\eta_t} / \partial t|_{t=0} = \dot{l}$.

We now show the LAN property of the moment restriction model. Let $\dot{p}_{\theta,\eta} = \partial p_{\theta,\eta} / \partial \theta$. Also, let $\mathcal{N}_{\theta_0} \subset \Theta$ be a neighborhood of θ_0. We impose the following conditions.

Assumption 4.1

(i) For any $\dot{l} \in {}_\eta\dot{\mathcal{P}}_P$, there exists a path $t \mapsto \eta_t$ such that

$$\int \left(\frac{\sqrt{p_{\theta_0,\eta_t}} - \sqrt{p_{\theta_0,\eta_0}}}{t} - \frac{1}{2}\dot{l}\sqrt{p_{\theta_0,\eta_0}} \right)^2 d\xi \to 0$$

as $t \to 0$.

(ii) $p_{\theta_0,\eta_t}(x)$ and $\dot{p}_{\theta_0,\eta_t}(x)$ are continuous at $t = 0$ for all $x \in \mathcal{X}$.

(iii) $m_\theta(x)$ is continuously differentiable in $\theta \in \mathcal{N}_{\theta_0}$ for all $x \in \mathcal{X}$.

(iv) $\int m_\theta m_\theta' dP_{\theta,\eta_t}$ *is finite positive-definite for all* $t \in [0, t_\epsilon)$ *and all* $\theta \in \mathcal{N}_{\theta_0}$.

(v) $\int \frac{\dot{p}_{\theta_0+th,\eta_t}}{p_{\theta_0+th,\eta_t}} \frac{\dot{p}_{\theta_0+th,\eta_t}}{p_{\theta_0+th,\eta_t}}' dP_{\theta_0+th,\eta_t}$ *is well-defined for all* $t \in [0, t_\epsilon)$ *and all* $h \in \mathbb{R}^p$ *and is continuous at* $t = 0$.

Condition (i) requires that p_{θ_0,η_t} is Hellinger differentiable at $t = 0$. This condition is satisfied for η_t specified by (4.5) although other specifications are also fine. Conditions (iii) and (iv) imply that $\dot{p}_{\theta,\eta_t}$ exists for all $t \in [0, t_\epsilon)$ and $\theta \in \mathcal{N}_{\theta_0}$.

Lemma 4.1

Suppose that Assumption 4.1 holds. Then, for any $h \in \mathbb{R}^p$ *and* $\dot{l} \in {}_\eta\dot{\mathcal{P}}_P$, *there exists a path* $t \mapsto \eta_t$ *such that*

$$\int \left(\frac{\sqrt{p_{\theta_0+th,\eta_t}} - \sqrt{p_{\theta_0,\eta_0}}}{t} - \frac{1}{2}(h'\dot{\ell}_{\theta_0,\eta_0} + \dot{l})\sqrt{p_{\theta_0,\eta_0}} \right)^2 d\xi \to 0$$

as $t \to 0$, *where* $\dot{\ell}_{\theta_0,\eta_0}$ *is given by (4.2).*

The result of Lemma 4.1 is a sufficient condition for the LAN property of our model with respect to the tangent set $\dot{\mathcal{P}}_P = \text{lin}\, \dot{\ell}_{\theta_0,\eta_0} + {}_\eta\dot{\mathcal{P}}_P$, where lin denotes the linear span (see Lemma 25.14 of van der Vaart 1998). For $g \in \dot{\mathcal{P}}_P$, let $P_{t,g}$ denote the one-dimensional parametric model whose score function is g. Then, we have

$$\log \prod_{i=1}^n \frac{dP_{1/\sqrt{n},g}}{dP}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g(X_i) - \frac{1}{2}\mathbb{E}[g(X)^2] + o_P(1)$$

for all $g \in \dot{\mathcal{P}}_P$. If η_0 is known, then the expansion is the same as (3.2).

The Hellinger differentiability of P_t is also a sufficient condition for the differentiability of the parameter. Let ψ be a functional on $\mathcal{P}$ that satisfies $\psi(P_{\theta,\eta}) = \theta$. The functional ψ is differentiable at P relative to the tangent set $\dot{\mathcal{P}}_P$ if there exists a linear continuous map $\dot{\psi}_P : \dot{\mathcal{P}}_P \to \mathbb{R}^p$ such that

$$\frac{\psi(P_{t,g}) - \psi(P)}{t} \to \dot{\psi}_P g$$

as $t \to 0$ for all $g \in \dot{\mathcal{P}}_P$. Here, because $\dot{\psi}_P$ is a linear functional in the Hilbert space $L_2(P)$, by Riesz representation theorem, there exists $\tilde{\psi}_P \in L_2(P)$ such that $\dot{\psi}_P g = \mathbb{E}[\tilde{\psi}_P(X)g(X)]$ (see also equation (2.2) of Newey 1994). By Lemma 25.25 of van der Vaart (1998), our functional is differentiable with $\tilde{\psi}_P = I^{-1}_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta_0}$. The function $\tilde{\psi}_P$ is called the efficient influence function and its variance determines the efficiency bound.

Finally, we define a class of regular estimators. A sequence of estimators T_n is regular for estimating θ_0 with respect to $\dot{\mathcal{P}}_P$ if there exists a probability measure L such that

$$\sqrt{n}(T_n - \psi(P_{1/\sqrt{n},g})) \overset{P_{1/\sqrt{n},g}}{\rightsquigarrow} L \tag{4.6}$$

for all $g \in \dot{\mathcal{P}}_P$, where $\overset{P_{1/\sqrt{n},g}}{\rightsquigarrow}$ denotes convergence in distribution under $P_{1/\sqrt{n},g}$. The bottom line of the definition is that L does not depend on g. Thus, the regularity requires that a small change in the underlying distribution does not change the distribution of the estimator too much.

Theorem 4.1

Let T_n be a regular estimator that satisfies (4.6) for all $g \in \dot{\mathcal{P}}_P$. Suppose that Assumption 4.1 holds and I_{θ_0,η_0} is nonsingular. Then, there exists a probability measure M such that

$$L = N(0, I_{\theta_0,\eta_0}^{-1}) * M$$

where $*$ denotes the convolution of probability measures.

The result follows immediately from Theorem 25.20 of van der Vaart (1998). The proof of Theorem 4.1 is given in the supplemental material.

Theorem 4.1 states that the lower bound of the asymptotic variance of regular estimators is I_{θ_0,η_0}^{-1} because convolution never decreases the variance. Of course, I_{θ_0,η_0}^{-1} is the same as the efficiency bound derived by Chamberlain (1987). Notice that the efficiency bound is the same as that of the parametric model $\theta \mapsto P_{\theta,\eta_0}$. Thus, knowing the true nuisance parameter does not change the efficiency bound in the formulation of our model. This also implies that the parametric model $\theta \mapsto P_{\theta,\eta_0}$ is a least favorable model of the moment restriction model.

The asymptotic variance of the optimally weighted GMM and generalized EL estimators (Kitamura and Stutzer 1997; Smith 1997; Imbens et al. 1998) coincides with I_{θ_0,η_0}^{-1}. Therefore, these estimators are best regular if they satisfy (4.6). Because P and $P_{1/\sqrt{n},g}$ are contiguous by the LAN property, the asymptotic distribution of a statistic under $P_{1/\sqrt{n},g}$ can be derived from that of the statistic under P. Let g be given by $g = h'\dot{\ell}_{\theta_0,\eta_0} + \dot{l}$ for some $h \in \mathbb{R}^p$ and $\dot{l} \in {}_\eta\dot{\mathcal{P}}_P$. Because all above estimators satisfy

$$\sqrt{n}(T_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{\theta_0,\eta_0}^{-1} \dot{\ell}_{\theta_0,\eta_0}(X_i) + o_P(1), \tag{4.7}$$

by noting that $I_{\theta_0,\eta_0}^{-1}\mathbb{E}[\dot{\ell}_{\theta_0,\eta_0}(X)g(X)] = h$, the Le Cam's third lemma yields

$$\sqrt{n}(T_n - \theta_0) \overset{P_{1/\sqrt{n},g}}{\rightsquigarrow} N(h, I_{\theta_0,\eta_0}^{-1}).$$

Moreover, because $\psi(P_{1/\sqrt{n},g}) = \theta_0 + h/\sqrt{n}$, we see that (4.6) is satisfied. In fact, every best regular estimator satisfies (4.7). The proof is almost the same as that of Lemma 8.14 of van der Vaart (1998).

Remark 4.1

Andrews and Mikusheva (2022) also consider a Hellinger-differentiable model and specify their tangent set. However, their purpose is quite different from ours. They assume that the true marginal distribution of $X_1, \ldots, X_n$, denoted by $P_{n,f}$, satisfies

$$\int \left[\sqrt{n}\left(dP_{n,f}^{1/2} - dP_0^{1/2}\right) - \frac{1}{2}f dP_0^{1/2}\right]^2 \to 0$$

for some P_0 and a score f. Then, the process $\frac{1}{\sqrt{n}}\sum_{i=1}^n m_\theta(X_i)$ weakly converges to a Gaussian process with mean function $\mathbb{E}_{P_0}[f(X)m_\theta(X)]$. They consider a Bayesian decision rule by introducing a nonparametric prior on the mean function rather than on the distribution of X.

5 Bayesian Empirical Likelihood

Given $X_1, \ldots, X_n$, the BEL posterior for θ is obtained by

$$\Pi_n^{EL}(\theta \in B | X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n dP_{\theta,\mathbb{P}_n}(X_i)\pi_\Theta(\theta)d\theta}{\int_\Theta \prod_{i=1}^n dP_{\theta,\mathbb{P}_n}(X_i)\pi_\Theta(\theta)d\theta}, \tag{5.1}$$

where $P_{\theta,\mathbb{P}_n}$ satisfies

$$\prod_{i=1}^n \frac{dP_{\theta,\mathbb{P}_n}}{d\mathbb{P}_n}(X_i) = \prod_{i=1}^n \frac{1}{1 + \hat{\lambda}_n(\theta)' m_\theta(X_i)}$$

with $\hat{\lambda}_n(\theta) = \arg\max_{\lambda \in \mathbb{R}^l} n^{-1}\sum_{i=1}^n \log(1 + \lambda' m_\theta(X_i))$. In what follows, we assume that $\pi_\Theta(\theta)$ is continuous and strictly positive at θ_0.

The BEL removes the nuisance parameter by replacing it with the empirical distribution, whereas a legitimate semiparametric Bayesian procedure removes it by integration. Or, the BEL uses the profile likelihood function rather than the marginal likelihood function. A validity of the Bayesian method that uses a profile likelihood is discussed by Severini (1999) for instance.

Owen (2001) considered a possibility to justify the BEL by its resemblance to a parametric Bayesian method (Section 9.4 of Owen 2001). As stated in Section 4, a least favorable model of the moment restriction model is given by $\theta \mapsto P_{\theta,\eta_0}$. Thus, the EL approximates the likelihood of the least favorable model (see also DiCiccio and Romano 1990, Bertail 2006 and Sueishi 2016 for related issues). The parametric Bayesian inference using the likelihood of the least favorable model is a valid Bayesian procedure in the sense that it is based on a genuine likelihood function and the Bayes rule. The resulting posterior is a legitimate conditional probability. Because the posterior of the BEL is asymptotically equivalent to that of the parametric Bayesian method, the BEL asymptotically yields a valid posterior distribution. Now we substantiate this claim.

Assumption 5.1

(i) $m_\theta(x)$ *is continuously differentiable with respect to* $\theta \in \mathcal{N}_{\theta_0}$ *for all* $x \in \mathcal{X}$.

(ii) $\mathbb{E}[\sup_{\theta\in\mathcal{N}_{\theta_0}} \|m_\theta(X)\|^3] < \infty$.

(iii) $\mathbb{E}[\sup_{\theta\in\mathcal{N}_{\theta_0}} \|\nabla m_\theta(X)\|] < \infty$.

(iv) $\mathbb{E}[m_\theta(X)m_\theta(X)']$ *is positive definite uniformly over* $\theta \in \mathcal{N}_{\theta_0}$.

Under Assumption 5.1, we obtain

$$\sup_{h\in K}\left|\log\prod_{i=1}^{n}\frac{dP_{\theta_n(h),\mathbb{P}_n}}{dP_{\theta_0,\mathbb{P}_n}}(X_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}h'\dot{\ell}_{\theta_0,\eta_0}(X_i) + \frac{1}{2}h'I_{\theta_0,\eta_0}h\right| = o_P(1) \tag{5.2}$$

for any compact set $K \subset \mathbb{R}^p$. See Lemma A.1 in the Appendix. The result implies that the EL ratio is asymptotically equivalent to the log likelihood ratio of the least favorable model.

Let $\mathcal{N}_0 \subset \mathbb{R}^l$ be a neighborhood of 0.

Assumption 5.2

(i) For any $\delta > 0$, *there exist* $\epsilon > 0$ *such that*

$$\Pr\left(\sup_{\theta:\|\theta-\theta_0\|>\delta}\frac{1}{n}\sum_{i=1}^{n}\log\frac{dP_{\theta,\mathbb{P}_n}}{dP_{\theta_0,\mathbb{P}_n}}(X_i) \le -\epsilon\right) \to 1.$$

(ii) $\mathbb{E}\left[\sup_{\theta\in\mathcal{N}_{\theta_0}}\sup_{\lambda\in\mathcal{N}_0}\left\|\frac{m_\theta(X)}{1+\lambda'm_\theta(X)}\right\|^2\right] < \infty$ *and* $\mathbb{E}\left[\sup_{\theta\in\mathcal{N}_{\theta_0}}\sup_{\lambda\in\mathcal{N}_0}\left\|\frac{\nabla m_\theta(X)}{1+\lambda'm_\theta(X)}\right\|^2\right] < \infty$.

(iii) $\mathbb{E}[\nabla m_{\theta_0}(X)]$ *is of full column rank.*

Assumption 5.2 (i) is essentially the same as (2.4) of Chib et al. (2018) (see also condition (B.3) in page 489 of Lehmann and Casella 1998 and Assumption 3 of Chernozhukov and Hong 2003). It requires that θ_0 is asymptotically the global maximizer of $n^{-1}\sum_{i=1}^{n}\log dP_{\theta,\mathbb{P}_n}(X_i)$. Under Assumptions 5.1 and 5.2, we have

$$\Pi_n^{EL}(\sqrt{n}\|\theta-\theta_0\|>M_n|X_1,\ldots,X_n)\xrightarrow{P}0 \tag{5.3}$$

for any sequence $\{M_n\}$ such that $M_n\to\infty$. That is, the BEL posterior converges to θ_0 at $n^{-1/2}$-rate. See Lemma A.2 in the Appendix. Notice that (5.3) is a necessary condition for the BvM theorem.

Theorem 5.1

Suppose that Assumptions 5.1 and 5.2 hold. Then we have

$$\sup_B\left|\Pi_n^{EL}(\sqrt{n}(\theta-\theta_0)\in B|X_1,\ldots,X_n)-N_{\Delta_n,I_{\theta_0,\eta_0}^{-1}}(B)\right|\xrightarrow{P}0,$$

where

$$\Delta_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}I_{\theta_0,\eta_0}^{-1}\dot{\ell}_{\theta_0,\eta_0}(X_i).$$

The supremum is taken over all measurable sets in Θ.

Under conditions (5.2) and (5.3), the result follows from the proof of Theorem 2.1 of Kleijn and van der Vaart (2012).

Theorem 5.1 shows that the limiting posterior of the BEL is the same as that of the parametric Bayesian method that uses the likelihood of the least favorable model, although the parametric Bayesian method is infeasible (see, for instance, Theorem 10.1 of van der Vaart 1998 for the parametric BvM theorem). This result provides an asymptotic justification for the BEL because the parametric Bayesian method is a valid Bayesian procedure.

The BEL is also valid as a frequentist inference method. Let $\hat{\theta}_n$ be a best regular estimator for θ_0. Because $\hat{\theta}_n$ satisfies $\sqrt{n}(\hat{\theta}_n-\theta_0)=\Delta_n+o_P(1)$ by the result of Section 4, Theorem 5.1 can be alternatively stated as

$$\sup_B\left|\Pi_n^{EL}(\theta\in B|X_1,\ldots,X_n)-N_{\hat{\theta}_n,(nI_{\theta_0,\eta_0})^{-1}}(B)\right|\xrightarrow{P}0.$$

Because the variance matrix of the normal distribution is the same as the asymptotic variance matrix of $\hat{\theta}_n$, a $1-\alpha$ credible set $\hat{B}_n$, which satisfies $\Pi_n^{EL}(\theta\in\hat{B}_n|X_1,\ldots,X_n)=1-\alpha$, is

asymptotically equivalent to the Wald-type $1-\alpha$ confidence set that is constructed on the basis of the asymptotic normality of $\hat{\theta}_n$.

Theorem 5.1 also implies that a center of the BEL posterior is asymptotically equivalent to best regular estimators of the moment restriction model. For instance, a point estimator that solves

$$\min_t \int \|t-\theta\| d\Pi_n^{EL}(\theta|X_1,\ldots,X_n)$$

is asymptotically efficient if $\int \|\theta\|\pi_\Theta(\theta)d\theta < \infty$. In particular, it is asymptotically equivalent to the EL estimator. Furthermore, if the prior and posterior distributions satisfy some additional conditions, then the posterior mean is also asymptotically equivalent to the EL estimator (see Theorem 2.3 of Kleijn and van der Vaart 2012, for instance).

Theorem 5.1 looks similar to Theorem 1 of Chib et al. (2018), who investigated the asymptotic properties of the BETEL posterior. However, the statement of their theorem is different from ours. Theorem 1 of Chib et al. (2018) states that

$$\sup_B \left| \Pi_n^{ETEL}(\sqrt{n}(\theta-\theta_0) \in B|X_1,\ldots,X_n) - N_{0,I^{-1}_{\theta_0,\eta_0}}(B) \right| \xrightarrow{P} 0,$$

where $\Pi_n^{ETEL}(\cdot|X_1,\ldots,X_n)$ denotes the posterior of the BETEL. Notice that the center of the normal distribution is 0, not Δ_n. Thus, the asymptotic equivalence between the Bayesian point estimator and the best regular estimator is not established.

Chernozhukov and Hong (2003) proposed the LTE, which is a quasi-Bayesian estimator that uses a general statistical criterion function in place of the parametric likelihood function. The BEL is obtained as a special case of the LTE if the EL is used as the criterion function. In fact, our theorem is similar to Theorem 1 and Proposition 2 of Chernozhukov and Hong (2003). However, they did not investigate the connection between the LTE and a genuine Bayesian estimator. Our finding gives a new insight into the LTE.

6 Semiparametric Baysian Method

This section considers the semiparametric Bayesian method that obtains the marginal posterior of θ by

$$\Pi_n(\theta \in B|X_1,\ldots,X_n) = \frac{\int_B \int_H \prod_{i=1}^n p_{\theta,\eta}(X_i) d\Pi_H(\eta)\pi_\Theta(\theta)d\theta}{\int_\Theta \int_H \prod_{i=1}^n p_{\theta,\eta}(X_i) d\Pi_H(\eta)\pi_\Theta(\theta)d\theta},$$

where Π_H is a prior of η. The semiparametric Bayesian method is a valid Bayesian procedure in that it obtains the marginal posterior based on the Bayes rule. We show that the posterior of the semiparametric Bayesian method converges to the same limit as that of the BEL. Thus, our goal is similar to that of Schennach (2005), who showed the existence of a Bayesian method whose limit is equivalent to the BETEL.

Schennach (2005) considered a model $\xi_N \mapsto P_{\xi_N}$ that has an N-dimensional nuisance parameter ξ_N and satisfies $\int m_\theta dP_{\xi_N} = 0$ for all $\theta \in \Theta$ and $\xi_N \in \Xi_N$. Given a conditional prior density $\pi_{\Xi_N|\Theta}(\xi_N|\theta)$, the posterior density of θ is proportional to

$$\int_{\Xi_N} \prod_{i=1}^{n} p_{\xi_N}(X_i)\pi_{\Xi_N|\Theta}(\xi_N|\theta)\pi_\Theta(\theta)d\xi_N. \tag{6.1}$$

Notice that the likelihood function is not an explicit function of θ. The parameter θ appears only in the priors. Schennach (2005) showed that the posterior obtained by (6.1) converges to that of the BETEL as $N \to \infty$ for a special choice of p_{ξ_N} and $\pi_{\Xi_N|\Theta}$.

There are two main differences between our approach and her approach. First, we take the limit with respect to the sample size n for fixed prior Π_H, whereas she takes the limit with respect to the dimension of the nuisance parameter N for fixed sample size n. Second, our result does not depend on the choice of priors, whereas the choice of conditional prior is the key to her result.

The proof of our result is based on Bickel and Kleijn (2012) and Chae (2015). Bickel and Kleijn (2012) showed the BvM theorem for general semiparametric models, while Chae (2015) considered the case where an adaptive estimation is possible. By virtue of our formulation of the model, the moment restriction model can be viewed as a special case of generic semiparametric models.

Let $d_H(P, P')$ denote the Hellinger distance between probability measures P and P', and let $d_\theta(Q_1, Q_2) = d_H(\frac{p_{\theta_0,\eta_0}}{p_{\theta,\eta_0}}Q_1, \frac{p_{\theta_0,\eta_0}}{p_{\theta,\eta_0}}Q_2)$ for any $Q_1, Q_2 \in \mathcal{Q}_\theta$; that is, d_θ is a weighted Hellinger distance between Q_1 and Q_2. Let $N(\epsilon, \mathcal{Q}_\theta, d_\theta)$ denote the ϵ-covering number of $\mathcal{Q}_\theta$ with respect to the metric d_θ. Because $\mathcal{Q}_\theta$ is convex, $N(\epsilon, \mathcal{Q}_\theta, d_\theta)$ gives a bound for $N_t(\epsilon, \mathcal{Q}_\theta, d_H; P, P_{\theta,\eta_0})$, which is the covering number for testing under misspecification (Kleijn and van der Vaart 2006). We define

$$B(\epsilon, P_{\theta,\eta_0}; P) = \left\{ \eta \in H : -\int \log \frac{dP_{\theta,\eta}}{dP_{\theta,\eta_0}} dP \leq \epsilon^2, \int \left(\log \frac{dP_{\theta,\eta}}{dP_{\theta,\eta_0}} \right)^2 dP \leq \epsilon^2 \right\}.$$

Assumption 6.1

For any $\epsilon > 0$, we have (i) $\inf_{\theta \in \mathcal{N}_{\theta_0}} \Pi_H(B(\epsilon, P_{\theta,\eta_0}; P)) > 0$ and (ii) $\sup_{\theta \in \mathcal{N}_{\theta_0}} N(\epsilon, \mathcal{Q}_\theta, d_\theta) < \infty$.

Assumption 6.1 is imposed for the convergence of the nonparametric posterior. Condition (i) requires that the prior Π_H put a mass on a neighborhood of η_0, while condition (ii) requires that the parameter space H is not too large.

It is known that the conditional posterior of η given θ concentrates on the point $\eta^*(\theta)$ that minimizes the KL divergence between $P_{\theta,\eta}$ and P. Here, because $\eta_0 = P$, our model satisfies

$$-\int \log \frac{dP_{\theta,\eta_0}}{dP} dP = \inf_{\eta \in H} -\int \log \frac{dP_{\theta,\eta}}{dP} dP.$$

That is, we have $\eta^*(\theta) = \eta_0$ for any θ. Thus, for any given θ, the conditional posterior of $P_{\theta,\eta}$ converges to P_{θ,η_0}. Assumption 6.1 implies that there exists a set $H_n \subset H$ such that

$$\sup_{\theta \in \mathcal{N}_{\theta_0}} \Pi_n(H_n^C | \theta, X_1, \ldots, X_n) \xrightarrow{P} 0 \tag{6.2}$$

and $\sup_{\theta \in \mathcal{N}_{\theta_0}} \sup_{\eta \in H_n} d_H(P_{\theta,\eta}, P_{\theta,\eta_0}) = o(1)$, where $\Pi_n(\cdot | \theta, X_1, \ldots, X_n)$ denotes the conditional posterior of η given θ. See Corollary 2.1 of Kleijn and van der Vaart (2006). We do not need to specify the convergence rate of the nonparametric posterior.

Remark 6.1

Under certain conditions, some existing nonparametric priors satisfy Assumption 6.1 (i). Let $K(P,\eta) = -\int \log \frac{d\eta}{dP} dP$ and $V(P,\eta) = \int (\log \frac{d\eta}{dP})^2 dP$. Consider the following four sets:

$$\begin{aligned}
A_1(\epsilon) &= \{\eta \in H : K(P,\eta) < \epsilon\} \\
A_2(\epsilon) &= \{\eta \in H : V(P,\eta) < \epsilon\} \\
A_3(\epsilon) &= \left\{\eta \in H : -\int \log \frac{1 + \lambda(\theta)' m_\theta}{1 + \lambda_\eta(\theta)' m_\theta} dP < \epsilon\right\} \\
A_4(\epsilon) &= \left\{\eta \in H : \int \left(\log \frac{1 + \lambda(\theta)' m_\theta}{1 + \lambda_\eta(\theta)' m_\theta}\right)^2 dP < \epsilon\right\}.
\end{aligned}$$

If a prior Π_H satisfies $\Pi_H(A_j(\epsilon)) > 0$ for any $\epsilon > 0$ and $j = 1, \ldots, 4$, then $\Pi_H(B(\epsilon, P_{\theta,\eta_0}; P)) > 0$ for any $\epsilon > 0$.

A nonparametric prior that satisfies $\Pi_H(A_1(\epsilon)) > 0$ for any $\epsilon > 0$ is called a KL prior. Some kernel mixture priors satisfy this condition when H is dominated by the Lebesgue measure (Wu and Ghosal 2008). For instance, we may use a Dirichlet process mixture, which is employed by Shin (2015). An appropriate choice of the kernel depends on the support of X.

A prior that satisfies $\Pi_H(A_2(\epsilon)) > 0$ is not known well. However, if $\frac{dP}{d\eta}$ is bounded for all $\eta \in H$, then we have $V(P,\eta) \leq 2K(P,\eta) \left\| \frac{dP}{d\eta} \right\|_\infty$ (see Lemmas B.1 and B.2 of Ghosal and van der Vaart (2017) for instance). Thus, the KL prior also satisfies $\Pi_H(A_2(\epsilon)) > 0$ for any $\epsilon > 0$.

By a Taylor expansion, the set $A_3(\epsilon)$ can be rewritten as

$$A_3(\epsilon) = \left\{ \eta \in H : -\frac{1}{2}(\lambda_\eta(\theta) - \lambda(\theta))' \left(\int \frac{m_\theta m_\theta'}{(1 + \bar{\lambda}_\eta(\theta)' m_\theta)^2} dP \right) (\lambda_\eta(\theta) - \lambda(\theta)) < \epsilon \right\},$$

where $\bar{\lambda}_\eta(\theta)$ is located between $\lambda_\eta(\theta)$ and $\lambda(\theta)$. Thus, $\Pi_H(A_3(\epsilon)) > 0$ is clearly satisfied. Moreover, under certain conditions, we have $\|\lambda_\eta(\theta) - \lambda(\theta)\| \to 0$ as $d_H(\eta, P) \to 0$. Thus, by using a similar expansion as in the case of $A_3(\epsilon)$, we see that the set $A_4(\epsilon)$ contains a set of probability measures that is close to P in terms of the Hellinger distance. Because the KL prior puts a mass on any Hellinger ball centered at P, we obtain $\Pi_H(A_4(\epsilon)) > 0$.

Let $\mathcal{M} \subset \mathcal{P}$ be a neighborhood of P with respect to the Hellinger distance. Also, let $m_{j,\theta}$ denote the j-th element of m_θ.

Assumption 6.2

(i) $m_\theta(x)$ *is twice continuously differentiable with respect to* $\theta \in \mathcal{N}_{\theta_0}$ *for all* $x \in \mathcal{X}$.

(ii) $\int \frac{m_\theta m_\theta'}{(1 + \lambda_\eta(\theta)' m_\theta)^2} d\eta$ *is positive definite uniformly over* $\theta \in \mathcal{N}_{\theta_0}$ *and* $\eta \in H_n$.

(iii) $\sup_{\theta \in \Theta} \sup_{\lambda \in \mathcal{N}_0} \int \left\| \frac{m}{1 + \lambda' m_\theta} \right\|^4 dQ < \infty$ *and* $\sup_{\theta \in \mathcal{N}_{\theta_0}} \sup_{\lambda \in \mathcal{N}_0} \int \left\| \frac{\nabla m_\theta}{1 + \lambda' m_\theta} \right\|^4 dQ < \infty$ *for all* $Q \in \mathcal{M}$.

(iv) $\mathbb{E}\left[\sup_{\theta \in \mathcal{N}_{\theta_0}} \sup_{\lambda \in \mathcal{N}_0} \left\| \frac{m_\theta(X)}{1 + \lambda' m_\theta(X)} \right\|^4 \right] < \infty$, $\mathbb{E}\left[\sup_{\theta \in \mathcal{N}_{\theta_0}} \sup_{\lambda \in \mathcal{N}_0} \left\| \frac{\nabla m_\theta(X)}{1 + \lambda' m_\theta(X)} \right\|^4 \right] < \infty$, *and* $\mathbb{E}\left[\sup_{\theta \in \mathcal{N}_{\theta_0}} \sup_{\lambda \in \mathcal{N}_0} \left\| \frac{\partial^2 m_{j,\theta}(X)/\partial\theta\partial\theta'}{1 + \lambda' m_\theta(X)} \right\| \right] < \infty$ *for* $j = 1, \ldots, l$.

Assumption 6.2 implies that

$$\sup_{h \in K} \sup_{\eta \in H_n} \left| \log \prod_{i=1}^n \frac{p_{\theta_n(h),\eta}}{p_{\theta_0,\eta}}(X_i) - \frac{1}{\sqrt{n}} \sum_{i=1}^n h' \dot{\ell}_{\theta_0,\eta_0}(X_i) + \frac{1}{2} h' I_{\theta_0,\eta_0} h \right| = o_P(1) \tag{6.3}$$

for any compact set $K \subset \mathbb{R}^p$. See Lemma A.4 in the Appendix. Combining (6.2) and (6.3), we can show the integral LAN property of our model. That is, for any bounded stochastic sequence $\{h_n\}$, we obtain

$$\log \frac{s_n(h_n)}{s_n(0)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n h_n' \dot{\ell}_{\theta_0,\eta_0}(X_i) - \frac{1}{2} h_n' I_{\theta_0,\eta_0} h_n + o_P(1), \tag{6.4}$$

where $s_n(h)$ is defined by (3.4). See Lemma A.5 in the Appendix.

As state in Section 5, a necessary condition for the BvM theorem is

$$\Pi_n(\sqrt{n}\|\theta-\theta_0\| > M_n|X_1,\ldots,X_n) \xrightarrow{P} 0 \tag{6.5}$$

for any $M_n \to \infty$. A difficulty to establish (6.5) in the semiparametric case is that global behavior of the likelihood function must be restricted in a certain way because the integral with respect to η is taken over the whole parameter space H to obtain the marginal distribution.

Assumption 6.3

(i) For any $\delta > 0$, there exists $\epsilon > 0$ such that

$$\Pr\left(\sup_{\theta:\|\theta-\theta_0\|>\delta}\sup_{\eta\in H}\frac{1}{n}\sum_{i=1}^{n}\log\frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(X_i) \le -\epsilon\right) \to 1.$$

(ii) $\mathbb{E}[\nabla m_{\theta_0}(X)]$ is of full column rank.

Assumption 6.3 (i) is similar to the assumption for Lemma 6.1 of Bickel and Kleijn (2012). It can also be viewed as a stronger version of Assumption 5.2 (i). The assumption restricts the behavior of the likelihood over the whole parameter space H. Lemma A.6 in the Appendix shows that (6.5) is satisfied under Assumptions 6.1–6.3.

Once (6.4) and (6.5) are established, the following theorem follows immediately from Theorem 5.1 of Bickel and Kleijn (2012).

Theorem 6.1

Suppose that Assumptions 6.1–6.3 hold. Then, for any measurable set $B \subset \Theta$, we have

$$\sup_B\left|\Pi_n(\sqrt{n}(\theta-\theta_0)\in B|X_1,\ldots,X_n) - N_{\Delta_n, I^{-1}_{\theta_0,\eta_0}}(B)\right| \xrightarrow{P} 0,$$

where

$$\Delta_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I^{-1}_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta_0}(X_i).$$

The supremum is taken over all measurable sets in Θ.

Comparing Theorems 5.1 and 6.1, we see that the limiting posterior distribution of the semiparametric Bayesian method is the same as that of the BEL regardless of the choice of parametric and nonparametric priors. Moreover, a center of the posterior of the semiparametric Bayesian method is also asymptotically efficient in the frequentist sense. Because the

computational burden of the semiparametric method is much heavier than that of the BEL, if one only cares about asymptotic properties, the BEL will be more attractive than the semiparametric Bayesian method.

Florens and Simoni (2021) obtained a similar asymptotic result for their semiparametric method. They used a Gaussian process as a conditional prior for the nuisance parameter and showed that their posterior for θ converges to the same limiting distribution as ours. Moreover, they showed that the limiting distribution is the same as the one obtained from the quasi-Bayesian method based on the limited information likelihood of Kim (2002). However, Florens and Simoni (2021) avoided obtaining the marginal posterior by integrating out the nuisance parameter. In fact, they did not specify the likelihood function of the semiparametric model. The posterior distribution for θ is obtained by conditioning on a statistic r_n, rather than by conditioning on the sample $X_1, \ldots, X_n$. Thus, our method to obtain the posterior distribution for θ is quite different from theirs.

7 Conclusion

This paper investigated large sample properties of the posterior of the BEL. We showed that the BEL posterior is asymptotically equivalent to the posterior of the parametric Bayesian method that uses the likelihood of the least favorable model. Moreover, the BEL posterior is asymptotically equivalent to the marginal posterior of the semiparametric Bayesian method that places both finite- and infinite dimensional priors.

This paper also showed the convolution theorem for the moment restriction model. It is commonly said that the EL estimator is asymptotically efficient because its asymptotic variance attains the efficiency bound of Chamberlain (1987). To the best of my knowledge, however, the regularity of the EL estimator has not been considered in the literature because a class of regular estimators depends on the tangent set of the semiparametric model. Our derivation of the efficiency bound is useful to understand the asymptotic efficiency of the EL estimator.

The novelty of this study is in the formulation of our model. Once the moment restriction model is expressed as a set of probability measures that are indexed by the finite-dimensional

parameter of interest and the infinite-dimensional nuisance parameter, we can apply existing theories of semiparametric estimation. The orthogonality of the score functions for θ_0 and η_0 further simplifies the asymptotic analysis.

The BEL also provides a valid inference procedure in the frequentist sense. The BvM theorem reveals that the BEL $1-\alpha$ credible set is asymptotically equivalent to the Wald-type $1-\alpha$ confidence set based on the EL. Moreover, some point estimators of the BEL are asymptotically efficient. The BEL point estimator may be computationally more tractable than the EL estimator because it can be obtained rather easily by MCMC, whereas finding the maximum or minimum is often difficult when the objective function has many local maxima or minima.

A Appendix

Throughout the Appendix, C denotes a generic positive constant which may vary according to context.

Proof of Proposition 2.1 We have $\inf_{Q\in\mathcal{Q}_\theta} -\int \log \frac{dQ}{d\eta} d\eta < \infty$ by Assumption 2.1 (ii). Thus, it follows from Corollary 2.6 of Borwein and Lewis (1991) that

$$\inf_{Q\in\mathcal{Q}_\theta} -\int \log \frac{dQ}{d\eta} d\eta = \gamma_\eta^*(\theta) - \int \phi_+^*(\gamma_\eta^*(\theta) + \lambda_\eta^*(\theta)' m_\theta) d\eta.$$

Because $\lim_{u\to\infty} -\log u/u = 0$, Theorem 4.8 of Borwein and Lewis (1991) implies that the unique solution $P_{\theta,\eta}$ satisfies

$$\frac{dP_{\theta,\eta}}{d\eta} = -\frac{1}{\gamma_\eta^*(\theta) + \lambda_\eta^*(\theta) m_\theta}.$$

Here, because $P_{\theta,\eta}$ satisfies $\int m_\theta dP_{\theta,\eta} = 0$, we have $\int \frac{\gamma_\eta^*(\theta)}{\gamma_\eta^*(\theta)+\lambda_\eta^*(\theta)' m_\theta} d\eta = 1$. Thus, we have $\gamma_\eta^*(\theta) = -1$ and hence

$$\frac{dP_{\theta,\eta}}{d\eta} = \frac{1}{1+\lambda_\eta(\theta)' m_\theta},$$

where $\lambda_\eta(\theta) = \arg\max_{\lambda\in\mathbb{R}^l} \int \log(1+\lambda' m_\theta) d\eta$. □

Proof of Lemma 4.1 The proof is based on Pollard (2010) although his definition of Hellinger differentiability is slightly different from ours.

The implicit function theorem implies that $\lambda_{\eta_t}(\theta)$ is continuously differentiable in θ for all t. Thus $\dot{p}_{\theta,\eta_t}$ is well defined and is continuous in θ. Let $\theta_t = \theta_0 + th$ and $s_{\theta_t,\eta_t} = \sqrt{p_{\theta_t,\eta_t}}$. Moreover, we define

$$\dot{s}_{\theta_t,\eta_t} = \frac{1}{2}\frac{\dot{p}_{\theta_t,\eta_t}}{p_{\theta_t,\eta_t}} s_{\theta_t,\eta_t},$$

where $\dot{p}_{\theta_t,\eta_t}/p_{\theta_t,\eta_t}$ is defined arbitrarily if $p_{\theta_t,\eta_t} = 0$. Thus, $\dot{s}_{\theta_t,\eta_t} = 0$ when $s_{\theta_t,\eta_t} = 0$.

For any $\delta > 0$, $p_{\theta_t,\eta_t} + \delta$ is bounded away from zero. Therefore, $\sqrt{p_{\theta_t,\eta_t} + \delta}$ satisfies

$$\sqrt{p_{\theta_t,\eta_t} + \delta} - \sqrt{p_{\theta_0,\eta_t} + \delta} = \frac{1}{2}\int_0^t h' \frac{\dot{p}_{\theta_u,\eta_t}}{\sqrt{p_{\theta_u,\eta_t} + \delta}} du.$$

If $p_{\theta_u,\eta_t} > 0$, the integrand of the right-hand side converges to $\dot{p}_{\theta_u,\eta_t}/\sqrt{p_{\theta_u,\eta_t}}$ as $\delta \to 0$. On the other hand, if $p_{\theta_u,\eta_t} = 0$, we have $\dot{p}_{\theta_u,\eta_t} = 0$ because p_{θ,η_t} is nonnegative for all θ. Thus, the integrand is also zero, and we obtain

$$s_{\theta_t,\eta_t} - s_{\theta_0,\eta_t} = \int_0^t h' \dot{s}_{\theta_u,\eta_t} du.$$

for all $x \in \mathcal{X}$. Moreover, by applying the Jensen's inequality to the uniform distribution on $[0, t]$, we have

$$\int \left(\frac{s_{\theta_t,\eta_t} - s_{\theta_0,\eta_t}}{t}\right)^2 d\xi \le \frac{1}{t}\int_0^t \int (h'\dot{s}_{\theta_u,\eta_t})^2 d\xi du \to \frac{1}{4} h' I_{\theta_0,\eta_0} h \tag{A.1}$$

as $t \to 0$.

Let $r_t = (s_{\theta_t,\eta_t} - s_{\theta_0,\eta_t})/t - h'\dot{s}_{\theta_0,\eta_0}$ and let

$$g_t = 2\left(\frac{s_{\theta_t,\eta_t} - s_{\theta_0,\eta_t}}{t}\right)^2 + 2(h'\dot{s}_{\theta_0,\eta_0})^2 - r_t^2.$$

For x such that $s_{\theta_0,\eta_0}(x) > 0$, s_{θ_t,η_0} is differentiable at $t = 0$. Therefore, we have $r_t \to 0$ and $g_t \to 4h'\dot{s}_{\theta_0,\eta_0}\dot{s}'_{\theta_0,\eta_0} h$ as $t \to 0$. On the other hand, for x such that $s_{\theta_0,\eta_0}(x) = 0$, we have $r_t = (s_{\theta_t,\eta_t} - s_{\theta_0,\eta_t})/t$ and $g_t \ge 0$. Thus, $\liminf_{t\to 0} g_t \ge 4h'\dot{s}_{\theta_0,\eta_0}\dot{s}'_{\theta_0,\eta_0} h$ for all $x \in \mathcal{X}$. Then, by the Fatou's lemma and (A.1), we obtain

$$h' I_{\theta_0,\eta_0} h \le \liminf_{t\to 0} \int g_t d\xi \le h' I_{\theta_0,\eta_0} h - \limsup_{t\to 0} \int r_t^2 d\xi$$

and hence $\int r_t^2 d\xi \to 0$. Because $\dot{\ell}_{\theta_0,\eta_0} = \dot{p}_{\theta_0,\eta_0}/p_{\theta_0,\eta_0}$ for x such that $s_{\theta_0,\eta_0}(x) > 0$, we obtain the desired result. □

Lemma A.1

Suppose that Assumption 5.1 holds. Then, we have

$$\sup_{h\in K}\left|\log\prod_{i=1}^{n}\frac{dP_{\theta_n(h),\mathbb{P}_n}}{dP_{\theta_0,\mathbb{P}_n}}(X_i)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}h'\dot{\ell}_{\theta_0,\eta_0}(X_i)+\frac{1}{2}h'I_{\theta_0,\eta_0}h\right|=o_P(1)$$

for any compact $K\subset\mathbb{R}^p$.

Proof We denote $m_{\theta,i}=m_\theta(X_i)$, $\bar{m}_\theta=n^{-1}\sum_{i=1}^n m_{\theta,i}$, and $\nabla\bar{m}_\theta=n^{-1}\sum_{i=1}^n\nabla m_\theta(X_i)$. Because $\sup_{h\in K}\|\bar{m}_{\theta_n(h)}\|=O_P(n^{-1/2})$, by modifying the proof of Lemma A.2 of Newey and Smith (2004), we have $\sup_{h\in K}\|\hat{\lambda}_n(\theta_n(h))\|=O_P(n^{-1/2})$. Thus, it follows that

$$\max_{1\le i\le n}\sup_{h\in K}|\hat{\lambda}_n(\theta_n(h))'m_{\theta_n(h),i}|=O_P(n^{-1/2})o_P(n^{1/2})=o_P(1)$$

by Assumption 5.1 (ii). Moreover, following Owen (1990) and the proof of Lemma 1 of Qin and Lawless (1994), we obtain

$$\sup_{h\in K}\left|\hat{\lambda}_n(\theta_n(h))-\left(\frac{1}{n}\sum_{i=1}^{n}m_{\theta_n(h),i}m'_{\theta_n(h),i}\right)^{-1}\bar{m}_{\theta_n(h)}\right|=o_P(n^{-1/2}).$$

Therefore, a Taylor expansion yields

$$\begin{aligned}
&\sum_{i=1}^{n}\log(1+\hat{\lambda}_n(\theta_n(h))'m_{\theta_n(h),i})\\
&\quad=\quad\hat{\lambda}_n(\theta_n(h))'\sum_{i=1}^{n}m_{\theta_n(h),i}-\frac{1}{2}\hat{\lambda}_n(\theta_n(h))'\sum_{i=1}^{n}m_{\theta_n(h),i}m'_{\theta_n(h),i}\hat{\lambda}_n(\theta_n(h))+o_P(1)\\
&\quad=\quad\frac{n}{2}\left\{\bar{m}_{\theta_0}+\frac{1}{\sqrt{n}}\nabla\bar{m}_{\bar{\theta}_n}h\right\}'\left(\frac{1}{n}\sum_{i=1}^{n}m_{\theta_n(h),i}m'_{\theta_n(h),i}\right)^{-1}\left\{\bar{m}_{\theta_0}+\frac{1}{\sqrt{n}}\nabla\bar{m}_{\bar{\theta}_n}h\right\}+o_P(1),
\end{aligned}$$

where $\bar{\theta}_n$ is located between θ_0 and $\theta_n(h)$. The small order term is uniform over $h\in K$. Similarly, we have

$$\sum_{i=1}^{n}\log(1+\hat{\lambda}_n(\theta_0)'m_{\theta_0,i})=\frac{n}{2}\bar{m}'_{\theta_0}\left(\frac{1}{n}\sum_{i=1}^{n}m_{\theta_0,i}m'_{\theta_0,i}\right)^{-1}\bar{m}_{\theta_0}+o_P(1).$$

Because $\dot{\ell}_{\theta_0,\eta_0}$ and I_{θ_0,η_0} are given by (4.2) and (4.3), respectively, the uniform law of large numbers yields the desired result. □

Lemma A.2

Suppose that Assumptions 5.1 and 5.2 hold. Then we have

$$\Pi_n^{EL}(\sqrt{n}\|\theta-\theta_0\|>M_n|X_1,\ldots,X_n)\xrightarrow{P}0$$

for any sequence $\{M_n\}$ such that $M_n\to\infty$.

Proof We define two sequences of events

$$A_n = \left\{ \sup_{\theta:\|\theta-\theta_0\|>\delta} \frac{1}{n}\sum_{i=1}^n \log \frac{dP_{\theta,\mathbb{P}_n}}{dP_{\theta_0,\mathbb{P}_n}}(X_i) \leq -\epsilon \right\}$$

and

$$B_n = \left\{ \int_\Theta \prod_{i=1}^n \frac{dP_{\theta,\mathbb{P}_n}}{dP_{\theta_0,\mathbb{P}_n}}(X_i)\pi_\Theta(\theta)d\theta \geq e^{-n\epsilon/2} \right\}.$$

Because of Lemma A.1, by modifying Lemma E.3 of Chib et al. (2018), we can show that $\Pr(B_n) \to 1$ for any $\epsilon > 0$. Therefore, by Assumption 5.2 (i), we obtain

$$\begin{aligned} \mathbb{E}\left[\Pi_n^{EL}(\|\theta-\theta_0\|>\delta|X_1,\ldots,X_n)\right] &\leq \mathbb{E}\left[\Pi_n^{EL}(\|\theta-\theta_0\|>\delta|X_1,\ldots,X_n)1_{A_n\cap B_n}\right] + o(1) \\ &\leq e^{n\epsilon/2}\mathbb{E}\left[\int_{\theta:\|\theta-\theta_0\|>\delta}\prod_{i=1}^n \frac{dP_{\theta,\mathbb{P}_n}}{dP_{\theta_0,\mathbb{P}_n}}(X_i)1_{A_n}\pi_\Theta(\theta)d\theta\right] + o(1) \\ &= o(1) \end{aligned} \tag{A.2}$$

for any $\delta > 0$.

Let $\Theta_n = \{\theta \in \Theta : M_n/\sqrt{n} < \|\theta-\theta_0\| \leq \delta\}$ with M_n such that $M_n \to \infty$ and $M_n/\sqrt{n} \to 0$. Now we show that there exists $C > 0$ such that

$$\Pr\left(\sup_{\theta\in\Theta_n} \frac{1}{n}\sum_{i=1}^n \log \frac{dP_{\theta,\mathbb{P}_n}}{dP_{\theta_0,\mathbb{P}_n}}(X_i) \leq -C\frac{M_n^2}{n} \right) \to 1. \tag{A.3}$$

Here, by Assumption 5.2 (i), there exists $C > 0$ such that

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n \log \frac{dP_{\theta,\mathbb{P}_n}}{dP_{\theta_0,\mathbb{P}_n}}(X_i) \leq -C\frac{M_n^2}{n} \right) \to 1$$

for any fixed θ. Thus, it is enough to consider the case where $\|\theta-\theta_0\| \leq \delta_n$ with $\delta_n = o(1)$.

By the implicit function theorem, $\hat{\lambda}_n(\theta)$ is continuously differentiable and its derivative is given by

$$\frac{\partial\hat{\lambda}_n(\theta)}{\partial\theta'} = \left(\frac{1}{n}\sum_{i=1}^n \frac{m_{\theta,i}m_{\theta,i}'}{(1+\hat{\lambda}_n(\theta)'m_{\theta,i})^2}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \frac{\nabla m_{\theta,i}}{1+\hat{\lambda}_n(\theta)'m_{\theta,i}} + \frac{1}{n}\sum_{i=1}^n \frac{\hat{\lambda}_n(\theta)m_{\theta,i}'\nabla m_{\theta,i}}{(1+\hat{\lambda}_n(\theta)'m_{\theta,i})^2}\right).$$

Moreover, we have $\sup_{\theta:\|\theta-\theta_0\|\leq\delta_n}\|\hat{\lambda}_n(\theta)\| = o(1)$. Hence, by the uniform law of large numbers, we have

$$\sup_{\theta:\|\theta-\theta_n\|\leq\delta_n}\left\|\frac{\partial\hat{\lambda}_n(\theta)}{\partial\theta'} - \mathbb{E}\left[m_{\theta_0}(X)m_{\theta_0}(X)'\right]^{-1}\mathbb{E}[\nabla m_{\theta_0}(X)]\right\| = o_P(1).$$

Because $\hat{\lambda}_n(\theta_0) = O_P(n^{-1/2})$ and $n^{-1}\sum_{i=1}^n \log(1+\hat{\lambda}_n(\theta_0)'m_{\theta_0,i}) = O_P(n^{-1})$, there exists

$C > 0$ such that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\log(1+\hat{\lambda}_n(\theta)'m_{\theta,i}) - \frac{1}{n}\sum_{i=1}^{n}\log(1+\hat{\lambda}_n(\theta_0)'m_{\theta_0,i}) \\
&= \hat{\lambda}_n(\theta)'\bar{m}_\theta - \frac{1}{2}\hat{\lambda}_n(\theta)'\frac{1}{n}\sum_{i=1}^{n}\frac{m_{\theta,i}m_{\theta,i}'}{(1+\bar{\lambda}'m_{\theta,i})^2}\hat{\lambda}_n(\theta) + O_P(n^{-1}) \\
&= \left(\hat{\lambda}_n(\theta_0) + \frac{\partial\hat{\lambda}_n(\bar{\theta})}{\partial\theta'}(\theta-\theta_0)\right)'\left(\bar{m}_{\theta_0} + \nabla\bar{m}_{\tilde{\theta}}(\theta-\theta_0)\right) \\
&\quad - \frac{1}{2}\left(\hat{\lambda}_n(\theta_0) + \frac{\partial\hat{\lambda}_n(\bar{\theta})}{\partial\theta'}(\theta-\theta_0)\right)'\frac{1}{n}\sum_{i=1}^{n}\frac{m_{\theta,i}m_{\theta,i}'}{(1+\bar{\lambda}'m_{\theta,i})^2}\left(\hat{\lambda}_n(\theta_0) + \frac{\partial\hat{\lambda}_n(\bar{\theta})}{\partial\theta'}(\theta-\theta_0)\right) + O_P(n^{-1}) \\
&\geq C\|\theta-\theta_0\|^2 + O_P(n^{-1/2}\|\theta-\theta_0\|) + O_P(n^{-1})
\end{aligned}$$

with probability approaching one, where $\bar{\lambda}$ is located between 0 and $\hat{\lambda}_n(\theta)$, and $\bar{\theta}$ and $\tilde{\theta}$ satisfy $\|\bar{\theta}-\theta_0\| = o(1)$ and $\|\tilde{\theta}-\theta_0\| = o(1)$, respectively. Therefore, we obtain (A.3).

Finally, using the Lemma E.3 of Chib et al. (2018) again, we have

$$\Pr\left(\int_\Theta \prod_{i=1}^{n}\frac{dP_{\theta,\mathbb{P}_n}}{dP_{\theta_0,\mathbb{P}_n}}(X_i)\pi_\Theta(\theta)d\theta \geq e^{-CM_n^2/2}\right) \to 1 \tag{A.4}$$

for any $M_n \to \infty$. Therefore, combining (A.3) and (A.4) and doing a similar calculation as in (A.2), we obtain $\Pi_n^{EL}(\theta \in \Theta_n | X_1, \ldots, X_n) \xrightarrow{P} 0$. □

Lemma A.3

Suppose that Assumption 6.2 holds. Then we have $\sup_{\theta\in\Theta_n}\sup_{\eta\in H_n}\|\lambda_\eta(\theta)\| = o(1)$ *and* $\sup_{\theta\in\Theta_n}\sup_{\eta\in H_n}\left\|\frac{\partial\lambda_\eta(\theta)}{\partial\theta'} - \mathbb{E}[m_{\theta_0}(X)m_{\theta_0}(X)']^{-1}\mathbb{E}[\nabla m_{\theta_0}(X)]\right\| = o(1)$ *for any set* Θ_n *such that* $\sup_{\theta\in\Theta_n}\|\theta-\theta_0\| = o(1)$.

Proof Because $dP_{\theta,\eta}/d\eta = 1/(1+\lambda_\eta(\theta)'m_\theta)$, it follows that

$$\begin{aligned}
\int|p_{\theta,\eta} - p_{\theta_0,\eta}|d\xi &= \int\left|\frac{d\eta}{d\xi}\frac{m_{\tilde{\theta}}'\frac{\partial\lambda_\eta(\tilde{\theta})}{\partial\theta'} + \lambda_\eta(\tilde{\theta})'\nabla m_{\tilde{\theta}}}{(1+\lambda_\eta(\tilde{\theta})'m_{\tilde{\theta}})^2}(\theta-\theta_0)\right|d\xi \\
&\leq C\left(\int\left\|\frac{m_{\tilde{\theta}}}{1+\lambda_\eta(\tilde{\theta})'m_{\tilde{\theta}}}\right\|dP_{\tilde{\theta},\eta} + \int\left\|\frac{\nabla m_{\tilde{\theta}}}{1+\lambda_\eta(\tilde{\theta})'m_{\tilde{\theta}}}\right\|dP_{\tilde{\theta},\eta}\right)\|\theta-\theta_0\|,
\end{aligned}$$

where $\tilde{\theta}$ is located between θ and θ_0. Because L_1 convergence of the density implies convergence in Hellinger distance, we obtain $\sup_{\theta\in\Theta_n}\sup_{\eta\in H_n}d_H(P_{\theta,\eta},P_{\theta_0,\eta}) = o(1)$. Thus, by the definition of H_n, we also have

$$\sup_{\theta\in\Theta_n}\sup_{\eta\in H_n}d_H(P_{\theta,\eta},P) = \sup_{\theta\in\Theta_n}\sup_{\eta\in H_n}d_H\left(\frac{\eta}{1+\lambda_\eta(\theta)'m_\theta},P\right) = o(1).$$

Because $P_{\theta,\eta} = P$ if and only if $\theta = \theta_0$ and $\eta = P$, it must be the case that $\sup_{\theta\in\Theta_n} \sup_{\eta\in H_n} \|\lambda_\eta(\theta)\| = o(1)$.

Next, because $\int m_\theta dP_{\theta,\eta} = 0$ for any $\theta \in \Theta$ and $\eta \in H$, we have

$$\frac{\partial \lambda_\eta(\theta)}{\partial \theta'} = \left(\int \frac{m_\theta m_\theta'}{1+\lambda_\eta(\theta)' m_\theta} dP_{\theta,\eta} \right)^{-1} \left(\int \nabla m_\theta dP_{\theta,\eta} + \lambda_\eta(\theta) \int \frac{m_\theta' \nabla m_\theta}{1+\lambda_\eta(\theta)' m_\theta} dP_{\theta,\eta} \right),$$

where the inverse matrix exists by Assumption 6.2 (ii). By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \left\| \int \nabla m_\theta dP_{\theta,\eta} - \int \nabla m_\theta dP \right\| &\leq \int \|\nabla m_\theta\| (\sqrt{dP_{\theta,\eta}} + \sqrt{dP})(\sqrt{dP_{\theta,\eta}} - \sqrt{dP}) \\ &\leq C \left(\int \|\nabla m_\theta\|^2 dP_{\theta,\eta} + \int \|\nabla m_\theta\|^2 dP \right) d_H(P_{\theta,\eta}, P). \end{aligned}$$

Similarly, we have

$$\begin{aligned} &\left\| \int \frac{m_\theta m_\theta'}{1+\lambda_\eta(\theta)' m_\theta} dP_{\theta,\eta} - \int \frac{m_\theta m_\theta'}{1+\lambda_\eta(\theta)' m_\theta} dP \right\| \\ &\leq C \left(\int \left\| \frac{m_\theta m_\theta'}{1+\lambda_\eta(\theta)' m_\theta} \right\|^2 dP_{\theta,\eta} + \int \left\| \frac{m_\theta m_\theta'}{1+\lambda_\eta(\theta)' m_\theta} \right\|^2 dP \right) d_H(P_{\theta,\eta}, P). \end{aligned}$$

Therefore, we obtain the second result. □

Lemma A.4

Suppose that Assumption 6.2 holds. Then we have

$$\sup_{h\in K} \sup_{\eta\in H_n} \left| \prod_{i=1}^n \log \frac{p_{\theta_n(h),\eta}}{p_{\theta_0,\eta}}(X_i) - \frac{1}{\sqrt{n}} \sum_{i=1}^n h' \dot{\ell}_{\theta_0,\eta_0}(X_i) + \frac{1}{2} h' I_{\theta_0,\eta_0} h \right| = o_P(1)$$

for any compact set $K \subset \mathbb{R}^p$.

Proof Let $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - \mathbb{E} f)$ denote the empirical process evaluated at a function f. Moreover, let $\ell_{\theta,\eta} = \log p_{\theta,\eta}$ and $\dot{\ell}_{\theta,\eta} = \partial \log p_{\theta,\eta}/\partial\theta$. Then we have

$$\dot{\ell}_{\theta,\eta} = -\frac{\nabla m_\theta' \lambda_\eta(\theta)}{1+\lambda_\eta(\theta)' m_\theta} - \frac{(\frac{\partial \lambda_\eta(\theta)}{\partial \theta'})' m_\theta}{1+\lambda_\eta(\theta)' m_\theta}.$$

Thus, for any $\theta_1, \theta_2 \in \mathcal{N}_{\theta_0}$, we have $|\ell_{\theta_1,\eta} - \ell_{\theta_2,\eta}| \leq \dot{\ell}_1 \|\theta_1 - \theta_2\|$, where

$$\dot{\ell}_1 = C \left(\sup_{\theta\in\mathcal{N}_{\theta_0}} \sup_{\eta\in H_n} \left\| \frac{\nabla m_\theta}{1+\lambda_\eta(\theta)' m_\theta} \right\| + \sup_{\theta\in\mathcal{N}_{\theta_0}} \sup_{\eta\in H_n} \left\| \frac{m_\theta}{1+\lambda_\eta(\theta)' m_\theta} \right\| \right),$$

which satisfies $\mathbb{E}[\dot{\ell}_1(X)^2] < \infty$. Thus, by applying Lemma 19.31 of van der Vaart (1998), we obtain

$$\sup_{h\in K} \sup_{\eta\in H_n} \mathbb{G}_n(\sqrt{n}(\ell_{\theta_n(h),\eta} - \ell_{\theta_0,\eta}) - h' \dot{\ell}_{\theta_0,\eta}) \xrightarrow{P} 0.$$

Moreover, for $\eta_1, \eta_2 \in H_n$, we obtain

$$\left\| \frac{\nabla m_{\theta_0}}{1+\lambda_{\eta_1}(\theta_0)' m_{\theta_0}} - \frac{\nabla m_{\theta_0}}{1+\lambda_{\eta_2}(\theta_0)' m_{\theta_0}} \right\| \leq \dot{\ell}_2 \|\lambda_{\eta_1}(\theta_0) - \lambda_{\eta_2}(\theta_0)\| \tag{A.5}$$

and

$$\left\| \frac{m_{\theta_0}}{1+\lambda_{\eta_1}(\theta_0)' m_{\theta_0}} - \frac{m_{\theta_0}}{1+\lambda_{\eta_2}(\theta_0)' m_{\theta_0}} \right\| \leq \dot{\ell}_3 \|\lambda_{\eta_1}(\theta_0) - \lambda_{\eta_2}(\theta_0)\|, \tag{A.6}$$

where

$$\dot{\ell}_2 = \sup_{\eta \in H_n} \left\| \frac{\nabla m'_{\theta_0} m_{\theta_0}}{(1+\lambda_\eta(\theta_0)' m_{\theta_0})^2} \right\| \quad \text{and} \quad \dot{\ell}_3 = \sup_{\eta \in H_n} \left\| \frac{m_{\theta_0} m'_{\theta_0}}{(1+\lambda_\eta(\theta_0)' m_{\theta_0})^2} \right\|.$$

Also, we have $\mathbb{E}[\dot{\ell}_2(X)^2] < \infty$ and $\mathbb{E}[\dot{\ell}_3(X)^2] < \infty$. The right-hand side of (A.5) and (A.6) depends on η_1 and η_2 only through $\lambda_{\eta_1}(\theta_0) - \lambda_{\eta_2}(\theta_0)$, which is a difference of finite-dimensional vectors. Thus it is straightforward to show that the classes of functions $\left\{ \frac{\nabla m_{\theta_0}}{1+\lambda_\eta(\theta_0)' m_{\theta_0}} : \eta \in H_n \right\}$ and $\left\{ \frac{m_{\theta_0}}{1+\lambda_\eta(\theta_0)' m_{\theta_0}} : \eta \in H_n \right\}$ are Donsker. Thus, by Lemma A.3, we obtain $\sup_{\eta \in H_n} \mathbb{G}_n(h' \dot{\ell}_{\theta_0,\eta} - h' \dot{\ell}_{\theta_0,\eta_0}) \xrightarrow{P} 0$.

The Taylor expansion yields

$$n\mathbb{E}\left[\log \frac{p_{\theta_n(h),\eta}}{p_{\theta_0,\eta}}(X) \right] = \sqrt{n} h' \mathbb{E}\left[\dot{\ell}_{\theta_0,\eta}(X) \right] + \frac{1}{2} h' \mathbb{E}\left[\ddot{\ell}_{\tilde{\theta}_n(h),\eta}(X) \right] h,$$

where $\tilde{\theta}_n(h)$ is located between $\theta_n(h)$ and θ_0. Also, $\ddot{\ell}_{\theta,\eta}$ is given by

$$\begin{aligned} \ddot{\ell}_{\theta,\eta} &= -\frac{\sum_{j=1}^l \left(\frac{\partial^2 m_{j,\theta}}{\partial \theta' \partial \theta} \right) \lambda_{\eta,j}(\theta)}{(1+\lambda_\eta(\theta)' m_\theta)} - 2\frac{\nabla m'_\theta \frac{\partial \lambda_\eta(\theta)}{\partial \theta'}}{(1+\lambda_\eta(\theta)' m_\theta)^2} + 2\frac{\nabla m'_\theta \lambda_\eta(\theta) \lambda_\eta(\theta)' \nabla m_\theta}{(1+\lambda_\eta(\theta)' m_\theta)^2} \\ &\quad - \frac{\sum_{j=1}^l m_{j,\theta} \frac{\partial^2 \lambda_{j,\eta}(\theta)}{\partial \theta' \partial \theta}}{(1+\lambda_\eta(\theta)' m_\theta)} + \frac{(\frac{\partial \lambda_\eta(\theta)}{\partial \theta'})' m_\theta m'_\theta (\frac{\partial \lambda_\eta(\theta)}{\partial \theta'})}{(1+\lambda_\eta(\theta)' m_\theta)^2} + \frac{(\frac{\partial \lambda_\eta(\theta)}{\partial \theta'})' m_\theta \lambda_\eta(\theta)' \nabla m_\theta}{(1+\lambda_\eta(\theta)' m_\theta)^2}, \end{aligned}$$

where $m_{j,\theta}$ and $\lambda_{j,\eta}(\theta)$ are the j-th element of m_θ and $\lambda_\eta(\theta)$, respectively. Assumption 6.2 is sufficient for the existence of $\partial^2 \lambda_{j,\eta}(\theta)/\partial\theta\partial\theta'$. Therefore, by Lemma A.3, we have $\sup_{h \in K} \sup_{\eta \in H_n} \left\| -\mathbb{E}[\ddot{\ell}_{\tilde{\theta}_n(h),\eta}(X)] - I_{\theta_0,\eta_0} \right\| = o(1)$ and the desired result follows. □

Lemma A.5

Suppose that Assumptions 6.1 and 6.2 hold. Then we have

$$\log \frac{s_n(h_n)}{s_n(0)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n h_n' \dot{\ell}_{\theta_0,\eta_0}(X_i) - \frac{1}{2} h_n' I_{\theta_0,\eta_0} h_n + o_P(1)$$

for any bounded random sequence $\{h_n\}$.

Proof Let $G_n(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h' \dot{\ell}_{\theta_0,\eta_0}(X_i) - \frac{1}{2} h' I_{\theta_0,\eta_0} h$. Then it follows from (6.2) and Lemma A.4 that for $\epsilon > 0$, we have

$$\begin{aligned}
\int_H \prod_{i=1}^n \log \frac{p_{\theta_n(h_n),\eta}}{p_{\theta_0,\eta_0}}(X_i) d\Pi_H(\eta) &\leq e^{\epsilon/2} \int_{H_n} \prod_{i=1}^n \log \frac{p_{\theta_n(h_n),\eta}}{p_{\theta_0,\eta_0}}(X_i) d\Pi_H(\eta) \\
&\leq e^{\epsilon + G_n(h_n)} \int_{H_n} \log \prod_{i=1}^n \frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}}(X_i) d\Pi_H(\eta) \\
&\leq e^{\epsilon + G_n(h_n)} \int_{H} \log \prod_{i=1}^n \frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}}(X_i) d\Pi_H(\eta)
\end{aligned}$$

with probability approaching one. Similarly, we have

$$\begin{aligned}
\int_H \prod_{i=1}^n \log \frac{p_{\theta_n(h_n),\eta}}{p_{\theta_0,\eta_0}}(X_i) d\Pi_H(\eta) &\geq \int_{H_n} \prod_{i=1}^n \log \frac{p_{\theta_n(h_n),\eta}}{p_{\theta_0,\eta_0}}(X_i) d\Pi_H(\eta) \\
&\geq e^{-\epsilon/2 + G_n(h_n)} \int_{H_n} \prod_{i=1}^n \log \frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}}(X_i) d\Pi_H(\eta) \\
&\geq e^{-\epsilon + G_n(h_n)} \int_{H} \prod_{i=1}^n \log \frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}}(X_i) d\Pi_H(\eta)
\end{aligned}$$

with probability approaching one. Because $\epsilon > 0$ is arbitrary, we obtain the desired result. □

Lemma A.6

Suppose that Assumptions 6.1–6.3 hold. Then we have

$$\Pi_n(\sqrt{n}\|\theta - \theta_0\| > M_n | X_1, \ldots, X_n) \xrightarrow{P} 0$$

for any sequence $\{M_n\}$ such that $M_n \to \infty$.

Proof We define two sequences of events

$$A_n = \left\{ \sup_{\theta : \|\theta - \theta_0\| > \delta} \sup_{\eta \in H} \frac{1}{n} \sum_{i=1}^n \log \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(X_i) \leq -\epsilon \right\}$$

and

$$B_n = \left\{ \int_\Theta S_n(\theta) \pi_\Theta(\theta) d\theta \geq e^{-n\epsilon/2} S_n(\theta_0) \right\},$$

where

$$S_n(\theta) = \int_H \prod_{i=1}^n \frac{p_{\theta,\eta}}{p_{\theta_0,\eta_0}}(X_i) d\Pi_H(\eta).$$

Lemma A.5 and Lemma 6.3 of Bickel and Kleijn (2012) imply that $\Pr(B_n) \to 1$ for any $\epsilon > 0$. Therefore, by Assumption 6.3 (i), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\Pi_n(\|\theta - \theta_0\| > \delta | X_1, \ldots, X_n)\right] \\
&\leq \mathbb{E}\left[\Pi_n(\|\theta - \theta_0\| > \delta | X_1, \ldots, X_n) 1_{A_n \cap B_n}\right] + o(1) \\
&\leq e^{n\epsilon/2}\mathbb{E}\left[S_n(\theta_0)^{-1} \int_H \int_{\theta:\|\theta-\theta_0\|>\delta} \prod_{i=1}^n \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(X_i) \prod_{i=1}^n \frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}}(X_i) 1_{A_n} \pi_\Theta(\theta) d\theta d\Pi_H(\eta)\right] + o(1) \\
&= o(1) \qquad \text{(A.7)}
\end{aligned}$$

for any $\delta > 0$. Moreover, (A.7) implies $\Pi_n(H_n | X_1, \ldots, X_n) \xrightarrow{P} 1$ because $\Pi_n(H_n | \theta, X_1, \ldots, X_n) \xrightarrow{P} 1$ for any $\theta \in \mathcal{N}_{\theta_0}$.

Let $\Theta_n = \{\theta \in \Theta : M_n/\sqrt{n} < \|\theta - \theta_0\| \leq \delta\}$ for M_n such that $M_n \to \infty$ and $M_n/\sqrt{n} \to 0$. Now, we show that there exists $C > 0$ such that

$$\Pr\left(\sup_{\theta \in \Theta_n} \sup_{\eta \in H_n} \frac{1}{n} \sum_{i=1}^n \log \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(X_i) \leq -C\frac{M_n^2}{n}\right) \to 1. \qquad \text{(A.8)}$$

Because of Assumption 6.3 (i), for any fixed θ, we can find $C > 0$ such that

$$\Pr\left(\sup_{\eta \in H_n} \frac{1}{n} \sum_{i=1}^n \log \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(X_i) \leq -C\frac{M_n^2}{n}\right) \to 1.$$

Thus, we only need to consider the case where $\|\theta - \theta_0\| \leq \delta_n$ with $\delta_n = o(1)$. Moreover, a similar argument as in the proof of Lemma A.4 yields

$$\sup_{\theta:\|\theta-\theta_0\|\leq\delta_n} \sup_{\eta \in H_n} \mathbb{G}_n(\ell_{\theta,\eta} - \ell_{\theta_0,\eta} - \dot{\ell}'_{\theta_0,\eta}(\theta - \theta_0)) \xrightarrow{P} 0$$

and $\sup_{\eta \in H_n} \mathbb{G}_n(\dot{\ell}_{\theta_0,\eta} - \dot{\ell}_{\theta_0,\eta_0}) \xrightarrow{P} 0$. Thus, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \log \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}}(X_i) &= \frac{1}{n}\sum_{i=1}^n \dot{\ell}_{\theta_0,\eta_0}(X_i)'(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)'\mathbb{E}[\ddot{\ell}_{\tilde{\theta},\eta}(X)](\theta - \theta_0) + O_P(n^{-1}) \\
&= \frac{1}{2}(\theta - \theta_0)'\mathbb{E}[\ddot{\ell}_{\tilde{\theta},\eta}(X)](\theta - \theta_0) + O_P(n^{-1/2}\|\theta - \theta_0\|) + O_P(n^{-1})
\end{aligned}$$

uniformly over $\|\theta - \theta_0\| \leq \delta_n$ and $\eta \in H_n$, where $\tilde{\theta}$ is located between θ and θ_0. Because $\| - \mathbb{E}[\ddot{\ell}_{\tilde{\theta},\eta}(X)] - I_{\theta_0,\eta_0}\| = o(1)$ by Lemma A.3, we obtain (A.8).

Finally, using the Lemma 6.3 of Bickel and Kleijn (2012) again, we have

$$\Pr\left(\int_\Theta S_n(\theta)\pi_\Theta(\theta)d\theta \geq e^{-CM_n^2/2} S_n(\theta_0)\right) \to 1 \qquad \text{(A.9)}$$

for any $C > 0$ and $M_n \to \infty$. Thus, combining (A.8) and (A.9) and doing a similar calculation as in (A.7), we obtain $\Pi_n(\Theta_n, H_n | X_1, \ldots, X_n) \xrightarrow{P} 0$, which also implies $\Pi_n(\Theta_n | X_1, \ldots, X_n) \xrightarrow{P}$

0. □

References

Andrews, I. and A. Mikusheva (2022). Optimal decision rules for weak GMM. *Econometrica 90*, 715–748.

Bedoui, A. and N. A. Lazar (2020). Bayesian empirical likelihood for ridge and lasso regressions. *Computational Statistics and Data Analysis 145*. 106917.

Begun, J. M., W. J. Hall, W.-M. Huang, and J. A. Wellner (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Annals of Statistics 11*, 432–452.

Bertail, P. (2006). Empirical likelihood in some semiparametric models. *Bernoulli 12*, 299–331.

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

Bornn, L., N. Shephard, and R. Solgi (2019). Moment conditions and Bayesian nonparametrics. *Journal of the Royal Statistical Society, Series B 81*, 5–43.

Borwein, J. M. and A. S. Lewis (1991). Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization 29*, 325–338.

Chae, M. (2015). *The Semiparametric Bernstein–von Mises Theorem for Models with Symmetric Error*. Ph. D. thesis, Seoul National University.

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

Chamberlain, G. and G. W. Imbens (2003). Nonparametric applications of Bayesian inference. *Journal of Business & Economic Statistics 21*, 12–18.

Chaudhuri, S., D. Mondal, and T. Yin (2017). Hamiltonian Monte Carlo sampling in Bayesian empirical likelihood computation. *Journal of the Royal Statistical Society, Series B 79*, 293–320.

Chen, X., H. Hong, and M. Shum (2007). Nonparametric likelihood ratio model selection tests between parametric likelihood and moment condition models. *Journal of Econometrics 141*, 109–140.

Cheng, Y. and Y. Zhao (2019). Bayesian jackknife empirical likelihood. *Biometrika 106*, 981–988.

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

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

Csiszár, I. (1995). Generalized projections for non-negative functions. *Acta Mathematica Hungarica 68*, 161–185.

DiCiccio, T. J. and J. P. Romano (1990). Nonparametric confidence limits by resampling methods and least favorable families. *International Statistical Review 58*, 59–76.

Dovonon, P. and Y. F. Atchadé (2020). Efficiency bounds for semiparametric models with singular score functions. *Econometric Reviews 39*, 612–648.

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

Fernández-Villaverde, J. (2010). The econometrics of DSGE models. *SERIEs 1*, 3–49.

Florens, J. and A. Simoni (2021). Gaussian processes and Bayesian moment estimation. *Journal of Business & Economic Statistics 39*, 482–492.

Ghosal, S. and A. van der Vaart (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

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

Imbens, G. W., R. H. Spady, and P. Johnson (1998). Information theoretic approaches to inference in moment condition models. *Econometrica 66*, 333–357.

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

Kitamura, Y. (2007). Empirical likelihood methods in econometrics: theory and practice. In R. Blundell, W. K. Newey, and T. Persson (Eds.), *Advances in Economics and Econometrics*, pp. 174–237. Cambridge University Press.

Kitamura, Y. and T. Otsu (2011). Bayesian analysis of moment condition models using nonparametric priors. Unpublised Manuscript, Yale University.

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

Kleijn, B. J. K. and A. W. van der Vaart (2006). Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics 34*, 837–877.

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

Komunjer, I. and G. Ragusa (2016). Existence and characterization of conditional density projections. *Econometric Theory 32*, 947–987.

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

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation*. Springer.

Monahan, J. F. and D. D. Boos (1992). Proper likelihood for Bayesian analysis. *Biometrika 79*, 271–278.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica 62*, 1349–1382.

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

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *Annals of Statistics 18*, 90–120.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC.

Pollard, D. (2010). Hellinger differentiability. `http://www.stat.yale.edu/~pollard/Courses/618.fall2010/Handouts/prelim_version_DQM.pdf`.

Qin, J. and J. Lawless (1994). Empirical likelihood and generalized estimating equations. *Annals of Statistics 22*, 300–325.

Ragusa, G. (2007). Bayesian likelihoods for moment condition models. University of California, Irvine.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics 9*, 130–134.

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

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *Annals of Statistics 35*, 634–672.

Severini, T. A. (1999). On the relationship between Bayesian and non-Bayesian elimination of nuisance parameters. *Statistica Sinica 9*, 713–724.

Severini, T. A. and G. Tripathi (2001). A simplified approach to computing efficiency bounds in semiparametric models. *Journal of Econometrics 2001*, 23–66.

Shin, M. (2015). *Bayesian GMM*. Ph. D. thesis, University of Pennsylvania.

Smith, R. J. (1997). Alternative semi-parametric likelihood approaches to generalised method of moments estimation. *Economic Journal 107*, 503–519.

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

Sueishi, N. (2016). A simple derivation of the efficiency bound for conditional moment restriction models. *Economics Letters 138*, 57–59.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer.

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

Wu, Y. and S. Ghosal (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics 2*, 298–331.

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