

Statistica Sinica Preprint No: SS-2017-0074.R2

Title	The semi-parametric Bernstein-von Mises theorem for regression models with symmetric errors
Manuscript ID	SS-2017-0074.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0074
Complete List of Authors	Minwoo Chae Yongdai Kim and Bas J. K. Kleijn
Corresponding Author	Yongdai Kim
E-mail	ydkim0903@gmail.com
Notice: Accepted version subject to English editing.	

The Semi-parametric Bernstein-von Mises Theorem for Regression Models with Symmetric Errors

Minwoo Chae¹, Yongdai Kim² and Bas J. K. Kleijn³

¹*Department of Mathematics, Applied Mathematics and Statistics
Case Western Reserve University*

²*Department of Statistics, Seoul National University*

³*Korteweg-de Vries Institute for Mathematics, University of Amsterdam*

Abstract: In a smooth semi-parametric model, the marginal posterior distribution for a finite dimensional parameter of interest is expected to be asymptotically equivalent to the sampling distribution of any efficient point-estimator. The assertion leads to asymptotic equivalence of credible and confidence sets for the parameter of interest and is known as the semi-parametric Bernstein-von Mises theorem. In recent years, it has received much attention and has been applied in many examples. We consider models in which errors with symmetric densities play a role; more specifically, it is shown that the marginal posterior distributions of regression coefficients in the linear regression and linear mixed effect models satisfy the semi-parametric Bernstein-von Mises assertion. As a consequence, Bayes estimators in these models achieve frequentist inferential optimality, as expressed *e.g.* through Hájek's convolution and asymptotic minimax theorems. For the prior on the space of error densities, we provide two well-known examples, the Dirichlet process mixture of normal densities and random series priors. Particularly, the result provides an efficient estimate of regression coefficients in the linear mixed effect model, for which no other efficient point-estimator was known previously.

Key words and phrases: Bernstein–von Mises theorem, linear mixed effect model, linear regression, semi-parametric efficiency, symmetric error.

1 Introduction

In this paper, we give an asymptotic, Bayesian analysis of models with errors that are distributed symmetrically. The observations $\mathbf{X} = (X_1, \dots, X_n)^T \in \mathbb{R}^n$ are modeled by,

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\epsilon}, \quad (1.1)$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$. Here the mean vector $\boldsymbol{\mu}$ is non-random and parametrized by a finite dimensional parameter θ , and the distribution of the error vector $\boldsymbol{\epsilon}$ is symmetric in the sense that $\boldsymbol{\epsilon}$ has the same distribution as $-\boldsymbol{\epsilon}$. Since the error has a symmetric but otherwise unknown distribution, the model is semi-parametric. Examples of models of the form (1.1) are the symmetric location model (where $\mu_i = \theta \in \mathbb{R}$), and the linear regression model (where $\mu_i = \theta^T Z_i$ for given covariates $Z_i \in \mathbb{R}^p$). Moreover, the form (1.1) includes models with dependent errors, like linear mixed effect models (Laird and Ware (1982)).

The main goal of this paper is to prove the semi-parametric Bernstein-von Mises (BvM) assertion for models of the form (1.1) with symmetric error distributions. Roughly speaking we show that the marginal posterior distribution of the parameter of interest is asymptotically normal, centered on an efficient estimator with variance equal to the inverse Fisher information matrix. As a result, statistical inference based on the posterior distribution satisfies frequentist criteria of optimality, see van der Vaart (1998) for i.i.d. cases and Bickel et al. (2005) for non-i.i.d. extensions.

Various sets of sufficient conditions for the semi-parametric BvM theorem based on the full LAN (local asymptotic normality) expansion (*i.e.* the LAN expansion with respect to both the finite and infinite dimensional parameters; McNeney and Wellner (2000)) have been developed in Shen

(2002); Castillo (2012); Bickel and Kleijn (2012); Castillo and Rousseau (2015). Because the models we consider are adaptive (Bickel (1982)), we can consider a simpler type of LAN expansion that involves only the parameter of interest, albeit that the expansion must be valid under data distributions that differ slightly from the one on which the expansion is centred. We call this property *misspecified LAN* and prove that it holds for the models of the form (1.1) and that, together with other regularity conditions, it implies the semi-parametric BvM assertion. The misspecified LAN condition is slightly weaker and easier to be checked than that given in Castillo (2012) and Castillo and Rousseau (2015) for adaptive models even though the condition in Castillo (2012) and Castillo and Rousseau (2015) covers more general nonadaptive cases.

While the BvM theorem for parametric Bayesian models is well established (*e.g.* Le Cam and Yang (1990); Kleijn and van der Vaart (2012)), the semi-parametric BvM theorem is still being studied very actively: initial examples Cox (1993); Freedman (1999) of simple semi-parametric problems with simple choices for the prior demonstrated failures of marginals posteriors to display BvM-type asymptotic behaviour. Subsequently, positive semi-parametric BvM results have been established in these and various other examples, including models in survival analysis (Kim and Lee (2004); Kim (2006)), multivariate normal regression models with growing numbers of parameters (Bontemps (2011); Johnstone (2010); Ghosal (1999)) and discrete probability measures (Boucheron and Gassiat (2009)). More delicate notions like finite sample properties and second-order asymptotics are considered in Panov and Spokoiny (2015); Spokoiny (2013); Yang et al. (2015). Furthermore, Castillo et al. (2015) have studied the BvM theorem in high dimensional models incorporating sparse priors.

Regarding models of the form (1.1), there is a sizable amount of literature on efficient point-estimation in the symmetric location problem (Beran (1978); Stone (1975); Sacks (1975)) and linear regression models (Bickel (1982)). By contrast, to date *no efficient point-estimator* for the regression coefficients in the linear mixed effect model has been found; the semi-parametric BvM theorem proved below, however, implies that the Bayes estimator is efficient! To the authors' best knowledge, this paper provides the first efficient semi-parametric estimator in the linear mixed effect model. While compactness assumptions imposed on the prior and true error density are rather strong, a numerical study given in section 5 supports the view that the Bayes estimator, which can be easily computed via MCMC methods, is superior to previous methods of estimation. Finally, it should be mentioned that an extension of the current paper to high-dimensional sparse setting has been studied in Chae et al. (2016).

This paper is organized as follows: section 2 proves the semi-parametric BvM assertion for all smooth adaptive models (*c.f.* the misspecified LAN expansion). In sections 3 and 4 we study the linear regression model and linear mixed effect model, respectively. For each, we consider two common choices for the nuisance prior, a Dirichlet process mixture and a series prior, and we show that both lead to validity of the BvM assertion. Results of numerical studies are presented in section 5. Proofs of main results can be found in the supplementary material.

Notation and conventions

For two real values a and b , $a \wedge b$ and $a \vee b$ are the minimum and maximum of a and b , respectively, and $a_n \lesssim b_n$ signifies that a_n is smaller than b_n

up to a constant multiple independent of n . Lebesgue measures are denoted by μ ; $|\cdot|$ represents the Euclidean norm on \mathbb{R}^d . The capitals P_η , $P_{\theta,\eta}$ etc. denote the probability measures associated with densities that we write in lower case, p_η , $p_{\theta,\eta}$ etc. (where it is always clear from the context which dominating measure μ is involved). The corresponding log densities are indicated with ℓ_η , $\ell_{\theta,\eta}$ etc. Hellinger and total-variational metrics are defined as $h^2(p_1, p_2) = \int (\sqrt{p_1} - \sqrt{p_2})^2 d\mu$ and $d_V(p_1, p_2) = \int |p_1 - p_2| d\mu$, respectively. The expectation of a random variable X under a probability measure P is denoted by PX . The notation P_0 always represents the true probability which generates the observation and $X^o = X - P_0X$ is the centered version of a random variable X . The indicator function for a set A is denoted 1_A . For a class of measurable functions \mathcal{F} , the quantities $N(\epsilon, \mathcal{F}, d)$ and $N_{[]}(\epsilon, \mathcal{F}, d)$ represent the ϵ -covering and -bracketing numbers (van der Vaart and Wellner (1996)) with respect to a (semi)metric d .

2 Misspecified LAN and the semi-parametric BvM theorem

2.1 Misspecified local asymptotic normality

In this section, we prove the semi-parametric BvM theorem for smooth adaptive models, *i.e.* those that satisfy the misspecified LAN expansion defined below. Consider a sequence of statistical models $\mathcal{P}^{(n)} = \{P_{\theta,\eta}^{(n)} : \theta \in \Theta, \eta \in \mathcal{H}\}$ on measurable spaces $(\mathcal{X}^{(n)}, \mathcal{A}^{(n)})$, parametrized by a finite dimensional parameter θ of interest and an infinite dimensional nuisance parameter η . Assume that Θ is a subset of \mathbb{R}^p , \mathcal{H} is a metric space equipped

2.1 Misspecified local asymptotic normality

with the associated Borel σ -algebra and $P_{\theta,\eta}^{(n)}$ has density $x \mapsto p_{\theta,\eta}^{(n)}(x)$ with respect to some σ -finite measures $\mu^{(n)}$ dominating $\mathcal{P}^{(n)}$.

Let $X^{(n)}$ be a $\mathcal{X}^{(n)}$ -valued random element following $P_0^{(n)}$ and assume that $P_0^{(n)} = P_{\theta_0,\eta_0}^{(n)}$ for some $\theta_0 \in \Theta$ and $\eta_0 \in \mathcal{H}$. We say that a sequence of statistical models $\mathcal{P}^{(n)}$ satisfies the *misspecified LAN expansion* if there exists a sequence of vector-valued (componentwise) $L_2(P_0^{(n)})$ -functions $(g_{n,\eta})$, a sequence (\mathcal{H}_n) of measurable subsets of \mathcal{H} and a sequence $(V_{n,\eta})$ of $p \times p$ -matrices such that,

$$\sup_{h \in K} \sup_{\eta \in \mathcal{H}_n} \left| \log \frac{p_{\theta_n(h),\eta}^{(n)}(X^{(n)})}{p_{\theta_0,\eta}^{(n)}} - \frac{h^T}{\sqrt{n}} g_{n,\eta}(X^{(n)}) + \frac{1}{2} h^T V_{n,\eta} h \right| = o_{P_0}(1), \quad (2.2)$$

for every compact $K \subset \mathbb{R}^p$, where $\theta_n(h)$ equals $\theta_0 + h/\sqrt{n}$. When we know η_0 , property (2.2) is nothing but the usual parametric LAN expansion, where we set $\mathcal{H}_n = \{\eta_0\}$. We refer to (2.2) as the *misspecified LAN expansion* because the base for the expansion is (θ_0, η) while rest-terms go to zero under P_0 , which corresponds to the point (θ_0, η_0) .

Note that the misspecified LAN expansion is slightly weaker than the LAN expansion used in Castillo (2012) and Castillo and Rousseau (2015) for adaptive models. In particular, the condition in Castillo (2012) and Castillo and Rousseau (2015) requires an Hilbert space structure on the space of $\eta - \eta_0$, where the LAN expansion and localizing sets \mathcal{H}_n rely on the Hilbert space structure through the associated norm. It is not easy to check that a given prior of η does have the required Hilbert space structure for $\eta - \eta_0$. In contrast, we establish the semi-parametric BvM theorem (Theorem 1) based on the misspecified LAN (2.2), which does not require an Hilbert space structure for $\eta - \eta_0$ and hence it is relatively easy to check the misspecified LAN for a given prior. For example, in Section

2.1 Misspecified local asymptotic normality

3.2 we verify the misspecified LAN for mixture of normal prior and basis expansion prior.

Although the misspecified LAN expansion (2.2) can be applied only to the adaptive cases, the verification of (2.2) is not easy due to misspecification and the required uniformity of convergence. LAN expansions have been shown to be valid even under misspecification: in Kleijn and van der Vaart (2012) for example, smoothness in misspecified parametric models is expressed through a version of local asymptotic normality under the true distribution of the data, with a likelihood expansion around points in the model where the Kullback-Leibler (KL)-divergence with respect to P_0 is minimal. In models with symmetric error, the point of minimal KL-divergence equals exactly θ_0 , provided that the misspecified η is close enough to η_0 in the sense of \mathcal{H}_n . This allows the usual LAN expansion at θ_0 for fixed η , that is, the left-hand side of (2.2) is expected to be of order $o_{P_0}(1)$. By choosing localizations \mathcal{H}_n appropriately, the family of score functions $\{\dot{\ell}_{\theta,\eta} : \eta \in \mathcal{H}_n\}$ is shown to be a Donsker class, which validates (2.2) in models with symmetric errors, where $\dot{\ell}_{\theta,\eta}(x) = \partial \ell_{\theta,\eta}(x) / \partial \theta$, $g_{n,\eta}(X^{(n)}) = \sum_{i=1}^n \dot{\ell}_{\theta_0,\eta}(X_i)$ and $V_{n,\eta} = n^{-1} P_0^{(n)}[g_{n,\eta} g_{n,\eta}^T]$. The score function is not necessarily the pointwise derivative of the log-likelihood, but in most examples (including the models considered in this paper), $g_{n,\eta} = \dot{\ell}_{\theta_0,\eta}^{(n)}$ where $\dot{\ell}_{\theta,\eta}^{(n)} = \partial \ell_{\theta,\eta}^{(n)} / \partial \theta$. From now on, since it conveys the natural meaning of derivative, we use the notation $\dot{\ell}_{\theta_0,\eta}^{(n)}$ instead of $g_{n,\eta}$.

2.2 The semi-parametric Bernstein-von Mises theorem

2.2 The semi-parametric Bernstein-von Mises theorem

We use a product prior $\Pi = \Pi_{\Theta} \times \Pi_{\mathcal{H}}$ on the Borel σ -algebra of $\Theta \times \mathcal{H}$ and denote the posterior distribution by $\Pi(\cdot|X^{(n)})$. Note that the misspecified LAN property gives rise to an expansion of the log-likelihood that applies only locally in sets $\Theta_n \times \mathcal{H}_n$, where $\Theta_n = \{\theta_0 + h/\sqrt{n} : h \in K\}$ (for some compact $K \in \mathbb{R}^p$ and appropriate $\mathcal{H}_n \subset \mathcal{H}$). So for the semi-parametric BvM theorem, the score function $\dot{\ell}_{\theta_0, \eta}^{(n)}$ as well as $V_{n, \eta}$ must ‘behave nicely’ on $\Theta_n \times \mathcal{H}_n$ and the posterior distribution must concentrate inside $\Theta_n \times \mathcal{H}_n$. Technically, these requirements are expressed by the following two conditions. For a matrix $A \in \mathbb{R}^{n_1 \times n_2}$, $\|A\|$ represents the operator norm of A , defined as $\sup_{x \neq 0} |Ax|/|x|$, and if A is a square matrix, $\rho_{\min}(A)$ and $\rho_{\max}(A)$ denote the minimum and maximum eigenvalues of A , respectively.

Condition A. (Equicontinuity and non-singularity)

$$\sup_{\eta \in \mathcal{H}_n} \left| \dot{\ell}_{\theta_0, \eta}^{(n)}(X^{(n)}) - \dot{\ell}_{\theta_0, \eta_0}^{(n)}(X^{(n)}) \right| = o_{P_0}(n^{1/2}), \quad (2.3)$$

$$\sup_{\eta \in \mathcal{H}_n} \|V_{n, \eta} - V_{n, \eta_0}\| = o(1), \quad (2.4)$$

$$0 < \liminf_{n \rightarrow \infty} \rho_{\min}(V_{n, \eta_0}) \leq \limsup_{n \rightarrow \infty} \rho_{\max}(V_{n, \eta_0}) < \infty. \quad (2.5)$$

Condition B. (Posterior localization)

$$P_0^{(n)} \Pi(\eta \in \mathcal{H}_n | X^{(n)}) \rightarrow 1, \quad (2.6)$$

$$P_0^{(n)} \Pi(\sqrt{n}|\theta - \theta_0| > M_n | X^{(n)}) \rightarrow 0, \quad \text{for every } M_n \uparrow \infty. \quad (2.7)$$

Conditions like (2.3) and (2.4) are to be expected in the context of semi-parametric estimation (see, *e.g.*, Theorem 25.54 of van der Vaart (1996)).

2.2 The semi-parametric Bernstein-von Mises theorem

Condition (2.3) amounts to *asymptotic equicontinuity* and is implied whenever scores form a Donsker class, a well-known sufficient condition in semi-parametric efficiency (see van der Vaart (1996)). Condition (2.4) is implied whenever the $L_2(P_0^{(n)})$ -norm of the difference between scores at (θ_0, η) and (θ_0, η_0) vanishes as η converges to η_0 in Hellinger distance, *c.f.* (S1.5) of supplementary material; it controls variations of the information matrix as η converges to η_0 with \mathcal{H}_n . Note that conditions (2.2)–(2.4) lead the following LAN assertion

$$\sup_{h \in K} \sup_{\eta \in \mathcal{H}_n} \left| \log \frac{p_{\theta_n(h), \eta}^{(n)}(X^{(n)})}{p_{\theta_0, \eta}^{(n)}} - \frac{h^T}{\sqrt{n}} \dot{\ell}_{\theta_0, \eta_0}^{(n)}(X^{(n)}) + \frac{1}{2} h^T V_{n, \eta_0} h \right| = o_{P_0}(1), \quad (2.8)$$

which is typically assumed in literature for semi-parametric BvM theorem. We separate (2.8) into three conditions because proofs of both (2.2) and (2.3) are demanding in a technical sense.

Condition (2.5) guarantees that the Fisher information matrix does not develop singularities as the sample size goes to infinity. Condition (2.6) formulates a requirement of posterior consistency in the usual sense, and sufficient conditions are well-known (Schwartz (1965); Barron et al. (1999); Walker (2004); Kleijn (2013)). Condition (2.7) requires $n^{-1/2}$ -rate of convergence for the marginal posterior distribution for the parameter of interest. Though (2.7) appears to be rather strong (Yang et al. (2015)), clearly, it is a necessary condition. It should be noted that conditions (2.2) and (2.7) can be replaced by

$$\begin{aligned} \sup_{\theta \in \tilde{\Theta}_n} \sup_{\eta \in \mathcal{H}_n} \left| \log \frac{p_{\theta, \eta}^{(n)}(X^{(n)})}{p_{\theta_0, \eta}^{(n)}} - (\theta - \theta_0)^T g_{n, \eta}(X^{(n)}) + \frac{n}{2} (\theta - \theta_0)^T V_{n, \eta} (\theta - \theta_0) \right| \\ = o_{P_0}(1 + n(|\theta - \theta_0|^2)) \end{aligned}$$

2.2 The semi-parametric Bernstein-von Mises theorem

and

$$P_0^{(n)}\Pi(\theta \in \tilde{\Theta}_n|X^{(n)}) \rightarrow 1$$

for some $\tilde{\Theta}_n \subset \Theta$, as is done by Castillo (2012) and Yang et al. (2015). That means, if the log-likelihood is uniformly approximated by a quadratic function in a neighborhood where the posterior contracts, then the posterior achieves $n^{-1/2}$ -rate of convergence.

We say the prior Π_Θ is *thick* at θ_0 if it has a strictly positive and continuous Lebesgue density in a neighborhood of θ_0 . The following theorem states the BvM theorem for semi-parametric models that are smooth in the sense of the misspecified LAN expansion.

Theorem 1. Consider statistical models $\{P_{\theta,\eta}^{(n)} : \theta \in \Theta, \eta \in \mathcal{H}\}$ with a product prior $\Pi = \Pi_\Theta \times \Pi_{\mathcal{H}}$. Assume that Π_Θ is thick at θ_0 and that (2.2) as well as Conditions A and B hold. Then,

$$\sup_B \left| \Pi(\sqrt{n}(\theta - \theta_0) \in B|X^{(n)}) - N_{\Delta_n, V_{n,\eta_0}^{-1}}(B) \right| \rightarrow 0, \quad (2.9)$$

in $P_0^{(n)}$ -probability, where $N_{\Delta_n, V_{n,\eta_0}^{-1}}$ is the normal distribution with mean

$$\Delta_n = \frac{1}{\sqrt{n}} V_{n,\eta_0}^{-1} \dot{\ell}_{\theta_0, \eta_0}^{(n)}(X^{(n)})$$

and variance V_{n,η_0}^{-1} .

Proof. Note first that (2.6) implies that $\Pi_{\mathcal{H}}(\mathcal{H}_n) > 0$ for large enough n . Let $\Pi_{\mathcal{H}_n}$ be the probability measure obtained by restricting $\Pi_{\mathcal{H}}$ to \mathcal{H}_n and next re-normalizing, and $\Pi_{\mathcal{H}_n}(\cdot|X^{(n)})$ be the corresponding posterior distribution. Then, for any measurable set B in Θ ,

$$\begin{aligned} \Pi(\theta \in B|X^{(n)}) &= \Pi(\theta \in B, \eta \in \mathcal{H}_n|X^{(n)}) + \Pi(\theta \in B, \eta \in \mathcal{H}_n^c|X^{(n)}) \\ &= \Pi_{\mathcal{H}_n}(\theta \in B|X^{(n)})\Pi(\eta \in \mathcal{H}_n|X^{(n)}) + \Pi(\theta \in B, \eta \in \mathcal{H}_n^c|X^{(n)}), \end{aligned}$$

2.2 The semi-parametric Bernstein-von Mises theorem

so we have,

$$\sup_B \left| \Pi(\theta \in B|X^{(n)}) - \Pi_{\mathcal{H}_n}(\theta \in B|X^{(n)}) \right| \rightarrow 0,$$

in $P_0^{(n)}$ -probability. Therefore it is sufficient to prove the BvM assertion with the priors $\Pi_{\mathcal{H}_n}$.

Particularly,

$$\Pi_{\mathcal{H}_n}(\sqrt{n}|\theta - \theta_0| > M_n|X^{(n)}) = \frac{\Pi(\sqrt{n}|\theta - \theta_0| > M_n, \eta \in \mathcal{H}_n|X^{(n)})}{\Pi(\eta \in \mathcal{H}_n|X^{(n)})}, \quad (2.10)$$

converges to 0 in $P_0^{(n)}$ -probability by (2.6) and (2.7). Using (2.2), (2.3) and (2.4), we obtain (2.8) for every compact $K \subset \mathbb{R}^p$. Let,

$$b_1(h) = \inf_{\eta \in \mathcal{H}_n} \log \frac{p_{\theta_n(h), \eta}^{(n)}(X^{(n)})}{p_{\theta_0, \eta}^{(n)}(X^{(n)})}, \quad \text{and} \quad b_2(h) = \sup_{\eta \in \mathcal{H}_n} \log \frac{p_{\theta_n(h), \eta}^{(n)}(X^{(n)})}{p_{\theta_0, \eta}^{(n)}(X^{(n)})}.$$

Then, trivially, we have,

$$e^{b_1(h)} \leq \frac{\int p_{\theta_n(h), \eta}^{(n)}(X^{(n)}) d\Pi_{\mathcal{H}_n}(\eta)}{\int p_{\theta_0, \eta}^{(n)}(X^{(n)}) d\Pi_{\mathcal{H}_n}(\eta)} \leq e^{b_2(h)}, \quad (2.11)$$

and the quantity,

$$\sup_{h \in K} \left| b_k(h) - \frac{h^T}{\sqrt{n}} \dot{\ell}_{\theta_0, \eta_0}^{(n)}(X^{(n)}) + \frac{1}{2} h^T V_{n, \eta_0} h \right|,$$

is bounded above by the left-hand side of (2.8) for $k = 1, 2$. As a result,

$$\sup_{h \in K} \left| \log \frac{\int p_{\theta_n(h), \eta}^{(n)}(X^{(n)}) d\Pi_{\mathcal{H}_n}(\eta)}{\int p_{\theta_0, \eta}^{(n)}(X^{(n)}) d\Pi_{\mathcal{H}_n}(\eta)} - \frac{h^T}{\sqrt{n}} \dot{\ell}_{\theta_0, \eta_0}^{(n)}(X^{(n)}) + \frac{1}{2} h^T V_{n, \eta_0} h \right| = o_{P_0}(1), \quad (2.12)$$

because $|c_2| \leq |c_1| \vee |c_3|$ for all real numbers c_1, c_2 and c_3 with $c_1 \leq c_2 \leq c_3$.

The remainder of the proof is (almost) identical to the proof for parametric models Le Cam and Yang (1990); Kleijn and van der Vaart (2012), replacing the parametric likelihood by $\theta \mapsto \int p_{\theta, \eta}^{(n)}(X^{(n)}) d\Pi_{\mathcal{H}_n}(\eta)$ as in Bickel and Kleijn (2012), details of which can be found in Theorem 3.1.1 of Chae (2015). \square

3 Semi-parametric BvM for linear regression models

3.1 Semi-parametric BvM theorem

Let \mathcal{H} be the set of all continuously differentiable densities η defined on $\mathbb{D} = (-r, r)$ (for some $r \in (0, \infty]$) such that $\eta(x) > 0$ and $\eta(x) = \eta(-x)$ for every $x \in \mathbb{D}$. Equip \mathcal{H} with the Hellinger metric. We consider a model for data satisfying,

$$X_i = \theta^T Z_i + \epsilon_i, \quad \text{for } i = 1, \dots, n, \quad (3.13)$$

where Z_i 's are p -dimensional non-random covariates and the errors ϵ_i are assumed to form an i.i.d. sample from a distribution with density $\eta \in \mathcal{H}$. We prove the BvM theorem for the regression coefficient θ .

Let $P_{\theta, \eta, i}$ denote the probability measure with density $x \mapsto \eta(x - \theta^T Z_i)$ and $\dot{\ell}_{\theta, \eta, i} = \partial \ell_{\theta, \eta, i} / \partial \theta$. Also let P_η be the probability measure with density $p_\eta = \eta$ and $s_\eta(x) = -\partial \ell_\eta(x) / \partial x$. Let $P_{\theta, \eta}^{(n)}$ represent the product measure $P_{\theta, \eta, 1} \times \dots \times P_{\theta, \eta, n}$ and let $\dot{\ell}_{\theta, \eta}^{(n)} = \sum_{i=1}^n \dot{\ell}_{\theta, \eta, i}$. With slight abuse of notation, we treat $p_{\theta, \eta, i}$, $\ell_{\theta, \eta, i}$ and $\dot{\ell}_{\theta, \eta, i}$ as either functions of x or the corresponding random variables when they are evaluated at $x = X_i$. For example, $\dot{\ell}_{\theta, \eta, i}$ represents either the function $x \mapsto \dot{\ell}_{\theta, \eta, i}(x) : \mathbb{D} \mapsto \mathbb{R}^p$ or the random vector $\dot{\ell}_{\theta, \eta, i}(X_i)$. We treat $p_{\theta, \eta}^{(n)}$, $\ell_{\theta, \eta}^{(n)}$ and $\dot{\ell}_{\theta, \eta}^{(n)}$ similarly.

Let $\theta_0 \in \Theta$ and $\eta_0 \in \mathcal{H}$ be the true regression coefficient and error density in the model (3.13). Define specialized KL-balls in $\Theta \times \mathcal{H}$ of the form,

$$B_n(\epsilon) = \left\{ (\theta, \eta) : \sum_{i=1}^n K(p_{\theta_0, \eta_0, i}, p_{\theta, \eta, i}) \leq n\epsilon^2, \sum_{i=1}^n V(p_{\theta_0, \eta_0, i}, p_{\theta, \eta, i}) \leq C_2 n\epsilon^2 \right\}, \quad (3.14)$$

3.1 Semi-parametric BvM theorem

where $K(p_1, p_2) = \int \log(p_1/p_2) dP_1$, $V(p_1, p_2) = \int (\log(p_1/p_2) - K(p_1, p_2))^2 dP_1$, and C_2 is some positive constant (see Ghosal and van der Vaart (2007a)).

Define the mean Hellinger distance h_n on $\Theta \times \mathcal{H}$ by,

$$h_n^2((\theta_1, \eta_1), (\theta_2, \eta_2)) = \frac{1}{n} \sum_{i=1}^n h^2(p_{\theta_1, \eta_1, i}, p_{\theta_2, \eta_2, i}). \quad (3.15)$$

Let $v_\eta = P_{\eta_0}[s_\eta s_{\eta_0}]$ and,

$$V_{n, \eta} = \frac{1}{n} P_0^{(n)} [\dot{\ell}_{\theta_0, \eta}^{(n)} \dot{\ell}_{\theta_0, \eta_0}^{(n)T}]. \quad (3.16)$$

It is easy to see that $V_{n, \eta} = v_\eta \mathbf{Z}_n$, where $\mathbf{Z}_n = n^{-1} \sum_{i=1}^n Z_i Z_i^T$.

We say that a sequence of real-valued stochastic processes $\{Y_n(t) : t \in T\}$, ($n \geq 1$), is *asymptotically tight* if it is asymptotically tight in the space of bounded functions on T with the uniform norm (van der Vaart and Wellner (1996)). A vector-valued stochastic process is asymptotic tight if each of its components is asymptotically tight.

Theorem 2. *Suppose that $\sup_{i \geq 1} |Z_i| \leq L$ for some constant $L > 0$, $\liminf_n \rho_{\min}(\mathbf{Z}_n) > 0$ and $v_{\eta_0} > 0$. The prior for (θ, η) is a product $\Pi = \Pi_\Theta \times \Pi_{\mathcal{H}}$, where Π_Θ is thick at θ_0 . Suppose also that there exist an $N \geq 1$, a sequence $\epsilon_n \rightarrow 0$ with $n\epsilon_n^2 \rightarrow \infty$, and partitions $\Theta = \Theta_{n,1} \cup \Theta_{n,2}$ and $\mathcal{H} = \mathcal{H}_{n,1} \cup \mathcal{H}_{n,2}$ such that $\eta_0 \in \mathcal{H}_{n,1}$ and*

$$\begin{aligned} \log N(\epsilon_n/36, \Theta_{n,1} \times \mathcal{H}_{n,1}, h_n) &\leq n\epsilon_n^2, \\ \log \Pi(B_n(\epsilon_n)) &\geq -\frac{1}{4}n\epsilon_n^2, \\ \log (\Pi_\Theta(\Theta_{n,2}) + \Pi_{\mathcal{H}}(\mathcal{H}_{n,2})) &\leq -\frac{5}{2}n\epsilon_n^2, \end{aligned} \quad (3.17)$$

for all $n \geq N$. For some $\bar{M}_n \uparrow \infty$, with $\epsilon_n \bar{M}_n \rightarrow 0$, let $\mathcal{H}_n = \{\eta \in \mathcal{H}_{n,1} : h(\eta, \eta_0) < \bar{M}_n \epsilon_n\}$ and assume that there exist a continuous $L_2(P_{\eta_0})$ -function

3.1 Semi-parametric BvM theorem

Q and an $\epsilon_0 > 0$ such that,

$$\sup_{|y| < \epsilon_0} \sup_{\eta \in \mathcal{H}^N} \left| \frac{\ell_\eta(x+y) - \ell_\eta(x)}{y} \right| \vee \left| \frac{s_\eta(x+y) - s_\eta(x)}{y} \right| \leq Q(x), \quad (3.18)$$

where $\mathcal{H}^N = \cup_{n=N}^\infty \mathcal{H}_n$. Furthermore, assume that the sequence of stochastic processes,

$$\left\{ \frac{1}{\sqrt{n}} \left(\dot{\ell}_{\theta, \eta}^{(n)} - P_0^{(n)} \dot{\ell}_{\theta, \eta}^{(n)} \right) : |\theta - \theta_0| < \epsilon_0, \eta \in \mathcal{H}^N \right\}, \quad (3.19)$$

indexed by (θ, η) is asymptotically tight. Then the assertion of the BvM theorem 1 holds for θ .

Since the observations are not i.i.d. we consider the mean Hellinger distance h_n as in Ghosal and van der Vaart (2007a). Conditions (3.17) are required for the convergence rate of $h_n((\theta, \eta), (\theta_0, \eta_0))$ to be ϵ_n , which in turn implies that the convergence rates of $|\theta - \theta_0|$ and $h(\eta, \eta_0)$ are ϵ_n (c.f. Lemma S1.1 of supplementary material). In fact, we only need to prove (3.17) with arbitrary rate ϵ_n because the so-called no-bias condition $\sup_{\eta \in \mathcal{H}_n} P_0 \dot{\ell}_{\theta_0, \eta}^{(n)} = o_{P_0}(n^{-1/2})$ holds trivially by the symmetry, which plays an important role to prove (2.2)-(2.4) as in frequentist literature (see Chapter 25 of van der Vaart (1998)). Condition (3.18), which is technical in nature, might be somewhat restrictive. However, (3.18) could be checked easily under a certain compactness assumption for the prior $\Pi_{\mathcal{H}}$. In Section 3.2, for example, we consider a mixture of normal densities for $\Pi_{\mathcal{H}}$ whose mixing distribution is supported on a compact set, and verify (3.18) without much difficulties. It should also be noted that condition (3.18) can be implied by a certain consistency of derivatives which is almost unknown in nonparametric Bayesian contexts; see Shen and Ghosal (2016).

For a random design, (3.19) is asymptotically tight if and only if the class of score functions forms a Donsker class, and sufficient conditions

3.1 Semi-parametric BvM theorem

for the latter are well established in empirical process theory. Since observations are not i.i.d. due to the non-randomness of covariates, (3.19) does not converge in distribution to a Gaussian process. Here, asymptotic tightness of (3.19) merely assures that the supremum of its norm is of order $O_{P_0}(1)$. Asymptotic tightness holds under a finite bracketing integral condition (where the definition of the bracketing number is extended to non-iid observations in a natural way). For sufficient conditions for asymptotic tightness with non-i.i.d. observations, readers are referred to Section 2.11 of van der Vaart and Wellner (1996).

In supplementary material, we prove Theorem 2 by checking the misspecified LAN condition as well as Conditions A and B. Conditions in Theorem 2 depend particularly on the choice of prior for the nuisance parameter η . In the following two subsections, we provide priors satisfying the conditions in Theorem 2: a symmetric Dirichlet mixture of normal distributions and a random series prior on a smoothness class. For a given density p on \mathbb{D} , its *symmetrization* \bar{p} is defined by $\bar{p} = (p + p^-)/2$, where $p^-(x) = p(-x)$ for all $x \in \mathbb{D}$. We can construct a prior on \mathcal{H} by putting a prior on $p \in \tilde{\mathcal{H}}$ and symmetrizing it, where $\tilde{\mathcal{H}}$ is the set of every density on \mathbb{D} whose symmetrization belongs to \mathcal{H} . Obviously, we have $\mathcal{H} \subset \tilde{\mathcal{H}}$. Let $\Pi_{\tilde{\mathcal{H}}}$ be a probability measure on $\tilde{\mathcal{H}}$ and $\Pi_{\mathcal{H}}$ be the corresponding probability measure on \mathcal{H} . Hellinger entropy bounds and prior concentration rates around KL neighborhoods are well known for various choices of $\Pi_{\tilde{\mathcal{H}}}$, so the following lemma is useful to prove (3.17).

Lemma 1. *For a subset $\tilde{\mathcal{H}}_0$ of $\tilde{\mathcal{H}}$ containing η_0 , suppose that there exists a function \tilde{Q} such that $\sup_{\eta \in \tilde{\mathcal{H}}_0} P_{\eta} \tilde{Q}^2 < \infty$, and for every x and sufficiently*

3.2 Symmetric Dirichlet mixtures of normal distributions

small y ,

$$\sup_{\eta \in \tilde{\mathcal{H}}_0} \left| \frac{\log \eta(x+y) - \log \eta(x)}{y} \right| \leq \tilde{Q}(x). \quad (3.20)$$

Furthermore, assume that for large enough n ,

$$\begin{aligned} \log N(\tilde{\epsilon}_n, \tilde{\mathcal{H}}_{n,1}, h) &\lesssim n\tilde{\epsilon}_n^2, \\ \log \Pi_{\tilde{\mathcal{H}}}(\{\eta \in \tilde{\mathcal{H}} : K(\eta_0, \eta) \leq \tilde{\epsilon}_n^2, V(\eta_0, \eta) \leq \tilde{\epsilon}_n^2\}) &\gtrsim -n\tilde{\epsilon}_n^2, \\ \log \Pi_{\tilde{\mathcal{H}}}(\tilde{\mathcal{H}}_{n,2}) &\leq -\frac{5}{2}n\tilde{\epsilon}_n^2 M_n^2, \end{aligned} \quad (3.21)$$

for some partition $\tilde{\mathcal{H}} = \tilde{\mathcal{H}}_{n,1} \cup \tilde{\mathcal{H}}_{n,2}$ with $\eta_0 \in \tilde{\mathcal{H}}_{n,1} \subset \tilde{\mathcal{H}}_0$ and sequences $\tilde{\epsilon}_n \rightarrow 0$, $M_n \rightarrow \infty$ with $\tilde{\epsilon}_n \gtrsim n^{-1/2} \log n$. If Θ is compact and $\sup_{i \geq 1} |Z_i| \leq L$, then, for any Π_{Θ} that is thick at θ_0 , the product prior $\Pi_{\Theta} \times \Pi_{\tilde{\mathcal{H}}}$ satisfies (3.17) with some $\mathcal{H}_{n,1} \subset \mathcal{H}_0$, $\Theta_{n,1} = \Theta$ and $\epsilon_n = \tilde{\epsilon}_n M_n$, where \mathcal{H}_0 is the set of symmetrizations of $p \in \tilde{\mathcal{H}}_0$.

3.2 Symmetric Dirichlet mixtures of normal distributions

Dirichlet process mixture priors are popular and the asymptotic behavior of the posterior distribution is well-studied. A random density η is said to follow a Dirichlet process mixture of normal densities (Lo (1984)) if $\eta(x) = \int \phi_{\sigma}(x-z)dF(z, \sigma)$, where $F \sim \text{DP}(\alpha, H)$ and ϕ_{σ} is the density of the normal distribution with mean 0 and variance σ^2 . Here, $\text{DP}(\alpha, H)$ denotes the Dirichlet process with precision $\alpha > 0$ and mean probability measure H on $\mathbb{R} \times (0, \infty)$ Ferguson (1973).

For given positive numbers σ_1, σ_2 , and M with $\sigma_1 < \sigma_2$, let \mathcal{F} be the set of all distribution functions supported on $[-M, M] \times [\sigma_1, \sigma_2]$, and let $\tilde{\mathcal{H}}_0$ be the set of all densities η on \mathbb{R} of the form $\eta(x) = \int \phi_{\sigma}(x-z)dF(z, \sigma)$

3.3 Random series prior

for $F \in \mathcal{F}$. Then it is easy to show that \mathcal{H}_0 , the symmetrization of $\tilde{\mathcal{H}}_0$, is the set of all $\eta \in \tilde{\mathcal{H}}_0$, where $F \in \mathcal{F}$ with $dF(z, \sigma) = dF(-z, \sigma)$. If $F \sim \text{DP}(\alpha, H)$, where H has a positive and continuous density supported on $[-M, M] \times [\sigma_1, \sigma_2]$, the corresponding Dirichlet process mixture prior and its symmetrization, denoted by $\Pi_{\tilde{\mathcal{H}}}$ and $\Pi_{\mathcal{H}}$, respectively, have full support on $\tilde{\mathcal{H}}_0$ and \mathcal{H}_0 relative to the Hellinger topology.

We note here that the class \mathcal{H}_0 (or $\tilde{\mathcal{H}}_0$), in spite of compact support of mixing distributions, is sufficiently flexible and the posterior convergence rate was also studied with this class of densities; see Ghosal and van der Vaart (2001) and Walker et al. (2007). Eliminating the compactness condition requires significantly different techniques even for deriving the posterior convergence rate (Ghosal and van der Vaart (2007b)). As mentioned earlier, the main challenge lies in condition (3.18), *i.e.* the consistency of derivatives, and we leave this problem for general mixtures as our future work.

Corollary 1. *Suppose that $\sup_{i \geq 1} |Z_i| \leq L$ and $\liminf_n \rho_{\min}(\mathbf{Z}_n) > 0$. With the symmetrized Dirichlet process mixture prior described above for η , the BvM theorem holds for the linear regression model provided that $\eta_0 \in \mathcal{H}_0$ and that Π_{Θ} is compactly supported and thick at θ_0 .*

3.3 Random series prior

Let W be a random function on $[-1/2, 1/2]$ defined as a series $W(\cdot) = \sum_{j=1}^{\infty} j^{-\alpha} C_j b_j(\cdot)$, where $b_1(t) = 1, b_{2j}(t) = \cos(2\pi jt), b_{2j+1}(t) = \sin(2\pi jt)$ and C_j 's are i.i.d. random variables drawn from a density supported on $[-M, M]$ that is continuous and bounded away from zero. We shall impose smoothness through the requirement that α be greater than 3, so that the

series is well defined as a continuous real-valued function on $[-1/2, 1/2]$ with the first and second derivatives that are bounded uniformly by a constant. Let \mathcal{W} be the set of all functions $w : [-1/2, 1/2] \rightarrow \mathbb{R}$ of the form $w(\cdot) = \sum_j a_j b_j(\cdot)$ for some sequence (a_1, a_2, \dots) with $j^\alpha |a_j| \leq M$ for all j . Let $\tilde{\mathcal{H}}_0$ denote the set of densities p_w , where $w \in \mathcal{W}$ and,

$$p_w(x) = \frac{e^{w(x)}}{\int_{-1/2}^{1/2} e^{w(y)} dy},$$

for every $x \in \mathbb{D} = (-1/2, 1/2)$. Let \mathcal{H}_0 denote the associated space of symmetrized \bar{p}_w . Let $\Pi_{\tilde{\mathcal{H}}}$ and $\Pi_{\mathcal{H}}$ be the laws of random densities p_W and \bar{p}_W , respectively.

Corollary 2. *Suppose that $\sup_{i \geq 1} |Z_i| \leq L$ and $\liminf_n \rho_{\min}(\mathbf{Z}_n) > 0$. If $\alpha > 3$, $\eta_0 \in \mathcal{H}_0$, $v_{\eta_0} > 0$, and Π_{Θ} is compactly supported and thick at θ_0 , then the random series prior $\Pi_{\mathcal{H}}$ for η leads to a posterior for θ that satisfies the BvM assertion (2.9) in the linear regression model.*

4 Efficiency in the linear mixed effect model

4.1 Semi-parametric BvM theorem

In this section, we consider the linear mixed effect model (Laird and Ware (1982)),

$$X_{ij} = \theta^T Z_{ij} + b_i^T W_{ij} + \epsilon_{ij}, \quad \text{for } i = 1, \dots, n \text{ and } j = 1, \dots, m_i,$$

where the covariates $Z_{ij} \in \mathbb{R}^p$ and $W_{ij} \in \mathbb{R}^q$ are non-random, the error ϵ_{ij} 's form an i.i.d. sequence drawn from a distribution with density f and the random effect coefficients b_i are i.i.d. from a distribution G . Under

4.1 Semi-parametric BvM theorem

the Gaussian assumption on f and G , the linear mixed effect model has been widely used for analyzing longitudinal data and repeated measures arising in many fields, see Diggle et al. (2002) and references therein. A lot of statistical softwares fitting this model have also been developed, among which the R package “lme4” is the most popular, see Bates et al. (2014). Relaxing the Gaussian assumption on f and G has also been studied, particularly focusing on robust estimation using Student’s t distributions, see Welsh and Richardson (1997); Pinheiro et al. (2001); Song et al. (2007). In Bayesian contexts, Wakefield et al. (1994) studied a Gibbs sampler algorithm with Student’s t distributions. It should be noted that without Gaussian assumption, it is a non-trivial problem to find the maximum likelihood estimator and confidence intervals while Bayesian computational methods are relatively easy to extend. In this regards, semi-parametric linear mixed effect models have been considered in Bayesian framework, see Bush and MacEachern (1996); Kleinman and Ibrahim (1998); Kyung et al. (2010). However, these works suffer from lack of theory.

We assume that both f and G are symmetric for the identifiability of θ . This might be a bit stronger than the zero-mean assumption, but a reasonable compromise between the model flexibility and theoretical development.

Thus, the nuisance parameter $\eta = (f, G)$ takes its values in the space $\mathcal{H} = \mathcal{F} \times \mathcal{G}$, where the first factor \mathcal{F} denotes the class of continuously differentiable densities supported on $\mathbb{D} = (-r, r)$ for some $r \in (0, \infty]$ with $f(x) > 0$ and $f(x) = f(-x)$ for all $x \in \mathbb{D}$ and \mathcal{G} is the class of symmetric distributions supported on $[-M_b, M_b]^q$ for some $M_b > 0$. As in the linear regression model, the compactness condition on the support of \mathcal{G} (and possibly \mathcal{F}), which is assumed for technical convenience, is rather strong. While we leave this technical problem as future work, the numerical results given

4.1 Semi-parametric BvM theorem

in Section 5 show the validity of Bayes estimator without such a strong assumption.

The true value of the nuisance is denoted by $\eta_0 = (f_0, G_0)$. We write $X_i = (X_{i1}, \dots, X_{im_i})^T$, and similarly, $Z_i \in \mathbb{R}^{p \times m_i}$ and $W_i \in \mathbb{R}^{q \times m_i}$. As in the linear regression model, we assume that,

$$|Z_{ij}| \leq L \quad \text{and} \quad |W_{ij}| \leq L, \quad \text{for all } i \text{ and } j. \quad (4.22)$$

Define,

$$p_{\theta, \eta, i}(x) = \int \prod_{j=1}^{m_i} f(x_j - \theta^T Z_{ij} - b_i^T W_{ij}) dG(b_i),$$

where $x = (x_1, \dots, x_{m_i})^T \in \mathbb{R}^{m_i}$. Quantities denoted by $p_{\theta, \eta}^{(n)}$, $\ell_{\theta, \eta, i}$, $\dot{\ell}_{\theta, \eta, i}$ and $\dot{\ell}_{\theta, \eta}^{(n)}$ are defined and used in the same way as in Section 3. The design matrix \mathbf{Z}_n is defined by $\mathbf{Z}_n = n^{-1} \sum_{i=1}^n Z_i Z_i^T$. For technical reasons and notational convenience, we assume that there exists an integer m such that $m_i = m$ for all i , but proofs can be extended to general cases without much hamper.

For $y = (y_1, \dots, y_m)^T \in \mathbb{R}^m$ and $w = (w_1, \dots, w_m) \in [-L, L]^{q \times m}$, define,

$$\psi_\eta(y|w) = \int \prod_{j=1}^m f(y_j - b^T w_j) dG(b),$$

and $\ell_\eta(y|w) = \log \psi_\eta(y|w)$. Let $s_\eta(y|w) = -\partial \ell_\eta(y|w) / \partial y \in \mathbb{R}^m$. Then it can be easily shown that $\dot{\ell}_{\theta, \eta, i}(x) = Z_i s_\eta(x - Z_i^T \theta | W_i) \in \mathbb{R}^p$. Furthermore, let $\Psi_\eta^w(\cdot)$ denote the probability measure on \mathbb{R}^m with density $y \mapsto \psi_\eta(y|w)$. The metric h_n on $\Theta \times \mathcal{H}$ is defined as in (3.15). With slight abuse of notation, we also use h_n as a metric on \mathcal{H} defined as $h_n(\eta_1, \eta_2) = h_n((\theta_0, \eta_1), (\theta_0, \eta_2))$.

Let,

$$d_w^2(\eta_1, \eta_2) = \int |s_{\eta_1}(y|w) - s_{\eta_2}(y|w)|^2 d\Psi_{\eta_0}^w(y).$$

4.1 Semi-parametric BvM theorem

Define $B_n(\epsilon)$ and $V_{n,\eta}$ as in (3.14) and (3.16), respectively. It can be easily shown that,

$$V_{n,\eta} = \frac{1}{n} \sum_{i=1}^n Z_i v_\eta(W_i) Z_i^T, \quad (4.23)$$

where $v_\eta(w)$ is the $m \times m$ matrix defined as,

$$v_\eta(w) = \int s_\eta(y|w) s_{\eta_0}(y|w)^T d\Psi_{\eta_0}^w(y).$$

To prove the BvM assertion in the linear mixed effect model, we need a condition to ensure that $\sup_{i \geq 1} h(\psi_{\eta_n}(\cdot|W_i), \psi_{\eta_0}(\cdot|W_i)) \rightarrow 0$ as $h_n(\eta_n, \eta_0) \rightarrow 0$. For this purpose, we define $N_{n,\epsilon}(u)$ to be the number of W_{ij} 's with $|W_{ij} - u| < \epsilon$, and assume that, for every (fixed) $\epsilon > 0$ and $u \in \mathbb{R}^q$,

$$N_{n,\epsilon}(u) = 0 \text{ for all } n, \text{ or } \liminf_n n^{-1} N_{n,\epsilon}(u) > 0. \quad (4.24)$$

Condition (4.24) is easily satisfied, for example when W_{ij} 's are i.i.d. realization from any distribution. The proof of following theorem is quite similar to that of Theorem 2 except for some technical details.

Theorem 3. *Suppose that $\liminf_n \rho_{\min}(\mathbf{Z}_n) > 0$, $\rho_{\min}(v_{\eta_0}(w)) > 0$ for every w , G_0 is thick at θ , Π_Θ is thick at θ_0 , and $w \mapsto v_{\eta_0}(w)$ is continuous. Also suppose that there exist a large integer N , a sequence (ϵ_n) , with $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \rightarrow \infty$, and sequences of partitions $\Theta = \Theta_{n,1} \cup \Theta_{n,2}$, $\mathcal{H} = \mathcal{H}_{n,1} \cup \mathcal{H}_{n,2}$ such that $\eta_0 \in \mathcal{H}_{n,1}$ and (3.17) holds for all $n \geq N$. For some $\bar{M}_n \uparrow \infty$, with $\epsilon_n \bar{M}_n \rightarrow 0$, let $\mathcal{H}_n = \{\eta \in \mathcal{H}_{n,1} : h_n(\eta, \eta_0) < \bar{M}_n \epsilon_n\}$. Assume that there exists a continuous function Q such that $\sup_w \int Q^3(x, w) \psi_{\eta_0}(x|w) d\mu(x) < \infty$, and,*

$$\sup_{\eta \in \mathcal{H}^N} \frac{|\ell_\eta(x+y|w) - \ell_\eta(x|w)|}{|y|} \vee \frac{|s_\eta(x+y|w) - s_\eta(x|w)|}{|y|} \leq Q(x, w), \quad (4.25)$$

4.1 Semi-parametric BvM theorem

for all x, w and small enough $|y|$, where $\mathcal{H}^N = \cup_{n=N}^{\infty} \mathcal{H}_n$. Also assume that the class of \mathbb{R}^2 -valued functions,

$$\left\{ w \mapsto \left(d_w(\eta_1, \eta_2), h(\psi_{\eta_1}(\cdot|w), \psi_{\eta_2}(\cdot|w)) \right) : \eta_1, \eta_2 \in \mathcal{H}^N \right\}, \quad (4.26)$$

is equicontinuous, and for sufficiently small $\epsilon_0 > 0$ the stochastic process,

$$\left\{ \frac{1}{\sqrt{n}} \left(\dot{\ell}_{\theta, \eta}^{(n)} - P_0^{(n)} \dot{\ell}_{\theta, \eta}^{(n)} \right) : |\theta - \theta_0| < \epsilon_0, \eta \in \mathcal{H}^N \right\}, \quad (4.27)$$

is asymptotically tight. Then, the BvM assertion (2.9) holds for the linear mixed effect model, provided that (4.22) and (4.24) hold.

Let $\tilde{\mathcal{F}}$ (resp. $\tilde{\mathcal{G}}$) be the set of every f (resp. G) whose symmetrization \bar{f} (resp. \bar{G}) belongs to \mathcal{F} (resp. \mathcal{G}), where $\bar{G} = (G + G^-)/2$ with $G^-(A) = G(-A)$ for every measurable set A . For the prior of η , we consider a product measure $\Pi_{\mathcal{F}} \times \Pi_{\mathcal{G}}$, where $\Pi_{\mathcal{F}}$ and $\Pi_{\mathcal{G}}$ are the symmetrized versions of probability measures $\Pi_{\bar{\mathcal{F}}}$ and $\Pi_{\bar{\mathcal{G}}}$ on $\tilde{\mathcal{F}}$ and $\tilde{\mathcal{G}}$, respectively. The following lemma plays the role of Lemma 1. Denote the Lévy-Prohorov metric between two probability measures P_1, P_2 is denoted by $d_W(P_1, P_2)$.

Lemma 2. Let $\mathcal{H}_0 = \mathcal{F}_0 \times \mathcal{G}_0 \subset \mathcal{H}$ for some $\mathcal{F}_0 \subset \mathcal{F}$ and $\mathcal{G}_0 \subset \mathcal{G}$ with $f_0 \in \mathcal{F}_0$ and $G_0 \in \mathcal{G}_0$. Assume that there exist a continuous function Q_0 and small enough $\delta_0 > 0$ such that,

$$\int \sup_w \sup_{\eta \in \mathcal{H}_0} Q_0(x, w)^2 \psi_{\eta}(x|w) d\mu(x) < \infty, \quad (4.28)$$

and,

$$\sup_{\eta \in \mathcal{H}_0} \frac{|\ell_{\eta}(x+y|w) - \ell_{\eta}(x|w)|}{|y|} \vee \left| \frac{\psi_{\eta_0}(x|w)}{\psi_{\eta}(x|w)} \right|^{\delta_0} \leq Q_0(x, w), \quad (4.29)$$

for all x, w and small enough $|y|$. Also assume that \mathcal{F}_0 is uniformly tight and,

$$\sup_{f \in \mathcal{F}_0} \sup_x f(x) \vee |\dot{f}(x)| < \infty, \quad (4.30)$$

4.2 Symmetric Dirichlet mixtures of normal distributions

where \dot{f} is the derivative of f . Then, on $\Theta \times \mathcal{H}_0$,

$$\sup_{n \geq 1} h_n((\theta_1, \eta_1), (\theta_2, \eta_2)) \rightarrow 0, \quad (4.31)$$

as $|\theta_1 - \theta_2| \vee h(f_1, f_2) \vee d_W(G_1, G_2) \rightarrow 0$, and,

$$\sup_{n \geq 1} \frac{1}{n} \sum_{i=1}^n K(p_{\theta_0, \eta_0, i}, p_{\theta, \eta, i}) \vee V(p_{\theta_0, \eta_0, i}, p_{\theta, \eta, i}) \rightarrow 0, \quad (4.32)$$

as $|\theta - \theta_0| \vee h(f, f_0) \vee d_W(G, G_0) \rightarrow 0$.

4.2 Symmetric Dirichlet mixtures of normal distributions

Let $\Pi_{\mathcal{F}}$ denote the symmetric Dirichlet mixtures of normal prior with a compact support defined in Section 3.2 and let \mathcal{F}_0 be the support of $\Pi_{\mathcal{F}}$ in Hellinger metric. Let \mathcal{G}_0 be the support of a prior $\Pi_{\mathcal{G}}$ on \mathcal{G} in the weak topology, and let $\mathcal{H}_0 = \mathcal{F}_0 \times \mathcal{G}_0$. The following corollary proves the BvM theorem for θ .

Corollary 3. *Assume that $\liminf_n \rho_{\min}(\mathbf{Z}_n) > 0$. With the prior $\Pi_{\mathcal{H}}$ described above, the BvM theorem holds for the linear mixed regression model provided that $\eta_0 \in \mathcal{H}_0$, G_0 is thick at 0, and Π_{Θ} is compactly supported and thick at θ_0 , provided (4.22) and (4.24) hold.*

It should be noted that the only condition for $\Pi_{\mathcal{G}}$ is that $G_0 \in \mathcal{G}_0$. Thus, we can consider both parametric and nonparametric priors for G . For example, the multivariate normal distribution truncated on $[-M_b, M_b]^q$ or the symmetrized DP(α, H_G) prior with a distribution H_G on $[-M_b, M_b]^q$ can be used for $\Pi_{\mathcal{G}}$.

4.3 Random series prior

Let $\Pi_{\mathcal{F}}$ be the random series prior defined in Section 3.3 and let \mathcal{F}_0 be the support of $\Pi_{\mathcal{F}}$. Since the distributions in \mathcal{F}_0 have compact supports, the distributions in \mathcal{G}_0 , the support of $\Pi_{\mathcal{G}}$, should have the same support for (4.25) to hold. Hence, we only consider truncated normal distributions truncated on $[-M_b, M_b]^q$ with positive definite covariance matrices. That is, $\mathcal{G}_0 = \{N_{M_b}(0, \Sigma) : 0 < \rho_1 \leq \rho_{\min}(\Sigma) \leq \rho_{\max}(\Sigma) \leq \rho_2 < \infty\}$ for some constants ρ_1 and ρ_2 , where $N_{M_b}(0, \Sigma)$ denotes the truncated normal distribution. Let $\Pi_{\mathcal{H}} = \Pi_{\mathcal{F}} \times \Pi_{\mathcal{G}}$.

Corollary 4. *Assume that $\liminf_n \rho_{\min}(\mathbf{Z}_n) > 0$ and $\rho_{\min}(v_{\eta_0}(w)) > 0$ for every w . With the prior $\Pi_{\mathcal{H}}$ described above, the BvM theorem holds for the linear mixed regression model provided that $\eta_0 \in \mathcal{H}_0$, and Π_{Θ} is compactly supported and thick at θ_0 provided (4.22) and (4.24) hold.*

5 Numerical study

In previous sections, we have proven the semi-parametric BvM theorems in regression models of the form (1.1). However, the required conditions are rather strong, *e.g.* the compactness condition on the mixing distribution of a DP prior is restrictive. The purpose of this section is to cover the discrepancy between theory and practice by performing numerical studies with practically convenient priors; and to illustrate the efficacy of Bayes estimators in linear mixed effect models. We focus on a DP mixture prior because it is computationally more efficient than a series prior such as a Gaussian process, and various tail properties can be covered by mixtures of normal distributions (Andrews and Mallows (1974); West (1987)). Algo-

rithmic details used in this section can be found in Chae (2015).

Firstly, to see the BvM phenomenon, we consider the coverage probability of credible intervals with the linear regression model (3.13). In simulations, a dataset is generated from model (3.13) with various error distributions. Then, the 95% credible (or confidence) set of θ is obtained using various methods. This procedure is repeated for $N = 500$ times, and the coverage probability and relative length of credible (or confidence) intervals for the first component of θ are reported.

We compare the results of 3 methods under 5 error distributions. In all simulations we let $Z_{ij} = (Z_{ij1}, Z_{ij2})^T$, where the Z_{ijk} 's are generated i.i.d. from the Bernoulli distribution with success probability $1/2$. The true parameter θ_0 is set to be $(-1, 1)^T$. For the error distribution, we consider the standard normal distribution (E1), the Student t -distributions with 2 degree of freedom (E2), the uniform(-3,3) distribution (E3), and two mixtures of normal distribution (E4 and E5). For the mixtures we take,

$$p(x) = \sum_{k=1}^K \pi_k \left(\phi_1(x - \mu_k) + \phi_1(x + \mu_k) \right),$$

with $K = 4$,

$$(\mu_1, \mu_2, \mu_3, \mu_4) = (0, 1.5, 2.5, 3.5); \quad (\pi_1, \pi_2, \pi_3, \pi_4) = (0.1, 0.2, 0.15, 0.05),$$

for E4, and $K = 4$,

$$(\mu_1, \mu_2, \mu_3, \mu_4) = (0, 1, 2, 4); \quad (\pi_1, \pi_2, \pi_3, \pi_4) = (0.05, 0.15, 0.1, 0.2),$$

for E5. These two densities (see Figure 5) have two and three modes, respectively.

For the estimators of θ , we consider one frequentist estimator (F) (the least square estimator), and two Bayesian estimators (B1 and B2). For the

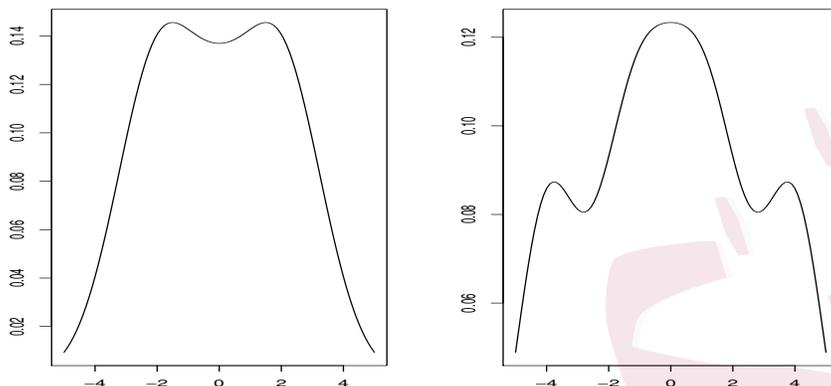


Figure 1: Density plots of error distribution in E4 (left) and E5 (right).

two Bayes estimators, we consider two different priors for the distribution of η : the normal distributions with mean 0 and variance σ^2 (B1), and a symmetrized Dirichlet location mixture of normal distributions with scale parameter σ (B2). A normal distribution is set to be the mean of the Dirichlet process. Independent diffuse normal and inverse gamma distributions are imposed for the prior of θ and σ^2 , respectively.

For each repetition, the sample size n is set to be 200. The coverage probabilities and relative lengths of credible (or confidence) intervals are summarized in Table 1. There is nearly no difference in the results of three methods if the true error distribution is normal. Under the model misspecification, however, the coverage probabilities of parametric methods F and B1 are far from 95% while the semi-parametric approaches are always close to 95%. In particular, credible (or confidence) intervals based on normal models underestimate (overestimate, resp.) the real confidence if the tail of the true distribution is heavier (lighter, resp.) than the Gaussian tail. Furthermore, if the model is misspecified, the lengths of credible (or

confidence) intervals of parametric methods are always larger than that of the semi-parametric method, explaining the BvM phenomenon.

Next, we provide simulation results to illustrate semi-parametric efficacy of the Bayes estimator in the linear mixed effect model. We specialize the model introduced in section 4 slightly: we only consider the random intercept model,

$$X_{ij} = \theta^T Z_{ij} + b_i + \epsilon_{ij}, \quad (5.33)$$

where the b_i 's are univariate random effects following a normal distribution with mean 0 and variance σ_b^2 .

The true parameters θ_0 and σ_{0b}^2 are set to be $(-1, 1)^T$ and 1, respectively, and a dataset is generated with the five error distributions and two covariates same as above. Then, the regression parameters θ are estimated using various methods. The performance of estimation methods is evaluated by mean squared error, $N^{-1} \sum_{k=1}^N |\hat{\theta}_n^{(k)} - \theta_0|^2$, where $\hat{\theta}_n^{(k)}$ is the estimate in the k -th simulation. We compare the performance of 3 estimators. For the estimators of θ , we consider one frequentist estimator (F) (the maximum likelihood estimator under the assumption of a normal error and normal random effect, which is equal to Henderson's best linear unbiased estimator (Henderson (1975))), and two Bayesian estimators (B1 and B2) based on normal and non-parametric priors on f , respectively.

For each of $N = 500$ repetition, we set $n = 20$ and $m_i = 5$ for all i . The mean squared errors and relative efficiencies of the three estimators are summarized in Table 2. As in the previous experiments, there is nearly no difference between three methods if the true error distribution is normal. Otherwise, B2 dominates the other two estimators. It is notable that the losses of efficiency for F and B1 compared to B2 are relatively large when

Table 1: Coverage probability (and relative length of credible (or confidence) intervals with respect to B2) of each methods F, B1 and B2 among $N = 500$ repetitions for each experiment E1–E5.

	F	B1	B2
E1	0.952 (0.999)	0.950 (0.999)	0.948 (1.000)
E2	0.932 (1.913)	0.930 (1.900)	0.952 (1.000)
E3	0.962 (1.393)	0.960 (1.384)	0.948 (1.000)
E4	0.936 (1.126)	0.932 (1.119)	0.946 (1.000)
E5	0.944 (1.145)	0.944 (1.137)	0.946 (1.000)

Table 2: Mean squared error (and relative efficiency with respect to B2) of each methods F, B1 and B2 among $N = 500$ repetitions for each experiment E1–E5.

	F	B1	B2
E1	0.027 (0.983)	0.027 (0.983)	0.028 (1.000)
E2	0.269 (3.056)	0.263 (2.987)	0.088 (1.000)
E3	0.068 (1.404)	0.068 (1.394)	0.049 (1.000)
E4	0.127 (1.176)	0.124 (1.156)	0.108 (1.000)
E5	0.190 (1.128)	0.188 (1.116)	0.168 (1.000)

there is tail mismatch between the prior and true distribution (E2 and E3). In particular, the losses become very large when the error distribution has a heavier tail than the normal distribution.

Supplementary Materials

The supplementary material contains technical proofs for Sections 3 and 4.

Acknowledgements BK thanks the *Statistics Department of Seoul National University, South Korea* for its kind hospitality.

References

- Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 36(1):99–102.
- Barron, A., Schervish, M. J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2):536–561.
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv:1406.5823*.
- Beran, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.*, 6(2):292–313.
- Bickel, P. and Kleijn, B. (2012). The semiparametric Bernstein–von Mises theorem. *Ann. Statist.*, 40(1):206–237.
- Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.*, 10(3):647–671.
- Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (2005). Semi-parametric inference and models. Technical report.

REFERENCES

- Bontemps, D. (2011). Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *Ann. Statist.*, 39(5):2557–2584.
- Boucheron, S. and Gassiat, E. (2009). A Bernstein–von Mises theorem for discrete probability distributions. *Electron. J. Stat.*, 3:114–148.
- Bush, C. A. and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika*, 83(2):275–285.
- Castillo, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields*, 152(1):53–99.
- Castillo, I. and Rousseau, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.*, 43(6):2353–2383.
- Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.*, 43(5):1986–2018.
- Chae, M. (2015). *The semiparametric Bernstein–von Mises theorem for models with symmetric error*. PhD thesis, Seoul National University. *arXiv:1510.05247*.
- Chae, M., Lin, L., and Dunson, D. B. (2016). Bayesian sparse linear regression with unknown symmetric error. *arXiv:1608.02143*.
- Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.*, 21(2):903–923.
- Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

REFERENCES

- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2):209–230.
- Freedman, D. (1999). Wald Lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist.*, 27(4):1119–1141.
- Ghosal, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, 5(2):315–331.
- Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263.
- Ghosal, S. and van der Vaart, A. W. (2007a). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223.
- Ghosal, S. and van der Vaart, A. W. (2007b). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723.
- Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2):423–447.
- Johnstone, I. M. (2010). High dimensional Bernstein–von Mises: simple examples. *Inst. Math. Stat. Collect.*, 6:87–98.
- Kim, Y. (2006). The Bernstein–von Mises theorem for the proportional hazard model. *Ann. Statist.*, 34(4):1678–1700.
- Kim, Y. and Lee, J. (2004). A Bernstein–von Mises theorem in the non-parametric right-censoring model. *Ann. Statist.*, 32(4):1492–1512.
- Kleijn, B. (2013). Criteria for Bayesian consistency. *arXiv:1308.1263*.

REFERENCES

- Kleijn, B. and van der Vaart, A. (2012). The Bernstein–von Mises theorem under misspecification. *Electron. J. Stat.*, 6:354–381.
- Kleinman, K. P. and Ibrahim, J. G. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics*, 54(3):921–938.
- Kyung, M., Gill, J., and Casella, G. (2010). Estimation in Dirichlet random effects models. *Ann. Statist.*, 38(2):979–1009.
- Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.
- Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer, New York.
- Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.*, 12(1):351–357.
- McNeney, B. and Wellner, J. A. (2000). Application of convolution theorems in semiparametric models with non-iid data. *J. Statist. Plann. Inference*, 91(2):441–480.
- Panov, M. and Spokoiny, V. (2015). Finite sample Bernstein–von Mises theorem for semiparametric problems. *Bayesian Anal.*, 10(3):665–710.
- Pinheiro, J. C., Liu, C., and Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *J. Comput. Graph. Statist.*, 10(2):249–276.
- Sacks, J. (1975). An asymptotically efficient sequence of estimators of a location parameter. *Ann. Statist.*, 3(2):285–298.

REFERENCES

- Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26.
- Shen, W. and Ghosal, S. (2016). Posterior contraction rates of density derivative estimation. *Unpublished manuscript*.
- Shen, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Amer. Statist. Assoc.*, 97(457):222–235.
- Song, P. X.-K., Zhang, P., and Qu, A. (2007). Maximum likelihood inference in robust linear mixed-effects models using multivariate t distributions. *Statist. Sinica*, 17(3):929–943.
- Spokoiny, V. (2013). Bernstein-von Mises theorem for growing parameter dimension. *arXiv:1302.3430*.
- Stone, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.*, 3(2):267–284.
- van der Vaart, A. W. (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *Ann. Statist.*, 24(2):862–878.
- van der Vaart, A. W. (1998). *Asymptotic Statistics.*, volume 3. Cambridge university press.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Verlag.
- Wakefield, J., Smith, A., Racine-Poon, A., and Gelfand, A. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Applied Statistics*, 43(1):201–221.

REFERENCES

- Walker, S. (2004). New approaches to Bayesian consistency. *Ann. Statist.*, 32(5):2028–2043.
- Walker, S. G., Lijoi, A., and Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.*, 35(2):738–746.
- Welsh, A. and Richardson, A. (1997). Approaches to the robust estimation of mixed models. In Maddala, G. S. and Rao, C. R., editors, *Handbook of Statistics*, volume 15, chapter 13, pages 343–384. Taylor & Francis.
- West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648.
- Yang, Y., Cheng, G., and Dunson, D. B. (2015). Semiparametric Bernstein-von Mises theorem: Second order studies. *arXiv:1503.04493*.

Department of Mathematics, Applied Mathematics and Statistics
Case Western Reserve University
E-mail: minwooo.chae@gmail.com

Department of Statistics
Seoul National University
E-mail: ydkim0903@gmail.com

Korteweg-de Vries Institute for Mathematics
University of Amsterdam
E-mail: b.kleijn@uva.nl