Statistica Sinica Preprint No: SS-2017-0074	
Title	The semi-parametric Bernstein-von Mises theorem for regression models with symmetric errors
Manuscript ID	SS-2017-0074
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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

Statistica Sinica

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

Minwoo Chae[1], Yongdai Kim[2] and Bas J. K. Kleijn[3]

[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 of a finite-dimensional parameter of interest is expected to be asymptotically equivalent to the sampling distribution of any efficient point estimator. This assertion leads to asymptotic equivalence of the credible and confidence sets of the parameter of interest, and is known as the semi-parametric Bernstein-von Mises theorem. In recent years, this theorem has received much attention and has been widely applied. Here, we consider models in which errors with symmetric densities play a role. Specifically, we show that the marginal posterior distributions of the regression coefficients in linear regression and linear mixed-effect models satisfy the semi-parametric Bernstein-von Mises assertion. As a result, Bayes estimators in these models achieve frequentist inferential optimality, as expressed, for example, in Hájek's convolution and asymptotic minimax theorems. For the prior on the space of error densities, we provide two well-known examples, namely, the Dirichlet process mixture of normal densities and random series priors. The results provide efficient estimates of the regression coefficients in the linear mixed-effect model, for which no efficient point estimators currently exist.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

1 Introduction

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

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. Here, the mean vector $\boldsymbol{\mu}$ is nonrandom 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 $-\boldsymbol{\epsilon}$. Because the error has a symmetric, but otherwise unknown distribution, the model is semi-parametric. Examples of models of the form given in (1.1) include the symmetric location model (where $\mu_i = \theta \in \mathbb{R}$), the linear regression model (where $\mu_i = \theta^T Z_i$ for given covariates $Z_i \in \mathbb{R}^p$), and models with dependent errors, such as linear mixed-effect models (Laird and Ware (1982)).

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

Various studies have developed sets of sufficient conditions for the semi-parametric BvM theorem based on the full local asymptotic normality (LAN) expansion (*i.e.*, the LAN expansion with respect to both the finite

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

and the infinite dimensional parameters; McNeney and Wellner (2000)), including those of Shen (2002), Castillo (2012), Bickel and Kleijn (2012), and Castillo and Rousseau (2015). Because the models we consider are adaptive (Bickel (1982)), we consider a simpler type of LAN expansion that involves only the parameter of interest. Nevertheless, the expansion must be valid under data distributions that differ slightly from the one on which the expansion is centered. We call this property a *misspecified LAN* and prove that it holds for models of the form given in (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 check than those of Castillo (2012) and Castillo and Rousseau (2015) for adaptive models, even though their conditions cover 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 actively studied. Here, examples (Cox (1993); Freedman (1999)) of simple semi-parametric problems with simple choices for the prior have demonstrated the failure of marginals posteriors to display BvM-type asymptotic behavior. Subsequently, positive semi-parametric BvM results have been established in these and other examples, including models of survival analyses (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, such as finite sample properties and second-order asymptotics, are considered in Panov and Spokoiny (2015), Spokoiny (2013), and Yang et al. (2015). Furthermore, Castillo et al. (2015) studied the BvM theorem in high-dimensional models incorporating sparse priors.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

With regard to models of the form given in (1.1), a sizable body of literature has examined efficient point estimations for the symmetric location problem (Beran (1978); Stone (1975); Sacks (1975)) and linear regression models (Bickel (1982)). In contrast, to date, *no efficient point estimators* exist for the regression coefficients in the linear mixed-effect model. However, the semi-parametric BvM theorem proved here implies that the Bayes estimator is efficient. To the best of our knowledge, this study provides the first efficient semi-parametric estimator for linear mixed-effect models. Although the compactness assumptions imposed on the prior and the true error density are rather strong, the numerical study given in section 5 supports the view that the Bayes estimator, which can be computed easily using MCMC methods, is superior to previous methods of estimation. Finally, note that an extension of the current work to a high-dimensional sparse setting is provided by Chae et al. (2016).

The remainder of 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 the linear mixed-effect model, respectively. In each case we consider two common choices for the nuisance prior, namely, a Dirichlet process mixture and a series prior, and we show that both lead to validity of the BvM assertion. The results of numerical studies are presented in section 5. Proofs of the main results can be found in the supplementary material.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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 and $|\cdot|$ represents the Euclidean norm on $\mathbb{R}^d$. The capitals P_η, $P_{\theta,\eta}$, and so on denote the respective probability measures associated with the densities that we write in lower case, p_η, $p_{\theta,\eta}$, and so on (where it is always clear from the context which dominating measure μ is involved). The corresponding log densities are indicated by ℓ_η, $\ell_{\theta,\eta}$, and so on. The Hellinger and total-variational metrics are defined as $h^2(p_1, p_2) = \int \left(\sqrt{p_1} - \sqrt{p_2}\right)^2 d\mu$ 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 that generates the observation, and $X^o = X - P_0 X$ is the centered version of the random variable X. The indicator function for a set A is denoted as 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, respectively (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 LAN

In this section, we prove the semi-parametric BvM theorem for smooth adaptive models, that is, those that satisfy the misspecified LAN expansion

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

defined below. Consider a sequence of statistical models $\mathscr{P}^{(n)} = \{P^{(n)}_{\theta,\eta} : \theta \in \Theta, \eta \in \mathcal{H}\}$ on measurable spaces $(\mathscr{X}^{(n)}, \mathscr{A}^{(n)})$, 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 with the associated Borel σ-algebra, and $P^{(n)}_{\theta,\eta}$ has density $x \mapsto p^{(n)}_{\theta,\eta}(x)$ with respect to some σ-finite measures $\mu^{(n)}$ dominating $\mathscr{P}^{(n)}$.

Let $X^{(n)}$ be a $\mathscr{X}^{(n)}$-valued random element following $P^{(n)}_0$, and assume that $P^{(n)}_0 = P^{(n)}_{\theta_0,\eta_0}$ for some $\theta_0 \in \Theta$ and $\eta_0 \in \mathcal{H}$. We say that a sequence of statistical models $\mathscr{P}^{(n)}$ satisfies the *misspecified LAN expansion* if there exists a sequence of vector-valued (component-wise) $L_2(P^{(n)}_0)$-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,

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

for every compact $K \subset \mathbb{R}^p$, where $\theta_n(h)$ is equal to $\theta_0 + h/\sqrt{n}$. 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), whereas the 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 conditions in Castillo (2012) and Castillo and Rousseau (2015) require a Hilbert space structure on the space of $\eta - \eta_0$, where the LAN expansion and the 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 has the required Hilbert space structure for $\eta - \eta_0$. In contrast, we establish the semi-parametric BvM theorem (Theorem 1)

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

based on the misspecified LAN (2.2), which does not require a Hilbert space structure for $\eta - \eta_0$. Hence, it is relatively easy to check the misspecified LAN for a given prior. For example, in Section 3.2 we verify the misspecified LAN for a mixture of a normal prior and a basis expansion prior.

Although the misspecified LAN expansion (2.2) can be applied only to adaptive cases, verifying (2.2) is not easy, owing to the misspecification and the required uniformity of convergence. LAN expansions have been shown to be valid even under misspecification. For example, Kleijn and van der Vaart (2012) expressed smoothness in misspecified parametric models using 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 errors, the point of minimal KL divergence is equal to θ_0, provided that the misspecified η is sufficiently close 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 $V_{n,\eta} = n^{-1} P_0^{(n)}[g_{n,\eta} g_{n,\eta_0}^T]$. The score function is not necessarily the pointwise derivative of the log-likelihood, but in most examples (including the models considered in this paper), $g_{n,\eta} = \dot{\ell}^{(n)}_{\theta_0,\eta}$, where $\dot{\ell}^{(n)}_{\theta,\eta} = \partial \ell^{(n)}_{\theta,\eta}/\partial\theta$. Henceforth, because it conveys the natural meaning of a derivative, we use the notation $\dot{\ell}^{(n)}_{\theta_0,\eta}$ instead of $g_{n,\eta}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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 $\mathcal{H}_n \subset \mathcal{H}$). Thus, for the semi-parametric BvM theorem, the score function $\dot{\ell}^{(n)}_{\theta_0,\eta}$ and $V_{n,\eta}$ 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 nonsingularity)

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

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

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

Condition B. (Posterior localization)

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

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

(1996)). 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 the scores at (θ_0, η) and (θ_0, η_0) vanishes as η converges to η_0 in the Hellinger distance (*c.f.*, (S1.5) in the supplementary material, which controls variations of the information matrix as η converges to η_0 with $\mathcal{H}_n$). Note that conditions (2.2)–(2.4) lead to the following LAN assertion:

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

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

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 the sufficient conditions are well known (Schwartz (1965); Barron et al. (1999); Walker (2004); Kleijn (2013)). Condition (2.7) requires an $n^{-1/2}$ rate of convergence for the marginal posterior distribution of the parameter of interest. Although (2.7) appears to be rather strong (Yang et al. (2015)), it is clearly a necessary condition. Note that conditions (2.2) and (2.7) can be replaced by

$$\sup_{\theta\in\tilde{\Theta}_n}\sup_{\eta\in\mathcal{H}_n}\left|\log\frac{p^{(n)}_{\theta,\eta}}{p^{(n)}_{\theta_0,\eta}}(X^{(n)})-(\theta-\theta_0)^Tg_{n,\eta}(X^{(n)})+\frac{n}{2}(\theta-\theta_0)^TV_{n,\eta}(\theta-\theta_0)\right|$$
$$=o_{P_0}\big(1+n(|\theta-\theta_0|^2)\big)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

and

$$P_0^{(n)}\Pi\big(\theta \in \widetilde{\Theta}_n | X^{(n)}\big) \to 1,$$

for some $\widetilde{\Theta}_n \subset \Theta$, as in Castillo (2012) and Yang et al. (2015). Thus, if the log-likelihood is uniformly approximated by a quadratic function in a neighborhood where the posterior contracts, then the posterior achieves an $n^{-1/2}$ 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 the neighborhood of θ_0. The following is 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) *and Conditions A and B hold. Then,*

$$\sup_B \Big|\Pi\big(\sqrt{n}(\theta - \theta_0) \in B | X^{(n)}\big) - N_{\Delta_n, V_{n,\eta_0}^{-1}}(B)\Big| \to 0, \tag{2.9}$$

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 sufficiently large n. Let $\Pi_{\mathcal{H}_n}$ be the probability measure obtained by restricting $\Pi_\mathcal{H}$ to $\mathcal{H}_n$ and then re-normalizing, and let $\Pi_{\mathcal{H}_n}(\cdot|X^{(n)})$ be the corresponding posterior distribution. Then, for any measurable set B in Θ,

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

Thus, we have

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

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

In particular,

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

converges to zero in $P_0^{(n)}$-probability from (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

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

Then, trivially, we have

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

and the quantity

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

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

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

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 $\theta \mapsto \int p^{(n)}_{\theta,\eta}(X^{(n)}) d\Pi_{\mathcal{H}_n}(\eta)$, as in Bickel and Kleijn (2012). For further detail, see Theorem 3.1.1 of Chae (2015). □

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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 that satisfies

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

where Z_i denotes the p-dimensional nonrandom 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 $\dot{\ell}_{\theta,\eta,i} = \partial \ell_{\theta,\eta,i}/\partial\theta$. In addition, 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 $P_{\theta,\eta,1} \times \cdots \times P_{\theta,\eta,n}$, and let $\dot{\ell}_{\theta,\eta}^{(n)} = \sum_{i=1}^n \dot{\ell}_{\theta,\eta,i}$. With a 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 as 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 the error density in the model (3.13), respectively. Define specialized KL balls in $\Theta \times \mathcal{H}$ of the form

$$B_n(\epsilon) = \Big\{(\theta, \eta) : \sum_{i=1}^n K(p_{\theta_0,\eta_0,i}, p_{\theta,\eta,i}) \le n\epsilon^2, \sum_{i=1}^n V(p_{\theta_0,\eta_0,i}, p_{\theta,\eta,i}) \le C_2 n\epsilon^2\Big\}, \tag{3.14}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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

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

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

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

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 a uniform norm (van der Vaart and Wellner (1996)). A vector-valued stochastic process is asymptotically 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 too that there exist an* $N \geq 1$, *a sequence* $\epsilon_n \to 0$ *with* $n\epsilon_n^2 \to \infty$, *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*

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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

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

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

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

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

Because the observations are not i.i.d., we consider the mean Hellinger distance h_n, as in Ghosal and van der Vaart (2007a). The conditions given in (3.17) are required for the convergence rate of $h_n\big((\theta,\eta),(\theta_0,\eta_0)\big)$ 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 in the supplementary material). In fact, we need only 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. This plays an important role in proving (2.2)–(2.4), as in the 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) can 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 $\Pi_{\mathcal{H}}$, the mixing distribution of which is supported on a compact set, and verify (3.18) without much difficulty. Note that condition (3.18) is 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 the sufficient conditions

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

for the latter are well established in empirical process theory. Because the observations are not i.i.d., owing to the nonrandomness of the covariates, (3.19) does not converge in distribution to a Gaussian process. Here, asymptotic tightness of (3.19) merely ensures 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-i.i.d. observations in a natural way). For the sufficient conditions for asymptotic tightness in non-i.i.d. observations, refer to Section 2.11 of van der Vaart and Wellner (1996).

In the supplementary material, we prove Theorem 2 by checking the misspecified LAN condition and Conditions A and B. The conditions in Theorem 2 depend particularly on the choice of prior for the nuisance parameter η. In the next two subsections, we provide priors that satisfy 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 $p \in \widetilde{\mathcal{H}}$ and then symmetrizing it, where $\widetilde{\mathcal{H}}$ is the set of every density on $\mathbb{D}$ with symmetrization in $\mathcal{H}$. Obviously, we have $\mathcal{H} \subset \widetilde{\mathcal{H}}$. Let $\Pi_{\widetilde{\mathcal{H}}}$ be a probability measure on $\widetilde{\mathcal{H}}$ and let $\Pi_{\mathcal{H}}$ 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 $\Pi_{\widetilde{\mathcal{H}}}$. Therefore, the following lemma is useful in the proof of (3.17).

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

sufficiently small y,

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

Furthermore, assume that for sufficiently large n,

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

for some partition $\widetilde{\mathcal{H}}=\widetilde{\mathcal{H}}_{n,1}\cup\widetilde{\mathcal{H}}_{n,2}$, with $\eta_0\in\widetilde{\mathcal{H}}_{n,1}\subset\widetilde{\mathcal{H}}_0$, and sequences $\widetilde{\epsilon}_n\to0$, $M_n\to\infty$, with $\widetilde{\epsilon}_n\gtrsim n^{-1/2}\log n$. If Θ is compact and $\sup_{i\ge1}|Z_i|\le L$, then, for any Π_Θ that is thick at θ_0, the product prior $\Pi_\Theta\times\Pi_{\mathcal{H}}$ satisfies (3.17) *for some $\mathcal{H}_{n,1}\subset\mathcal{H}_0$, $\Theta_{n,1}=\Theta$, and $\epsilon_n=\widetilde{\epsilon}_nM_n$, where $\mathcal{H}_0$ is the set of symmetrizations of $p\in\widetilde{\mathcal{H}}_0$.*

3.2 Symmetric Dirichlet mixtures of normal distributions

Dirichlet process mixture priors are popular and the asymptotic behavior of the posterior distribution is well documented. 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 $F\sim\mathrm{DP}(\alpha,H)$ and ϕ_σ is the density of the normal distribution with mean zero and variance σ^2. Here, $\mathrm{DP}(\alpha,H)$ 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 $[-M,M]\times[\sigma_1,\sigma_2]$. Furthermore, let $\widetilde{\mathcal{H}}_0$ be the set of all densities η on $\mathbb{R}$ of the form $\eta(x)=\int\phi_\sigma(x-$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

$z)dF(z,\sigma)$, for $F \in \mathcal{F}$. Then, it is easy to show that $\mathcal{H}_0$, the symmetrization of $\widetilde{\mathcal{H}}_0$, is the set of all $\eta \in \widetilde{\mathcal{H}}_0$, where $F \in \mathcal{F}$ with $dF(z,\sigma) = dF(-z,\sigma)$. If $F \sim \mathrm{DP}(\alpha, H)$, 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 symmerization, denoted by $\Pi_{\widetilde{\mathcal{H}}}$ and $\Pi_{\mathcal{H}}$, respectively, have full support on $\widetilde{\mathcal{H}}_0$ and $\mathcal{H}_0$, relative to the Hellinger topology.

Note that the class $\mathcal{H}_0$ (or $\widetilde{\mathcal{H}}_0$) is sufficiently flexible, in spite of the compact support of the mixing distributions. The posterior convergence rate has also been studied for 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), that is, the consistency of the derivatives. We leave this problem for general mixtures for 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 $b_1(t) = 1, b_{2j}(t) = \cos(2\pi jt)$, and $b_{2j+1}(t) = \sin(2\pi jt)$. In addition, C_j denotes i.i.d. random variables drawn from a density supported on $[-M, M]$, which is continuous and bounded away from zero. We impose smoothness through the requirement that α be greater than

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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

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

where the covariates $Z_{ij} \in \mathbb{R}^p$ and $W_{ij} \in \mathbb{R}^q$ are nonrandom, the errors ϵ_{ij} 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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

the Gaussian assumption on f and G, the linear mixed-effect model has been widely used for analyzing longitudinal data and repeated measures in many fields; see Diggle et al. (2002), and the references therein. Many statistical sofeware packages have been developed to fit this model, 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, focusing particularly on robust estimations using Student's t distributions; see Welsh and Richardson (1997), Pinheiro et al. (2001), and Song et al. (2007). In Bayesian contexts, Wakefield et al. (1994) studied a Gibbs sampler algorithm using Student's t distributions. Note that without the Gaussian assumption, it is a nontrivial problem to find the maximum likelihood estimator and confidence intervals, whereas Bayesian computational methods are relatively easy to extend. In this regard, semi-parametric linear mixed-effect models have been considered in Bayesian frameworks; see Bush and MacEachern (1996), Kleinman and Ibrahim (1998), and Kyung et al. (2010). However, these works suffer from a 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 offers a reasonable compromise between 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 $r \in (0, \infty]$. Here, $f(x) > 0$ 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

Although we leave this technical problem as future work, the numerical results given in Section 5 show the validity of the Bayes estimator without such a strong assumption.

The true value of the nuisance parameter is denoted by $\eta_0 = (f_0, G_0)$. We write $X_i = (X_{i1}, \ldots, X_{im_i})^T$ 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,

$$|Z_{ij}| \le L \quad \text{and} \quad |W_{ij}| \le L, \quad \text{for all } i \text{ and } j. \tag{4.22}$$

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 $x = (x_1, \ldots, x_{m_i})^T \in \mathbb{R}^{m_i}$. 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. However, the proofs can be extended to general cases relatively easily.

For $y = (y_1, \ldots, y_m)^T \in \mathbb{R}^m$ and $w = (w_1, \ldots, w_m) \in [-L, L]^{q \times m}$, 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 is easily shown that $\dot{\ell}_{\theta,\eta,i}(x) = Z_i s_\eta\left(x - Z_i^T \theta | W_i\right) \in \mathbb{R}^p$. 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 a 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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

Define $B_n(\epsilon)$ and $V_{n,\eta}$ as in (3.14) and (3.16), respectively. Then, it is easily shown that

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

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

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

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

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

Condition (4.24) is easily satisfied, for example, when W_{ij} is an i.i.d. realization from any distribution. The proof of the following theorem is quite similar to that of Theorem 2, except for a few 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 0,* Π_Θ *is thick at* θ_0, *and* $w \mapsto v_{\eta_0}(w)$ *is continuous. In addition, suppose there exist a large integer* N; *a sequence* (ϵ_n), *with* $\epsilon_n \downarrow 0$; *and* $n\epsilon_n^2 \to \infty$; *and sequences of 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* (3.17) *holds for all* $n \geq N$. *For some* $\overline{M}_n \uparrow \infty$, *with* $\epsilon_n \overline{M}_n \to 0$, *let* $\mathcal{H}_n = \{\eta \in \mathcal{H}_{n,1} : h_n(\eta, \eta_0) < \overline{M}_n \epsilon_n\}$. *Assume 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,*

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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

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

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

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

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 $\widetilde{\mathcal{F}}$ (resp. $\widetilde{\mathcal{G}}$) be the set of every f (resp. G) the symmetrization of which, $\bar{f}$ (resp. $\overline{G}$), belongs to $\mathcal{F}$ (resp. $\mathcal{G}$), where $\overline{G} = (G + G^-)/2$ 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 the probability measures $\Pi_{\widetilde{\mathcal{F}}}$ and $\Pi_{\widetilde{\mathcal{G}}}$ on $\widetilde{\mathcal{F}}$ and $\widetilde{\mathcal{G}}$, respectively. The following lemma plays the role of Lemma 1. Denote the Lévy–Prohorov metric between two probability measures, P_1 and P_2, as $d_W(P_1, P_2)$.

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 a sufficiently small $\delta_0 > 0$, 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 sufficiently small $|y|$. Furthermore, assume that $\mathcal{F}_0$ is uniformly tight, and

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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

$$\sup_{n\geq 1} h_n\big((\theta_1,\eta_1),(\theta_2,\eta_2)\big) \to 0, \tag{4.31}$$

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

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

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

4.2 Symmetric Dirichlet mixtures of normal distributions

Let $\Pi_{\mathcal{F}}$ denote the symmetric Dirichlet mixtures of a normal prior with the compact support defined in Section 3.2, and let $\mathcal{F}_0$ be the support of $\Pi_{\mathcal{F}}$ in the 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 $\liminf_n \rho_{\min}(\mathbf{Z}_n) > 0$. 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, assuming* (4.22) *and* (4.24) *hold.*

Note 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, we can use the multivariate normal distribution truncated on $[-M_b, M_b]^q$ or the symmetrized $\mathrm{DP}(\alpha, H_G)$ prior with a distribution H_G on $[-M_b, M_b]^q$ for $\Pi_{\mathcal{G}}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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 $\Pi_{\mathcal{F}}$. Because the distributions in $\mathcal{F}_0$ have compact supports, the distributions in $\mathcal{G}_0$ (*i.e.*, the support of $\Pi_{\mathcal{G}}$) should have the same support in order for (4.25) to hold. Hence, we only consider 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* $\liminf_n \rho_{\min}(\mathbf{Z}_n) > 0$ *and* $\rho_{\min}(v_{\eta_0}(w)) > 0$ *for all* w. *With the prior* $\Pi_{\mathcal{H}}$ *described above, the BvM theorem holds for the linear mixed-regression model, assuming provided that* $\eta_0 \in \mathcal{H}_0$ *and* Π_Θ *is compactly supported and thick at* θ_0, *assuming* (4.22) *and* (4.24) *hold.*

5 Numerical study

In previous sections, we have proved the semi-parametric BvM theorems in regression models of the form (1.1). However, the required conditions are rather strong; for example, 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 using practical and convenient priors, as well as to illustrate the efficacy of the Bayes estimators for 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);

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

West (1987)). Details of the algorithms used in this section can be found in Chae (2015).

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

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

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

with $K = 4$, $(\mu_1, \mu_2, \mu_3, \mu_4) = (0, 1.5, 2.5, 3.5)$, and $(\pi_1, \pi_2, \pi_3, \pi_4) = (0.1, 0.2, 0.15, 0.05)$ for E4, and $K = 4$, $(\mu_1, \mu_2, \mu_3, \mu_4) = (0, 1, 2, 4)$, and $(\pi_1, \pi_2, \pi_3, \pi_4) = (0.05, 0.15, 0.1, 0.2)$ 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 two Bayes estimators, we consider two different priors for the distribution of η: a normal distribution with mean zero and variance σ^2 (B1), and a symmetrized Dirichlet location mixture of normal distributions with scale

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

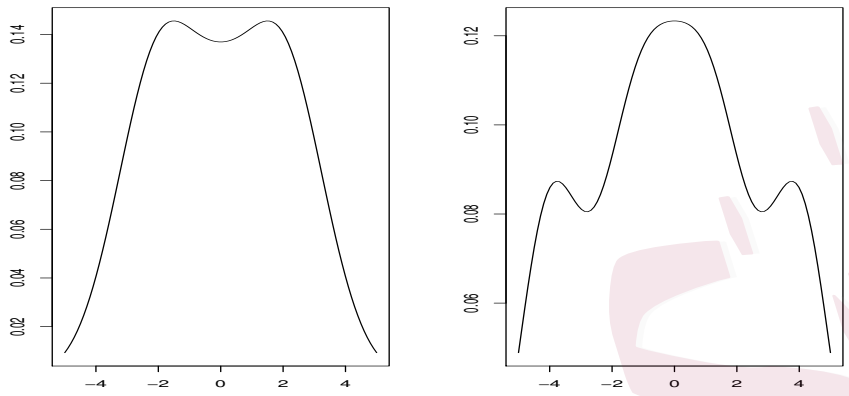

Figure 1: Density plots of the error distributions in E4 (left) and E5 (right).

parameter σ (B2). A normal distribution is set as the mean of the Dirichlet process. Independent diffuse normal and inverse gamma distributions are imposed for the priors of θ and σ^2, respectively.

For each repetition, the sample size n is set as 200. The coverage probabilities and the relative lengths of the credible (or confidence) intervals are summarized in Table 1. There is nearly no difference between the results of three methods if the true error distribution is normal. However, under the model misspecification, the coverage probabilities of the parametric methods F and B1 are far from 95%, whereas the semi-parametric approaches are always close to 95%. In particular, the credible (or confidence) intervals based on the 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 the credible (or confidence) intervals of the parametric methods are always larger than those of the semi-parametric methods, which explaining the BvM phenomenon.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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

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

where b_i denotes the univariate random effects that follow a normal distribution with mean zero and variance σ_b^2.

The true parameters θ_0 and σ_{0b}^2 are set as $(-1, 1)^T$ and 1, respectively, and data sets are generated using the five error distributions and two covariates, as above. Then, the regression parameters θ are estimated using various methods. The performance of the estimation methods is evaluated using the 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 three estimators. For the estimators of θ, we consider one frequentist estimator (F) (the maximum likelihood estimator under the assumption of a normal error and a 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 nonparametric priors on f, respectively.

For each of the $N = 500$ repetitions, 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 the three methods if the true error distribution is normal. Otherwise, B2 outperforms the other two estimators. Note that the loss of efficiency when using F or B1 compared with using B2 is relatively large when there is tail mismatch between the prior and the true distribution (E2 and E3). In particular, the loss becomes very large when the error

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

Table 1: Coverage probability (and the relative lengths of the credible (or confidence) intervals with respect to B2) of each method F, B1, and B2, using $N = 500$ 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 method F, B1, and B2, using $N = 500$ 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)

distribution has a heavier tail than that of the normal distribution.

Supplementary Material

The supplementary material contains technical proofs for Sections 3

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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). Semiparametric inference and models. Technical report.

Bontemps, D. (2011). Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *Ann. Statist.*, 39(5):2557–2584.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2):209–230.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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*.

Kleijn, B. and van der Vaart, A. (2012). The Bernstein–von Mises theorem under misspecification. *Electron. J. Stat.*, 6:354–381.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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.

Walker, S. (2004). New approaches to Bayesian consistency. *Ann. Statist.*, 32(5):2028–2043.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0074

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