Journal of the Korean Statistical Society
https://doi.org/10.1007/s42952-020-00091-4

Online ISSN 2005-2863
Print ISSN 1226-3192

RESEARCH ARTICLE

Bayesian high-dimensional semi-parametric inference beyond sub-Gaussian errors

Kyoungjae Lee[1] · **Minwoo Chae**[2] · **Lizhen Lin**[3]

Received: 2 March 2020 / Accepted: 30 October 2020

Abstract
We consider a sparse linear regression model with unknown symmetric error under the high-dimensional setting. The true error distribution is assumed to belong to the locally β-Hölder class with an exponentially decreasing tail, which does not need to be sub-Gaussian. We obtain posterior convergence rates of the regression coefficient and the error density, which are nearly optimal and adaptive to the unknown sparsity level. Furthermore, we derive the semi-parametric Bernstein-von Mises (BvM) theorem to characterize asymptotic shape of the marginal posterior for regression coefficients. Under the sub-Gaussianity assumption on the true score function, strong model selection consistency for regression coefficients are also obtained, which eventually asserts the frequentist's validity of credible sets.

Keywords High-dimensional semi-parametric model · Posterior convergence rate · Bernstein-von Mises theorem · Strong model selection consistency

1 Introduction

We consider the linear regression model

$$Y = X\theta + \epsilon, \tag{1}$$

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s42952-020-00091-4) contains supplementary material, which is available to authorized users.

✉ Kyoungjae Lee
leekjstat@gmail.com

[1] Department of Statistics, Inha University, Incheon, South Korea

[2] Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, South Korea

[3] Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, Notre Dame, USA

Published online: 12 November 2020

where $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ is a vector of response variables, $X = (x_{ij}) \in \mathbb{R}^{n \times p}$ is the $n \times p$ matrix of covariates whose i-th row is $x_i^T = (x_{i1}, \ldots, x_{ip})$, $\theta \in \mathbb{R}^p$ is the p-dimensional regression coefficient and $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$ is the vector of random errors with $\epsilon_i \overset{i.i.d.}{\sim} \eta$ for $i = 1, \ldots, n$. Statistical inference with the model (1) in high-dimensional settings has received increasing attention in recent years. For the estimability of θ under large p, certain sparsity condition is often imposed which assumes most components of θ are nearly zero. Under the sparsity assumption, regularization methods have been at the center of statistical research due to their computational tractability, ease of interpretation, elegant theory and good performance in practice. Some pioneering references include Tibshirani (1996), Fan and Li (2001), Tibshirani et al. (2005), Zou and Hastie (2005), Zou (2006), Candes and Tao (2007) and Zhang and Zhang (2014). We also refer to the monograph Bühlmann and van de Geer (2011) for reviews with abundant examples.

In a Bayesian framework, the sparsity can be expressed through a prior on θ for which there are two well-known classes: spike-and-slab and continuous shrinkage priors. The former has been considered as the gold standard for sparse priors supported by rich theory, see Castillo et al. (2015), Ročková and George (2018), Martin et al. (2017) and Chae et al. (2019b). Continuous shrinkage priors have been developed as computationally efficient alternatives of spike-and-slab prior, see Polson and Scott (2010), Carvalho et al. (2010), Armagan et al. (2013a, b) and Bhattacharya et al. (2015).

With regard to the high-dimensional regression model (1), there are three fundamental problems attracting statistical interest: (i) recovery of θ; (ii) selection of nonzero coefficients; and (iii) quantifying the uncertainty of inference. Note that even for Bayesian methods, it is common to analyse the performance of those methods from a frequentist's perspective by assuming a true data-generating distribution. Under the assumption that errors are i.i.d. from the standard Gaussian, Castillo et al. (2015) investigated the posterior convergence rate, strong model selection consistency and Bernstein-von Mises (BvM) theorem. Slightly different sets of conditions and priors also lead to similar results, see Shin et al. (2015), Song and Liang (2017), Yang et al. (2016a), Martin et al. (2017) and Yang (2017). Although some of their results, *e.g.* the recovery of θ, tend to be robust to the misspecification of error distribution, Gaussian models have certain limitations; for example, they are vulnerable to outliers. Some theoretical justification for this can be found in Castillo et al. (2015) and Bühlmann and van de Geer (2011).

Another problem of a misspecified Gaussian model arises in model selection. It should be noted that the sub-Gaussianity of the score function is a very important condition for consistent model selection, see Kim and Jeon (2016) and Chae et al. (2019b). Although it is not clear whether this is a necessary condition, empirical results given in Rossell and Rubio (2017) show that a Gaussian model might lead to inconsistency in model selection when true error distributions are heavy-tailed. There are a few works concerning Bayesian variable selection beyond the Gaussian assumption, which however often suffered from lack of theory in high-dimensional setting. See Rossell and Rubio (2017) and references therein for recent advances on Bayesian variable selection without Gaussianity.

Uncertainty quantification, in particular its theoretical justification, is perhaps the most difficult task. In Bayesian methods, the uncertainty of parameters based on posteriors is typically expressed through a credible set, which has frequentist's validity in a smooth parametric model by the BvM theorem, see *e.g.* van der Vaart (1998). Although the BvM theorem cannot be fully extended to high- or infinite-dimensional models, in some models with carefully chosen priors, credible sets can provide valid confidence satisfying certain frequentist's criteria of optimality, often called as non- or semi-parametric BvM theorem, see Castillo and Nickl (2013), Castillo and Nickl (2014), Castillo and Rousseau (2015), Panov and Spokoiny (2015) and Chae et al. (2019a). If the model is misspecified, however, the credible set loses the frequentist's validity even in a very simple parametric model (Kleijn and van der Vaart 2012). Some adjusting techniques are known (Yang et al. 2016b), but they are not applicable more generally.

In this paper, we study frequentist's property of Bayesian methods for model (1) by investigating large sample behavior of the posterior distributions. We assume a symmetric error density η rather than assuming a Gaussian error density. The symmetric assumption might be slightly restrictive in practice, but a good compromise for the theoretical analysis. In fact, a zero mean or median condition might be more realistic, but without symmetric assumption, uncertainty quantification is challenging in a semi-parametric Bayesian framework.

Asymptotic properties of the posterior distribution in a high-dimensional semi-parametric regression model has been extensively studied in Chae et al. (2019b) under a rather strong assumption on η. In particular, they assumed that η is a mixture of Gaussians with a compactly supported mixing distribution, still falling into a sub-Gaussian framework. In this paper, we use the result of Shen et al. (2013) to eliminate this strong assumption. Specifically, the true error density will be assumed to be in a locally β-Hölder class with an exponentially decreasing tail. This is a much weaker assumption than that given in Chae et al. (2019b). In particular, the true error density need to be neither a mixture of Gaussians nor sub-Gaussian. For the prior, a spike-and-slab and a symmetrized Dirichlet process (DP) mixture priors are imposed on θ and η, respectively. Asymptotic results given in this paper provide reasonable sufficient conditions for the frequentist's validity on (i) recovery of θ, (ii) variable selection, and (iii) uncertainty quantification.

It would be worthwhile to mention some technical contributions of this paper. First of all, our results allow error densities whose tails are thicker than sub-Gaussian for which well-known concentration bounds such as the Hoeffding's inequality make the proof simpler. Although the results are limited to exponentially decaying tails, it is highly expected that recent advances on heavy tail distributions (Canale and De Blasi 2017) are also applicable. Secondly, we provide simpler proof for posterior convergence rates compared to that of Chae et al. (2019b). To derive the posterior convergence rates, they used the misspecified LAN (local asymptotic normality) and some bounded conditions for empirical process, which turn out to be not necessary using our techniques.

The rest of the paper is organized as follows. In Sect. 2, we define the model and prior with some preliminary materials. In Sect. 3, main results on posterior convergence

rates, asymptotic shape and selection property are presented. Concluding remarks follow in Sect. 4, and technical proofs are given in the Supplementary Material.

2 Preliminaries

2.1 Notations

For any positive sequences a_n and b_n, $a_n = o(b_n)$ implies that $a_n/b_n \longrightarrow 0$ as $n \to \infty$. We denote $a_n \lesssim b_n$, or equivalently $a_n = O(b_n)$, if $a_n \leq Cb_n$ for all sufficiently large n and some constant $C > 0$, which is an absolute constant or at least does not depend on n and p. For any $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the largest integer which is smaller than or equal to x. For any constants a and b, we denote $a \vee b$ as the maximum of a and b. We denote the indicator function for some set A as $I_A(\cdot)$ and $I(\cdot \in A)$. For any $\theta \in \mathbb{R}^p$, the support of θ is denoted by S_θ, which is the nonzero index of θ, i.e. $S_\theta = \{1 \leq i \leq p : \theta_i \neq 0\}$. We denote the cardinality of S_θ as $s_\theta = |S_\theta|$. For any index set $S \subseteq \{1, \ldots, p\}$ and $n \times p$ matrix X, let $\theta_S = (\theta_i)_{i \in S} \in \mathbb{R}^{|S|}$, $\widetilde{\theta}_S = (\theta_i I(i \in S))_{1 \leq i \leq p} \in \mathbb{R}^p$ and $X_S = (X_j)_{j \in S} \in \mathbb{R}^{n \times |S|}$, where X_j is the j-th column of X. For any $y \in \mathbb{R}$ and density η, we denote $\ell_\eta(y) = \log \eta(y)$, $\dot{\ell}_\eta(y) = \partial \ell_\eta(y)/\partial y$, $\ddot{\ell}_\eta(y) = \partial^2 \ell_\eta(y)/(\partial y)^2$ and $\dddot{\ell}_\eta(y) = \partial^3 \ell_\eta(y)/(\partial y)^3$, whenever they exist. Similarly, for any $x \in \mathbb{R}^p$ and $\theta \in \mathbb{R}^p$, let $\ell_{\theta,\eta}(x, y) = \ell_\eta(y - x^T\theta)$, $\dot{\ell}_{\theta,\eta}(x, y) = \dot{\ell}_\eta(y - x^T\theta)x$ and $\ddot{\ell}_{\theta,\eta}(x, y) = \ddot{\ell}_\eta(y - x^T\theta)xx^T$. Let $\mathbb{E}_{\theta_0,\eta_0}$ be the expectation under $\mathbb{P}_{\theta_0,\eta_0}$ and $\mathbb{P}_{\theta,\eta}$ be the probability measure corresponding to the model (1). We denote $\mathbb{E}_{\eta_0} = \mathbb{E}_{0,\eta_0}$ for simplicity of exposition. For given a sequence of random variables Y_n, $Y_n = o_{P_0}(1)$ means that Y_n converges to zero in $\mathbb{P}_{\theta_0,\eta_0}$-probability as $n \to \infty$. For given a real function $f : \mathbb{R}^p \times \mathbb{R} \mapsto \mathbb{R}$ and the data $D_n = ((Y_i, x_i))_{i=1}^n$ from the model (1), we define $L_n(\theta, \eta) = \sum_{i=1}^n \ell_{\theta,\eta}(x_i, Y_i)$, $R_n(\theta, \eta) = \prod_{i=1}^n \eta(Y_i - x_i^T\theta)/\eta_0(Y_i - x_i^T\theta_0)$,

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(x_i, Y_i),$$

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{f(x_i, Y_i) - \mathbb{E}_{\theta_0,\eta_0}\left[f(x_i, Y_i)\right]\right\} \text{ and}$$

$$V_{n,\eta} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\theta_0,\eta_0}\left[\dot{\ell}_{\theta_0,\eta}\dot{\ell}^T_{\theta_0,\eta_0}(x_i, Y_i)\right].$$

Note that $V_{n,\eta} = \nu_\eta \Sigma$, where $\nu_\eta = \mathbb{E}_{\eta_0}(\dot{\ell}_\eta \dot{\ell}_{\eta_0})$, $\Sigma = n^{-1}X^TX$. Let $N_{n,\eta,S}$ be the $|S|$-dimensional normal distribution with mean $V^{-1}_{n,\eta,S}G_{n,\eta,S}$ and variance $V^{-1}_{n,\eta,S}$, where $G_{n,\eta,S}$ is the $|S|$-dimensional projection of $\mathbb{G}_n\dot{\ell}_{\theta_0,\eta}$, $V_{n,\eta,S} = \nu_\eta \Sigma_S$ and $\Sigma_S = n^{-1}X_S^TX_S$. For simplicity, we denote $G_{n,\eta_0,S}$, $V_{n,\eta_0,S}$ and $N_{n,\eta_0,S}$ as $G_{n,S}$, $V_{n,S}$ and $N_{n,S}$, respectively. For given positive real numbers a and b, we denote $IG(a, b)$ as an inverse gamma distribution whose shape and scale parameters are a and b, respectively. For given positive integer p, $\mu_0 \in \mathbb{R}^p$ and $p \times p$ positive definite matrix Σ_0, we denote $N_p(\mu_0, \Sigma_0)$ as a p-dimensional normal distribution with mean μ_0 and covariance matrix Σ_0.

For any $\theta \in \mathbb{R}^p$, denote the vector ℓ_q-norm as $\|\theta\|_q := \left(\sum_{j=1}^p |\theta_i|^q\right)^{1/q}$. For any pair of densities η_1 and η_2 with respect to a probability measure μ, define the total variation and Hellinger distance as $d_V(\eta_1, \eta_2) := \int |\eta_1 - \eta_2| d\mu$ and $d_H^2(\eta_1, \eta_2) := \int (\sqrt{\eta_1} - \sqrt{\eta_2})^2 d\mu$, respectively. For any pairs of vectors $\theta^1, \theta^2 \in \mathbb{R}^p$ and densities η_1, η_2, we define the mean Hellinger distance as

$$d_n^2\big((\theta^1, \eta_1), (\theta^2, \eta_2)\big) = \frac{1}{n} \sum_{i=1}^n d_H^2(p_{\theta^1, \eta_1, i}, p_{\theta^2, \eta_2, i}),$$

where $p_{\theta,\eta,i}(y) = \eta(y - x_i^T\theta)$.

2.2 Prior

As mentioned earlier, we consider the following model

$$\begin{aligned} Y_i &= x_i^T\theta + \epsilon_i, \\ \epsilon_i &\overset{i.i.d.}{\sim} \eta, \quad i = 1, \dots, n, \end{aligned}$$

where $\theta \in \mathbb{R}^p$ and η is a symmetric density. We impose prior distributions on θ and η to conduct Bayesian inference. Let $\Theta = \mathbb{R}^p$ and $\mathcal{H}$ be the class of symmetric and continuously differentiable densities equipped with the Hellinger metric. We use a product prior $\Pi = \Pi_\Theta \times \Pi_\mathcal{H}$ for (θ, η), where Π_Θ and $\Pi_\mathcal{H}$ are Borel probability measures on Θ and $\mathcal{H}$, respectively.

For the prior Π_Θ on the coefficient vector θ, we select (i) the number of nonzero components s from a prior $\pi_p(s)$ on $\{0, \dots, p\}$, (ii) a random set $S \subseteq \{1, \dots, p\}$ whose cardinality is $s = |S|$ from the uniform prior, and (iii) the nonzero values θ_S from a prior g_S on $\mathbb{R}^{|S|}$ in turn. Specifically, we consider the following prior distribution on (S, θ):

$$(S, \theta) \mapsto \pi_p(|S|) \frac{1}{\binom{p}{|S|}} g_S(\theta_S)\, \delta_0(\theta_{S^c}),$$

where δ_0 is the Dirac measure at 0. This type of prior has been studied by George and Foster (2000), Scott and Berger (2010), Castillo and van der Vaart (2012) and Castillo et al. (2015). For the prior π_p and g_S, we assume that

$$A_1 p^{-A_3} \pi_p(s-1) \le \pi_p(s) \le A_2 p^{-A_4} \pi_p(s-1), \quad s = 1, \dots, p \tag{2}$$

$$g_S(\theta_S) = \left(\frac{\lambda}{2}\right)^{|S|} \exp(-\lambda \|\theta_S\|_1), \quad \frac{\sqrt{n}}{p} \le \lambda \le \sqrt{n \log p}, \tag{3}$$

for some positive constants A_1, A_2, A_3 and A_4. Note that the prior g_S is the product of the Laplace distribution $g(\theta) = \lambda \exp(-\lambda|\theta|)/2$, i.e., $g_S(\theta_S) = \prod_{i \in S} g(\theta_i)$.

For the prior $\Pi_{\mathcal{H}}$ on the error density η, we consider the location mixture of a symmetrized DP,

$$\begin{aligned}\eta(x) &= \int \phi_\sigma(x-z)d\bar{F}(z),\\ F &\sim DP(\alpha),\\ \sigma^2 &\sim G\end{aligned}$$

where $\phi_\sigma(x) := (\sqrt{2\pi}\sigma)^{-1}\exp\{-x^2/(2\sigma^2)\}, \bar{F} := (F+F^-)/2, dF^-(z) := dF(-z)$ and $DP(\alpha)$ is the Dirichlet process with a finite positive measure α. For the base measure α and the prior on σ^2, we further assume that

$$\bar{\alpha} \in \mathcal{M}[-C'n, C'n], \tag{4}$$

$$\bar{\alpha}([-x,x]^c) \le \exp(-C''x^{a_1}) \text{ for all sufficiently large } x > 0, \tag{5}$$

$$G(\sigma^2 \le x) \le \exp(-C''x^{-a_2}) \text{ for all sufficiently small } x > 0, \tag{6}$$

$$G(\sigma^2 \ge x) \le x^{-a_3} \text{ for all sufficiently large } x > 0, \tag{7}$$

$$G(s < \sigma^{-2} < s(1+t)) \ge a_6 s^{a_4} t^{a_5} \exp(-C''s^{\kappa/2}) \text{ for any } s > 0 \text{ and } t \in (0,1), \tag{8}$$

for some positive constants $a_1, \ldots, a_6, C', C''$ and κ, where $\bar{\alpha} = \alpha/\alpha([-C'n, C'n])$ and $\mathcal{M}[a,b]$ is the set of probability measures on (a, b). We assume that $\bar{\alpha}$ has a positive density function on $(-C'n, C'n)$.

We need additional assumptions to achieve a distributional approximation and model selection consistency. Specifically, we assume that

$$\bar{\alpha} \in \mathcal{M}\left[-C'(\log n)^{\frac{2}{\tau}}, C'(\log n)^{\frac{2}{\tau}}\right], \tag{9}$$

$$G \in \mathcal{M}[0, C'\log n], \tag{10}$$

and $\bar{\alpha} = \alpha/\alpha([-C'(\log n)^{\frac{2}{\tau}}, C'(\log n)^{\frac{2}{\tau}}])$ has a positive density function on $(-C'(\log n)^{\frac{2}{\tau}}, C'(\log n)^{\frac{2}{\tau}})$, where $\tau > 0$ will be used to define true parameter class (condition (D2)) in Sect. 2.3.

Remark 1 The above prior conditions are mild which include popular prior choices. If we choose $\bar{\alpha}$ as a truncated normal distribution on interval $[-n, n]$, conditions (4) and (5) are satisfied with $a_1 = 2$. If we consider $\sigma^{m_0} \sim IG(a_0, b_0)$ for some positive constants a_0, b_0 and m_0, conditions (6)–(8) are satisfied with $a_2 = m_0/2$ and $\kappa = m_0$. For conditions (9) and (10), it suffices to consider the truncated normal and inverse-gamma distribution on $(-C'(\log n)^{\frac{2}{\tau}}, C'(\log n)^{\frac{2}{\tau}})$ and $(0, C'\log n)$, respectively, for some large constant $C' > 0$. As n grows to infinity, the above supports in (9) and (10) are getting close to the whole supports, $\mathbb{R}$ and $\mathbb{R}^+ = (0, \infty)$.

2.3 True parameter class

We focus on the "large p and small n" setting, i.e. $p \geq n$, throughout the paper. We assume $\theta_0 \in \mathbb{R}^p$ to be a s_0-sparse vector, which means the number of nonzero elements of θ_0 is equal to s_0. The support of θ_0 is denoted by $S_0 = S_{\theta_0}$. Further conditions on the sparsity s_0 and the magnitude of θ_0 will be introduced in the main theorems in Sect. 3. We introduce here conditions (D1)–(D5) for the true error density η_0:

(D1) (Locally β-Hölder class) for given positive constants β, τ_0 and a real-valued function L, $\eta_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R})$ where $\mathcal{C}^{\beta,L,\tau_0}(\mathbb{R})$ is the class of every density η whose kth order derivative $\eta^{(k)}$ exists up to $k \leq \lfloor\beta\rfloor$ and for $k_1 = \lfloor\beta\rfloor$,

$$\left|\eta^{(k_1)}(x+y) - \eta^{(k_1)}(x)\right| \leq L(x)\exp(\tau_0 y^2)|y|^{\beta-\lfloor\beta\rfloor}, \quad \forall x, y \in \mathbb{R}.$$

(D2) (Light tail) There exist positive constants a, b and τ such that

$$\eta_0(x) \leq \exp(-b|x|^{\tau}), \quad |x| > a.$$

(D3) There exists a constant $\upsilon > 0$ such that $\mathbb{E}_{\eta_0}\left(|\eta_0^{(k)}|/\eta_0\right)^{(2\beta+\upsilon)/k} < \infty$ and $\mathbb{E}_{\eta_0}\left(L/\eta_0\right)^{(2\beta+\upsilon)/\beta} < \infty$ for $1 \leq k \leq \lfloor\beta\rfloor$.

(D4) (Symmetry) $\eta_0(x) = \eta_0(-x)$ and $\eta_0(x) > 0$ for all $x \in \mathbb{R}$.

(D5) there exist positive constants $\gamma_1, \gamma_2, \gamma_3, b', C_{\eta_0}$ and $\tau' < \tau$ such that for any $y \in \mathbb{R}$,

$$|\dot{\ell}_{\eta_0}(y)| \leq C_{\eta_0}(|y|^{\gamma_1} + 1), \tag{11}$$

$$|\ddot{\ell}_{\eta_0}(y)| \leq C_{\eta_0}(|y|^{\gamma_2} + 1), \tag{12}$$

$$|\dddot{\ell}_{\eta_0}(y)| \leq C_{\eta_0}(|y|^{\gamma_3} + 1), \tag{13}$$

and, for any small $|x|$,

$$\frac{\eta_0(y+x)}{\eta_0(y)} \leq C_{\eta_0} e^{b'|y|^{\tau'}}. \tag{14}$$

Now we describe the above conditions in more details. Condition (D1), locally β -Hölder class, has been extensively studied in Kruijer et al. (2010), Shen et al. (2013), Canale and De Blasi (2017) and Bochkina and Rousseau (2017). This class is much more general than the Hölder class because it only requires the local smoothness by adopting $L(x)$ instead of a constant $L > 0$. Furthermore, due to condition (D2), it is essentially weaker than the condition in Kruijer et al. (2010), which assumes $\log \eta_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R})$ (Shen et al. 2013).

Condition (D2) ensures that the true density has an exponentially light tail. It is mainly required to prove the prior thickness condition for the density part and use the Hanson-Wright inequality for the strong model selection consistency. The technical details for the former issue can be found in Shen et al. (2013) (Lemma 2,

Theorem 3 and Proposition 1). Recently, Bochkina and Rousseau (2017) adopted much weaker tail condition, $\int_x^\infty y^2\eta_0(y)dy \le C(1+x)^{-\tau}$ for some constants $C > 0$ and $\tau > 0$, which includes some polynomially decreasing tail densities. However, they considered only the densities on $\mathbb{R}^+$, and it is unclear whether their techniques are applicable to the densities on $\mathbb{R}$.

Condition (D3) is needed for the prior thickness condition for the density (Shen et al. 2013) and implicitly controls the tail behavior of η_0. It has been commonly used in literature including Kruijer et al. (2010), Shen et al. (2013) and Bochkina and Rousseau (2017).

Condition (D4) is not needed for proving the optimal convergence results (Theorems 1 and 2). However, the symmetric assumption will play an important role in proving the BvM theorem (Theorem 3). Based on current techniques in this paper, this assumption is also needed to prove Corollaries 1 and 2, although we suspect that this can be weakened.

Condition (D5) is required only for the BvM theorem and selection consistency results. This condition is closely related to the tail of η_0 and satisfied for a wide range of densities. For example, if $\eta_0(y) \propto \exp(-a|y|^b)$ for some constants $a, b > 0$ and every large enough $|y|$, condition (D5) is met. In fact, it holds unless η_0 has an extremely thin tail.

2.4 Design matrix

We consider a fixed design matrix $X \in \mathbb{R}^{n\times p}$ and assume that every element of the design matrix is bounded by $\sqrt{\log p}$ up to some constant, i.e. $\sup_{i,j}|x_{ij}| \le M\sqrt{\log p}$ for some constant $M > 0$. The upper bound for entries of the design matrix is introduced due to technical reasons, and some recent works (Narisetty et al. 2019; Song and Liang 2017) on high-dimensional Bayesian inference also used similar conditions. In this paper, we require this condition mainly to (i) derive the posterior convergence rate for $\|X(\theta-\theta_0)\|_2$ using Corollary 3.2 of Chae et al. (2019b) and (ii) obtain upper bounds for $\dot{\ell}_{\theta,\eta}(x,y)$ (or its derivatives) based on $\dot{\ell}_\eta(y)$ (or its derivatives).

In high-dimensional linear regression model (1), certain *regularity conditions* have been imposed on the design matrix X for the estimability of θ. In this paper, we define the *uniform compatibility number* by

$$\phi^2(s) = \inf\left\{ s_\theta \cdot \frac{\theta^T\Sigma\theta}{\|\theta\|_1^2} \ : \ \theta\in\mathbb{R}^p,\ 0 < s_\theta \le s \right\}$$

and the *restricted eigenvalue* by

$$\psi^2(s) = \inf\left\{ \frac{\theta^T\Sigma\theta}{\|\theta\|_2^2} \ : \ \theta\in\mathbb{R}^p,\ 0 < s_\theta \le s \right\}$$

for any $1 \le s \le p$, where $\Sigma = n^{-1}X^TX$, $s_\theta = |S_\theta|$ and S_θ is the support of θ. These quantities have been commonly used in literature (van de Geer and Bühlmann 2009;

Bickel et al. 2009; Castillo et al. 2015), and the bounded below conditions have been introduced for consistent estimation. Note that the infimum, which is used to define compatibility number (or restricted eigenvalue), is often taken over all $S \subseteq \{1, \ldots, p\}$ and $\theta \in \mathbb{R}^p$ such that $\|\theta_{S^c}\|_1 \leq c\|\theta_S\|_1$ for some constant $c > 0$. However, our definitions for $\phi^2(s)$ and $\psi^2(s)$ focus on sparse vectors θ such that $0 < s_\theta \leq s$. For example, Castillo et al. (2015) uses similar definitions. Bounded below assumption on $\psi(s)$ is required for the convergence rate under ℓ_2 norm and BvM result, while the same assumption on $\phi(s)$ is required for the convergence rate under ℓ_1 norm. If the restricted eigenvalue $\psi^2(s)$ is bounded away from zero, it implies that Σ_S is positive definite for any $|S| = s$. Note that $\psi(s) \leq \phi(s)$ because $\|\theta\|_1^2 \leq s_\theta \|\theta\|_2^2$ by the Cauchy-Schwartz inequality. Thus, the restricted eigenvalue conditions is stronger than the uniform compatibility number condition.

Because $p \geq n$, if we consider, for example, a random design matrix $X = (x_{ij})$, where x_{ij}'s are random samples from the standard normal, $\sup_{i,j} |x_{ij}| \leq M\sqrt{\log p}$ and the restricted eigenvalue condition are met with high probability tending to 1 as $p \to \infty$. Furthermore, by Lemma 6.1 in Narisetty and He (2014), these conditions are also satisfied with high probability tending to 1 if the rows of X are independent isometric sub-Gaussian random vectors.

3 Main results

3.1 Posterior convergence rates

The first theorem is about the model dimension which states that the posterior distribution puts most of its mass on moderately small dimensional models. We denote the posterior distribution based on D_n as $\Pi(\cdot \mid D_n)$.

Theorem 1 *Assume that conditions (2)–(8) hold, $\lambda\|\theta_0\|_1 = O(s_0 \log p)$ and $\log p \leq n^2$. Then, for any η_0 satisfying (D1)–(D4), there exists a constant $K_{\dim} > 1$ not depending on n and p such that*

$$\mathbb{E}_{\theta_0,\eta_0} \Pi\left(s_\theta > K_{\dim}\left\{s_0 \vee n^{\kappa^*/(2\beta+\kappa^*)}(\log n)^{2t-1}\right\} \mid D_n\right) = o(1)$$

where $\kappa^* := (\kappa \vee 1)$ and $t > \{\kappa^*(1 + \tau^{-1} + \beta^{-1}) + 1\}/(2 + \kappa^*\beta^{-1})$.

Since we use a Laplace prior for nonzero coefficients, the condition $\lambda\|\theta_0\|_1 = O(s_0 \log p)$ might seem to a bit restrictive. Note that this condition can be avoided in Gaussian models by utlizing explicit form of the log-likelihood, see Castillo et al. (2015), van der Pas et al. (2016) and Gao et al. (2015). To use the same technique in our semi-parametric model, quadratic approximation of the log-likelihood should be preceded, for which empirical process techniques can be applied. However, quadratic approximation is highly difficult when models have many nonzero coefficients. Therefore, the proof of Theorem 1 heavily relies on the prior, requiring an additional condition $\lambda\|\theta_0\|_1 = O(s_0 \log p)$. An empirical Bayes approach proposed in Martin et al. (2017) might be an alternative way to

relax this condition. However, the choice of the least squared estimators as the center of the prior may yield another problems when errors have heavier tails than the sub-Gaussian tail. Since we believe the condition $\lambda\|\theta_0\|_1 = O(s_0 \log p)$ is not too restrictive under the large λ regime, we leave the problem of relaxing this condition as future work.

For a given $t > \{\kappa^*(1+\tau^{-1}+\beta^{-1})+1\}/(2+\kappa^*\beta^{-1})$, let $s_n := 2K_{\dim}\{s_0 \vee n^{\kappa^*/(2\beta+\kappa^*)}(\log n)^{2t-1}\}$. Theorem 1 effectively reduces the meaningful parameter space when s_n is not too big and makes the theoretical development easier. Theorem 2 describes a result on posterior convergence rate under the mean Hellinger distance. The obtained rate has the term s_n defined above, where s_0 and $n^{\kappa^*/(2\beta+\kappa^*)}(\log n)^{2t-1}$ come from the coefficient and density estimation, respectively.

Theorem 2 *Assume that conditions (2)–(8) hold, $\lambda\|\theta_0\|_1 = O(s_0 \log p)$ and $s_n \log p = o(n)$. Then, for any η_0 satisfying (D1)–(D4),*

$$\mathbb{E}_{\theta_0,\eta_0} \Pi\left(d_n\big((\theta,\eta),(\theta_0,\eta_0)\big) > K_{\mathrm{Hel}}\sqrt{\frac{s_n \log p}{n}} \;\middle|\; D_n \right) = o(1),$$

for some constant $K_{\mathrm{Hel}} > 0$ not depending on n and p.

Remark 2 The symmetric condition (D4) is not directly used in the proof of Theorems 1 and 2. Hence, they can be easily re-stated without (D4). We did not try to re-state them because it entails a redefinition of the prior and a lot of minor changes. We need the symmetric assumption for the BvM theorem, particularly for proving that the score function has zero expectation, that is, $\mathbb{E}_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta} = 0$ for symmetric η. This will play an important role in the proof of the misspecified LAN, see Lemma 11 of the Supplement. Finally, we note that the current proof of Corollaries 1 and 2 below relies on the symmetric assumption (D4), but it might be possible to prove them without it.

Based on Theorem 2, the posterior convergence rate of η and θ can be achieved as follows. The proof of Corollary 2 is straightforward by Theorem 2 and similar arguments used in the proof of Corollary 3.2 of Chae et al. (2016), so we omit the proof here.

Corollary 1 *Under the conditions of Theorem 2, we have*

$$\mathbb{E}_{\theta_0,\eta_0} \Pi\left(d_H(\eta,\eta_0) > K_{\mathrm{eta}}\sqrt{\frac{s_n \log p}{n}} \;\middle|\; D_n \right) = o(1),$$

for some constant $K_{\mathrm{eta}} > 0$ not depending on n and p, and for any η_0 satisfying (D1)–(D4) with $2\beta + \upsilon \geq 2$.

Corollary 2 *Under the conditions of Theorem 2 and $s_n \log p/\phi(s_n) = o(\sqrt{n})$, we have*

$$\mathbb{E}_{\theta_0,\eta_0}\Pi\left(\|\theta-\theta_0\|_1 > K_{\text{theta}}\frac{s_n}{\phi(s_n)}\sqrt{\frac{\log p}{n}}\;\middle|\; D_n\right) = o(1),$$

$$\mathbb{E}_{\theta_0,\eta_0}\Pi\left(\|\theta-\theta_0\|_2 > K_{\text{theta}}\frac{1}{\psi(s_n)}\sqrt{\frac{s_n\log p}{n}}\;\middle|\; D_n\right) = o(1),$$

$$\mathbb{E}_{\theta_0,\eta_0}\Pi\left(\|X(\theta-\theta_0)\|_2 > K_{\text{theta}}\sqrt{s_n\log p}\;\middle|\; D_n\right) = o(1),$$

for some constant $K_{\text{theta}} > 0$ *not depending on* n *and* p, *and for any* η_0 *satisfying (D1)–(D4).*

If we assume that $\phi(s_n)$ is bounded away from zero, the condition $s_n \log p/\phi(s_n) = o(\sqrt{n})$ in Corollary 2 becomes $s_n \log p = o(\sqrt{n})$. Similar condition was made by Chae et al. (2019b) to convert the convergence rate of the mean Hellinger distance $d_n((\theta,\eta),(\theta_0,\eta_0))$ to that of $\|\theta-\theta_0\|_1$. Note that Chae et al. (2019b) assumed $s_n\sqrt{\log p}/\phi(s_n) = o(\sqrt{n})$ under the bounded design matrix assumption, $\sup_{i,j}|x_{ij}| \le M$. If we assume $\sup_{i,j}|x_{ij}| \le M$, the condition in Corollary 2 is also relaxed to $s_n\sqrt{\log p}/\phi(s_n) = o(\sqrt{n})$. It is a quite natural condition to obtain a meaningful convergence rate tending to zero under the ℓ_1-norm.

The posterior convergence rate in Theorem 2 is *nearly* optimal if the hyperparameter κ is set equal to 1. Note that $\kappa^* = 1$ in this case. For example, if $s_0 \ge n^{1/(2\beta+1)}(\log n)^{2t-1}$, the posterior convergence rate with respect to the mean Hellinger distance is $\sqrt{s_0 \log p/n}$, leading to the same marginal convergence rate for θ in ℓ_2-norm. Note that the minimax rate is $\sqrt{s_0\log(p/s_0)/n}$ (Ye and Zhang 2010). If $s_0 < n^{1/(2\beta+1)}(\log n)^{2t-1}$, the marginal convergence for η rate with respect to the Hellinger distance is $n^{-\beta/(2\beta+1)}\times\sqrt{(\log n)^{2t-1}\log p}$ which is the minimax rate up to a logarithmic factor and the same as that of Shen et al. (2013). In conclusion, the global rate for the whole parameter (θ,η) is determined by the slower one among the two rates for θ and η, where both of them are close to the minimax rate provided that $\log p$ is negligible relative to n.

3.2 Bernstein von-Mises theorem

In this subsection, we study the distributional limit of the marginal posterior distribution for θ. Assume for a moment that p is moderately slowly increasing and the model is not sparse. Since we are working with a smooth semi-parametric model, it is highly expected that asymptotic shape of the map $\theta \mapsto L_n(\theta,\eta) - L_n(\theta_0,\eta)$ is quadratic around θ_0 for every η. Since the posterior mass is concentrated around (θ_0,η_0), we only need to consider η's that are sufficiently close to η_0. The assertion leads to the semi-parametric BvM theorem which guarantees the asymptotic efficiency of Bayes estimator.

With a sparse model considered in this paper, the marginal posterior distribution cannot converge to a single normal distribution unless posterior puts most of

its mass on a single model. To be more specific, note that the marginal posterior distribution of θ is given as

$$d\Pi(\theta \mid D_n) = \sum_{S \subseteq \{1,\ldots,p\}} w_S \, dQ_S(\theta_S) \, d\delta_0(\theta_{S^c}),$$

where

$$w_S \propto \frac{\pi_p(|S|)}{\binom{p}{|S|}} \int \int \exp\left(L_n(\widetilde{\theta}_S, \eta) - L_n(\theta_0, \eta_0)\right) d\Pi_{\mathcal{H}}(\eta) g_S(\theta_S) d\theta_S$$

and

$$Q_S(\theta_S \in B) = \frac{\int_B \int \exp\left(L_n(\widetilde{\theta}_S, \eta) - L_n(\theta_0, \eta_0)\right) d\Pi_{\mathcal{H}}(\eta) g_S(\theta_S) d\theta_S}{\int \int \exp\left(L_n(\widetilde{\theta}_S, \eta) - L_n(\theta_0, \eta_0)\right) d\Pi_{\mathcal{H}}(\eta) g_S(\theta_S) d\theta_S}$$

for every measurable set $B \subseteq \mathbb{R}^{|S|}$. Here, Q_S can be understood as the conditional posterior distribution of θ_S given $S_\theta = S$. If $|S|$ is not too large, Q_S is expected to be asymptotically normal as in the semi-parametric BvM theorem described in the previous paragraph. As a consequence, if there is a limit distribution of the marginal posterior for θ, it should be a mixture of the form

$$d\Pi^{\infty}(\theta \mid D_n) = \sum_{S \subseteq \{1,\ldots,p\}} w_S \, n^{-\frac{|S|}{2}} \, dN_{n,S}(h_S) \, d\delta_0(\theta_{S^c}),$$

where $h_S = \sqrt{n}(\theta_S - \theta_{0,S})$, $n^{-\frac{|S|}{2}}$ is the determinant of the Jacobian matrix, and $N_{n,S}$ is defined in Sect. 2.1. Theorem 3 says that the semi-parametric BvM theorem holds under slightly stronger condition than those needed for the posterior convergence rate results.

Theorem 3 (Bernstein von-Mises) *Assume that the prior conditions* (2), (3), (5)–(10) *hold with* $a_2 = 3$, $\lambda \|\theta_0\|_1 = O(s_0 \log p)$ *and* $\lambda s_n \log p = o(\sqrt{n})$. *Further assume that* $s_n^6\{(\log p)^{11} \vee s_n^{\frac{5}{12}}(\log p)^{8+\frac{11}{12}}\} = o(n^{1-\zeta})$ *holds for some constant* $\zeta > 0$ *and* $\psi(s_n)$ *is bounded away from zero. Then, we have*

$$\mathbb{E}_{\theta_0,\eta_0}\left[d_V\left(\Pi(\cdot|D_n), \Pi^{\infty}(\cdot|D_n)\right)\right] = o(1)$$

for any η_0 *satisfying (D1)–(D5).*

The bounded condition on $\psi(s_n)$ ensures that the quadratic term of log-likelihood ratio does not vanish. The condition $\lambda s_n \log p = o(\sqrt{n})$ is required to wash out the prior effect and is a quite mild condition if we consider the small λ regime such as $\lambda = \sqrt{n}/p$. To prove the BvM theorem, Castillo et al. (2015) also used similar condition, $\lambda s_n \sqrt{\log p} = o(\|X\|)$, where $\|X\|$ is the maximum ℓ_2-norm of the

columns of matrix X. Note that if $\log p = o(n)$ and x_{ij}'s are random samples from $N(0, 1)$, it coincide with $\lambda s_n \log p = o(\sqrt{n})$ with high probability tending to 1, as $p \to \infty$.

The condition $s_n^6\{(\log p)^{11} \vee s_n^{\frac{5}{12}}(\log p)^{8+\frac{11}{12}}\} = o(n^{1-\zeta})$ for some constant $\zeta > 0$, are sufficient conditions for

$$\int \sup_{\eta \in \mathcal{H}_n^*} \left(\dot{\ell}_\eta(y) - \dot{\ell}_{\eta_0}(y)\right)^2 dP_{\eta_0}(y)$$

converging to zero at a certain rate, where $\mathcal{H}_n^*$ is a neighborhood of η_0 to which the posterior distribution contracts. See Lemmas 6 and 12 in Supplementary Material for details. To satisfy this condition, a certain level of smoothness of η_0 is essential. For example, suppose we consider the prior $\sigma^6 \sim IG(a_0, b_0)$ for some positive constants a_0 and b_0, i.e., $a_2 = 3$ and $\kappa = 6$. Then, the condition $\left(s_n \log p\right)^{6+\frac{5}{12}}(\log p)^{\frac{5}{2}} = o(n^{1-\zeta})$ is satisfied when $(s_0 \log p)^{6+\frac{5}{12}}(\log p)^{\frac{5}{2}} = o(n^{1-\zeta})$ and $n^{\frac{7}{4\beta+12}}(\log p)^{\frac{107}{12}} = o(n^{1-\zeta})$, which hold for $\beta > 16.25$ provided that $\log p$ is negligible relative to n. Chae et al. (2019b) assumed $(s_0 \log p)^6 = o(n^{1-\zeta})$ for some constant $\zeta > 0$ to establish the semi-parametric BvM theorem. Our condition is slightly stronger due to relaxation on the tail condition of η_0.

3.3 Strong model selection consistency

Theorem 4 states that the posterior probability of S_θ for the strict supersets of the true model S_0 tends to zero. It implies that the posterior probability is asymptotically concentrated on the union of some strict subset of S_0 and possibly other coordinates of S_0^c.

Theorem 4 (No superset) *Under the conditions of Theorem 3 and $\tau \geq 2\gamma_1$, we have*

$$\mathbb{E}_{\theta_0,\eta_0} \Pi(S_\theta \supsetneq S_0 \mid D_n) = o(1)$$

for any η_0 satisfying (D1)–(D5), provided that $A_4 > K_{\text{sel}}$ for some constant K_{sel} depending only on η_0.

Since we assume that $\eta_0(y) \lesssim \exp(-b|y|^\tau)$ and $|\dot{\ell}_{\eta_0}(y)| \lesssim |y|^{\gamma_1} + C$, the condition $\tau \geq 2\gamma_1$ implies that $\dot{\ell}_{\eta_0}(y_i - x_i^T\theta_0)$ is a sub-Gaussian random variable. The sub-Gaussian assumption enables us to use the Hanson-Wright inequality (Hanson and Wright 1971; Wright 1973), which is one of the key properties for proving Theorem 4. Note that a normal distribution and a location-scale mixture of normal with compact mixing distribution satisfy the above condition. One important consequence of Theorem 4 is that if we assume that

$$\min\left\{|\theta_{0,j}| : \theta_{0,j} \neq 0,\ 1 \leq j \leq p\right\} \geq \frac{K_{\text{theta}}}{\psi(s_n)}\sqrt{\frac{s_n \log p}{n}}, \tag{15}$$

Corollary 2 and Theorem 4 guarantee the strong model selection consistency, which means $\mathbb{E}_{\theta_0,\eta_0}\Pi(S_\theta = S_0 \mid D_n) \longrightarrow 1$ as $n \to \infty$. The above condition (15) is called the

beta-min condition commonly assumed to obtain the model selection consistency (Castillo et al. 2015; Song and Liang 2017). The following corollary asserts that one can achieve the selection consistency and efficiently capture the uncertainty of the nonzero coordinates under the beta-min condition.

Corollary 3 (Selection) *Let* $\widehat{\theta}_{S_0} = n^{-1/2} V_{n,S_0}^{-1} G_{n,S_0} + \theta_{0,S_0}$, $\widehat{\Sigma}_{S_0} = n^{-1} V_{n,S_0}^{-1}$, *and* $\delta_{S_0^c}$ *be the Dirac measure at* $0 \in \mathbb{R}^{|S^c|}$. *Denote* $\theta \sim N_{|S_0|}(\widehat{\theta}_{S_0}, \widehat{\Sigma}_{S_0}) \otimes \delta_{S_0^c}$ *if* $\theta_{S_0} \sim N_{|S_0|}(\widehat{\theta}_{S_0}, \widehat{\Sigma}_{S_0})$ *and* $\theta_{S_0^c} = 0$, *independently. Under the conditions of Theorem* 4 *and* (15), *we have*

$$\mathbb{E}_{\theta_0,\eta_0}\left[d_V\left(\Pi(\cdot|D_n), N_{|S_0|}(\widehat{\theta}_{S_0}, \widehat{\Sigma}_{S_0}) \otimes \delta_{S_0^c} \right) \right] = o(1)$$

for any η_0 *satisfying (D1)–(D5), provided that* $A_4 > K_{\text{sel}}$ *for some constant* K_{sel} *depending only on* η_0.

Remark 3 Yang (2017) proved the asymptotic normality for an individual coordinate θ_i without the beta-min condition. However, her results focus on the posterior distribution of an individual coordinate under the normal error distribution and cannot be extended to the posterior distribution of the whole θ.

4 Discussion

In this paper, we study asymptotic properties of posterior distributions for high-dimensional linear regression models under unknown symmetric error. We extend the previous works on Bayesian asymptotic theory to deal with much more general error densities beyond the sub-Gaussian class. To the best of our knowledge, this is the first work that has proved posterior convergence rates and BvM theorem for high-dimensional linear regression model without the sub-Gaussian assumption. For the BvM theorem and selection consistency, the conditions, $s_n^6 (\log p)^{11} = o(n^{1-\zeta})$ and $\left(s_n \log p\right)^{6+\frac{5}{12}} (\log p)^{\frac{5}{2}} = o(n^{1-\zeta})$, are needed, which requires that the true error distribution is smooth enough.

Note that algorithms for sampling a DP mixture and a spike-and-slab prior can be suitably combined to generate MCMC samples from the posterior distribution in our semiparametric model, see Sect. 4 of Chae et al. (2019b). Although our theoretical analysis is limited to error densities with exponentially decaying tails, results of numerical experiments in Chae et al. (2019b) demonstrate that a semi-parametric estimator performs much better in prediction, model selection and uncertainty quantification than a parametric counterpart when the tail of error density is polynomially decaying. In particular, with a location-scale mixture of Gaussians with a conjugate DP prior, the selection consistency and BvM phenomena seem to hold while a location mixture only does not provide satisfactory results.

Future work will focus on the theoretical development of a location-scale mixtures with heavy-tailed components such as the Student's t distributions. This will likely entail new techniques for Bayesian asymptotic, see Chae and Walker (2017) for example.

Funding KL and LL were supported by NSF Grants IIS 1663870 and DMS Career 1654579. KL was also supported by INHA UNIVERSITY Research Grant. MC was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea government (MSIT) (No. 2020R1F1A1A01054718).

References

Armagan, A., Dunson, D. B., & Lee, J. (2013a). Generalized double pareto shrinkage. *Statistica Sinica*, *23*(1), 119–143.

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., & Strawn, N. (2013b). Posterior consistency in linear models under shrinkage priors. *Biometrika*, *100*(4), 1011–1018.

Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2015). Dirichlet-laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, *110*(512), 1479–1490.

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, *37*(4), 1705–1732.

Bochkina, N., & Rousseau, J. (2017). Adaptive density estimation based on a mixture of gammas. *Electronic Journal of Statistics*, *11*(1), 916–962.

Bühlmann, P., & van de Geer, S. (2011). Statistics for high-dimensional data: Methods, theory and applications. Springer Series in Statistics. Springer, Berlin. Retrieved from https://books.google.com/books?id=S6jYXmh988UC.

Canale, A., & De Blasi, P. (2017). Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli*, *23*(1), 379–404.

Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, *35*(6), 2313–2351.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480.

Castillo, I., & Nickl, R. (2013). Nonparametric Bernstein-von Mises theorems in Gaussian white noise. *The Annals of Statistics*, *41*(4), 1999–2028.

Castillo, I., & Nickl, R. (2014). On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *The Annals of Statistics*, *42*(5), 1941–1969.

Castillo, I., & Rousseau, J. (2015). A Bernstein-von Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics*, *43*(6), 2353–2383.

Castillo, I., Schmidt-Hieber, J., & van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, *43*(5), 1986–2018.

Castillo, I., & van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, *40*(4), 2069–2101.

Chae, M., Kim, Y., & Kleijn, B. J. K. (2019a). The semi-parametric Bernstein-von Mises theorem for regression models with symmetric errors. *Statistica Sinica*, *29*(3), 1465–1487.

Chae, M., Lin, L., & Dunson, D. B. (2016). Bayesian sparse linear regression with unknown symmetric error. arXiv e-prints arXiv:1608.02143.

Chae, M., Lin, L., & Dunson, D. B. (2019b). Bayesian sparse linear regression with unknown symmetric error. *Information and Inference*, *8*(3), 621–653.

Chae, M., & Walker, S. G. (2017). A novel approach to Bayesian consistency. *Electronic Journal of Statistics*, *11*(2), 4723–4745.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360.

Gao, C., van der Vaart, A. W. & Zhou, H. H. (2015). A general framework for Bayes structured linear models. The Annals of Statistics, To appear.

George, E. I., & Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, *87*(4), 731–747.

Hanson, D. L., & Wright, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, *42*(3), 1079–1083.

Kim, Y., & Jeon, J. J. (2016). Consistent model selection criteria for quadratically supported risks. *The Annals of Statistics*, *44*(6), 2467–2496.

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

Kruijer, W., Rousseau, J., & van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, *4*, 1225–1257.

Martin, R., Mess, R., & Walker, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, *23*(3), 1822–1847.

Narisetty, N. N., & He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, *42*(2), 789–817.

Narisetty, N. N., Shen, J., & He, X. (2019). Skinny Gibbs: A consistent and scalable Gibbs sampler for model selection. *Journal of the American Statistical Association*, *114*(527), 1205–1217.

Panov, M., & Spokoiny, V. (2015). Finite sample Bernstein-von Mises theorem for semiparametric problems. *Bayesian Analysis*, *10*(3), 665–710.

Polson, N. G., & Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian Statistics*, *9*, 501–538.

Ročková, V., & George, E. I. (2018). The spike-and-slab Lasso. *Journal of the American Statistical Association*, *113*(521), 431–444.

Rossell, D., & Rubio, F. J. (2017). Tractable Bayesian variable selection: Beyond normality. Journal of the American Statistical Association (just-accepted).

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, *38*(5), 2587–2619.

Shen, W., Tokdar, S. T., & Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, *100*(3), 623–640.

Shin, M., Bhattacharya, A., & Johnson, V. E. (2015). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. arXiv:150707106.

Song, Q., & Liang, F. (2017). Nearly optimal Bayesian shrinkage for high dimensional regression. ArXiv e-prints arXiv:1712.08964.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *58*(1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108.

van de Geer, S. A., & Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, *3*, 1360–1392.

van der Pas, S., Salomond, J. B., & Schmidt-Hieber, J. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electronic Journal of Statistics*, *10*(1), 976–1000.

van der Vaart, A. (1998). Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. Retrieved from https://books.google.com/books?id=UEuQEM5RjWgC.

Wright, F. T. (1973). A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *The Annals of Probability*, *1*(6), 1068–1070.

Yang, D. (2017). Posterior asymptotic normality for an individual coordinate in high-dimensional linear regression. arXiv preprint arXiv:170402646.

Yang, Y., Wainwright, M. J., & Jordan, M. I. (2016a). On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, *44*(6), 2497–2532.

Yang, Y., Wang, H. J., & He, X. (2016b). Posterior inference in Bayesian quantile regression with asymmetric laplace likelihood. *International Statistical Review*, *84*(3), 327–344.

Ye, F., & Zhang, C. H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, *11*(Dec), 3519–3540.

Zhang, C. H., & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 217–242.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Preflight Results

Document Overview

Title: Bayesian high-dimensional semi-parametric inference beyond sub-Gaussian tails
Author: Kyoungjae Lee
Creator: Springer
Producer: Acrobat Distiller 10.1.8 (Windows)

Preflight Information

Profile: Convert to PDF/A-1b
Version: Qoppa jPDFPreflight v2020R2.01
Date: Jun 22, 2021 6:55:27 AM

Legend: (X) - Can NOT be fixed by PDF/A-1b conversion.
(!X) - Could be fixed by PDF/A-1b conversion. User chose to be warned in PDF/A settings.

Page 8 Results
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 9 Results
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 11 Results
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 12 Results
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 13 Results
(X) Font widths must be the same in both the font dictionary and the embedded font file.