SCIENCE CHINA
Mathematics

• ARTICLES •

February 2023 Vol. 66 No. 2: 409–442
https://doi.org/10.1007/s11425-020-1912-6

Nearly optimal Bayesian shrinkage for high-dimensional regression

Qifan Song* & Faming Liang

Department of Statistics, Purdue University, West Lafayette, IN 47906*, USA*
Email: qfsong@purdue.edu, fmliang@purdue.edu

Received October 11, 2020; accepted September 26, 2021; published online October 14, 2022

Abstract During the past decade, shrinkage priors have received much attention in Bayesian analysis of high-dimensional data. This paper establishes the posterior consistency for high-dimensional linear regression with a class of shrinkage priors, which has a heavy and flat tail and allocates a sufficiently large probability mass in a very small neighborhood of zero. While enjoying its efficiency in posterior simulations, the shrinkage prior can lead to a nearly optimal posterior contraction rate and the variable selection consistency as the spike-and-slab prior. Our numerical results show that under the posterior consistency, Bayesian methods can yield much better results in variable selection than the regularization methods such as LASSO and SCAD. This paper also establishes a BvM-type result, which leads to a convenient way of uncertainty quantification for regression coefficient estimates.

Keywords Bayesian variable selection, absolutely continuous shrinkage prior, heavy tail, posterior consistency, high-dimensional inference

MSC(2020) 62J07, 62F15

Citation: Song Q F, Liang F M. Nearly optimal Bayesian shrinkage for high-dimensional regression. Sci China Math, 2023, 66: 409–442, https://doi.org/10.1007/s11425-020-1912-6

1 Introduction

The dramatic improvement in data collection and acquisition technologies during the last two decades has enabled scientists to collect a great amount of high-dimensional data. Due to their intrinsic nature, many of the high-dimensional data, such as omics data and single nucleotide polymorphism (SNP) data, have a much smaller sample size than their dimension (also known as small-n-large-p). Toward an appropriate understanding of the system underlying the small-n-large-p data, variable selection plays a vital role. In this paper, we consider the problem of variable selection for the high-dimensional linear regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{y}$ is an n-dimensional response vector, $\boldsymbol{X}$ is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is the vector of regression coefficients, σ is the standard deviation, and $\boldsymbol{\varepsilon}$ follows $N(0, I_n)$. This problem has received much attention in the recent literature. Methods have been developed from both frequentist and Bayesian perspectives.

[*] Corresponding author

The frequentist methods are usually regularization-based, which enforce the model sparsity through imposing a penalty on the negative log-likelihood function. For example, the least absolute shrinkage and selection operator (LASSO) [59] employs an L_1-penalty, elastic net [75] employs a combination of L_1- and L_2-penalties, [20] employs a smoothly clipped absolute deviation (SCAD) penalty, [71] employs a minimax concave penalty (MCP), and rLASSO [56] employs a reciprocal L_1-penalty. In general, these penalty functions encourage model sparsity, and tend to shrink the coefficients of false predictors to exactly zero. Under appropriate conditions, consistency can be established for both variable selection and parameter estimation.

The Bayesian methods encourage sparsity of the *posteriori* model through choosing appropriate prior distributions. A classical choice is the spike-and-slab prior,

$$\beta_j \sim r h(\beta_j) + (1-r)\delta_0(\beta_j),$$

where $\delta_0(\cdot)$ is the degenerated "spike distribution" at zero, $h(\cdot)$ is an absolutely continuous "slab distribution", and r is the prior mixing proportion. Generally, it can be equivalently represented as the following hierarchical prior:

$$\xi \sim \pi(\xi), \quad \boldsymbol{\beta}_\xi \sim h_\xi(\boldsymbol{\beta}_\xi), \quad \boldsymbol{\beta}_{\xi^c} \equiv 0 \tag{1.2}$$

for some multivariate density function h_ξ, where ξ denotes a subset model, and $\boldsymbol{\beta}_\xi$ and $\boldsymbol{\beta}_{\xi^c}$ denote the coefficient vectors of the covariates included in and excluded from the model ξ, respectively. The theoretical properties of the prior (1.2) have been thoroughly investigated [12, 35, 36, 40, 44, 46, 53, 57, 70]. Under proper choices of π and h_ξ, the spike-and-slab prior achieves a (nearly-) optimal contraction rate and the model selection consistency.

Alternative to the hierarchical priors, some shrinkage priors have been proposed for (1.1) motivated by the equivalence between the regularization estimator and the maximum *a posteriori* (MAP) estimator (see, e.g., the discussion in [59]). Examples of such priors include the Laplace prior [32, 48], the horseshoe prior [11], the structuring shrinkage prior [30], the double Pareto shrinkage prior [2], the Dirichlet Laplace prior [8], and the elliptical Laplace prior [22]. Compared with the hierarchical prior, the shrinkage prior is conceptually much simpler. The former involves specification of priors for a large set of models, while the latter avoids this issue as for which only a single model is considered. Consequently, for the hierarchical prior, a trans-dimensional Markov chain Monte Carlo (MCMC) sampler is required for simulating of the posterior in a huge space of submodels, and this has constituted the major obstacle for the use of Bayesian methods in high-dimensional variable selection. For the shrinkage prior, there is only a single model used in posterior simulations, and thus some gradient-based MCMC algorithms, such as stochastic gradient Langevin dynamics (SGLD) [68], Hamiltonian Monte Carlo [18, 47], Riemann manifold Hamiltonian Monte Carlo [28], and stochastic gradient Hamiltonian Monte Carlo [15], can be easily used in simulations. This is extremely attractive for the problems where both n and p are very large, for which mini-batch data can be conveniently used to accelerate simulations.

Despite the popularity and potential advantages of shrinkage priors, few works have been done to study their theoretical properties. There is a lack of general guarantee of posterior consistency for Bayesian shrinkage priors, especially under the high-dimensional setting. Bayesian community already realized that the Laplace distribution is not a good shrinkage prior for high-dimensional linear regression. Bhattacharya et al. [8] and Castillo et al. [12] showed that the L_2-contraction rate of Bayesian LASSO is suboptimal, and one can also show that the posterior of Bayesian LASSO is inconsistent in the L_1 sense under regularity conditions. To tackle this issue, many other types of shrinkage priors have been proposed (see, e.g., [1, 3, 8, 11, 27, 29, 30]). In the literature, there have been rich theoretical results on the Bayesian shrinkage prior for the case of slowly increasing p (i.e., $p = o(n)$) [3, 9, 24] and normal mean models [8, 27, 62, 64]. For the high-dimensional case, i.e., $p > n$, the non-invertibility and eigen-structure of the Gram matrix $\boldsymbol{X}^\top \boldsymbol{X}$ complicate the analysis. Hence, the results derived from low-dimensional models or normal mean models do not trivially apply to regression problems. It is worth noting that most of the Bayesian works for the normal mean models [8, 13, 62] aim to achieve a minimax contraction rate of $O(\sqrt{s\log(n/s)})$. A recent preprint [55] shows that for the normal mean problem, *any monotone estimator $\widehat{\boldsymbol{\beta}}$ which asymptotically guarantees no false discovery has at best the L_2-estimation error rate* $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2 = O_p(\sqrt{s\log n})$. This frequentist assertion implies that the existing rate-minimax Bayesian approaches cannot consistently recover the underlying sparsity structure for normal mean models (see also [63, Theorem 3] and [8, Theorem 3.4]). For high-dimensional regression models, the variable selection consistency remains an unresolved issue for Bayesian shrinkage priors.

In this paper, we lay down a general theoretical foundation for Bayesian high-dimensional linear regression with shrinkage priors. Instead of focusing on certain types of shrinkage priors, we investigate sufficient conditions of posterior consistency for general shrinkage priors. We show that if the prior density has a dominating peak around zero and a heavy and flat tail, then its theoretical properties are as good as the spike-and-slab prior: its contraction rate is nearly optimal, variable selection is consistent, and posterior follows a BvM (Bernstein von Mises)-type phenomenon. Specifically, we consider two types of shrinkage priors for high-dimensional linear regression, namely, polynomially decaying priors and two-Gaussian mixture priors [23]. Empirical studies show that the Bayesian method with a consistent shrinkage prior can lead to more accurate results in variable selection than the regularization methods. The general theoretical framework and technical tools developed in this paper have inspired a series of follow-up works (see, e.g., the R2-D2 shrinkage prior [74], the beta prime prior [4] and Bayesian additive nonparametric regression [67]).

Finally, we note that there are some other Bayesian works which deal with high-dimensional problems with shrinkage priors. For example, Pati et al. [49] employed a Dirichlet-Laplace (DL) prior in dealing with high-dimensional factor models, but their results only allow the magnitude of true parameters to increase very slowly with n; Bhadra et al. [7] studied the prediction risk, instead of the posterior properties of $\boldsymbol{\beta}$ for high-dimensional regression with a horseshoe prior; Ročková and George [51] established for high-dimensional linear regression the same posterior convergence rate as ours with a two-group Laplace prior, but failed to establish consistency of variable selection.

The rest of this paper is organized as follows. In Section 2, we present the main theoretical results, where we lay down the theory of posterior consistency for high-dimensional linear regression with shrinkage priors. In Section 3, we study posterior consistency for several commonly used shrinkage priors. In Section 4, we discuss some important practical issues on Bayesian computation, and illustrate the performance of Bayesian variable selection using a toy example. In Section 5, we present some simulation studies and a real data example. In Section 6, we conclude the paper with a brief discussion. In Appendix A, we give the proofs of the main theorems.

2 Main theoretical results

Notation. In what follows, we rewrite the dimension p of the model (1.1) by p_n to indicate that the number of covariates can increase with the sample size n. We use superscript $*$ to indicate true parameter values, e.g., $\boldsymbol{\beta}^*$ and σ^*. For simplicity, we assume that the true standard deviation σ^* is unknown but fixed, and it does not change as n grows. For vectors, we let $\|\cdot\|$ or $\|\cdot\|_2$ denote the L_2-norm; let $\|\cdot\|_1$ denote the L_1-norm; let $\|\cdot\|_\infty$ denote the L_∞-norm, i.e., the maximum absolute value among all the entries of the vector; let $\|\cdot\|_0$ denote the L_0-norm, i.e., the number of nonzero entries. As in (1.2), we let $\xi\subset\{1,2,\ldots,p_n\}$ denote a subset model, and let $|\xi|$ denote the size of the model ξ. We let s denote the size of the true model, i.e., $s=\|\boldsymbol{\beta}^*\|_0=|\xi^*|$. We let $\boldsymbol{X}_\xi$ denote the sub-design matrix corresponding to the model ξ, and let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and smallest eigenvalues of a square matrix, respectively. We let $1(\cdot)$ denote the indicator function. For two positive sequences a and b, $a\prec b$ means $\lim a/b=0$, $a\asymp b$ means

$$0<\liminf a/b\leqslant\limsup a/b<\infty,$$

and $a\preccurlyeq b$ means $a\prec b$ or $a\asymp b$. We use $\{\epsilon_n\}$ to denote the Bayesian contraction rate which satisfies $\epsilon_n\prec 1$.

2.1 Posterior consistency

The posterior distribution for the model (1.1) follows a general form

$$\pi(\boldsymbol{\beta}, \sigma^2 \mid D_n) \propto f(\boldsymbol{\beta}, \sigma^2; D_n)\pi(\boldsymbol{\beta}, \sigma^2),$$

where $f(\boldsymbol{\beta}, \sigma^2; D_n) \propto \sigma^{-n} \exp(-\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2/2\sigma^2)$ is the likelihood function of the observed data $D_n = (\boldsymbol{X}, \boldsymbol{y})$, and $\pi(\boldsymbol{\beta}, \sigma^2)$ denotes the prior density of $\boldsymbol{\beta}$ and σ^2. Consider a general shrinkage prior: σ^2 is subject to an inverse gamma prior $\sigma^2 \sim \mathrm{IG}(a_0, b_0)$, where a_0 and b_0 denote the prior-hyperparameters, and conditioned on σ^2, $\boldsymbol{\beta}$ has the independent prior for each entry, with an absolutely continuous density function of the form

$$\pi(\boldsymbol{\beta} \mid \sigma^2) = \prod_j [g_\lambda(\beta_j/\sigma)/\sigma], \tag{2.1}$$

where λ is some tuning parameter(s). It is easy to derive that

$$\log \pi(\boldsymbol{\beta}, \sigma^2 \mid D_n) = C + \sum_{j=1}^{p_n} \log g_\lambda\left(\frac{\beta_j}{\sigma}\right) - (n/2 + p_n/2 + a_0 + 1)\log(\sigma^2) - \frac{2b_0 + \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2\sigma^2} \tag{2.2}$$

for some additive constant C.

The shape and scale of the density function g_λ play a crucial role for posterior consistency. Intuitively, we may decompose the parameter space $\mathbb{R}^{p_n}$ into three subsets: the neighborhood set $B_1 = \{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leqslant \epsilon_n\}$, the "overfitting" set $B_2 = \{\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) - \boldsymbol{\varepsilon}\| \lesssim \sigma^*\sqrt{n}\}\backslash B_1$ and the rest B_3. Heuristically, the likelihood $f(\boldsymbol{\beta}_2) \gtrsim f(\boldsymbol{\beta}_1) \gtrsim f(\boldsymbol{\beta}_3)$ for any $\boldsymbol{\beta}_i \in B_i$, $i = 1, 2, 3$. Therefore, to drive the posterior mass toward the set B_1, it is sufficient to require that $\pi(B_1) \gg \pi(B_2)$ and the ratio $\pi(B_1)/\pi(B_3)$ is not too tiny. In other words, the prior distribution should (1) assign at least a minimum probability mass around $\boldsymbol{\beta}^*$, and (2) assign a tiny probability mass on the overfitting set. However, under the high-dimensional setting, the "overfitting" set is geometrically intractable (and it expands to infinity) due to the arbitrariness of the eigen-structure of the design matrix. Therefore, analytically, it is difficult to directly study the prior on the "overfitting" set. One possible way to control the prior on the "overfitting" set is to impose a strong prior concentration for each β_j such that the most of the prior mass is allocated on the "less-complicated" models under a certain complexity measure. Under regular identifiability conditions, the overfitting models are always complicated, so the prior probability mass on the "overfitting" models should be small, but it is worth noting that the overfitting models are a subset of all the complicated models and the strong prior concentration is only a sufficient condition. When the geometry of the overfitting set is easier to handle, e.g., under $p_n = o(n)$ or in the normal mean models, the overfitting set can be a neighboring set of $\boldsymbol{\beta}^*$, potentially annulus-shaped. In this case, it is absolutely unnecessary to require a strong prior concentration on the neighboring set of $\boldsymbol{\beta}^*$, i.e., we only need to impose conditions on the local shape of the prior around $\boldsymbol{\beta}^*$ (see [14, 24, 63]). This is also the key difference between high-dimensional models and slowly increasing models/normal mean models.

Before rigorously studying the properties of the posterior distribution, we first state some regularity conditions on the eigen-structure of the design matrix $\boldsymbol{X}$:

A$_1$(1) All the covariates are uniformly bounded. For simplicity, we assume that $\boldsymbol{x}_j \in [-1, 1]^n$ for $j = 1, 2, \ldots, p_n$, where $\boldsymbol{x}_j$ denotes the j-th column of $\boldsymbol{X}$.

A$_1$(2) The dimensionality is high: $p_n \succcurlyeq n$.

A$_1$(3) There exist some integer $\bar{p}$ (depending on n and p_n) and a fixed constant λ_0 such that $\bar{p} \succ s$ and $\lambda_{\min}(\boldsymbol{X}_\xi^\top \boldsymbol{X}_\xi) \geqslant n\lambda_0$ for any subset model $|\xi| \leqslant \bar{p}$.

Remark 2.1. A$_1$(1) implies that $\lambda_{\max}(\boldsymbol{X}^\top\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X}) \leqslant np$. A$_1$(3) has often been used in the literature to overcome the non-identifiability issue of $\boldsymbol{\beta}$ (see, e.g., [46, 56, 71]). This condition is also equivalent to the lower bounded compatibility number condition used in [12]. In general, $\bar{p}$ should be much smaller than n. For example, for an $n \times n$ design matrix with all the entries i.i.d. distributed, the Marchenko-Pastur law states that the empirical distribution of the eigenvalues of the corresponding sample covariance matrix converges to $\mu(x) \propto \sqrt{(2-x)/x}1$ $(x \in [0, 2])$. The random matrix theory typically allows $\bar{p} \asymp n/\log p_n$ with a high probability when the rows of $\boldsymbol{X}$ are independent isotropic sub-Gaussian random vectors; please refer to [46, Lemma 6.1] and [66, Theorem 5.39].

The next set of assumptions concerns the sparsity of $\boldsymbol{\beta}^*$ and the magnitude of nonzero entries of $\boldsymbol{\beta}^*$.

$\mathrm{A}_2(1)$ $s \log p_n \prec n$, where s is the size of the true model.

$\mathrm{A}_2(2)$ $\max\{|\beta_j^*/\sigma^*|\} \leqslant \gamma_3 E_n$ for some fixed $\gamma_3 \in (0,1)$, and E_n is nondecreasing with respect to n.

Remark 2.2. The condition $\mathrm{A}_2(1)$ is regularly used in the literature of high-dimensional statistics, which restricts the size of the true model to be of the order $o(n/\log p_n)$. The condition $\mathrm{A}_2(2)$ constrains the growth of the nonzero true regression coefficients such that $\max\{|\beta_j^*|\} \preccurlyeq E_n$. Together with the second condition in (2.3), it ensures that the prior probability around the true model does not decay too fast, which echoes the heuristics discussed in the previous paragraph that the shrinkage prior shall assign at least a minimum probability mass around $\boldsymbol{\beta}^*$. Note that such an upper bound condition is fairly common in the literature of Bayesian asymptotics. For example, Ghosal et al. [25] established a general posterior convergence rate, which requires that the prior mass over a small f-divergence ball of the true density p_0 is not too small. For linear regression models, Armagan et al. [3, Theorem 1], Bhattacharya et al. [8, Theorem 3.1] and Yang et al. [70, Condition (7a)] imposed a similar upper bound condition on $\boldsymbol{\beta}^*$. A similar condition has also been used in [26, 35, 37]. We note that it is also possible to establish posterior consistency without such an upper bound condition for certain types of shrinkage priors. Noticeable examples include [51] which used a two-component mixture Laplace prior, [12, 22] which used a Dirac-Laplace prior, and [44] which used a g-prior centered at the least-square estimator. More discussions on this issue can be found after Corollary 3.2.

The next theorem provides sufficient conditions for posterior consistency. Hereafter, we let $\epsilon_n = M\sqrt{s \log p_n / n}$ denote the contraction rate, where M is a fixed positive constant.

Theorem 2.3 (Posterior consistency)**.** *Consider the linear regression model* (1.1)*, where the design matrix* $\boldsymbol{X}$ *and the true* $\boldsymbol{\beta}^*$ *satisfy the conditions* A_1 *and* A_2*,* σ^2 *is subject to an inverse gamma prior* $\mathrm{IG}(a_0, b_0)$*, and the prior of* $\boldsymbol{\beta}$ *is given by* (2.1)*. If* g_λ *satisfies the conditions*

$$
\begin{aligned}
&1 - \int_{-a_n}^{a_n} g_\lambda(x)dx \leqslant p_n^{-(1+u)},\\
&-\log\Big(\inf_{x \in [-E_n, E_n]} g_\lambda(x)\Big) = O(\log p_n),
\end{aligned}
\tag{2.3}
$$

where $u > 0$ *is a constant,* $a_n \asymp \sqrt{s \log p_n/n}/p_n$*, and the constant* M *is sufficiently large, then the following posterior consistency holds*:

$$
\begin{aligned}
&P^*(\pi(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \geqslant c_1 \sigma^* \epsilon_n \mid D_n) \geqslant \mathrm{e}^{-c_2 n \epsilon_n^2}) \leqslant \mathrm{e}^{-c_3 n \epsilon_n^2},\\
&P^*(\pi(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \geqslant c_1 \sqrt{s} \epsilon_n \sigma^* \mid D_n) \geqslant \mathrm{e}^{-c_2 n \epsilon_n^2}) \leqslant \mathrm{e}^{-c_3 n \epsilon_n^2}
\end{aligned}
\tag{2.4}
$$

for some positive constants c_1*,* c_2 *and* c_3*.*

The proof of this theorem is given in Appendix A. The results in (2.4) imply that

$$
\lim_{n \to \infty} \mathrm{E}(\pi(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \geqslant c_1 \sigma^* \epsilon_n \mid D_n)) = 0
$$

and

$$
\lim_{n \to \infty} \mathrm{E}(\pi(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \geqslant c_1 \sigma^* \sqrt{s} \epsilon_n \mid D_n)) = 0,
$$

i.e., the L_2- and L_1-contraction rates of the posterior distribution of $\boldsymbol{\beta}$ are $O(\sqrt{s \log p_n / n})$ and $O(s\sqrt{\log p_n/n})$, respectively. These contraction rates are nearly optimal by recalling that the minimax L_2-contraction rate is $O(\sqrt{s \log(p_n/s)/n})$ [50], and they are not worse than the rates achieved with the spike-and-slab prior [12]. In other words, there is no performance loss due to the use of shrinkage priors.

The conditions (2.3) in the above theorem are consistent with our heuristic arguments in previous paragraphs. The first equation of (2.3) concerns the prior concentration, which requires that the prior density of β_j/σ has a dominating peak inside a tiny interval $\pm a_n$. Such a steep prior peak plays the role of "spike" as in spike-and-slab prior modeling. In the literature, Castillo et al. [12] assigned on the spike a prior probability $\pi(\xi_j = 1) = O(p_n^{-u})$ with $u > 1$, Narisetty and He [46] employed an SSVS (stochastic search variable selection)-type prior [23] under which the prior probability $\pi(\xi_j = 1) \propto 1/p_n$, and Yang et al. [70] assigned on the spike a prior probability $\pi(\xi_j = 1) = O(p_n^{-u})$ with $u > 0$. All these prior specifications are comparable to our condition $\pi(|\beta_j/\sigma| > a_n) = O(p_n^{-(1+u)})$ with $u > 0$. Note that [46] and [70] seem to require less prior concentration, and they both imposed additional prior concentration conditions to bound the model size such that $\pi(|\xi| > O(n/\log p_n)) = 0$. It is worth noting that all our theorems require the prior distribution to have a tiny scale by imposing a very small bound on a_n. The scale of the shrinkage prior affects the convergence rate of the posterior through its logarithm only. In other words, no matter how small the scale of the prior distribution is, it does not affect much the convergence rate of the posterior as long as $\log(1/a_n)$ is of order $\log(p_n)$. One established example is the horseshoe prior (see [62, Theorem 3.3] for the convergence theory of the posterior). The second equation of (2.3), as discussed previously, essentially requires that the prior density around the true nonzero regression coefficient $\boldsymbol{\beta}_j^*/\sigma^*$ is at least $\exp\{-O(\log p_n)\}$, i.e., $g_\lambda(\boldsymbol{\beta}_j^*/\sigma^*) \geqslant \exp\{-c\log p_n\}$ for some positive constant c. Finally, we note that this prior concentration condition is only sufficient. In practice, a moderate degree of concentration can often lead to satisfactory results.

Other than the regression coefficients, similar results to (2.4) can be derived for the fitting error $\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^*\|$.

Theorem 2.4. *If the conditions of Theorem* 2.3 *hold, then*

$$P^*(\pi(\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^*\| \geqslant c_1\sigma^*\sqrt{n}\epsilon_n \mid D_n) \geqslant \mathrm{e}^{-c_2 n\epsilon_n^2}) \leqslant \mathrm{e}^{-c_3 n\epsilon_n^2} \tag{2.5}$$

for some positive constants c_1, c_2 *and* c_3.

Remark 2.5. Theorem 2.4 actually holds without Condition $\mathrm{A}_1(3)$. To intuitively understand the redundancy of Condition $\mathrm{A}_1(3)$, let us consider the fitted error under any selected subset model $\xi \supseteq \xi^*$, i.e., $\boldsymbol{X}_\xi(\boldsymbol{X}_\xi^\top\boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^\top\boldsymbol{\varepsilon}$. Without any assumption on the eigen-structure of $\boldsymbol{X}$, this term can be bounded in probability since the eigenvalues of $\boldsymbol{X}_\xi(\boldsymbol{X}_\xi^\top\boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^\top$ are 0 or 1. However, to prove Theorem 2.3, we need to bound the estimation error $(\boldsymbol{X}_\xi^\top\boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^\top\boldsymbol{\varepsilon}$, and hence an eigen-structure assumption such as Condition $\mathrm{A}_1(3)$ is necessary.

To conclude this subsection, we state that an appropriate shrinkage prior can lead to almost the same posterior consistency result as the spike-and-slab prior.

2.2 Variable selection consistency

In this subsection, we perform a theoretical study on how to retrieve the sparse structure of $\boldsymbol{\beta}^*$ with a shrinkage prior. To achieve this goal, it is necessary to "sparsify" the continuous posterior distribution induced by the continuous prior. In the literature, this is usually done by (1) hard (or adaptive) thresholding on β_j or on the shrinkage weight $1/(1+\lambda_j^2)$ [11, 34, 38, 58], or (2) decoupling shrinkage and selection methods [31, 69]. Note that the approaches in the latter class intend to incorporate the dependency between covariates into the sparse posterior summary. All the aforementioned approaches depend solely on the magnitude of the Bayesian estimates of β_j's, without accounting for the degree of the prior concentration.

We propose to use a prior-dependent hard thresholding method, which sets $\tilde{\beta}_j = \beta_j 1(|\beta_j| > \eta_n)$ for some threshold η_n. This induces a sparse pseudo posterior $\pi(\tilde{\boldsymbol{\beta}} \mid D_n)$, which thereafter can be used to assess the model uncertainty and conduct variable selection as if it is induced by a spike-and-slab prior. The correlation structure of $\pi(\tilde{\boldsymbol{\beta}} \mid D_n)$ will reflect the dependency knowledge in $\boldsymbol{X}$.

First of all, Theorem 2.3 trivially implies that

$$\mathrm{E}\pi(|\beta_j - \beta_j^*| \geqslant c_1\sigma^*\epsilon_n \text{ for all } j = 1, \ldots, p_n \mid D_n) = o_p(1).$$

Therefore, if $\min_{j\in\xi^*}|\beta_j^*| > 2c_1\sigma^*\epsilon_n$ and $\eta_n = c_1\sigma^*\epsilon_n$, then

$$\mathrm{E}\pi(1(\tilde{\beta}_j = 0) \neq 1(\beta_j^* = 0) \text{ for all } j \mid D_n) = o_p(1)$$

and $\pi(\tilde{\boldsymbol{\beta}})$ can consistently select the true model. However, one potential issue of using $c_1\sigma^*\epsilon_n$ for thresholding is that it greatly alters the theoretical characteristic of $\pi(\boldsymbol{\beta}\mid D_n)$ in the sense that the L_2-contraction rate of $\pi(\tilde{\boldsymbol{\beta}}\mid D_n)$ can be as large as $s\sqrt{\log p_n/n}$ but not $\sqrt{s\log p_n/n}$.

This motivates us to consider another choice of η_n. As discussed previously, (2.3) implies a "spike" between $[-a_n,a_n]$ for the prior of β/σ, which plays the same role as the Dirac measure in the spike-and-slab prior. Hence, from the point of view of prior specification, a_n distinguishes between zero and nonzero coefficients, and it is natural to consider $\tilde{\beta}_j=\beta_j1(|\beta_j/\sigma|>a_n)$. The posterior $\pi(\tilde{\boldsymbol{\beta}},\sigma^2\mid D_n)$ thus implies the selection rule as $\xi(\boldsymbol{\beta},\sigma^2)=\{j;|\beta_j/\sigma|>a_n\}$. This hard-thresholding rule of Bayesian variable selection can be viewed as a counterpart of the selection rule $\{j:|\beta_j/\sigma|>0\}$ used in spike-and-slab modeling. It is also closely related with the idea of "generalization dimension" [8, 51]. [8, Theorem 3.4] defines $\mathrm{supp}_\delta(\boldsymbol{\beta})=\{j:|\beta_j/\sigma|\geqslant\delta\}$ as the set of variables selected based on a nonsparse posterior sample $\boldsymbol{\beta}$, where $\sigma=1$ is known, $p_n=n$ ($\boldsymbol{X}=I$), and δ satisfies the condition

$$\pi(|\beta_j|\geqslant\delta)\leqslant C\log(n/s)/\Gamma(n^{-1-u})\asymp\log(n/s)/(n^{1+u})$$

for some $u>0$. This choice of δ matches our threshold a_n, which is the quantile of the prior distribution satisfying $\pi(|\beta_j/\sigma|\geqslant a_n)\leqslant p_n^{-1-u}$ for some $u>0$.

The following theorem establishes the variable selection consistency of the above hard-thresholding rule, while Bhattacharya et al. [8] and Ročková and George [51] proved only that the selected model has a bounded size.

Theorem 2.6 (Variable selection consistency). *Suppose that the conditions of Theorem* 2.3 *hold under* $a_n\prec\sqrt{\log p_n}/(\sqrt{n}p_n)$ *and* $u>1$. *Let* l_n *be a measure of flatness of the function* $g_\lambda(\cdot)$, *i.e.,*

$$l_n=\max_{j\in\xi^*}\sup_{\substack{x_1,x_2\in\beta_j^*/\sigma^*\pm c_0\epsilon_n\\|x_1|,|x_2|\geqslant a_n}}\frac{g_\lambda(x_1)}{g_\lambda(x_2)},$$

where c_0 *is some large constant. If* $\min_{j\in\xi^*}|\beta_j^*|>M_1\sqrt{\log p_n/n}$ *for some sufficiently large* M_1 *and* $s\log l_n\prec\log p_n$, *then*

$$P^*\{\pi[\xi(\boldsymbol{\beta},\sigma^2)=\xi^*|D_n]>1-o(1)\}>1-o(1). \tag{2.6}$$

This theorem is a simple corollary of Theorem A.7 in Appendix A. It requires a smaller value of a_n and a larger value of u, i.e., a narrower and more concentrated prior peak, compared with Theorem 2.3. Besides the prior concentration and the tail thickness, the condition $s\log l_n\prec\log p_n$ also requires tail flatness such that the prior density around the true value β^*/σ^* is not rugged. This flatness facilitates an analytic study for the posterior $\pi(\xi(\boldsymbol{\beta},\sigma^2)\mid D_n)$. Generally speaking, for smooth g_λ, the flatness measure approximately follows $\log l_n\asymp\max_{j\in\xi^*}\epsilon_n[\log g_\lambda]'(\beta_j^*/\sigma^*)\to 0$, where $[\log g_\lambda]'$ is the first derivative of $\log g_\lambda$. In the extreme situation, we can utilize an exactly flat tail such that $\log l_n\equiv 0$. An example could be $g_\lambda(x)\propto\exp\{-p_\lambda(x)\}1_{x\in[-E_n,E_n]}$, where $p_\lambda(x)$ has a shape like a non-concave penalty function such as SCAD. If $\log l_n$ is not exactly 0, then the condition $s\log l_n\prec\log p_n$ imposes an additional constraint on the sparsity s other than $s\prec n/\log p_n$. More discussions on l_n can be found in Section 3.

The result of this theorem also implies a stronger posterior contraction for the false covariates such that $|\beta_j/\sigma|$ is bounded in posterior by a_n.

2.3 Shape approximation of the posterior distribution

Another important aspect of Bayesian asymptotics is the shape of the posterior distribution. The general theory on the posterior shape is the BvM theorem. It claims that the posterior distribution of the parameter θ in a regular finite-dimensional model is approximately a normal distribution as $n\to\infty$, i.e.,

$$\|\pi(\cdot\mid D_n)-N(\cdot;\hat{\theta}_{\mathrm{MLE}},(n\hat{I})^{-1})\|_{\mathrm{TV}}\to 0, \tag{2.7}$$

regardless of the choice of the prior $\pi(\theta)$, where $\pi(\cdot\mid D_n)$ is the posterior distribution given data D_n, $N(\cdot;\mu,\Sigma)$ denotes a (multivariate) normal distribution, $\hat{\theta}_{\mathrm{MLE}}$ stands for the maximum likelihood estimator of θ, I is Fisher's information matrix, and $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance between two measures. The BvM theorem provides an important link between the frequentist limiting distribution and the posterior distribution, and it can be viewed as a frequentist justification for Bayesian credible intervals. To be specific, the Bayesian credible intervals are asymptotically equivalent to the Wald confidence intervals, and also have the long-run relative frequency interpretation.

The BvM theorem generally holds for fixed-dimensional problems. For linear regression with known σ^*, the posterior normality always holds under an (improper) uniform prior, as

$$\pi(\boldsymbol{\beta} \mid D_n) \sim N((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}, \sigma^{*2}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}),$$

as long as $p \leqslant n$ and the matrix $\boldsymbol{X}$ is of full rank.

Under the scenario $p_n \gg n$, all the false coefficients are bounded in posterior by a threshold value by Theorem 2.6. Combining this with the fact that $f(\boldsymbol{\beta}_{\xi^*}, \boldsymbol{\beta}_{(\xi^*)^c}; \boldsymbol{X}, \boldsymbol{y}) \approx f(\boldsymbol{\beta}_{\xi^*}, \boldsymbol{\beta}_{(\xi^*)^c} = 0; \boldsymbol{X}_{\xi^*}, \boldsymbol{y})$ when $\|\boldsymbol{\beta}_{(\xi^*)^c}\|_\infty$ is sufficiently small, we have

$$\pi(\boldsymbol{\beta} \mid D_n) \propto L(\boldsymbol{\beta}_{\xi^*}, \boldsymbol{\beta}_{(\xi^*)^c} \mid \boldsymbol{X}, \boldsymbol{y})\pi(\boldsymbol{\beta}_{\xi^*}, \boldsymbol{\beta}_{(\xi^*)^c}) \approx L(\boldsymbol{\beta}_{\xi^*}; \boldsymbol{X}_{\xi^*}, \boldsymbol{y})\pi(\boldsymbol{\beta}_{\xi^*})\pi(\boldsymbol{\beta}_{(\xi^*)^c}).$$

If $\pi(\boldsymbol{\beta}_{\xi^*})$ is sufficiently flat around $\boldsymbol{\beta}^*_{\xi^*}$ and acts like a uniform prior, then the low-dimensional term $L(\boldsymbol{\beta}_{\xi^*}; \boldsymbol{X}_{\xi^*}, \boldsymbol{y})\ \pi(\boldsymbol{\beta}_{\xi^*})$ follows a normal BvM approximation. More rigorously, we have the next theorem.

Theorem 2.7 (Shape approximation). *Assume the conditions of Theorem* 2.6 *hold,* $\lim s \log l_n = 0$ *and*

$$a_n \prec (1/p_n)\sqrt{1/(ns\log p_n)}.$$

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_{\xi^*}, \sigma^2)^\top$. *Then with dominating probability,* $\pi(\boldsymbol{\beta}, \sigma^2 \mid D_n)$ *converges in total variation to*

$$\phi(\boldsymbol{\beta}_{\xi^*}; \hat{\boldsymbol{\beta}}_{\xi^*}, \sigma^2(\boldsymbol{X}_{\xi^*}^\top \boldsymbol{X}_{\xi^*})^{-1}) \prod_{j\notin\xi^*} \pi(\beta_j \mid \sigma^2)\mathrm{ig}\bigg(\sigma^2, \frac{n-s}{2}, \frac{\hat{\sigma}^2(n-s)}{2}\bigg), \tag{2.8}$$

where $\phi(\cdot)$ *is a multivariate normal density function,* $\mathrm{ig}(\cdot)$ *is an inverse gamma density function,* $\pi(\beta_j \mid \sigma^2)$ *is the conditional prior distribution of* β_j, *and* $\hat{\boldsymbol{\beta}}_{\xi^*}$ *and* $\hat{\sigma}^2$ *are, respectively, the maximum likelihood estimates* (*MLEs*) *of* $\boldsymbol{\beta}_{\xi^*}$ *and* σ^2 *given data* $(\boldsymbol{y}, \boldsymbol{X}_{\xi^*})$.

Refer to Theorem A.8 for the proof of this theorem. Its condition is slightly stronger than that of Theorem 2.6. It requires that a_n is smaller and the prior log-density $\log g_\lambda(\cdot)$ is almost constant around the true value of β^*_j/σ^*. The following corollary can be easily derived from the above theorem.

Corollary 2.8. *Under the condition of Theorem* 2.7, *for any* $j \in \xi^*$, *the marginal posterior of* β_j *converges to the normal distribution* $\phi(\beta_j, \hat{\beta}_j, \sigma^{*2}\sigma_j)$, *where* $\hat{\beta}_j$ *is the* j*-th entry of* $\hat{\boldsymbol{\beta}}_{\xi^*}$, $\sigma_j = [(\boldsymbol{X}_{\xi^*}^\top \boldsymbol{X}_{\xi^*})^{-1}]_{j,j}$. *Furthermore, if* $s \prec \sqrt{n}$, *the posterior* $\pi(\boldsymbol{\beta}_{\xi^{*c}}, \boldsymbol{\beta}_{\xi^*}, \sigma^2 \mid D_n)$ *converges in total variation to*

$$\prod_{j\notin\xi^*} \pi(\beta_j \mid \sigma^2)\phi(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, (n\hat{I}))$$

with probability approaching 1, *where* $\boldsymbol{\theta} = (\boldsymbol{\beta}_{\xi^*}, \sigma^2)^\top$, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_{\xi^*}, \hat{\sigma}^2)^\top$, *and* $(n\hat{I})^{-1} = \mathrm{diag}(\hat{\sigma}^2(\boldsymbol{X}_{\xi^*}^\top \boldsymbol{X}_{\xi^*})^{-1}, 2\hat{\sigma}^4/n)$. *In other words, the BvM theorem holds for the parameter component* $(\boldsymbol{\beta}_{\xi^*}, \sigma^2)$.

Theorem 2.7 is comparable to the result developed under the spike-and-slab prior [12]. Under the spike-and-slab prior, the posterior density of $\boldsymbol{\beta}$ can be rewritten as a mixture,

$$\pi(\boldsymbol{\beta} \mid D_n) = \sum_{\xi\subset\{1,\ldots,p\}} \pi(\xi \mid D_n)\pi(\boldsymbol{\beta}_\xi \mid \boldsymbol{X}_\xi, \boldsymbol{y})1\{\boldsymbol{\beta}_{\xi^c} = 0\}, \tag{2.9}$$

where $\pi(\boldsymbol{\beta}_\xi \mid \boldsymbol{X}_\xi, \boldsymbol{y}) \propto h_\xi(\boldsymbol{\beta}_\xi)f(\boldsymbol{\beta}_\xi; \boldsymbol{X}_\xi, \boldsymbol{y})$, and $h_{|\xi|}(\cdot)$ is defined in (1.2). If $\pi(\xi^* \mid D_n) \to 1$, $\pi(\boldsymbol{\beta} \mid D_n)$ converges to $\pi(\boldsymbol{\beta}_{\xi^*} \mid \boldsymbol{X}_{\xi^*}, \boldsymbol{y})1\{\boldsymbol{\beta}_{\xi^{*c}} = 0\}$. Furthermore, if $\pi(\boldsymbol{\beta}_{\xi^*})$ is sufficiently flat and BvM holds for the low-dimensional term $\pi(\boldsymbol{\beta}_{\xi^*} \mid \boldsymbol{X}_{\xi^*}, \boldsymbol{y})$, then it leads to a posterior normal approximation as

$$\pi(\boldsymbol{\beta} \mid D_m) \approx N(\boldsymbol{\beta}_{\xi^*}; \hat{\boldsymbol{\beta}}_{\xi^*}, (\boldsymbol{X}_{\xi^*}^\top \boldsymbol{X}_{\xi^*})^{-1}) \otimes \delta_0(\boldsymbol{\beta}_{(\xi^*)^c}), \tag{2.10}$$

where $\otimes$ denotes the multiplication of the measure.

Theorem 2.7 and Corollary 2.8 extend the BvM-type result from the spike-and-slab prior to the shrinkage prior. They show that the marginal posterior distribution for the true covariates follows the BvM theorem as if under the low-dimensional setting, the marginal posterior for the false covariates can be approximated by its prior distribution. Since the prior distribution is already highly concentrated, the posterior of the false covariates being almost the same as the prior does not contradict our contraction results. Note that the Bayesian procedure can be viewed as a process of updating the probabilistic knowledge of parameters. The concentrated prior distribution reflects our prior belief that almost all the predictors are inactive, and (2.8) implies that the Bayesian procedure correctly identifies the true model ξ^* and updates the distribution of $\boldsymbol{\beta}_{\xi^*}$ using the data, but it obtains no evidence to support $\beta_j \neq 0$ for any $j \notin \xi^*$ and thus does not update their concentrated prior distributions.

Let $\mathrm{CI}_i(\alpha)$ denote the posterior quantile credible interval of the i-th covariate. If $\pi(\beta \mid \sigma^2)$ is a symmetric distribution, then Corollary 2.8 implies that

$$\begin{aligned} &\lim P^*(\beta_i^* \in \mathrm{CI}_i(\alpha)) = 1-\alpha, \quad \text{if } i \in \xi^*, \\ &\lim P^*(0 \in \mathrm{CI}_i(\alpha)) = 1, \quad \text{if } i \notin \xi^* \end{aligned} \tag{2.11}$$

for any $1 > \alpha > 0$. This result implies that for the false covariates, the Bayesian credible interval is super-efficient: asymptotically, it can be very narrow (as the prior is highly concentrated), but has always 100% probability coverage. This is much different from the confidence interval.

It is important to note that both Theorem 2.7 and its counterpart (2.10) rely on the selection consistency (and the beta-min condition), which drives Bayesian post-selection inference. Therefore, the frequentist coverage of the Bayesian credible interval (the first equation of (2.11)) does not hold uniformly for all the nonzero β_i values, but only holds for those bounded away from 0. If the beta-min condition is violated, one can rewrite the posterior with the shrinkage prior as a mixture distribution similar to (2.9). Hence, the corresponding posterior inference will be model-average-based.

The above asymptotic studies are completely different from the frequentist sampling distribution-based inference tools such as de-biased LASSO [61, 73]. The de-biased LASSO method establishes asymptotic normality as

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \xrightarrow{\mathrm{d}} N(0, \sigma^{*2} S \boldsymbol{X}^\top \boldsymbol{X} S^\top / n) \tag{2.12}$$

for any $\boldsymbol{\beta}^*$, even when it is arbitrarily close to zero, and S is some surrogate inverse matrix of the sample covariance. Different from our posterior consistency result, the asymptotic distribution on the right-hand side of (2.12) is a divergent distribution when $p_n \gg n$.

In the literature, there is a different line of research about the validity of Bayesian credible intervals, which do not require the selection consistency (see, e.g., [6, 63]). These works are usually based on the first-order Bayesian convergence rate only. As a consequence, these credible intervals/balls involve an unknown multiplicative constants (e.g., c_1 and M) that appear in the posterior convergence rate (2.4) and their coverage always converges to 1, rather than the nominal level $1 - \alpha$.

We conjecture that if the consistent point estimation and the inference of credible intervals are made simultaneously, the credible intervals will be super-efficient for the false covariates due to the sparsity constraints (i.e., the prior distribution) imposed on the regression coefficients. These constraints ensure posterior consistency and thus reduce the variability of the coefficients of the false covariates. Based on this understanding, it seems that under the framework of consistent high-dimensional Bayesian analysis, a separate post-selection inference procedure (without sparsity constraints) is necessary to induce the correct second-order inference. For example, it can be done in a sequential manner (referring the idea to [42] and [60]): attempting to add each of the unselected variables to the selected model, and calculating the corresponding credible interval for the unselected variable.

3 Consistent shrinkage priors

In the previous section, we establish general theory for shrinkage priors based on abstract conditions. In this section, we apply the theory to several types of shrinkage priors, and study the corresponding posterior asymptotics.

The condition (2.3) requires certain balance between the prior concentration and the tail thickness. First of all, it is easy to see that the Laplace prior fails to satisfy the condition (2.3) unless the tuning parameter

$$\lambda_n \sim p_n \log p_n \Big/ \sqrt{\frac{s \log p_n}{n}}$$

and the true coefficients are as tiny as $|\beta_j^*| = O(\sqrt{s\log p_n/n}/p_n)$ for all $j \in \xi^*$. Therefore, we first consider the prior specification that has a heavier tail than the exponential distribution.

3.1 Polynomial-tailed distributions

We assume that the prior density of $\boldsymbol{\beta}$ has the form $\pi(\boldsymbol{\beta} \mid \sigma^2) = \prod_{i=1}^{p_n} \frac{1}{\lambda_n \sigma} g(\beta_i/\lambda_n\sigma)$, where λ_n is a scale hyperparameter, and the density $g(\cdot)$ is symmetric and polynomial-tailed, i.e., $g(x) \asymp x^{-r}$ as $|x| \to \infty$ for some positive $r > 1$. Under the above prior specification, we adapt Theorem 2.3 as follows.

Theorem 3.1. *Assume that the conditions* A_1 *and* A_2 *hold for the linear regression model, and a polynomial-tailed prior is used. If* $\log(E_n) = O(\log p_n)$*, and the scale parameter* λ_n *satisfies* $\lambda_n \leqslant a_n p_n^{-(u+1)/(r-1)}$ *and* $-\log\lambda_n = O(\log p_n)$ *for some* $u > 0$*, then*

- *the posterior consistency* (2.4) *holds when* $a_n \asymp \sqrt{s\log p_n/n}/p_n$;
- *the model selection consistency* (2.6) *holds when* $a_n \prec \sqrt{\log p_n}/\sqrt{n}p_n$, $\min_{j\in\xi^*}|\beta_j^*| \geqslant M_1\sqrt{\log p_n/n}$ *for sufficiently large* M_1, $s\log l_n \prec \log p_n$ *and* $u > 1$;
- *the posterior approximation* (2.8) *holds if* $a_n \prec \sqrt{1/(ns\log p_n)}/p_n$, $\min_{j\in\xi^*}|\beta_j^*| \geqslant M_1\sqrt{\log p_n/n}$ *for sufficiently large* M_1, $s\log l_n \prec 1$ *and* $u > 1$.

Note that polynomially decaying distributions that we most commonly use satisfy $g(x) = Cx^{-r}L(x)$, where $\lim_x L(x) = 1$ with the rate

$$|L(x) - 1| = O(x^{-t}) \quad \text{for some } t \geqslant 0. \tag{3.1}$$

It is not difficult to see that if $\min_{j\in\xi^*}|\beta_j^*| > M_2\epsilon_n$ for some large M_2, and $\lambda_n = O(\epsilon_n)$, then $s\log l_n \asymp s\epsilon_n/\min_{j\in\xi^*}|\beta_j^*|$. Therefore, Theorem 3.1 can be refined as follows.

Corollary 3.2. *Consider the polynomial-tailed prior distributions satisfying* (3.1)*. Assume that the condition* A_1 *holds,* $s\log p_n \prec n$*, and* $\log(\max_{j\in\xi^*}|\beta_j^*|) = O(\log p_n)$*. Let the choice of* λ_n *satisfy* $-\log\lambda_n = O(\log p_n)$*. Then*

- *if* $\lambda_n = O\{\sqrt{s\log p_n/n}/p_n^{(u+r)/(r-1)}\}$ *with* $u > 0$*, then posterior consistency holds with a nearly optimal contraction rate*;
- *if* $s\sqrt{s\log p_n/n}/\min_{j\in\xi^*}|\beta_j^*| \prec \log p_n$, $\lambda_n \prec \sqrt{\log p_n/n}/p^{(u+r)/(r-1)}$ *with* $u > 1$*, and* $\min_{j\in\xi^*}|\beta_j^*| \geqslant M_1\sqrt{s\log p_n/n}$ *for sufficiently large* M_1*, then the variable selection consistency holds*;
- *if* $s\sqrt{s\log p_n/n}/\min_{j\in\xi^*}|\beta_j^*| \prec 1$, $\lambda_n \prec \sqrt{1/n\log p_n}/p_n^{(u+r)/(r-1)}$ *with* $u > 1$*, and* $\min_{j\in\xi^*}|\beta_j^*| \geqslant M_1\sqrt{s\log p_n/n}$ *for sufficiently large* M_1*, then the posterior shape approximation holds.*

Theorem 3.1 and Corollary 3.2 show that a nearly optimal contraction rate can be achieved for high-dimensional linear regression by adopting a polynomial-tailed prior with an appropriate value of λ_n. As suggested by Corollary 3.2, it is sufficient to choose the scale parameter as $\log\lambda_n \sim -c\log p_n$ for some $c \ll (u+r)/(r-1)$, since $n = O(p_n)$ and $s = o(p_n)$. Compared with the choice $\lambda_n = (s/p)\sqrt{\log(p/s)}$ under normal mean models [27,62], we note that a stronger prior concentration is required for regression models. Our results allow the maximum magnitude of nonzero coefficients to increase up to a polynomial of p_n. In contrast, the DL prior allows $|\beta_j^*|$ to increase with a logarithmic order of n only [8]. It is worth noting that the boundedness condition on $|\beta_j^*|$ is not necessary for a polynomially decaying prior under normal mean models, i.e., when $\boldsymbol{X} = I$ [27,54,62]. However, under general regression settings, such a condition may be necessary due to the dependency among covariates. One should also notice that the selection consistency or posterior normality requires the stronger beta-min condition (i.e., minimal β^* is greater than the order of $\sqrt{s \log p_n / n}$) and an additional condition on the true sparsity s (e.g., if $\min_{j \in \xi^*} |\beta_j^*| > C$ for some constant C, the selection consistency and the posterior normality require $s^3 \prec n \log^2 p_n$ and $s^3 \prec n / \log p_n$, respectively). The reason we need such unpleasant conditions is that the polynomially decaying prior modeling utilizes only one scale hyperparameter. Although this simplifies the modeling part, we lose control on the shape or tail flatness of the prior distribution. If we utilize both scale and shape hyperparameters in prior modeling, the conditions can be improved, as seen in Subsection 3.2.

For the convenience of posterior sampling, one way to construct a polynomially decaying prior is to design a hierarchical scale mixture Gaussian distribution as

$$\beta_j \sim N(0, \lambda_j^2 \sigma^2), \quad \lambda_j^2 \sim \pi_{s_n}(\lambda_j^2), \quad \text{independently for all } j, \tag{3.2}$$

where s_n is the scale hyperparameter of the mixing distribution $\pi_{s_n}(\cdot)$, i.e., $\pi_{s_n}(\cdot) = \pi_1(\cdot / s_n)/s_n$. Equivalently, $\sqrt{s_n}$ is the scale parameter of the marginal prior of β_j. The scale mixture Gaussian distribution can also be viewed as a local-global shrinkage prior, where λ_j^2's are local shrinkage parameters, and s_n is a deterministic global shrinkage parameter. As shown in the next lemma, the tail behavior of the marginal distribution of β_j is determined by the tail behavior of π_1.

Lemma 3.3. *If the mixing distribution $\pi_{s_n}(\cdot)$ is a polynomial-tailed distribution satisfying $\pi_1(\lambda^2) = C\lambda^{-2\tilde{r}} \tilde{L}(\lambda^2)$ and $|\tilde{L}(\lambda^2) - 1| = O((\lambda^2)^{-\tilde{t}})$, then the marginal prior distribution of β_j induced by* (3.2) *is polynomial-tailed with order $2\tilde{r} - 1$ and satisfies $|L(x) - 1| = O(x^{-2\tilde{t}})$, where L is defined in* (3.1).

The proof of this lemma is trivial and hence omitted in this paper.

Combining the above lemma and Corollary 3.2, it is sufficient to assign λ_j^2 a polynomial-tailed distribution and properly choose the scale parameter s_n such that $\sqrt{s_n}$ is decreasing and satisfies the conditions in Corollary 3.2. Ghosh and Chakrabarti [27] studied the posterior convergence of the normal mean models with a scale mixture Gaussian prior (3.2) and achieved a minimax contraction rate. However, their result is only applicable to the case where the polynomial order $\tilde{r}$ of $\pi_1(\lambda_j^2)$ is between 1.5 and 2. Our result is more general and valid for any $\tilde{r} > 1$.

In what follows, we list some examples of polynomially decaying prior distributions which can be represented as a scale mixture Gaussian distribution. All these priors satisfy the condition (3.1):

- the student's t-distribution, for which the mixing distribution of λ^2 is an inverse gamma distribution $\mathrm{IG}(a_1, s_n)$ with $a_1 > 0$;
- the normal-exponential-gamma (NEG) distribution [29], for which the mixing distribution is $\pi(\lambda^2) = \nu s_n^{-1}(1 + s_n^{-1}\lambda^2)^{-\nu-1}$ with $\nu > 0$;
- the generalized double Pareto distribution [3] with the density $g(x) = (2\lambda_n)^{-1}\,(1+|x|/(a_1\lambda_n))^{-(a_1+1)}$, for which the mixing distribution can be represented as a gamma mixture of exponential distributions with $a_1 > 0$;
- the generalized beta mixture of Gaussian distributions [1], for which the mixing distribution is an inverted beta distribution: $\lambda_j^2 / s_n \sim$ inverted $\mathrm{Beta}(a_1, b_1)$ with $a_1 > 0$. Note that the horseshoe prior is a special case of generalized beta mixture Gaussian distributions with $a_1 = b_1 = 1/2$.

In addition, Theorem 3.1 implies a simple way to remedy the inconsistency of Bayesian LASSO by imposing a heavy tail prior on the hyperparameter: $\beta/\sigma \sim \mathrm{DE}(\lambda_j)$, $\lambda_j^{-1} \sim \pi_{s_n}$, where $\mathrm{DE}(\lambda)$ denotes the double exponential distribution $\lambda \exp\{-\lambda x\}/2$, and the mixing distribution π_{s_n} of λ_j^{-1} has a polynomial tail with the scale parameter s_n.

In the above analysis, we choose the scale parameters λ_n or s_n to decrease deterministically as n increases. Hence, in practice, certain tuning procedures are recommended as described in Section 4. Such hyperparameter tuning occurs in most Bayesian procedures under the spike-and-slab prior as well. Note that such a tuning procedure usually requires multiple simulations under different levels of λ_n. In the literature, an adaptive Bayesian way to choose λ_n is to assign a hyper-prior on λ_n. van der Pas et al. [64] studied the horseshoe prior for the normal mean models, and they showed that the posterior consistency remains if λ_n is subject to a hyper-prior which is truncated on $[1/n, 1]$. However, the results derived for normal mean models may not be trivially applicable to regression models. Note that there is a $\sqrt{n}$ difference between regression models and normal mean models, in terms of the L_2-norm for the columns in the design matrix. The result of [64] suggests to truncate the prior of λ_n on $[n^{-3/2}, n^{-1/2}]$ for regression models. A toy example shown in Figure 4 indicates that such truncation still leads to many false discoveries. Another popular choice is to impose the global shrinkage parameter on a half Cauchy prior $\lambda_n \sim \mathcal{C}^+(0,1)$. The numerical results show that this hierarchical prior leads to insufficient prior shrinkage and less accurate posterior concentration. Finally, our posterior shape approximation result relies on the fact that β_j's are *a priori* independent conditioned on σ^2. If a hyper-prior on λ_n is used, then the conditional *a priori* independence does not hold any more, and the BvM result (2.8) fails.

3.2 Two-component mixture Gaussian distributions

Another prior that has been widely used in the Bayesian linear regression analysis is the two-component mixture Gaussian distribution (see, e.g., [23, 46])

$$\beta_j/\sigma \sim (1-\xi_j)N(0,\sigma_0^2) + \xi_j N(0,\sigma_1^2), \quad \xi_j \sim \text{Bernoulli}(m_1). \tag{3.3}$$

The component $N(0,\sigma_0^2)$ has a very small σ_0 and can be viewed as an approximation to the point mass at 0. In the literature, the interest in this prior has been focused only on the consistency of variable selection, i.e.,

$$\pi(\{j : \xi_j = 1\} = \xi^* \mid D_n).$$

Here, we treat it as an absolutely continuous prior and study the posterior properties of $\boldsymbol{\beta}$ in the next theorem.

Theorem 3.4. *Suppose that the two-component mixture Gaussian prior* (3.3) *is used for the high-dimensional linear regression model* (1.1)*, and the following conditions hold: the conditions* A_1 *and* A_2*,* $E_n^2/\sigma_1^2 + \log\sigma_1 \asymp \log p_n$*,* $m_1 = 1/p_n^{1+u}$ *and* $\sigma_0 \leqslant a_n/\sqrt{2(1+u)\log p_n}$ *for some* $u > 0$*. Then*

- *the posterior consistency* (2.4) *holds when* $a_n \asymp \sqrt{s\log p_n/n}/p_n$*;*
- *the model selection consistency* (2.6) *holds when* $a_n \prec \sqrt{\log p_n}/\sqrt{n}p_n$*,* $sE_n\sqrt{s\log p_n/n}/\sigma_1^2 \prec \log p_n$*,* $\min_{j\in\xi^*}|\beta_j^*| \geqslant M_1\sqrt{\log p_n/n}$ *for sufficiently large* M_1 *and* $u > 1$*;*
- *the posterior approximation* (2.8) *holds when* $a_n \prec \sqrt{1/(ns\log p_n)}/p_n$*,* $sE_n\sqrt{s\log p_n/n}/\sigma_1^2 \prec 1$*,* $\min_{j\in\xi^*}|\beta_j^*| \geqslant M_1\sqrt{\log p_n/n}$ *for sufficiently large* M_1 *and* $u > 1$*.*

The two-normal mixture distribution contains three hyperparameters m, σ_0^2 and σ_1^2. Hence, we have more control on the prior shape compared with the polynomially decaying priors, and the theoretic properties are improved slightly compared with Corollary 3.2. Specifically, Theorem 3.4 allows us to choose $\sigma_1 = E_n = p_n^c$ for some $c > 1$ and thus $sE_n\sqrt{s\log p_n/n}/\sigma_1^2 \prec 1$ always holds, i.e, there will be no additional conditions on the upper bound of the model size s; Theorem 3.4 only requires that $\min_{j\in\xi^*}|\beta_j^*|$ is larger than the order of $\sqrt{\log p/n}$.

4 Bayesian computation and an illustrative example

In this section, we first discuss some important practical issues, including the posterior computation, the model selection and the hyperparameter tuning, and then we use some toy examples to illustrate the performance of the shrinkage priors. For convenience, we call the Bayesian method, whose consistency is guaranteed by Theorem 2.3 with a shrinkage prior, a Bayesian consistent shrinkage (BCS) method in what follows. In particular, we use the student-t prior, as an example of the shrinkage prior, and compare it with the Laplace prior.

The scale mixture Gaussian priors (3.2), under a proper hierarchical representation, usually lead to posterior conjugate Gibbs updates. For example, for the student-t prior, the posterior distribution can be updated in the following way:

$$
\begin{aligned}
&\sigma^2 \mid \boldsymbol{\beta}, \lambda_1, \ldots, \lambda_{p_n} \sim \mathrm{IG}\biggl(a_0 + \frac{n+p_n}{2}, b_0 + \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2} + \sum_j \frac{\beta_j^2}{2\lambda_j^2}\biggr),\\
&\boldsymbol{\beta} \mid \sigma^2, \lambda_1, \ldots, \lambda_{p_n} \sim N(K^{-1}\boldsymbol{X}^\top \boldsymbol{y}/\sigma^2, K^{-1}),\\
&f(\lambda_j^2 \mid \boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\lambda_j} \exp\biggl\{-\frac{\beta_j^2}{2\lambda_j^2\sigma^2}\biggr\}\pi(\lambda_j^2), \quad j = 1, \ldots, p_n,
\end{aligned}
\tag{4.1}
$$

where $K = (\boldsymbol{X}^\top\boldsymbol{X} + \Lambda)/\sigma^2$, $\Lambda = \mathrm{diag}(1/\lambda_j^2)$, and $\pi(\lambda_j^2)$ denotes the density function of an inverse gamma distribution, i.e., $\lambda_j^2 \sim \mathrm{IG}(a_1, s_n)$.

The step of updating $\boldsymbol{\beta}$ is computationally difficult due to the inverse of a $p_n \times p_n$ matrix. However, the special structure of the covariance matrix K^{-1} allows for a blockwise update of $\boldsymbol{\beta}$ [34]. For example, if we partition $\boldsymbol{\beta}$ into two blocks $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$, and partition $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ and $\Lambda = \mathrm{diag}(\Lambda_1, \Lambda_2)$ accordingly, then the conditional distribution of $\boldsymbol{\beta}^{(1)}$ is given by

$$
\boldsymbol{\beta}^{(1)} \mid \boldsymbol{\beta}^{(2)} \sim N((\boldsymbol{X}_1^\top\boldsymbol{X}_1 + \Lambda_1)^{-1}\boldsymbol{X}_1^\top(\boldsymbol{y} - \boldsymbol{X}_2\boldsymbol{\beta}^{(2)}), \sigma^2(\boldsymbol{X}_1^\top\boldsymbol{X}_1 + \Lambda_1)^{-1}),
\tag{4.2}
$$

which requires only an inverse of a lower-dimensional matrix. The computational complexity of updating $\boldsymbol{\beta}$ in (4.1) is $O(p_n^3)$, while that in (4.2) is $O((d^3 + n(p_n - d))p_n/d)$, where d is the block size and the term $n(p_n - d)$ comes from computing the product $\boldsymbol{X}_2\boldsymbol{\beta}^{(2)}$. The optimal order of d is $O(\sqrt[3]{np_n})$, which yields a computational complexity of $O(n^{2/3}p_n^{5/3})$ for one update of the entire vector $\boldsymbol{\beta}$. Further improvement in computation is possible when we incorporate the idea of the skinny Gibbs sampler [45].

Posterior model selection based on BCS has been discussed in Sections 2 and 3 from the theoretical aspect. However, in practice, the selection rule $\xi(\boldsymbol{\beta}) = \{j : |\beta_j/\sigma| > a_n\}$ cannot be directly used since a_n is not an explicit hyperparameter of the prior distribution. Recall that a_n represents the boundary of the prior spike region, and it is implicitly defined through the condition (2.3) as $\pi(|\beta_j/\sigma| > a_n) = p_n^{-1-u}$. Since u is unknown, we suggest to choose the threshold a in the rule $\pi(|\beta_j/\sigma| > a) = 1/p_n$, i.e., let $u = 0$. This rule means that we set the expected *a priori* model size to be 1. Such a rule has often been in the literature of Bayesian model selection (see, e.g., [46]). Obviously, $a_n \leqslant a$, and thus it leads to a conservative selection. However, if $a \ll \min_{j\in\xi^*}|\beta_j^*|$, it is not difficult to see that the Bayesian selection consistency remains, when $\min_{j\in\xi^*}|\beta_j^*|$ satisfies the beta-min condition. In the simulation studies of this paper, we choose the Bayesian estimator for the model as $\hat{\xi} = \{j : q_j \triangleq \pi(|\beta_j/\sigma| > a|D_n) > t\}$, where $t = 0.5$ and q_j plays the role of the posterior inclusion probability. It is worth mentioning that one may also use a data-driven method to determine the value of t, and make the variable selection rule more robust across different sparsity regimes. For example, we can conduct a multiple hypothesis test based on the marginal inclusion probabilities q_j's for the hypotheses $H_{j0} : \beta_j = 0$ versus $H_{j1} : \beta_j \neq 0$, $j = 1, \ldots, p_n$ based on posterior summaries. This can be done using an empirical Bayesian approach as developed in [19, 41].

Another important practical issue is how to select hyperparameters. The theory developed in Sections 2 and 3 provides only sufficient conditions for the asymptotic order of hyperparameters. For example, by Theorem 3.2, one can set the scale parameter $\lambda_n = 1/[\sqrt{n\log p_n}p_n^\gamma]$ with any sufficiently large value of γ for the student-t prior. Asymptotically, an excessively large value of γ does not affect the rate of convergence, but affects only the multiplicative constants, such as M and c_1, in the statement of Theorem 2.3. However, in finite-sample applications, it is crucial to select a properly scaled parameter such that the posterior is neither over- nor under-shrunk. In this work, we let $\lambda_n = 1/[\sqrt{n\log p_n}p_n^{\hat{\gamma}}]$ and choose $\hat{\gamma}$ to minimize the posterior mean of a "BIC-like score":

$$
\int \mathrm{bic}(\boldsymbol{\beta}, \sigma^2) d\pi(\boldsymbol{\beta}, \sigma^2 \mid D_n, \gamma),
$$

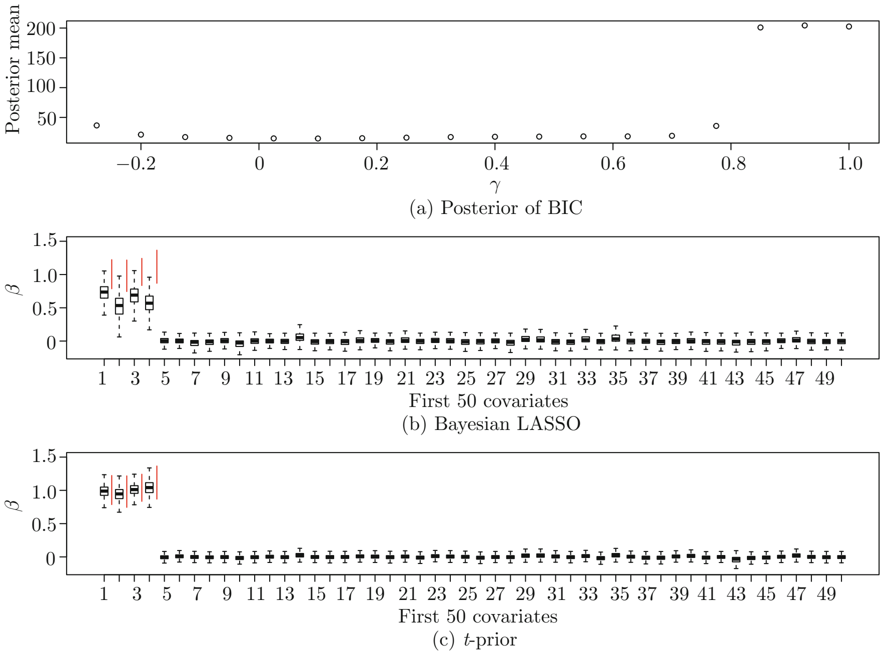
where $\mathrm{bic}(\boldsymbol{\beta}) = n\log(\|Y - \boldsymbol{X}^\top\tilde{\boldsymbol{\beta}}\|^2/n) + \|\tilde{\boldsymbol{\beta}}\|_0 \log n$, $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_{p_n})$, $\tilde{\beta}_j = \beta_j 1(|\beta_j/\sigma| > a)$, and $\pi(\boldsymbol{\beta}, \sigma^2 \mid D_n, \gamma)$ is the posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$ given the hyperparameter γ. In practice, one can run multiple posterior simulations with different values of γ, and then choose the one that yields the smallest posterior sample mean of the "BIC-like" score. Since the multiple runs can be made in parallel on a high-performance computer, such a parameter tuning strategy does not add much on computational time. Since investigating the theoretical properties of tuning parameter selection is beyond the scope of this work, such study will be conducted elsewhere.

We illustrate the performance of BCS using a simulated example, where $p = 200$, $n = 120$, and the nonzero coefficients are $(\beta_1, \beta_2, \beta_3, \beta_4) = (1, 1, 1, 1)$. For the Laplace prior, we set the hyperparameter $\lambda = \sqrt{n \log(p_n)}$ at which the LASSO estimator is known to be consistent (see, e.g., [72]). For the student-t prior, we set the degree of freedom to 3 with the scale parameter

$$s_n = \lambda_n^2 = 1/[n \log p_n p_n^{-2\gamma}],$$

where γ ranges from -0.25 to 1.1, and the best $\hat{\gamma}$ is selected as described in the above. For both priors, we let σ^2 be subject to an inverse gamma distribution with $a_0 = b_0 = 1$.

The numerical results are summarized in Figure 1. The first plot shows the posterior sample mean of the BIC-like score with different values of γ. It shows that when γ is larger than 0.8, the tuning parameter λ_n is too small, the posterior begins to miss true covariates due to over-shrinkage, and thus the posterior mean of the BIC-like score rapidly increases to a very large value. The second and third plots are the posterior boxplots of $\pi(\beta_j \mid D_n)$ of Bayesian LASSO, and BCS under the optimal setting of $\hat{\gamma}$. To make the boxplots more visible, we only include the coefficients of the first 50 covariates, including four true covariates. The comparison shows that BCS leads to a consistent inference of the model in the sense that the coefficients of the false covariates are shrunk to zero, and the coefficients of the true covariates are distributed around their true values. In contrast, Bayesian LASSO over-shrink the coefficients of true covariates, and under-shrink the coefficients of false covariates. This is due to the fact that the Laplace prior fails to achieve the balance between the prior concentration and the tail thickness. But it is worth noting that the posterior Bayesian LASSO can still separate the true and false covariates, and thus it can be used for the model selection.

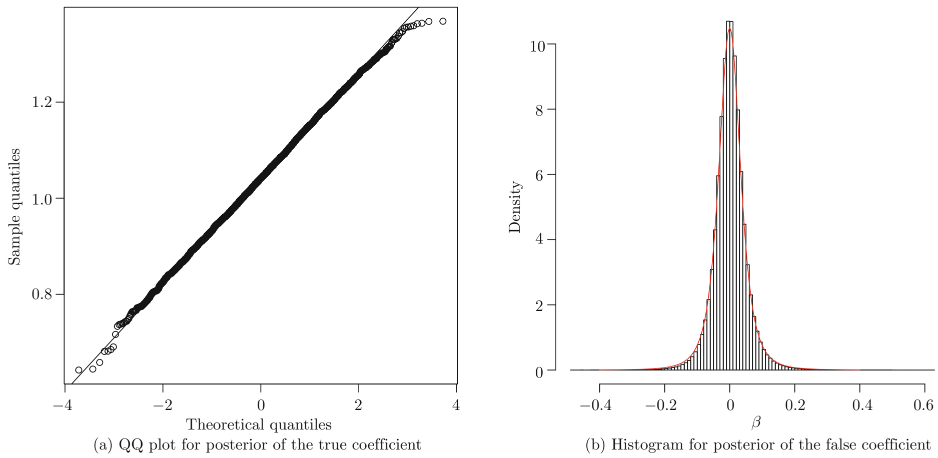


(a) Posterior of BIC

(b) Bayesian LASSO

(c) t-prior

Figure 1 (Color online) (a) The posterior mean of the BIC-like score for different values of γ; (b) box-plots of the posterior samples by Bayesian LASSO; (c) box-plots of the posterior samples by BCS



(a) QQ plot for posterior of the true coefficient

(b) Histogram for posterior of the false coefficient

Figure 2 (Color online) Shape of the posterior distribution by BCS: (a) shows the QQ plot for one true covariate, and (b) shows the histogram of the posterior samples of β_j/σ, $j \notin \xi^*$ (i.e., false covariates), where the red curve represents the density of the student-t prior

In addition, we draw in Figure 1 four red vertical segments which represent the 99% oracle confidence intervals of the true coefficients by assuming that the true model is known. In Figure 2, we examine the shape of posterior samples resulted from BCS. The plots are consistent with the established BvM theorem, i.e., the equation (2.8).

Figure 1 shows that for this example a wide range of γ, from -0.1 to 0.6, yields similar posterior means for the BIC-like score, which implies that the true model is correctly selected under γ within this range. The BIC-like score posterior mean criterion tends to select a smaller value of γ within this range, since a smaller γ reduces the shrinkage effect on the true covariates. But further experiments can show that the performance of BCS is actually quite stable with any γ in this range. This also implies that BCS is tolerant to stochastic tuning errors.

As discussed previously, the Bayesian interval estimates obtained by BCS will be super-efficient for false covariates. Their coverages highly rely on the selection consistency, and have completely different performance compared with frequentist confidence intervals. The frequentist de-biased LASSO estimator is defined as

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{LASSO}} + \frac{1}{n} S \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{LASSO}}),$$

where S is the surrogate inverse matrix of the sample covariance. This de-bias step applies an OLS (ordinary least squares)-type bias correction to the LASSO estimator. In the ideal case where $p_n \leqslant n$ and $\frac{1}{n}S = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, the de-biased LASSO estimator reduces to the OLS estimator. Therefore, the marginal confidence intervals of all the covariates, including both true and false, have the same length scale.

5 Numerical studies

This section examines the performance of BCS in variable selection and uncertainty assessment for the regression coefficient estimates. The method is tested on two simulation examples and a real data example.

In the simulation study, two design matrices are considered for the model (1.1): $(n,p) = (80, 201)$ and $(n,p) = (100, 501)$, where the intercept term has been included. The true values of the parameters are

$$\sigma^* = 1, \quad \boldsymbol{\beta} = (0, 1, 1.5, 2, 0, \ldots, 0)^\top,$$

where the first 0 corresponds to the intercept term. The design matrices are generated from the multivariate normal distribution $N(0, \Sigma)$ with the covariance structure, (1) independent covariates:

$\Sigma = I$, or (2) pairwisely dependent covariates: $\Sigma_{ii} = 1.0$ for all i, $\Sigma_{i,j} = 0.5$ for $i \neq j$. The methods under comparison include BCS, Bayesian LASSO, LASSO and SCAD. For Bayesian LASSO, we set the scale parameters to

$$\lambda = \sqrt{n \log p_n}.$$

For BCS, the tuning parameter s_n is selected by the posterior mean of the BIC-like score as discussed in Section 4. For the setting of the Gibbs sampler, we set the total iteration number to $N = 40{,}000$ in addition to 5,000 iterations for the burn-in process. The posterior samples are collected at every 40 iterations. The R-packages *glmnet* [21] and *ncvreg* [10] are used for implementing LASSO and SCAD, where the tuning parameter λ is chosen to minimize the 10-fold cross-validation error. For LASSO, this is to set $\lambda =$ `lambda.min` in *glmnet*. Since LASSO is known to select many false variables, we have also tried to set $\lambda =$ `lambda.1se`, which is to choose the largest value of λ such that the cross-validation error is within one standard deviation of the minimum cross-validation error. The R-package *hdi* [17] is used for implementing de-biased LASSO. All the results reported below are based on 112 simulated replicates.

5.1 Simulation I: $n = 80$, $p = 201$

We evaluate accuracy of the estimates obtained from various methods in the L_1-error, which is defined as $\sum_{j\in\xi^*} |\beta_j^* - \hat{\beta}_j|$ for the true covariates and $\sum_{j\notin\xi^*} |\hat{\beta}_j|$ for the false ones. For the Bayesian methods, the posterior mean is used as the point estimator, although which is not the optimal choice for minimizing the L_1-error. We evaluate the accuracy of variable selection using the average number of selected true covariates $|\hat{\xi} \cap \xi^*|$ (where the perfect value is 3), and the average number of selected false covariates $|\hat{\xi} \cap (\xi^*)^c|$ (where the perfect value is 0). For each covariate, we also compare the marginal credible intervals produced by the Bayesian methods and the confidence intervals produced by de-biased LASSO under a nominal level of 95%. For simplicity, the credible intervals are constructed based on the empirical quantiles from posterior samples instead of the highest density region.

The results are summarized in Tables 1 and 2 for the case of independent covariates and the case of dependent covariates, respectively. First of all, we can see that BCS works extremely well in identifying true models, whose performance is almost perfect. As seen in Section 4, Bayesian LASSO can also distinguish between the true and false covariates from posterior samples when the coefficients of the true covariates are sufficiently large. However, due to over-shrinkage, it does not work well when they are small. Hence, Bayesian LASSO mis-identifies some true covariates for this example. Both LASSO and SCAD tend to select dense models, although the true covariates can be selected. As mentioned previously, this is due to an inherent drawback of the regularization methods. The regularization shrinks the true regression coefficients toward zero. To compensate the shrinkage effect, some false covariates have to be included. For LASSO, the comparison shows that the choice of $\lambda =$ `lambda.1se` alleviates the "overselection" issue, and leads to the less estimation error for zero β_j's and larger estimation bias for nonzero β_j's. BCS also shrinks the true regression coefficients, but it can still perform well in variable selection. This is due to the fact that BCS accounts for the uncertainty of coefficient estimates in variable selection: BCS is sample-based, for which different false covariates might be selected to compensate the shrinkage effect at different iterations, and thus the chance of selecting false covariates can be largely eliminated by averaging over different iterations.

Regarding the parameter estimation, we note that SCAD yields somehow better results than BCS. However, a direct comparison of these two methods is unfair, as the BCS tells us something more beyond point estimation, e.g., credible interval. Also, BCS leads to much accurate variable selection as reported above. Among the Bayesian methods, we can see that BCS performs much better than Bayesian LASSO, which indicates the importance of posterior consistency. We note that it is unfair to directly compare L_1-estimation errors of $\boldsymbol{\beta}_{(\xi^*)^c}$ for shrinkage estimators (BCS or Bayesian LASSO) and sparse estimators (LASSO or SCAD), since the shrinkage estimators never shrink any coefficients to exactly zero. For example, in Table 1, the L_1-error of BCS is 2.3, which is much larger than those by LASSO and SCAD. However, it actually implies that $\hat{\beta}_j \approx 2.3/200 \approx 0.011$ for each zero β_j, as BCS selects almost no false predictors. Hence, it represents fairly successful shrinkage for the false predictors.

Table 1 Comprehensive comparison of BCS, Bayesian LASSO (Bay-LASSO), LASSO with `lambda.min` (LASSO$_1$), LASSO with `lambda.1se` (LASSO$_2$), SCAD and de-biased LASSO for the datasets with independent covariates, $n = 80$ and $p = 201$

	Methods					
	BCS	Bay-LASSO	de-biased LASSO	LASSO$_1$	LASSO$_2$	SCAD
L_1-error of $\boldsymbol{\beta}_{\xi^*}$	0.3380	2.1115	0.3503	0.6678	1.0850	0.2811
Standard error	0.0149	0.0281	0.2537	0.0211	0.0275	0.0133
L_1-error of $\boldsymbol{\beta}_{(\xi^*)^c}$	2.3137	4.5533	23.3050	0.8402	0.1650	0.2180
Standard error	0.0758	0.0360	0.2537	0.0950	0.0319	0.0324
$\lvert\hat{\xi} \cap \xi^*\rvert$	3.0000	2.3036	–	3.0000	3.0000	3.0000
Standard error	–	0.0505	–	–	–	–
$\lvert\hat{\xi} \cap (\xi^*)^c\rvert$	0	0	–	14.1610	3.1964	4.3304
Standard error	–	–	–	1.2841	0.5041	0.5176
Coverage of ξ^*	0.9067	0.0595	0.9613	–	–	–
Average length	0.4996	0.8471	0.5798	–	–	–
Coverage of $(\xi^*)^c$	1.0000	1.0000	0.9492	–	–	–
Average length	0.1371	0.3322	0.5490	–	–	–

Table 2 Comprehensive comparison of BCS, Bayesian LASSO (Bay-LASSO), LASSO with `lambda.min` (LASSO$_1$), LASSO with `lambda.1se` (LASSO$_2$), SCAD and de-biased LASSO for the datasets with dependent covariates, $n = 80$ and $p = 201$

	Methods					
	BCS	Bay-LASSO	de-biased LASSO	LASSO$_1$	LASSO$_2$	SCAD
L_1-error of $\boldsymbol{\beta}_{\xi^*}$	0.5040	2.7798	0.4469	0.8735	0.9516	0.3593
Standard error	0.0342	0.0302	0.0192	0.0265	0.0242	0.0170
L_1-error of $\boldsymbol{\beta}_{(\xi^*)^c}$	0.3805	4.8558	24.7950	1.3638	0.4509	0.1587
Standard error	0.0782	0.0302	0.2448	0.1083	0.0274	0.0198
$\lvert\hat{\xi} \cap \xi^*\rvert$	2.9000	1.7500	–	3.0000	3.0000	3.0000
Standard error	0.0300	0.0546	–	–	–	–
$\lvert\hat{\xi} \cap (\xi^*)^c\rvert$	0.0100	0	–	16.1790	6.7232	2.3125
Standard error	0.0100	–	–	0.9801	0.3418	0.2568
Coverage of ξ^*	0.9000	0.0327	0.8988	–	–	–
Average length	0.6970	0.9279	0.6046	–	–	–
Coverage of $(\xi^*)^c$	1.0000	1.0000	0.9543	–	–	–
Average length	0.0373	0.3804	0.5418	–	–	–

For the interval estimation, de-biased LASSO produces high quality confidence intervals. For both true and false covariates, it produces about the same length confidence intervals, and the coverage rates of these confidence intervals are about the same as the nominal level. This observation is consistent with our previous discussion. For the true covariates, BCS yields almost 95% converge; in contrast, Bayesian LASSO yields a very low coverage due to the effect of over-shrinkage. For the false covariates, both BCS and Bayesian LASSO produce 100% coverage with very narrow credible intervals. Hence, they do not have the correct long-run frequency coverage for false predictors. These discoveries agree with our theoretical results. The de-biased LASSO yields wider intervals for the false covariates, as it cannot incorporate the model sparsity information into the construction of confidence intervals.

The performance of BCS for the cases of independent and dependent covariates is quite consistent, except that the proposed method tends to select a smaller value of γ for the independent case and, as a consequence, the posterior L_1-error of the false covariates tends to be larger than for the dependent case. This is reasonable, as the high spurious correlation requires a higher penalty for the multiplicity adjustment.

5.2 Simulation II: $n = 100$, $p = 501$

The results are summarized in Tables 3 and 4 for the independent and dependent covariates, respectively.

As in the case of $n = 80$ and $p = 201$, BCS performs much better than the regularization methods in variable selection, and performs much better than Bayesian LASSO in all the aspects of variable selection, parameter estimation and interval estimation.

Before moving forward to the real application in the next section, we mention that we also conduct simulations, under the same data generation scheme, for the two-Gaussian mixture prior specification. While the two-Gaussian mixture prior also achieves near-perfect model selection performance, we find that its shrinkage effect on $\boldsymbol{\beta}_{(\xi^*)^c}$ and its interval estimation coverage performance are inferior to those of t shrinkage prior (although they are much better than Bayesian LASSO inference results). One potential reason is that the hyperparameters m_1, σ_1^2 and σ_0^2 are not optimally tuned. Our empirical experience shows that the value of m_1 has a large effect on model selection performance, and the values of σ_1^2 and σ_0^2 affect the level of the posterior shrinkage and the posterior normality asymptotics. However, tuning all three hyperparameters simultaneously is much more difficult in practice, than tuning only one hyperparameter of the t-shrinkage methods, and hence is not recommended.

Table 3 Comprehensive comparison of BCS, Bayesian LASSO (Bay-LASSO), LASSO with `lambda.min` (LASSO$_1$), LASSO with `lambda.1se` (LASSO$_2$), SCAD and de-biased LASSO for the datasets with independent covariates, $n = 100$ and $p = 501$

	Methods					
	BCS	Bay-LASSO	de-biased LASSO	LASSO$_1$	LASSO$_2$	SCAD
L_1-error of $\boldsymbol{\beta}_{\xi^*}$	0.2789	2.3863	0.3177	0.7173	0.9645	0.2616
Standard error	0.0115	0.0310	0.0145	0.0229	0.0253	0.0107
L_1-error of $\boldsymbol{\beta}_{(\xi^*)^c}$	4.4011	8.7190	50.3010	0.9736	0.2158	0.3080
Standard error	0.0312	0.0602	0.4636	0.0900	0.0436	0.0402
$\lvert\hat{\xi} \cap \xi^*\rvert$	3.0000	2.1964	–	3.0000	3.0000	3.0000
Standard error	–	0.0436	–	–	–	–
$\lvert\hat{\xi} \cap (\xi^*)^c\rvert$	0.0268	0	–	20.554	4.7411	7.0178
Standard error	0.0153	–	–	1.6070	0.8629	0.8042
Coverage of ξ^*	0.9285	0.0208	0.9494	–	–	–
Average length	0.4300	0.7412	0.4985	–	–	–
Coverage of $(\xi^*)^c$	1.0000	1.0000	0.9517	–	–	–
Average length	0.1506	0.2841	0.6038	–	–	–

Table 4 Comprehensive comparison of BCS, Bayesian LASSO (Bay-LASSO), LASSO with `lambda.min` (LASSO$_1$), LASSO with `lambda.1se` (LASSO$_2$), SCAD and de-biased LASSO for the datasets with dependent covariates, $n = 100$ and $p = 501$

	Methods					
	BCS	Bay-LASSO	de-biased LASSO	LASSO$_1$	LASSO$_2$	SCAD
L_1-error of $\boldsymbol{\beta}_{\xi^*}$	0.3960	3.1087	0.3888	0.8742	1.0228	0.3338
Standard error	0.0260	0.0300	0.0155	0.0282	0.0223	0.0158
L_1-error of $\boldsymbol{\beta}_{(\xi^*)^c}$	0.4288	9.2585	54.2889	1.3656	0.5331	0.1424
Standard error	0.1076	0.0694	0.4754	0.1045	0.0342	0.0196
$\lvert\hat{\xi} \cap \xi^*\rvert$	2.9464	1.4554	–	3.0000	3.0000	3.0000
Standard error	0.0213	0.0566	–	–	–	–
$\lvert\hat{\xi} \cap (\xi^*)^c\rvert$	0.0089	0	–	21.4280	9.7324	6.4732
Standard error	0.0089	–	–	1.3218	0.4742	0.8065
Coverage of ξ^*	0.9107	0.0060	0.9077	–	–	–
Average length	0.5783	0.7498	0.5263	–	–	–
Coverage of $(\xi^*)^c$	1.0000	1.0000	0.9316	–	–	–
Average length	0.0219	0.2870	0.6142	–	–	–

5.3 A real data example

We analyze a reduced gene expression dataset on Bardet-Biedl syndrome from [52]. The reduced dataset is available in the R-package *flare* [39], which contains 120 samples with 201 gene expression levels. The scientific community has discovered that TRIM32 is the causal gene to Bardet-Biedl syndrome [16]. In this example, we treat the expression level of gene TRIM32 as the response variable and the expression levels of the other 200 genes as predictors. Therefore, the selected set of genes from this regression will cover the regulators of gene TRIM32 by the consistency property of BCS.

We apply both de-biased LASSO and BCS to this regression problem. De-biased LASSO identifies gene 153 as the only significant covariate according to the Bonferroni-adjusted p-values, and produces a 95% confidence interval of $[0.024, 0.072]$ for this gene. For BCS, the optimal value $\hat{\gamma} = 0.58$ is selected, and the posterior exceedance probability $q_j \triangleq \pi(|\beta_j/\sigma| > a \mid D_n)$ is used to quantify the significance of each covariate, where a is as defined in Section 4. BCS also identifies gene 153 as the most significant covariate with $q_{153} = 0.54$. Figure 3 shows the posterior distribution of the regression coefficient of gene 153 under the choice of $\hat{\gamma} = 0.58$ as well as the confidence intervals produced by the two methods. The 95% highest posterior density (HPD) credible interval produced by BCS is $[-0.018, 0.018] \cup [0.064, 0.131]$, which is the union of two intervals representing the evidence against and for the true covariate, respectively. Note that if the true model is exactly gene 153, its OLS estimator will be 0.109. The de-biased LASSO confidence interval (represented by the dashed segment in Figure 3) seems a compromise between the two intervals, and it does not contain the OLS value 0.109.

6 Conclusion

In this paper, we have studied the posterior asymptotics under absolutely continuous priors for high-dimensional linear regression. We first prove that if the prior distribution is heavy-tailed and allocates a sufficiently large probability mass in a very small neighborhood of zero, then the posterior consistency holds with a nearly optimal contraction rate. More specifically, we find that any polynomial-tailed distribution with a scale parameter, which decreases as p_n increases, can be used as an appropriate prior to derive valid Bayesian inference for high-dimensional regression models. Note that it is not necessary for the continuous prior distribution to have an infinite density at zero as in the DL or horseshoe priors.

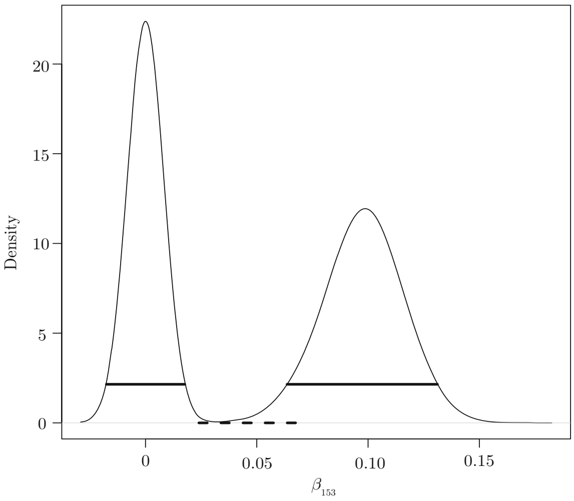


Figure 3 Histogram of the posterior samples of the regression coefficient of gene 153, where the black line shows the posterior HPD interval, and the dashed line shows the de-biased LASSO confidence interval

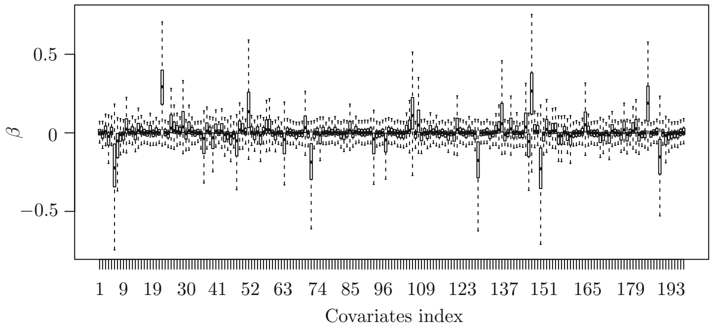


Figure 4 Boxplots of $\{\beta_j\}_{j\notin\xi^*}$ simulated from a posterior distribution with a horseshoe prior for the same dataset used in Figure 1, where the global shrinkage parameter is truncated into $[n^{-3/2}, n^{-1/2}]$

In the literature, the local-global shrinkage prior has been widely studied, especially for the normal mean problem. Such a prior follows $\beta_j \sim N(0, \sigma^2\lambda_j^2\tau^2)$, where λ_j^2 controls the local shrinkage, and τ^2 controls the degree of global shrinkage. Our work verifies that a sufficient condition that ensures consistency of the local-global shrinkage is to let the local shrinkage parameter λ_j^2 follow some polynomial-tailed distribution, and let the global shrinkage parameter τ^2 deterministically decrease in the order $-\log(\tau^2) = O(\log p_n)$. In this work, we suggest a BIC-like score posterior mean criterion for tuning the global shrinkage parameter. Although it works well for our examples, it is still of great interest to the Bayesian community if an adaptive or full Bayesian approach can be developed for choosing, rather than tuning, the global shrinkage parameter. Such analysis has been conducted by van der Pas et al. [64] under normal mean models. However, there is a significant difference between normal mean models and regression models. For the former, one can directly analyze the marginal posterior $\pi(\beta_j \mid D_n)$ as β_j's are (conditionally) independent. For the latter, one needs to take into account the dependency among covariates. Empirically, the result of [64] seems not applicable to regression problems. Figure 4 shows the boxplots of the regression coefficients drawn from a posterior $\pi(\beta_j|D_n)$ constructed with a horseshoe prior for the same dataset used in the toy example of Section 4, where λ_j is subject to a half-Cauchy prior, and τ is subject to a uniform prior truncated on $[n^{-3/2}, n^{-1/2}]$. The plot shows that the horseshoe prior leads to many false discoveries for this example. Therefore, we would note that adaptively choosing the global shrinkage parameter is nontrivial due to spurious multicollinearity caused by the curse of dimensionality.

In this paper, we have also studied the selection consistency based on the sparsified posterior, as well as the posterior shape approximation. We prove that if the tail of the prior distribution is sufficiently flat, then selection is consistent and the BvM-type result holds. This further implies that for the true covariates, the credible intervals are asymptotically equivalent to the oracle confidence intervals, and for the false covariates, the credible intervals are super-efficient.

The theory established in this paper implies that a consistent shrinkage prior shares almost the same posterior asymptotic behavior with the golden standard spike-and-slab prior (see, e.g., [12]). However, the shrinkage prior is more efficient in computation. In this paper, we use a student-t prior in all the numerical studies, and the Gibbs sampler is conveniently used in sampling from posterior distributions. The computation shall be further improved if a stochastic gradient MCMC algorithm is employed for simulations. However, for the spike-and-slab prior, a trans-dimensional MCMC sampler has to be used for simulations.

Acknowledgements Qifan Song was supported by National Science Foundation of USA (Grant No. DMS-1811812). Faming Liang was supported by National Science Foundation of USA (Grant No. DMS-2015498) and National Institutes of Health of USA (Grant Nos. R01GM117597 and R01GM126089). The authors thank the referees for their constructive comments on the paper.

References

1 Armagan A, Dunson D B, Clyde M. Generalized beta mixtures of Gaussians. Adv Neural Information Process Syst, 2011, 24: 523–531

2 Armagan A, Dunson D B, Lee J. Generalized double Pareto shrinkage. Statist Sinica, 2013, 23: 119–143

3 Armagan A, Dunson D B, Lee J, et al. Posterior consistency in linear models under shrinkage priors. Biometrika, 2013, 100: 1011–1018

4 Bai R, Ghosh M. On the beta prime prior for scale parameters in high-dimensional Bayesian regression models. Statist Sinica, 2021, 31: 843–865

5 Barron A R. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In: Bayesian Statist, vol. 6. Oxford: Clarendon Press, 1999, 27–52

6 Belitser E, Ghosal S. Empirical Bayes oracle uncertainty quantification for regression. Ann Statist, 2020, 48: 3113–3137

7 Bhadra A, Datta J, Li Y, et al. Prediction risk for the Horseshoe regression. J Mach Learn Res, 2019, 20: 1–39

8 Bhattacharya A, Pati D, Pillai N S, et al. Dirichlet-Laplace priors for optimal shrinkage. J Amer Statist Assoc, 2015, 110: 1479–1490

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

10 Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann Appl Stat, 2011, 5: 232–253

11 Carvalho C M, Polson N G, Scott J G. The horseshoe estimator for sparse signals. Biometrika, 2010, 97: 465–480

12 Castillo I, Schmidt-Hieber J, van der Vaart A. Bayesian linear regression with sparse priors. Ann Statist, 2015, 43: 1986–2018

13 Castillo I, van der Vaart A. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. Ann Statist, 2012, 40: 2069–2101

14 Chen J H, Chen Z H. Extended Bayesian information criteria for model selection with large model spaces. Biometrika, 2008, 95: 759–771

15 Chen T Q, Fox E B, Guestrin C. Stochastic gradient Hamiltonian Monte Carlo. J Mach Learn Res, 2014, 15: 1683–1691

16 Chiang A P, Beck J S, Yen H J, et al. Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (BBS11). Proc Natl Acad Sci USA, 2006, 103: 6287–6292

17 Dezeure R, Bühlmann P, Meier L, et al. High-dimensional inference: Confidence intervals, p-values and R-software hdi. Statist Sci, 2015, 30: 533–558

18 Duane S, Kennedy A D, Pendleton B J, et al. Hybrid Monte Carlo. Phys Lett B, 1987, 195: 216–222

19 Efron B. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. J Amer Statist Assoc, 2004, 99: 96–104

20 Fan J Q, Li R Z. Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc, 2001, 96: 1348–1360

21 Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Statist Softw, 2010, 33: 1–22

22 Gao C, Ada W, van der Vaart A W, et al. A general framework for Bayes structured linear models. Ann Statist, 2020, 48: 2848–2878

23 George E I, McCulloch R E. Variable selection via Gibbs sampling. J Amer Statist Assoc, 1993, 88: 881–889

24 Ghosal S. Asymptotic normality of posterior distributions in high-dimensional linear models. Bernoulli, 1999, 5: 315–331

25 Ghosal S, Ghosh J K, van der Vaart A W. Convergence rates of posterior distributions. Ann Statist, 2000, 28: 500–531

26 Ghosal S, van der Vaart A W. Convergence rates of posterior distributions for noniid observations. Ann Statist, 2007, 35: 192–223

27 Ghosh P, Chakrabarti A. Posterior concentration properties of a general class of shrinkage priors around nearly black vectors. arXiv:1412.8161, 2014

28 Girolami M, Galderhead B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J R Stat Soc Ser B Stat Methodol, 2011, 73: 123–214

29 Griffin J E, Brown P J. Bayesian hyper-Lassos with non-convex penalization. Aust N Z J Stat, 2011, 53: 423–442

30 Griffin J E, Brown P J. Structuring shrinkage: Some correlated priors for regression. Biometrika, 2012, 99: 481–487

31 Hahn P R, Carvalho C M. Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. J Amer Statist Assoc, 2015, 110: 435–448

32 Hans C. Bayesian lasso regression. Biometrika, 2009, 96: 835–845

33 Inglot T. Inequalities for quantiles of the chi-square distribution. Probab Math Statist, 2010, 30: 339–351

34 Ishwaran H, Rao J S. Spike and slab variable selection: Frequentist and Bayesian strategies. Ann Statist, 2005, 33: 730–773

35 Jiang W X. Bayesian variable selection for high dimensional generalized linear models: Convergence rate of the fitted

densities. Ann Statist, 2007, 35: 1487–1511

36 Johnson V E, Rossel D. Bayesian model selection in high-dimensional settings. J Amer Statist Assoc, 2012, 107: 649–660

37 Kleijn B J K, van der Vaart A W. Misspecification in infinite-dimensional Bayesian statistics. Ann Statist, 2006, 34: 837–877

38 Li H N, Pati D. Variable selection using shrinkage priors. Comput Statist Data Anal, 2017, 107: 107–119

39 Li X G, Zhao T, Yuan X M, et al. The flare package for high dimensional linear regression and precision matrix estimation in R. J Mach Learn Res, 2015, 16: 553–557

40 Liang F M, Song Q F, Yu K. Bayesian subset modeling for high-dimensional generalized linear models. J Amer Statist Assoc, 2013, 108: 589–606

41 Liang F M, Zhang J. Estimating the false discovery rate using the stochastic approximation algorithm. Biometrika, 2008, 95: 961–977

42 Lockhart R, Taylor J, Tibshirani R J, et al. A significance test for the Lasso. Ann Statist, 2014, 42: 413–468

43 Luo S K, Song R, Witten D. Sure screening for Gaussian graphical models. arXiv:1407.7819, 2014

44 Martin R, Mess R, Walker S G. Empirical Bayes posterior concentration in sparse high-dimensional linear models. Bernoulli, 2017, 23: 1822–1847

45 Narisetty N N. Statistical analysis of complex data: Bayesian model selection and functional data depth. PhD Thesis. Ann Arbor: University of Michigan, 2016

46 Narisetty N N, He X M. Bayesian variable selection with shrinking and diffusing priors. Ann Statist, 2014, 42: 789–817

47 Neal R M. MCMC using Hamiltonian dynamics. In: Handbook of Markov Chain Monte Carlo. New York: Chapman and Hall/CRC, 2011, 113–162

48 Park T, Casella G. The Bayesian Lasso. J Amer Statist Assoc, 2008, 103: 681–686

49 Pati D, Bhattacharya A, Pillai N S, et al. Posterior contraction in sparse Bayesian factor models for massive covariance matrices. Ann Statist, 2014, 42: 1102–1130

50 Raskutti G, Wainwright M J, Yu B. Minimax rates of estimation for high-dimensional linear regression over L_q-balls. IEEE Trans Inform Theory, 2011, 57: 6976–6994

51 Ročková V, George E I. The spike-and-slab LASSO. J Amer Statist Assoc, 2018, 113: 431–444

52 Scheetz T E, Kim K Y A, Swiderski R E, et al. Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proc Natl Acad Sci USA, 2006, 103: 14429–14434

53 Scott J G, Berger J O. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Ann Statist, 2010, 38: 2587–2619

54 Song Q. Bayesian shrinkage towards sharp minimaxity. Electron J Stat, 2020, 14: 2714–2741

55 Song Q, Cheng G. Optimal false discovery control of minimax estimator. arXiv:1812.10013, 2018

56 Song Q, Liang F. High-dimensional variable selection with reciprocal L_1-regularization. J Amer Statist Assoc, 2015, 110: 1607–1620

57 Song Q, Liang F. A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. J R Stat Soc Ser B Stat Methodol, 2015, 77: 947–972

58 Tang X, Xu X, Ghosh M, et al. Bayesian variable selection and estimation based on global-local shrinkage priors. Sankhya A, 2018, 80: 215–246

59 Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol, 1996, 58: 267–288

60 Tibshirani R, Taylor J, Lockhart R, et al. Exact post-selection inference for sequential regression procedures. J Amer Statist Assoc, 2016, 111: 600–620

61 van de Geer S, Bühlmann P, Ritov Y, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann Statist, 2014, 42: 1166–1202

62 van der Pas S L, Kleijn B J K, van der Vaart A W. The horseshoe estimator: Posterior concentration around nearly black vectors. Electron J Stat, 2014, 8: 2585–2618

63 van der Pas S L, Szabo B, van der Vaart A W. Uncertainty quantification for the horseshoe (with discussion). Bayesian Anal, 2017, 12: 1221–1274

64 van der Pas S L, Szabo B, van der Vaart A W. Adaptive posterior contraction rates for the horseshoe. Electron J Stat, 2017, 11: 3196–3225

65 van der Vaart A W, Wellner J A. Weak Convergence and Empirical Processes. New York: Springer, 1996

66 Vershynin R. Introduction to the non-asymptotic analysis of random matrices. In: Compressed Sensing Theory and Applications. Cambridge: Cambridge University Press, 2012, 210–268

67 Wei R, Reich B J, Hoppin J A, et al. Sparse Bayesian additive nonparametric regression with application to health effects of pesticides mixtures. Statist Sinica, 2020, 30: 55–79

68 Welling M, Yee W T. Bayesian learning via stochastic gradient Langevin dynamics. In: Proceedings of the 28th International Conference on Machine Learning. Bellevue: ICML, 2011, 681–688

69 Xu Z, Schmidt D F, Makalic E, et al. Bayesian sparse global-local shrinkage regression for grouped variables.

arXiv:1709.04333, 2017

70 Yang Y, Wainwright M J, Jordan M I. On the computational complexity of high-dimensional Bayesian variable selection. Ann Statist, 2016, 44, 2497–2532

71 Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. Ann Statist, 2010, 38: 894–942

72 Zhang C-H, Huang J. The sparsity and bias of the LASSO selection in high-dimensional regression. Ann Statist, 2008, 36: 1567–1594

73 Zhang C-H, Zhang S. Confidence intervals for low dimensional parameters in high dimensional linear models. J R Stat Soc Ser B Stat Methodol, 2014, 76: 217–242

74 Zhang Y, Naughton B, Bondell H D, et al. High dimensional linear regression via the R2-D2 shrinkage prior. J Amer Statist Assoc, 2022, in press

75 Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol, 2005, 67: 301–320

76 Zubkov A M, Serov A A. A complete proof of universal inequalities for the distribution function of the binomial law. Theory Probab Appl, 2013, 57: 539–544

Appendix A Proof of the main theorem

First, we restate the result from [65, Lemma 2.2.11] for the sake of readability.

Lemma A.1 (Bernstein's inequality). *If $Z_1,\ldots,Z_n$ are independent random variables with mean zero and satisfy that $\mathrm{E}|Z_i|^m \leqslant m!M^{m-2}v_i/2$ for every $m>1$ and some constants M and v_i, then*

$$P\left(\left|\sum Z_i\right| > z\right) \leqslant 2\exp\{-z^2/2(v+Mz)\}$$

for $v \geqslant \sum v_i$.

As mentioned in [43], the conditions in Lemma A.1 are satisfied by the centered one-degree chi-square distribution.

Lemma A.2. *If X follows the χ_1^2 distribution, there exists some constant C such that for any $m\in\mathbb{N}$, we have $\mathrm{E}|X-\mathrm{E}(X)|^m \leqslant Cm!2^m$. Therefore, given any constant scale λ,*

$$\mathrm{E}|\lambda X - \mathrm{E}(\lambda X)|^m \leqslant m!(2\lambda)^{m-2}(4C\lambda^2).$$

The following lemma (see [76]) gives an upper bound for the tail probability of the binomial distribution.

Lemma A.3. *For a binomial random variable $X\sim B(n,v)$, for any $1<k<n-1$,*

$$\Pr(X \geqslant k+1) \leqslant 1-\Phi(\mathrm{sign}(k-nv)\sqrt{2nH(v,k/n)}),$$

where Φ is the cumulative distribution function of the standard Gaussian distribution and

$$H(v,k/n) = (k/n)\log(k/nv) + (1-k/n)\log[(1-k/n)/(1-v)].$$

We also restate [5, Lemma 6].

Lemma A.4. *Let B_n and C_n be two subsets of the parameter space Θ, and ϕ_n be the test function satisfying $\phi_n(D_n)\in[0,1]$ for any realization D_n of the data generation. If $\pi(B_n)\leqslant b_n$, $E_{\theta^*}\phi_n(D_n)\leqslant b_n'$ and $\sup_{\theta\in C_n} E_\theta(1-\phi_n(D_n))\leqslant c_n$, where $E_\theta(\cdot)$ denotes the expectation with respect to the data generation with the true parameter value being θ. Furthermore, if*

$$P^*\left\{\frac{m(D_n)}{f^*(D_n)} \geqslant a_n\right\} \geqslant 1-a_n',$$

where $f^*=f_{\theta^*}$ is the true density function, and

$$m(D_n) = \int_\Theta \pi(\theta)f_\theta(D_n)d\theta$$

is the margin probability of D_n, then

$$P^*\left(\pi(C_n\cup B_n \mid D_n) \geqslant \frac{b_n+c_n}{a_n\delta_n}\right) \leqslant \delta_n + b_n' + a_n'$$

for any δ_n.

Theorem A.5. *Consider a linear regression model* (1.1) *with the design matrix satisfying conditions* A_1 *and* A_2*. The prior of* σ^2 *follows an inverse gamma distribution* $\mathrm{IG}(a,b)$*, and the prior density of* $\boldsymbol{\beta}$ *is given by*

$$\pi(\boldsymbol{\beta} \mid \sigma^2) = \prod_{i=1}^{p_n} \frac{1}{\sigma} g_\lambda(\beta_i/\sigma).$$

If there exists a positive constant u *such that*

$$1 - \int_{-a_n}^{a_n} g_\lambda(x)dx \leqslant p_n^{-(1+u)} \quad \text{and} \quad -\log\Big(\inf_{x\in[-E_n,E_n]} g_\lambda(x)\Big) = O(\log p_n) \tag{A.1}$$

hold for $a_n \asymp \sqrt{s\log p_n/n}/p_n$*, then the posterior consistency holds asymptotically, i.e.,*

$$P^*\{\pi[A_n \mid D_n] > \exp(-c_1 n\epsilon_n^2)\} \leqslant \exp(-c_2 n\epsilon_n^2),$$

where

$$\begin{aligned} A_n = {} & \{\text{at least } \tilde{p} \text{ entries of } |\boldsymbol{\beta}/\sigma| \text{ are larger than } a_n\} \cup \{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \geqslant (3+\sqrt{\lambda_0})\sigma^*\epsilon_n\} \\ & \cup \{\sigma^2/\sigma^{*2} > (1+\epsilon_n)/(1-\epsilon_n) \text{ or } \sigma^2/\sigma^{*2} < (1-\epsilon_n)/(1+\epsilon_n)\} \end{aligned}$$

with $\tilde{p} \asymp s$ *and* $\epsilon_n = M\sqrt{s\log p_n/n}$ *for some large constant* M*.*

Proof. We apply Lemma A.4 to prove this theorem. Define $C_n = A_n \backslash B_n$, where

$$B_n = \{\text{at least } \tilde{p} \text{ entries of } |\boldsymbol{\beta}/\sigma| \text{ are larger than } a_n\},$$

$\tilde{p} \leqslant \bar{p} - s$, $\tilde{p} \prec n\epsilon_n^2$, and its specific choice will be given below. The proof consists of three parts.

Firstly, we show the existence of a testing function ϕ_n such that

$$E_{(\boldsymbol{\beta}^*,\sigma^{*2})}(\phi_n) \leqslant \exp(-c_3 n\epsilon_n^2) \quad \text{and} \quad \sup_{(\boldsymbol{\beta},\sigma^2)\in C_n} E_{(\boldsymbol{\beta},\sigma^2)}(1-\phi_n) \leqslant \exp(-c_3' n\epsilon_n^2) \tag{A.2}$$

for some positive constants c_3 and c_3'.

Secondly, we show that for some $c_4 > 0$,

$$\pi(B_n) < \mathrm{e}^{-c_4 n\epsilon_n^2}. \tag{A.3}$$

Thirdly, we show that

$$\lim_n P^*\left\{\frac{m(D_n)}{f^*(D_n)} \geqslant \exp(-c_5 n\epsilon_n^2)\right\} > 1 - \exp\{-c_5' n\epsilon_n^2\} \tag{A.4}$$

for some positive $0 < c_5 < \min(c_3', c_4)$. Therefore, the proof can be concluded by Lemma A.4.

Part I. We consider the testing function $\phi_n = \max\{\phi_n', \tilde{\phi}_n\}$, where

$$\begin{aligned} \phi_n' &= \max_{\{\xi\supseteq\xi^*,|\xi|\leqslant\tilde{p}+s\}} 1\{|\boldsymbol{y}^\top(I-H_\xi)\boldsymbol{y}/(n-|\xi|)\sigma^{*2} - 1| \geqslant \epsilon_n\}, \\ \tilde{\phi}_n &= \max_{\{\xi\supseteq\xi^*,|\xi|\leqslant\tilde{p}+s\}} 1\{\|(\boldsymbol{X}_\xi^\top\boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^\top\boldsymbol{y} - \boldsymbol{\beta}_\xi^*)\| \geqslant \sigma^*\epsilon_n\} \end{aligned}$$

and $H_\xi = \boldsymbol{X}_\xi(\boldsymbol{X}_\xi^\top\boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^\top$ is the hat matrix corresponding to ξ.

For any ξ that satisfies $\xi \supseteq \xi^*$ and $|\xi| \leqslant \tilde{p} + s$, we have

$$\begin{aligned} & E_{(\boldsymbol{\beta}^*,\sigma^{*2})} 1\{|\boldsymbol{y}^\top(I-H_\xi)\boldsymbol{y}/(n-|\xi|)\sigma^{*2} - 1| \geqslant \epsilon_n\} \\ & \quad = \Pr(|\chi^2_{n-|\xi|} - (n-|\xi|)| \geqslant (n-|\xi|)\epsilon_n) \leqslant \exp(-\hat{c}_3 n\epsilon_n^2) \end{aligned} \tag{A.5}$$

for some small constant $\hat{c}_3$, where χ_p^2 denotes a chi-square distribution with degree of freedom p, and the last inequality follows from the Bernstein inequality (see Lemmas A.1 and A.2) and the facts that $\epsilon \prec 1$ and $s + \tilde{p} \prec n$.

Following the similar arguments to those in the proof of [3, Lemma 1], we have that for any ξ satisfying $\xi \supseteq \xi^*$ and $|\xi| \leqslant \tilde{p} + s \prec n\epsilon_n^2$,

$$\begin{aligned}&E_{(\boldsymbol{\beta}^*,\sigma^{*2})}1\{\|(\boldsymbol{X}_\xi^\top \boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^\top \boldsymbol{y} - \boldsymbol{\beta}_\xi^*\| \geqslant \sigma^*\epsilon_n \mid \boldsymbol{\beta}^*, \sigma^{*2}\}\\ &\quad = E_{(\boldsymbol{\beta}^*,\sigma^{*2})}1\{\|(\boldsymbol{X}_\xi^\top \boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^\top \boldsymbol{\varepsilon}\| \geqslant \epsilon_n\} \leqslant \Pr(\chi^2_{|\xi|} \geqslant n\lambda_0\epsilon_n^2)\\ &\quad \leqslant \exp(-\tilde{c}_3 n\epsilon_n^2) \qquad\qquad (\text{A.6})\end{aligned}$$

for some $\tilde{c}_3 > 0$. Note that the last inequality holds due to the Bernstein inequality and the large value of M.

Combining (A.5) and (A.6), we obtain that

$$\begin{aligned}E_{(\boldsymbol{\beta}^*,\sigma^{*2})}\phi_n &\leqslant E_{(\boldsymbol{\beta}^*,\sigma^{*2})} \sum_{\{\xi\supseteq\xi^*,|\xi|\leqslant\tilde{p}+s\}} (1\{|\boldsymbol{y}^\top(I - H_\xi)\boldsymbol{y}/(n-|\xi|)\sigma^{*2} - 1| \geqslant \epsilon_n\}\\ &\quad + 1\{\|(\boldsymbol{X}_\xi^\top \boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^\top \boldsymbol{y} - \boldsymbol{\beta}_\xi^*)\| \geqslant \epsilon_n\})\\ &< (\tilde{p}+s)\binom{p_n}{\tilde{p}+s}[\exp(-c_3 n\epsilon_n^2) + \exp(-c_3' n\epsilon_n^2)]. \qquad (\text{A.7})\end{aligned}$$

We set $\tilde{p} = \lfloor \min\{\hat{c}_3, \tilde{c}_3\}n\epsilon_n^2/(2\log p_n)\rfloor$. (Since $\bar{p}\log p_n \succ n\epsilon_n^2$, $\tilde{p}$ always exists.) Hence, we have

$$\log(\tilde{p}+s) + (\tilde{p}+s)\log p_n < (2\min\{\hat{c}_3, \tilde{c}_3\}n\epsilon_n^2)/3,$$

which leads to $E_{(\boldsymbol{\beta}^*,\sigma^{*2})}\phi_n \leqslant \exp(-c_3 n\epsilon_n^2)$ for some fixed c_3.

Now we study $\sup_{(\boldsymbol{\beta},\sigma^2)\in C_n} E_{(\boldsymbol{\beta},\sigma^2)}(1-\phi_n)$. Let $C_n \subset \hat{C}_n \cup \tilde{C}_n$, where

$$\begin{aligned}\hat{C}_n &= \{\sigma^2/\sigma^{*2} > (1+\epsilon_n)/(1-\epsilon_n) \text{ or } \sigma^2/\sigma^{*2} < (1-\epsilon_n)/(1+\epsilon_n)\}\\ &\quad \cap \{\text{at most } \tilde{p} \text{ entries of } |\boldsymbol{\beta}/\sigma| \text{ are larger than } a_n\},\\ \tilde{C}_n &= \{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| > (3+\sqrt{\lambda_0})\sigma^*\epsilon_n, \sigma^2/\sigma^{*2} \leqslant (1+\epsilon_n)/(1-\epsilon_n)\\ &\quad \text{and at most } \tilde{p} \text{ entries of } |\boldsymbol{\beta}/\sigma| \text{ are larger than } a_n\}.\end{aligned}$$

Then we have

$$\begin{aligned}\sup_{(\boldsymbol{\beta},\sigma^2)\in C_n} E_{(\boldsymbol{\beta},\sigma^2)}(1-\phi_n) &= \sup_{(\boldsymbol{\beta},\sigma^2)\in C_n} E_{(\boldsymbol{\beta},\sigma^2)} \min\{1-\phi_n', 1-\tilde{\phi}_n\}\\ &\leqslant \max\Big\{\sup_{(\boldsymbol{\beta},\sigma^2)\in \hat{C}_n} E_{(\boldsymbol{\beta},\sigma^2)}(1-\phi_n'), \sup_{(\boldsymbol{\beta},\sigma^2)\in \tilde{C}_n} E_{(\boldsymbol{\beta},\sigma^2)}(1-\tilde{\phi}_n)\Big\}.\end{aligned}$$

Let $\tilde{\xi} = \tilde{\xi}(\boldsymbol{\beta}) = \{k : |\beta_k/\sigma| > a_n\} \cup \xi^*$ and $\tilde{\xi}^c = \{1,\ldots,p_n\}\backslash\tilde{\xi}$. Hence, for any $(\boldsymbol{\beta},\sigma^2) \in \tilde{C}_n \cup \hat{C}_n$, $|\tilde{\xi}(\boldsymbol{\beta})| \leqslant \tilde{p} + s \leqslant \bar{p}$, and $\|\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}\| \leqslant \sqrt{np}\|\boldsymbol{\beta}_{\tilde{\xi}^c}\| \leqslant \sqrt{n}\sqrt{\lambda_0'}\sigma\epsilon_n$ given a large value of M. It holds that

$$\begin{aligned}&\sup_{(\boldsymbol{\beta},\sigma^2)\in C_n'} E_{\boldsymbol{\beta}}(1-\phi_n')\\ &\quad = \sup_{(\boldsymbol{\beta},\sigma^2)\in C_n'} E_{(\boldsymbol{\beta},\sigma^2)} \min_{\xi\supseteq\xi^*,|\xi|\leqslant\tilde{p}+s} 1\{|\boldsymbol{y}^\top(I-H_\xi)\boldsymbol{y}/(n-|\tilde{\xi}|)\sigma^{*2} - 1| \leqslant \epsilon_n\}\\ &\quad \leqslant \sup_{(\boldsymbol{\beta},\sigma^2)\in C_n'} E_{(\boldsymbol{\beta},\sigma^2)}1\{|\boldsymbol{y}^\top(I-H_{\tilde{\xi}})\boldsymbol{y}/(n-|\tilde{\xi}|)\sigma^{*2} - 1| \leqslant \epsilon_n\}\\ &\quad = \sup_{(\boldsymbol{\beta},\sigma^2)\in C_n'} \Pr\{|\sigma^2(\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}/\sigma + \boldsymbol{\varepsilon})^\top(I-H_{\tilde{\xi}})(\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}/\sigma + \boldsymbol{\varepsilon})/[(n-|\tilde{\xi}|)\sigma^{*2}] - 1| \leqslant \epsilon_n\}\\ &\quad \leqslant \sup_{(\boldsymbol{\beta},\sigma^2)\in C_n'} \Pr\{\sigma^2(\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}/\sigma + \boldsymbol{\varepsilon})^\top(I-H_{\tilde{\xi}})(\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}/\sigma + \boldsymbol{\varepsilon})/[(n-|\tilde{\xi}|)\sigma^{*2}] \in [1-\epsilon_n, 1+\epsilon_n]\}\\ &\quad \leqslant \sup_{(\boldsymbol{\beta},\sigma^2)\in C_n'} \Pr\{(\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}/\sigma + \boldsymbol{\varepsilon})^\top(I-H_{\tilde{\xi}})(\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}/\sigma + \boldsymbol{\varepsilon})/(n-|\tilde{\xi}|) \notin [1-\epsilon_n, 1+\epsilon_n]\}\\ &\quad \leqslant \sup_{(\boldsymbol{\beta},\sigma^2)\in C_n'} \Pr\{|\chi^2_{n-|\tilde{\xi}|}(k) - (n-|\tilde{\xi}|)| \geqslant (n-|\tilde{\xi}|)\epsilon_n\}\\ &\quad \leqslant \exp(-\hat{c}_3' n\epsilon_n^2)\end{aligned}$$

for some $\hat{c}_3' > 0$. Note that $(\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}/\sigma + \boldsymbol{\varepsilon})^\top (I - H_{\tilde{\xi}})(\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}/\sigma + \boldsymbol{\varepsilon})$ follows a noncentral χ^2 distribution $\chi_{n-|\tilde{\xi}|}(k)$ with the noncentral parameter

$$k = \boldsymbol{\beta}_{\tilde{\xi}^c}^\top \boldsymbol{X}_{\tilde{\xi}^c}^\top (I - H_{\tilde{\xi}})\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}/\sigma^2 \leqslant (\sqrt{n}\sqrt{\lambda_0'}\epsilon_n/4)^2.$$

Since the noncentral χ^2 distribution is a sub-exponential, the last inequality follows from the Bernstein inequality as well. Also, we have

$$\begin{aligned}
&\sup_{(\boldsymbol{\beta},\sigma^2)\in\tilde{C}_n} E_{(\boldsymbol{\beta},\sigma^2)}(1-\tilde{\phi}_n)\\
&\quad= \sup_{(\boldsymbol{\beta},\sigma^2)\in\tilde{C}_n} E_{(\boldsymbol{\beta},\sigma^2)} \min_{|\xi|\leqslant\tilde{p}+s} 1\{\|(\boldsymbol{X}_\xi^\top\boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^\top\boldsymbol{y} - \boldsymbol{\beta}_\xi^*\| \leqslant \sigma^*\epsilon_n\}\\
&\quad\leqslant \sup_{(\boldsymbol{\beta},\sigma^2)\in\tilde{C}_n} E_{(\boldsymbol{\beta},\sigma^2)} 1\{\|(\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{y} - \boldsymbol{\beta}_{\tilde{\xi}}^*\| \leqslant \sigma^*\epsilon_n\}\\
&\quad= \sup_{(\boldsymbol{\beta},\sigma^2)\in\tilde{C}_n} \Pr\{\|(\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{y} - \boldsymbol{\beta}_{\tilde{\xi}}^*\| \leqslant \sigma^*\epsilon_n \mid \boldsymbol{\beta},\sigma^2\}\\
&\quad= \sup_{(\boldsymbol{\beta},\sigma^2)\in\tilde{C}_n} \Pr\{\|(\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^\top\sigma\boldsymbol{\varepsilon} + \boldsymbol{\beta}_{\tilde{\xi}} + (\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c} - \boldsymbol{\beta}_{\tilde{\xi}}^*\| \leqslant \sigma^*\epsilon_n\}\\
&\quad\leqslant \sup_{(\boldsymbol{\beta},\sigma^2)\in\tilde{C}_n} \Pr\{\|(\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^\top\sigma\boldsymbol{\varepsilon}\| \geqslant \|\boldsymbol{\beta}_{\tilde{\xi}} - \boldsymbol{\beta}_{\tilde{\xi}}^*\| - (\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c} - \sigma^*\epsilon_n\}\\
&\quad= \sup_{(\boldsymbol{\beta},\sigma^2)\in\tilde{C}_n} \Pr\{\|(\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{\varepsilon}\| \geqslant [\|\boldsymbol{\beta}_{\tilde{\xi}} - \boldsymbol{\beta}_{\tilde{\xi}}^*\| - \sigma^*\epsilon_n - (\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}]/\sigma\}\\
&\quad\leqslant \sup_{(\boldsymbol{\beta},\sigma^2)\in\tilde{C}_n} \Pr\{\|(\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{\varepsilon}\| \geqslant \epsilon_n\} \leqslant \exp(-\tilde{c}_3 n\epsilon_n^2),
\end{aligned}$$

where the above inequalities hold asymptotically because

$$\|\boldsymbol{\beta}_{\tilde{\xi}} - \boldsymbol{\beta}_{\tilde{\xi}}^*\| \geqslant \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| - p_n(\sqrt{\lambda_0}\epsilon_n\sigma/p_n), \quad \sigma^*/\sigma \geqslant \sqrt{(1-\epsilon_n)/(1+\epsilon_n)}$$

and

$$\begin{aligned}
\|(\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}})^{-1}\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}/\sigma\| &\leqslant \sqrt{\lambda_{\max}((\boldsymbol{X}_{\tilde{\xi}}^\top\boldsymbol{X}_{\tilde{\xi}})^{-1})}\|\boldsymbol{X}_{\tilde{\xi}^c}\boldsymbol{\beta}_{\tilde{\xi}^c}\|\\
&\leqslant \sqrt{1/n\lambda_0}\sqrt{n\lambda_0'}\epsilon_n\sigma \leqslant \epsilon_n.
\end{aligned}$$

Hence, (A.2) is proved.

Part II. Define $N = |\{i : |\beta_i/\sigma| \geqslant a_n\}|$, and thus $N \sim \text{Binomial}(p_n, v_n)$, where

$$v_n = \int_{|x|\geqslant a_n} g_\lambda(x)dx$$

and $g_\lambda(x)$ is the prior density function of β_i/σ. Thus $\pi(B_n) = \Pr(\text{Binomial}(p_n, v_n) \geqslant \tilde{p})$. By Lemma A.3, we have

$$\begin{aligned}
&\pi(B_n) \leqslant 1 - \Phi(\sqrt{2p_n H[v_n, (\tilde{p}-1)/p_n]}) \leqslant \frac{\exp\{-p_n H[v_n, (\tilde{p}-1)/p_n]\}}{\sqrt{2\pi}\sqrt{2p_n H[v_n, (\tilde{p}-1)/p_n]}},\\
&p_n H[v_n, (\tilde{p}-1)/p_n] = (\tilde{p}-1)\log[(\tilde{p}-1)/(p_n v_n)] + (p_n - \tilde{p} + 1)\log[(p_n - \tilde{p} + 1)/(p_n - p_n v_n)].
\end{aligned}$$

Therefore, to prove (A.3), it is sufficient to show that

$$p_n H[v_n, (\tilde{p}-1)/p_n] \geqslant O(n\epsilon_n^2).$$

Since $1/(p_n v_n) \geqslant O(p_n^u)$, $\tilde{p}\log p_n^u \asymp n\epsilon_n^2$ (if M is sufficiently large), $(\tilde{p}-1)\log[(\tilde{p}-1)/(p_n v_n)] \asymp n\epsilon_n^2$ and

$$(p_n - \tilde{p} + 1)\log[(p_n - \tilde{p} + 1)/(p_n - p_n v_n)] \approx \tilde{p} - \tilde{p}^2/p_n \prec n\epsilon_n^2.$$

Hence, we have

$$p_n H[v_n, (\tilde{p}-1)/p_n] = O(n\epsilon_n^2).$$

Part III. Now we prove (A.4). Because

$$m(D_n)/f^*(D_n) = \int \frac{(\sigma^*)^n \exp\{-\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2/2\sigma^2\}}{\sigma^n \exp\{-\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*\|^2/2\sigma^{*2}\}} \pi(\boldsymbol{\beta},\sigma^2) d\boldsymbol{\beta} d\sigma^2,$$

it is sufficient to show that

$$\begin{aligned} &P^*(\pi(\{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2/2\sigma^2 + n\log(\sigma/\sigma^*) < \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*\|^2/2\sigma^{*2} + c_5 n\epsilon_n^2/2\}) \geqslant \mathrm{e}^{-c_5 n\epsilon_n^2/2}) \\ &\quad \geqslant 1-\exp\{-c_5' n\epsilon_n^2\} \end{aligned}$$

for some sufficiently small positive c_5.

Note that

$$P^*(\Omega = \{\|\boldsymbol{\varepsilon}\|^2 \leqslant n(1+\hat{c}_5) \text{ and } \|\boldsymbol{\varepsilon}^\top \boldsymbol{X}\|_\infty \leqslant \hat{c}_5 n\epsilon_n\}) \geqslant 1-\exp\{-c_5' n\epsilon_n^2\}$$

for some $\hat{c}_5$, by the properties of the chi-square distribution and the normal distribution. On the event of Ω, it is easy to see that $\{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2/2\sigma^2 + n\log(\sigma/\sigma^*) < \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*\|^2/2\sigma^{*2} + c_5 n\epsilon_n^2/2\}$ is a super-set of $\{\sigma \in [\sigma^*, \sigma^* + \eta_1\epsilon_n^2]$ and $\|(\boldsymbol{\beta}^*-\boldsymbol{\beta})/\sigma\|_1 < 2\eta_2\epsilon_n\}$ for some small constants η_1 and η_2.

In addition, we have

$$\begin{aligned} &-\log \pi(\{\sigma \in [\sigma^*, \sigma^*+\eta_1\epsilon_n^2] \text{ and } \|(\boldsymbol{\beta}^*-\boldsymbol{\beta})/\sigma\|_1 < 2\eta_2\epsilon_n\}) \\ &\quad = -\log \pi(\{0 \leqslant \sigma^2 - \sigma^{*2} \leqslant \eta_1\epsilon_n^2\}) - \log \pi(\{\|(\boldsymbol{\beta}^*-\boldsymbol{\beta})/\sigma\|_1 < 2\eta_2\epsilon_n\}). \end{aligned} \tag{A.8}$$

Given the fact that the inverse gamma density is always bounded away from zero around σ^{*2}, hence the first term in (A.8) satisfies

$$\begin{aligned} -\log \pi(\{0 \leqslant \sigma^2 - \sigma^{*2} \leqslant \eta_1\epsilon_n^2\}) &\leqslant -\log(\eta_1\epsilon_n^2) - \log\Big(\min_{\sigma\in[\sigma^*,\sigma^*+\eta_1\epsilon_n^2]} \pi(\sigma^2)\Big) \\ &< \text{constant} + \log(1/\epsilon_n^2) \leqslant \delta_1 n\epsilon_n^2, \end{aligned}$$

where δ_1 can be an arbitrary constant if we choose M to be sufficiently large.

For the second term in (A.8),

$$\begin{aligned} \{\|(\boldsymbol{\beta}^*-\boldsymbol{\beta})/\sigma\|_1 < 2\eta_2\epsilon_n\} \supset &\{|\beta_j/\sigma| \leqslant \eta_2\epsilon_n/p_n \text{ for all } j \notin \xi^*\} \\ &\cap \{\beta_j/\sigma \in [\beta_j^*/\sigma - \eta_2\epsilon_n/s, \beta_j^*/\sigma + \eta_2\epsilon_n/s] \text{ for all } j \in \xi^*\} \end{aligned}$$

and

$$\begin{aligned} &\pi(\{|\beta_j/\sigma| \leqslant \eta_2\epsilon_n/p_n \text{ for all } j \notin \xi^*\}) \\ &\quad \geqslant \pi(\{|\beta_j/\sigma| \leqslant a_n \text{ for all } j \notin \xi^*\}) \geqslant (1-p_n^{-1-u})^{p_n} \to 1, \end{aligned} \tag{A.9}$$

given a large value of M. For those $\beta_j^* \neq 0$ and

$$\pi(\{\beta_j/\sigma \in [\beta_j^*/\sigma \pm \eta_2\epsilon_n/s] \text{ for all } j \in \xi^*\}) \geqslant \Big[2\eta_2\epsilon_n \inf_{x\in[-E,E]} g_\lambda(x)/s\Big]^s, \tag{A.10}$$

the inequality holds because $|\beta_j^*/\sigma| + \eta\epsilon_n/s \leqslant E$ which is implied by $\sigma^2 < \sigma^{*2} + \eta'\epsilon_n^2$ and $|\beta_j^*/\sigma^*| \leqslant \gamma E$. By (A.9), (A.10) and the condition (A.1), (A.4) holds. □

Theorem A.6. *If all the conditions of Theorem* A.5 *except the condition* A$_1$(3) *hold, then the posterior prediction for the observed data is consistent, i.e.,*

$$P^*\{\pi[\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^*\| \geqslant c_0\sqrt{n}\sigma^*\epsilon_n \mid D_n] < 1-\exp(-c_1 n\epsilon_n^2)\} \leqslant \exp(-c_2 n\epsilon_n^2)$$

for some c_0, c_1 and c_2.

Proof. Define

$$\begin{aligned} A_n &= \{\text{at least } \tilde{p} \text{ entries of } |\boldsymbol{\beta}/\sigma| \text{ are larger than } a_n\} \\ &\quad \cup \{\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^*\| \geqslant c_0\sqrt{n}\sigma^*\epsilon_n\} \cup \{\sigma^2/\sigma^{*2} > (1+\epsilon_n)/(1-\epsilon_n) \text{ or } \sigma^2/\sigma^{*2} < (1-\epsilon_n)/(1+\epsilon_n)\}, \\ B_n &= \{\text{at least } \tilde{p} \text{ entries of } |\boldsymbol{\beta}/\sigma| \text{ are larger than } a_n\} \end{aligned}$$

and

$$C_n = A_n \backslash B_n,$$

where $\tilde{p} \leqslant \bar{p} - s$ and $\tilde{p} \prec n\epsilon_n^2$.

We still follow the three-step proof as in Theorem A.5. Since the proof is quite similar, the details are omitted here. The only difference is that we now consider a slightly different testing function as

$$\begin{aligned} \phi_n' &= \max_{\{\xi \supseteq \xi^*, |\xi| \leqslant \tilde{p}+s\}} 1\{|\boldsymbol{y}^\top (I - H_\xi)\boldsymbol{y}/(n - |\xi|)\sigma^{*2} - 1| \geqslant \epsilon_n\}, \\ \tilde{\phi}_n &= \max_{\{\xi \supseteq \xi^*, |\xi| \leqslant \tilde{p}+s\}} 1\{\|\boldsymbol{X}_\xi(\boldsymbol{X}_\xi^\top \boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^\top \boldsymbol{y} - \boldsymbol{X}_\xi \boldsymbol{\beta}_\xi^*)\| \geqslant c_0\sigma^*\sqrt{n}\epsilon_n/3\}. \end{aligned}$$

Note that in the proof of Theorem A.5, we need to bound the singular value of $(\boldsymbol{X}_\xi^\top \boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^\top$ via the condition $\mathrm{A}_1(3)$. However, in the proof of Theorem A.6, only the matrix $\boldsymbol{X}_\xi(\boldsymbol{X}_\xi^\top \boldsymbol{X}_\xi)^{-1}\boldsymbol{X}_\xi^\top$ gets involved, and its eigenvalues are always bounded by 1. Thus the condition $\mathrm{A}_1(3)$ is redundant. □

Theorem A.7. *Assume the conditions of Theorem* A.5 *hold, and let* $\xi = \{j : |\beta_j/\sigma| > a_n\}$ *denote a posterior subset model. If the following conditions also hold*:

$$\begin{aligned} &\limsup \sqrt{n}a_n p_n \sigma^*/\sqrt{\log p_n} < k, \\ &\min_{j \in \xi^*} |\beta_j^*| \geqslant M_1\sqrt{\log p_n/n} \textit{ for some large } M_1, \\ &u > 1 + c/2 + k^2/2\sigma^{*2} + 2\sqrt{c'}k, \\ &l_n = \max_{j \in \xi^*} \sup_{\substack{x_1, x_2 \in \beta_j^*/\sigma^* \pm c_0\epsilon_n \\ |x_1|, |x_2| \geqslant a_n}} \frac{g_\lambda(x_1)}{g_\lambda(x_2)} \quad \textit{and} \quad s\log l_n \prec \log p_n \end{aligned}$$

for some constants $c' > 1$, c *and sufficiently large* c_0, *then*

$$P^*\{\pi(\xi = \xi^* \mid \boldsymbol{X}, \boldsymbol{y}) > 1 - o(1)\} > 1 - o(1).$$

Proof. For any $\boldsymbol{\beta}_\xi$ which is a subvector of $\boldsymbol{\beta}$ corresponding to ξ, we define

$$\begin{aligned} \mathrm{SSE}(\boldsymbol{\beta}_\xi) &= \min_{\boldsymbol{\beta}_{\xi^c}} \|\boldsymbol{Y} - \boldsymbol{X}_\xi\boldsymbol{\beta}_\xi - \boldsymbol{X}_{\xi^c}\boldsymbol{\beta}_{\xi^c}\|^2 \\ &= (\boldsymbol{Y} - \boldsymbol{X}_\xi\boldsymbol{\beta}_\xi)^\top (I - \boldsymbol{X}_{\xi^c}(\boldsymbol{X}_{\xi^c}^\top \boldsymbol{X}_{\xi^c})^{-1}\boldsymbol{X}_{\xi^c}^\top)(\boldsymbol{Y} - \boldsymbol{X}_\xi\boldsymbol{\beta}_\xi). \end{aligned}$$

By the consistency result in Theorem A.5, let A_n' be the set

$$\{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leqslant c_1\epsilon_n\} \cap \{|\sigma^2 - \sigma^{*2}| \leqslant c_2\epsilon_n\} \cap \{\text{at most } c_3\sqrt{n\epsilon_n^2/\log p_n} \text{ entries of } \boldsymbol{\beta}/\sigma \text{ are larger than } a_n\}$$

and Ω_n be the event

$$\{\pi(A_n' \mid D_n) > 1 - \exp\{-c_4 n\epsilon_n^2\}\}.$$

Then we have

$$P^*(\Omega_n) > 1 - \mathrm{e}^{-c_5 n\epsilon_n^2}$$

for some c_1 to c_5. All the following analysis is conditioned on the event Ω_n, and we can ignore the set $(A_n')^c$ in all the following posterior probability calculation.

Let

$$E_1 = \{\|\boldsymbol{\beta}_1/\sigma - \boldsymbol{\beta}_1^*/\sigma^*\|_\infty \leqslant c_1\epsilon_n, \|\boldsymbol{\beta}_1/\sigma\|_\infty \geqslant a_n, |\sigma^2 - \sigma^{*2}| \leqslant c_2\epsilon_n\},$$

where $\|\cdot\|_{\min}$ is the smallest absolute value of the entries of a vector. We define

$$\underline{\pi(\boldsymbol{\beta}_1 \mid \sigma^2)} = \inf_{(\boldsymbol{\beta}_1,\sigma^2)\in E_1} \pi(\boldsymbol{\beta}_1,\sigma^2)/\pi(\sigma^2), \quad \overline{\pi(\boldsymbol{\beta}_1 \mid \sigma^2)} = \sup_{(\boldsymbol{\beta}_1,\sigma^2)\in E_1} \pi(\boldsymbol{\beta}_1,\sigma^2)/\pi(\sigma^2).$$

First, we study the posterior probability $\pi(\xi=\xi^* \mid \boldsymbol{X},\boldsymbol{y})$ up to the normalizing constant. For simplicity of notation, we use the subscript "1" to denote the true model ξ^*, and the subscript "2" to denote the rest $(\xi^*)^c$. Then

$$\begin{aligned}
&\int \frac{1}{\sigma^n}\exp\bigg\{-\frac{\|\boldsymbol{y}-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2\|^2}{2\sigma^2}\bigg\}\pi(\boldsymbol{\beta},\sigma)I(\|\boldsymbol{\beta}_2/\sigma\|_\infty\leqslant a_n, \|\boldsymbol{\beta}_1/\sigma\|_{\min}\geqslant a_n)d\sigma^2 d\boldsymbol{\beta}\\
&\quad\geqslant \pi(\|\boldsymbol{\beta}_2/\sigma\|_\infty\leqslant a_n)\int_{E_1}\inf_{\|\boldsymbol{\beta}_2/\sigma\|_\infty\leqslant a_n}\frac{1}{\sigma^n}\exp\bigg\{-\frac{\|\boldsymbol{y}-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2\|^2}{2\sigma^2}\bigg\}\pi(\boldsymbol{\beta}_1,\sigma^2)d\sigma^2 d\boldsymbol{\beta}_1. \qquad \text{(A.11)}
\end{aligned}$$

The integral in the above inequality satisfies

$$\begin{aligned}
&\int_{E_1}\inf_{\|\boldsymbol{\beta}_2/\sigma\|_\infty\leqslant a_n}\frac{1}{\sigma^n}\exp\bigg\{-\frac{\|\boldsymbol{y}-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2\|^2}{2\sigma^2}\bigg\}\pi(\boldsymbol{\beta}_1,\sigma^2)d\sigma^2 d\boldsymbol{\beta}_1\\
&\quad\geqslant \underline{\pi(\boldsymbol{\beta}_1\mid\sigma^2)}\int_{E_1}\inf_{\|\boldsymbol{\beta}_2/\sigma\|_\infty\leqslant a_n}\frac{1}{\sigma^n}\exp\bigg\{-\frac{\|\boldsymbol{y}-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2\|^2}{2\sigma^2}\bigg\}\pi(\sigma^2)d\sigma^2 d\boldsymbol{\beta}_1\\
&\quad= \underline{\pi(\boldsymbol{\beta}_1\mid\sigma^2)}\int_{E_1}\inf_{\|\boldsymbol{\beta}_2/\sigma\|_\infty\leqslant a_n}\frac{1}{\sigma^n}\exp\bigg\{-\frac{\mathrm{SSE}(\boldsymbol{\beta}_2)+(\boldsymbol{\beta}_1-\hat{\boldsymbol{\beta}}_1)^\top\boldsymbol{X}_1^\top\boldsymbol{X}_1(\boldsymbol{\beta}_1-\hat{\boldsymbol{\beta}}_1)}{2\sigma^2}\bigg\}\pi(\sigma^2)d\boldsymbol{\beta}_1 d\sigma^2\\
&\quad\approx \underline{\pi(\boldsymbol{\beta}_1\mid\sigma^2)}\int_{E_1}\inf_{\|\boldsymbol{\beta}_2/\sigma\|_\infty\leqslant a_n}\frac{1}{\sigma^n}\exp\bigg\{-\frac{\mathrm{SSE}(\boldsymbol{\beta}_2)}{2\sigma^2}\bigg\}\pi(\sigma^2)(2\pi)^{s/2}\sqrt{|\sigma^2(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}|})d\sigma^2\\
&\quad\approx \underline{\pi(\boldsymbol{\beta}_1\mid\sigma^2)}\sqrt{|(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}|}\sqrt{2\pi}^s\inf_{\|\boldsymbol{\beta}_2\|_\infty\leqslant a_n(\sigma^*+c_2\epsilon_n)}\frac{\Gamma(a_0+(n-s)/2)}{(\mathrm{SSE}(\boldsymbol{\beta}_2)/2+b_0)^{a_0+(n-s)/2}}, \qquad \text{(A.12)}
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_1=(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top(\boldsymbol{y}-\boldsymbol{X}_2\boldsymbol{\beta}_2)$. The first approximation holds because most probability mass of the normal density is in the region of $\{\|\boldsymbol{\beta}_1-\hat{\boldsymbol{\beta}}_1\|\leqslant C\sqrt{s/n}\}$, which is a subset of E_1 in probability, if c_1 is large. Similarly, the second approximation holds since the distribution $\mathrm{IG}(a_0+(n-s)/2, \mathrm{SSE}(\boldsymbol{\beta}_2)/2+b_0)$ puts most of its probability mass inside the region $\{|\sigma^2-\sigma^{*2}|\leqslant c_2\epsilon_n\}$.

Next, we study the posterior probability $\pi(\xi=\xi' \mid \boldsymbol{X},\boldsymbol{y})$ for any $\xi'\supset\xi^*$ up to the normalizing constant. Similarly, we use the subscript "1" to denote the true model ξ^*, the subscript "2" to denote $(\xi'\backslash\xi^*)$, and the subscript "3" to denote the rest $(\xi')^c$. It holds that

$$\begin{aligned}
&\int \frac{1}{\sigma^n}\exp\bigg\{-\frac{\|\boldsymbol{y}-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2-\boldsymbol{X}_3\boldsymbol{\beta}_3\|^2}{2\sigma^2}\bigg\}\pi(\boldsymbol{\beta},\sigma)I(\|\boldsymbol{\beta}_2/\sigma\|_{\min}>a_n, \|\boldsymbol{\beta}_3/\sigma\|_\infty\leqslant a_n)d\sigma^2 d\boldsymbol{\beta}\\
&\quad\lesssim \pi(\|\boldsymbol{\beta}_2/\sigma\|_{\min}>a_n, \|\boldsymbol{\beta}_3/\sigma\|_\infty\leqslant a_n)\\
&\qquad\times\sup_{\|\boldsymbol{\beta}_3/\sigma\|_\infty\leqslant a_n,\boldsymbol{\beta}_2}\int_{E_1}\frac{1}{\sigma^n}\exp\bigg\{-\frac{\|\boldsymbol{y}-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2-\boldsymbol{X}_3\boldsymbol{\beta}_3\|^2}{2\sigma^2}\bigg\}\pi(\boldsymbol{\beta}_1,\sigma)d\sigma^2 d\boldsymbol{\beta}_1\\
&\quad\leqslant \pi(\|\boldsymbol{\beta}_2/\sigma\|_{\min}>a_n, \|\boldsymbol{\beta}_3/\sigma\|_\infty\leqslant a_n)\\
&\qquad\times\sup_{\|\boldsymbol{\beta}_3/\sigma\|_\infty\leqslant a_n,\boldsymbol{\beta}_2}\int_{E_1}\frac{1}{\sigma^n}\exp\bigg\{-\frac{\mathrm{SSE}((\boldsymbol{\beta}_2,\boldsymbol{\beta}_3)^\top)+(\boldsymbol{\beta}_1-\tilde{\boldsymbol{\beta}}_1)^\top\boldsymbol{X}_1^\top\boldsymbol{X}_1(\boldsymbol{\beta}_1-\tilde{\boldsymbol{\beta}}_1)}{2\sigma^2}\bigg\}\pi(\boldsymbol{\beta}_1,\sigma)d\sigma^2 d\boldsymbol{\beta}_1\\
&\quad\leqslant \pi(\|\boldsymbol{\beta}_2/\sigma\|_{\min}>a_n, \|\boldsymbol{\beta}_3/\sigma\|_\infty\leqslant a_n)\overline{\pi(\boldsymbol{\beta}_1\mid\sigma^2)}\sqrt{|(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}|}(2\pi)^{s/2}\\
&\qquad\times\sup_{\|\boldsymbol{\beta}_3\|_\infty\leqslant a_n(\sigma^*+c_2\epsilon_n)}\frac{\Gamma(a_0+(n-s)/2)}{(\mathrm{SSE}(\boldsymbol{\beta}_3)/2+b_0)^{a_0+(n-s)/2}}, \qquad \text{(A.13)}
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_1=(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top(\boldsymbol{y}-\boldsymbol{X}_2\boldsymbol{\beta}_2-\boldsymbol{X}_3\boldsymbol{\beta}_3)$.

Therefore, combining the above results, we obtain that for any $\xi'\supset\xi^*$,

$$\frac{\pi(\xi=\xi'\mid\boldsymbol{X},\boldsymbol{y})}{\pi(\xi=\xi^*\mid\boldsymbol{X},\boldsymbol{y})}\lesssim\frac{\overline{\pi(\boldsymbol{\beta}_1\mid\sigma^2)}}{\underline{\pi(\boldsymbol{\beta}_1\mid\sigma^2)}}[p_n^{-(1+u)}/(1-p_n^{-(1+u)})]^{|\xi'\backslash\xi^*|}$$

$$\times \frac{\sup_{\|\boldsymbol{\beta}_{(\xi^*)^c}\|_\infty \leqslant a_n(\sigma^*+c_2\epsilon_n)} (\mathrm{SSE}(\boldsymbol{\beta}_{(\xi^*)^c})/2+b_0)^{a_0+(n-s)/2}}{\inf_{\|\boldsymbol{\beta}_{(\xi')^c}\|_\infty \leqslant a_n(\sigma^*+c_2\epsilon_n)} (\mathrm{SSE}(\boldsymbol{\beta}_{(\xi')^c})/2+b_0)^{a_0+(n-s)/2}}. \tag{A.14}$$

It is easy to see that with probability larger than $1-4p_n \cdot p_n^{-c_6'}$,

$$\|\boldsymbol{X}^\top A\boldsymbol{\varepsilon}\|_\infty \leqslant \sqrt{2c_6' n\log p_n}$$

for any idempotent matrix A and $c_6'>1$, and thus,

$$\begin{aligned} \mathrm{SSE}(\boldsymbol{\beta}_{(\xi^*)^c}) &= (\boldsymbol{y}-\boldsymbol{X}_{(\xi^*)^c}\boldsymbol{\beta}_{(\xi^*)^c})^\top (I-P_{\boldsymbol{X}_{\xi^*}})(\boldsymbol{y}-\boldsymbol{X}_{(\xi^*)^c}\boldsymbol{\beta}_{(\xi^*)^c}) \\ &\leqslant \sigma^{*2}\boldsymbol{\varepsilon}^\top(I-P_{\boldsymbol{X}_{\xi^*}})\boldsymbol{\varepsilon} + \|\boldsymbol{X}_{(\xi^*)^c}\boldsymbol{\beta}_{(\xi^*)^c}\|^2 - 2\sigma^*\boldsymbol{\varepsilon}^\top(I-P_{\boldsymbol{X}_{\xi^*}})\boldsymbol{X}_{(\xi^*)^c}\boldsymbol{\beta}_{(\xi^*)^c} \\ &\leqslant \sigma^{*2}\boldsymbol{\varepsilon}^\top(I-P_{\boldsymbol{X}_{\xi^*}})\boldsymbol{\varepsilon} + \|\boldsymbol{X}_{(\xi^*)^c}\boldsymbol{\beta}_{(\xi^*)^c}\|^2 + 2\sigma^*\sqrt{2c_6' n\log p_n}\|\boldsymbol{\beta}_{(\xi^*)^c}\|_1, \\ \mathrm{SSE}(\boldsymbol{\beta}_{(\xi')^c}) &= (\sigma^*\boldsymbol{\varepsilon}-\boldsymbol{X}_{(\xi')^c}\boldsymbol{\beta}_{(\xi')^c})^\top (I-P_{\boldsymbol{X}_{\xi'}})(\sigma^*\boldsymbol{\varepsilon}-\boldsymbol{X}_{(\xi')^c}\boldsymbol{\beta}_{(\xi')^c}) \\ &\geqslant \sigma^{*2}\boldsymbol{\varepsilon}^\top(I-P_{\boldsymbol{X}_{\xi'}})\boldsymbol{\varepsilon} - 2(\sigma^*\boldsymbol{\varepsilon})^\top(I-P_{\boldsymbol{X}_{\xi'}})\boldsymbol{X}_{(\xi')^c}\boldsymbol{\beta}_{(\xi')^c} \\ &\geqslant \sigma^{*2}\boldsymbol{\varepsilon}^\top(I-P_{\boldsymbol{X}_{\xi'}})\boldsymbol{\varepsilon} - 2\sigma^*\sqrt{2c_6' n\log p_n}\|\boldsymbol{\beta}_{(\xi')^c}\|_1. \end{aligned} \tag{A.15}$$

Let $\tilde{p}_n \triangleq c_3\sqrt{n\epsilon_n^2/\log p_n}$. By the properties of the quantiles of the chi-square distribution (see, e.g., [33]) and Lemma A.2, with probability larger than $1-p_n^{-c_6}$ for any constant c_6,

$$\begin{aligned} &\boldsymbol{\varepsilon}^\top(I-P_{\boldsymbol{X}_{\xi^*}})\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^\top(I-P_{\boldsymbol{X}_{\xi'}})\boldsymbol{\varepsilon} \leqslant \sigma^{*2}\{c_7|\xi'\backslash\xi^*|\log p_n\}, \\ &\boldsymbol{\varepsilon}^\top(I-P_{\boldsymbol{X}_{\xi^*}})\boldsymbol{\varepsilon} \in \sigma^{*2}(n-s)[1-c_8, 1+c_8] \end{aligned} \tag{A.16}$$

hold for all ξ' with $1<|\xi'\backslash\xi^*| \leqslant \tilde{p}_n$, when n is sufficiently large, any $c_7>c_6+2$ and $c_8>0$.

Combining (A.15), (A.16) and the fact that $\sqrt{n}a_np_n\sigma^* < k\sqrt{\log p_n}$ for large n, we have

$$\frac{\sup_{\|\boldsymbol{\beta}_{(\xi^*)^c}\|_\infty \leqslant a_n(\sigma^*+c_2\epsilon_n)} (\mathrm{SSE}(\boldsymbol{\beta}_{(\xi^*)^c})/2+b_0)^{a_0+(n-s)/2}}{\inf_{\|\boldsymbol{\beta}_{(\xi')^c}\|_\infty \leqslant a_n(\sigma^*+c_2\epsilon_n)} (\mathrm{SSE}(\boldsymbol{\beta}_{(\xi')^c})/2+b_0)^{a_0+(n-s)/2}} \leqslant \exp\{c_9|\xi'\backslash\xi^*|\log p_n\}, \tag{A.17}$$

where c_9 is any constant satisfying

$$c_9 > c_7/2(1-c_8) + (k^2/2\sigma^{*2} + 2\sqrt{2c_6'}k)/(1-c_8).$$

Furthermore, it is easy to see that

$$\overline{\pi(\boldsymbol{\beta}_1 \mid \sigma^2)}/\underline{\pi(\boldsymbol{\beta}_1 \mid \sigma^2)} \leqslant l_n^s.$$

Combining it with (A.14) and (A.17), we obtain

$$\frac{\pi(\xi=\xi' \mid \boldsymbol{X},\boldsymbol{y})}{\pi(\xi=\xi^* \mid \boldsymbol{X},\boldsymbol{y})} \leqslant l_n^s[p_n^{-(1+u)}/(1-p_n^{-(1+u)})]^{|\xi'\backslash\xi^*|}\exp\{c_9|\xi'\backslash\xi^*|\log p_n\}. \tag{A.18}$$

By the condition

$$1+u > 2+c_6/2+k^2/2\sigma^{*2}+2\sqrt{2c_6'}k,$$

we can choose proper values of c_7, c_8 and c_9 such that $1+u-c_9=u'>1$ and

$$\frac{\pi(\xi=\xi' \mid \boldsymbol{X},\boldsymbol{y})}{\pi(\xi=\xi^* \mid \boldsymbol{X},\boldsymbol{y})} \leqslant l_n^s p_n^{-u'|\xi'\backslash\xi^*|} \quad \text{for any } \xi' \supset \xi^*.$$

Therefore, since $u'>1$, we have

$$\sum_{\xi'\supset\xi^*, 1<|\xi'\backslash\xi^*|\leqslant\tilde{p}_n} \frac{\pi(\xi=\xi' \mid \boldsymbol{X},\boldsymbol{y})}{\pi(\xi=\xi^* \mid \boldsymbol{X},\boldsymbol{y})} \leqslant l_n^s[(1+p_n^{-u'})^{p_n}-1] \asymp l_n^s p_n^{-(u'-1)}. \tag{A.19}$$

Finally, we study the posterior probability $\pi(\xi=\xi' \mid \boldsymbol{X},\boldsymbol{y})$ for any ξ' such that $|\xi'\backslash\xi^*| \leqslant \tilde{p}_n$ and ξ' does not include ξ^*, up to the normalizing constant. Similarly, we use the subscript "4" to denote the model $(\xi^*\cap\xi')$, the subscript "5" to denote $(\xi^*\backslash\xi')$, the subscript "2" to denote $(\xi'\backslash\xi^*)$, and the subscript "3" to denote the rest $(\xi'\cup\xi^*)^c$. Define

$$E_4=\{\|\boldsymbol{\beta}_4/\sigma-\boldsymbol{\beta}_4^*/\sigma^*\|_\infty\leqslant c_1\epsilon_n,\|\boldsymbol{\beta}_4/\sigma\|_\infty\geqslant a_n,|\sigma^2-\sigma^{*2}|\leqslant c_2\epsilon_n\}$$

and

$$\underline{\pi}=\inf_{x\in[-E_n,E_n]}g_\lambda(x).$$

Then

$$\begin{aligned}&\frac{\pi(\xi=\xi'\mid\boldsymbol{X},\boldsymbol{y})}{\pi(\xi=\xi'\cup\xi^*\mid\boldsymbol{X},\boldsymbol{y})}\\&\quad=\frac{\int_{|\sigma^2-\sigma^{*2}|\leqslant c_2\epsilon_n}\frac{1}{\sigma^n}\exp\{-\frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2}{2\sigma^2}\}\pi(\boldsymbol{\beta},\sigma)I(\|(\boldsymbol{\beta}_2,\boldsymbol{\beta}_4)/\sigma\|_{\min}>a_n,\|(\boldsymbol{\beta}_3,\boldsymbol{\beta}_5)/\sigma\|_\infty\leqslant a_n)d\sigma^2d\boldsymbol{\beta}}{\int_{|\sigma^2-\sigma^{*2}|\leqslant c_2\epsilon_n}\frac{1}{\sigma^n}\exp\{-\frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2}{2\sigma^2}\}\pi(\boldsymbol{\beta},\sigma)I(\|(\boldsymbol{\beta}_2,\boldsymbol{\beta}_4,\boldsymbol{\beta}_5)/\sigma\|_{\min}>a_n,\|\boldsymbol{\beta}_3/\sigma\|_\infty\leqslant a_n)d\sigma^2d\boldsymbol{\beta}}\\&\quad\lesssim\max_{\{\boldsymbol{\beta}_2,\boldsymbol{\beta}_3,\|\boldsymbol{\beta}_4\|_{\min}\geqslant a_n,|\sigma^2-\sigma^{*2}|\leqslant c_2\epsilon_n\}}\frac{p_n^{(1+u)|\xi^*\backslash\xi'|}}{\sqrt{2\pi}\underline{\pi}^{|\xi^*\backslash\xi'|}\sqrt{|\sigma^2(\boldsymbol{X}_5^\top\boldsymbol{X}_5)^{-1}|}}\\&\qquad\times\frac{\max_{\|\boldsymbol{\beta}_5/\sigma\|_\infty\leqslant a_n}\exp\{-\frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2}{2\sigma^2}\}}{\max_{\|\boldsymbol{\beta}_5/\sigma\|_{\min}\geqslant a_n}\exp\{-\frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2}{2\sigma^2}\}}.\end{aligned}\tag{A.20}$$

It is not difficult to see that in probability, uniformly for all ξ', $\boldsymbol{\beta}_2,\boldsymbol{\beta}_3$, $\|\boldsymbol{\beta}_4\|_{\min}\geqslant a_n$ and $|\sigma^2-\sigma^{*2}|\leqslant c_2\epsilon_n$, we have

$$\begin{aligned}&\max_{\|\boldsymbol{\beta}_5/\sigma\|_\infty\leqslant a_n}\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2-\max_{\|\boldsymbol{\beta}_5/\sigma\|_{\min}\geqslant a_n}\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2\\&\quad\geqslant\max_{\|\boldsymbol{\beta}_5/\sigma\|_\infty\leqslant a_n}\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2-\|\boldsymbol{y}-\boldsymbol{X}_5\boldsymbol{\beta}_5^*-\boldsymbol{X}_2\boldsymbol{\beta}_2-\boldsymbol{X}_3\boldsymbol{\beta}_3-\boldsymbol{X}_4\boldsymbol{\beta}_4\|^2\geqslant M'|\xi^*\backslash\xi'|\log p_n\end{aligned}$$

for some M' if M_1 (which appeared in the beta-min condition) is sufficiently large, and the condition $\mathrm{A}_1(3)$ holds.

Given sufficiently large M', uniformly for all ξ', (A.20) reduces to

$$\frac{\pi(\xi=\xi'\mid\boldsymbol{X},\boldsymbol{y})}{\pi(\xi=\xi'\cup\xi^*\mid\boldsymbol{X},\boldsymbol{y})}\leqslant p_n^{-M''|\xi^*\backslash\xi'|}$$

for some $M''>1$. This further implies that

$$\frac{\pi(\xi\text{ does not includes }\xi^*,|\xi\backslash\xi^*|\leqslant\tilde{p}_n\mid\boldsymbol{X},\boldsymbol{y})}{\pi(\xi\supset\xi^*,|\xi\backslash\xi^*|\leqslant\tilde{p}_n\mid\boldsymbol{X},\boldsymbol{y})}\leqslant(1+p_n^{-M''})^s-1=o(1).\tag{A.21}$$

Combining (A.19) and (A.21), we conclude that with probability $1-o(1)$, $\pi(\xi=\xi^*\mid\boldsymbol{X},\boldsymbol{y})>1-o(1)$. □

Theorem A.8 (BvM theorem). *Under the conditions of Theorem A.7, $a_n\prec(1/p_n)\sqrt{1/(ns\log p_n)}$ and $\lim_{n\to\infty}s\log l_n=0$, we have*

$$\left\|\pi(\boldsymbol{\beta},\sigma^2\mid\boldsymbol{X},\boldsymbol{y})-\phi(\boldsymbol{\beta}_{\xi^*};\hat{\boldsymbol{\beta}}_{\xi^*},\sigma^2(\boldsymbol{X}_{\xi^*}^\top\boldsymbol{X}_{\xi^*})^{-1})\prod_{j\notin\xi^*}\pi(\beta_j\mid\sigma^2)\mathrm{ig}(\sigma^2,(n-s)/2,\hat{\sigma}^2(n-s)/2)\right\|_{\mathrm{TV}}\to0$$

in probability, where ϕ denotes the density function of a multivariate normal distribution, ig *denotes the density function of an inverse gamma distribution, and $\hat{\boldsymbol{\beta}}_{\xi^*}$ and $\hat{\sigma}^2$ are the MLEs of $\boldsymbol{\beta}_{\xi^*}$ and σ^2, respectively, given data $(\boldsymbol{y},\boldsymbol{X}_{\xi^*})$.*

Proof. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_{\xi^*}, \sigma^2)^\top$, $\boldsymbol{\theta}' = \boldsymbol{\beta}_{(\xi^*)^c}$ and let $\pi_0(\boldsymbol{\theta})$ denote the normal-inverse gamma distribution $\phi(\boldsymbol{\beta}_{\xi^*}; \hat{\boldsymbol{\beta}}_{\xi^*}, \sigma^2(\boldsymbol{X}_{\xi^*}^\top \boldsymbol{X}_{\xi^*})^{-1})\mathrm{ig}(\sigma^2, (n-s)/2, \hat{\sigma}^2(n-s)/2)$, and

$$\begin{aligned}\pi_1(\boldsymbol{\theta}) &= C\frac{1}{\sigma^n}\exp\bigg\{-\frac{\|\boldsymbol{y}-\boldsymbol{X}_{\xi^*}\boldsymbol{\beta}_{\xi^*}\|^2}{2\sigma^2}\bigg\}\pi(\sigma^2),\\ \pi_2(\boldsymbol{\theta}) &= \prod_{j\in\xi^*}\frac{\pi(\beta_j \mid \sigma^2)}{\pi(\beta_j^* \mid \sigma^{*2})},\\ \pi_3(\boldsymbol{\theta},\boldsymbol{\theta}') &= \exp\bigg\{-\frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 - \|\boldsymbol{y}-\boldsymbol{X}_{\xi^*}\boldsymbol{\beta}_{\xi^*}\|^2}{2\sigma^2}\bigg\}\prod_{j\notin\xi^*}\pi(\beta_j \mid \sigma^2),\end{aligned}$$

where C normalizes π_1. Thus, we have the posterior $\pi(\boldsymbol{\beta}, \sigma^2|\boldsymbol{X}, \boldsymbol{y}) \propto \pi_1\pi_2\pi_3$.

It is trivial to see that π_1 is exactly a normal-inverse gamma distribution, i.e.,

$$\sigma^2 \sim \mathrm{IG}((n-s)/2 + a_0, \hat{\sigma}^2(n-s)/2 + b_0),$$

and the conditional distribution of $\boldsymbol{\beta}_{\xi^*}$ follows

$$\boldsymbol{\beta}_{\xi^*}|\sigma^2 \sim N(\hat{\boldsymbol{\beta}}_{\xi^*}, \sigma^2(\boldsymbol{X}_{\xi^*}^\top\boldsymbol{X}_{\xi^*})^{-1}),$$

where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_{\xi^*}, \hat{\sigma}^2)$. Furthermore, as long as $n - s \to \infty$, it is not difficult to show that

$$\|\mathrm{IG}((n-s)/2, \hat{\sigma}^2(n-s)/2) - \mathrm{IG}((n-s)/2 + a_0, \hat{\sigma}^2(n-s)/2 + b_0)\|_{\mathrm{TV}} \to 0$$

with dominating probability, i.e., $\|\pi_1(\boldsymbol{\theta}) - \pi_0(\boldsymbol{\theta})\|_{\mathrm{TV}} = o_p(1)$.

Let $\Omega_1 = \{\|\boldsymbol{\beta}_{\xi^*} - \boldsymbol{\beta}_{\xi^*}^*\| \leqslant \sigma^*\epsilon_n$ and $|\sigma^2 - \sigma^{*2}| < c_4\epsilon_n\}$. By the conditions of the theorem, if $\boldsymbol{\theta} \in \Omega_1$, then $|\pi_2 - 1| \leqslant |l_n^s - 1| \to 0$. Therefore,

$$\begin{aligned}\int_{\Omega_1}|\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta}) - \pi_0(\boldsymbol{\theta})|d\boldsymbol{\theta} &\leqslant \int_{\Omega_1}|\pi_1\pi_2 - \pi_1|d\boldsymbol{\theta} + \int_{\Omega_1}|\pi_1(\boldsymbol{\theta}) - \pi_0(\boldsymbol{\theta})|d\boldsymbol{\theta}\\ &\leqslant \max_{\Omega_1}|\pi_2(\boldsymbol{\theta}) - 1| + \int_{\Omega_1}|\pi_1(\boldsymbol{\theta}) - \pi_0(\boldsymbol{\theta})|d\boldsymbol{\theta} = o_p(1).\end{aligned}$$

Let $\boldsymbol{\varepsilon}(\boldsymbol{\beta}_{\xi^*}) = \boldsymbol{y} - \boldsymbol{X}_{\xi^*}\boldsymbol{\beta}_{\xi^*}$ and

$$\Omega_2 = \{(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Omega_1, \|\beta_j/\sigma\| \leqslant a_n, \forall\, j \notin \xi^*\}.$$

For any $(\boldsymbol{\theta}, \boldsymbol{\theta}')^\top \in \Omega_2$,

$$\|\boldsymbol{\varepsilon}(\boldsymbol{\beta}_{\xi^*})\| \in [\|\sigma^*\boldsymbol{\varepsilon}\| \pm \sigma^*\sqrt{|\xi^*|}\sqrt{n}\epsilon_n]$$

and

$$\begin{aligned}|\|\boldsymbol{\varepsilon}(\boldsymbol{\beta}_\xi^*)\|^2 - \|\boldsymbol{\varepsilon}(\boldsymbol{\beta}_\xi^*) - \boldsymbol{X}_{\xi^{*c}}\boldsymbol{\beta}_{\xi^{*c}}\|^2| &\leqslant \|\boldsymbol{X}_{\xi^{*c}}\boldsymbol{\beta}_{\xi^{*c}}\|^2 + 2\boldsymbol{\varepsilon}(\boldsymbol{\beta}_\xi^*)^\top\boldsymbol{X}_{\xi^{*c}}\boldsymbol{\beta}_{\xi^{*c}}\\ &\leqslant na_n^2p_n^2 + 2(\boldsymbol{\varepsilon} + \boldsymbol{X}_{\xi^*}(\boldsymbol{\beta}_{\xi^*}^* - \boldsymbol{\beta}_{\xi^*}))^\top\boldsymbol{X}_{\xi^{*c}}\boldsymbol{\beta}_{\xi^{*c}}\\ &\leqslant na_n^2p_n^2 + O(\sqrt{n\log p_n}a_np_n) + O(\sqrt{n}\epsilon_n\sqrt{n}a_np_n)\end{aligned}$$

in probability. Since $na_np_n \prec 1/\epsilon_n$, we have

$$|\|\boldsymbol{\varepsilon}(\boldsymbol{\beta}_\xi^*)\|^2 - \|\boldsymbol{\varepsilon}(\boldsymbol{\beta}_\xi^*) - \boldsymbol{X}_{\xi^{*c}}\boldsymbol{\beta}_{\xi^{*c}}\|^2| = o_p(1).$$

Therefore,

$$\begin{aligned}&\int_{\Omega_2}\bigg[\pi_3(\boldsymbol{\theta},\boldsymbol{\theta}') - \prod_{j\notin\xi^*}\pi(\beta_j \mid \sigma^2)\bigg]d\boldsymbol{\theta}'\\ &\leqslant \int_{\Omega_2}\bigg|\exp\bigg\{-\frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 - \|\boldsymbol{y}-\boldsymbol{X}_{\xi^*}\boldsymbol{\beta}_{\xi^*}\|^2}{2\sigma^2}\bigg\} - 1\bigg|\prod_{j\notin\xi^*}\pi(\beta_j \mid \sigma^2)d\boldsymbol{\theta}'\\ &\leqslant |\exp[o_p(1)/(2\sigma^{*2} - c_4\epsilon_n)] - 1|\int_{\Omega_3}\prod_{j\notin\xi^*}\pi(\beta_j \mid \sigma)d\boldsymbol{\theta}' = o_p(1).\end{aligned}$$

Combining the above inequalities, we have

$$\begin{aligned}
&\int_{\Omega_2}\left|\pi_1\pi_2\pi_3(\boldsymbol{\theta},\boldsymbol{\theta}')-\pi_0(\boldsymbol{\theta})\prod_{j\notin\xi^*}\pi(\beta_j\mid\sigma^2)\right|d\boldsymbol{\theta}'d\boldsymbol{\theta}\\
&\leqslant\int_{\Omega_2}\pi_1\pi_2(\boldsymbol{\theta})\left|\pi_3(\boldsymbol{\theta},\boldsymbol{\theta}')-\prod_{j\notin\xi^*}\pi(\beta_j\mid\sigma^2)\right|d\boldsymbol{\theta}'d\boldsymbol{\theta}+\int_{\Omega_2}|\pi_1\pi_2(\boldsymbol{\theta})-\pi_0(\boldsymbol{\theta})|\prod_{j\notin\xi^*}\pi(\beta_j\mid\sigma^2)d\boldsymbol{\theta}'d\boldsymbol{\theta}\\
&\leqslant o_p(1)\int_{\Omega_1}\pi_1\pi_2(\boldsymbol{\theta})d\boldsymbol{\theta}+\int_{\Omega_1}|\pi_1\pi_2(\boldsymbol{\theta})-\pi_0(\boldsymbol{\theta})|d\boldsymbol{\theta}=o_p(1).
\end{aligned}$$

By Theorems A.5 and A.7, with high probability,

$$\int_{\Omega_2^c}\pi(\boldsymbol{\theta},\boldsymbol{\theta}'\mid D_n)\to 0.$$

Also it is not difficult to verify that

$$\int_{\Omega_2^c}\pi_0(\boldsymbol{\theta})\prod_{j\notin\xi^*}\pi(\beta_j\mid\sigma^2)d\boldsymbol{\theta}'d\boldsymbol{\theta}=o_p(1).$$

Therefore, we conclude that

$$\int\left|\pi(\boldsymbol{\theta},\boldsymbol{\theta}'\mid D_n)-\pi_0(\boldsymbol{\theta})\prod_{j\notin\xi^*}\pi(\beta_j\mid\sigma^2)\right|d\boldsymbol{\theta}'d\boldsymbol{\theta}=o_p(1).$$

This completes the proof. □

Proof of Theorem 3.1. It is sufficient to show that $g(\beta_i/\lambda_n)/\lambda_n$ satisfies the condition (A.1). Assume that $\underline{c}x^{-r}<g(x)<\bar{c}x^{-r}$ for sufficiently large x. Then

$$\int_{a_n}^{\infty}g(x/\lambda_n)/\lambda_n dx=\int_{a_n/\lambda_n}^{\infty}g(x)dx\leqslant\bar{c}\frac{1}{r-1}\{a_n/\lambda_n\}^{-(r-1)}.$$

Given $\lambda_n\leqslant a_np_n^{-(u+1)/(r-1)}$ for some $u>0$,

$$\bar{c}\frac{1}{r-1}\{a_n/\lambda_n\}^{-(r-1)}\leqslant c\frac{1}{r-1}p_n^{-1-u}\prec\frac{1}{2}p_n^{-1-u'},$$

where $0<u'<u$. Hence,

$$1-\int_{-a_n}^{a_n}g(x/\lambda_n)/\lambda_n dx\leqslant p_n^{-(1+u')},$$

i.e., the first inequality of (A.1) holds. It holds that

$$\begin{aligned}
-\log\Big(\inf_{x\in[-E_n,E_n]}g(x/\lambda_n)/\lambda_n\Big)&=-\log\Big(\inf_{x\in[-E_n/\lambda,E_n/\lambda_n]}g(x)/\lambda_n\Big)\\
&\leqslant-\log(\underline{c}(E_n/\lambda_n)^{-r}/\lambda_n)=-\log\underline{c}+(r+1)\log(1/\lambda_n)+r\log(E_n).
\end{aligned}$$

Given that $\log(E_n)\asymp\log p_n$ and $-\log\lambda_n=O(\log p_n)$, the second inequality of (A.1) holds. □

Proof of Theorem 3.4. We first verify the condition (A.1). Let $g_\lambda(x)=m_0\phi(x;0,\sigma_0^2)+m_1\phi(x;0,\sigma_1^2)$. Then

$$\begin{aligned}
1-\int_{-a_n}^{a_n}g_\lambda(x)dx&=2[m_0(1-\Phi(a_n/\sigma_0))+m_1(1-\Phi(a_n/\sigma_1))]\\
&\leqslant m_1+2m_0(1-\Phi(a_n/\sigma_0))\leqslant m_1+\frac{\sqrt{2}}{a_n\sqrt{\pi}/\sigma_0}\exp\{-a_n^2/2\sigma_0^2\}\leqslant 1/p_n^{1+u'}
\end{aligned}$$

for some $0 < u' \leqslant u$. By the conditions, we also have

$$\begin{aligned}-\log\Big(\inf_{x\in[-E_n,E_n]} g_\lambda(x)\Big) &\leqslant -\log\Big(m_1 \inf_{x\in[-E_n,E_n]} \phi(x/\sigma_1)\Big)\\ &= C + (1+u)\log p_n + E_n^2/(2\sigma_1^2) + \log\sigma_1 \asymp \log p_n.\end{aligned}$$

Next, we study the flatness of l_n. When $E \geqslant x \geqslant a_n$,

$$\begin{aligned}\frac{(1-m_1)\sigma_1\exp\{-x^2/2\sigma_0^2\}}{m_1\sigma_0\exp\{-x^2/2\sigma_1^2\}} &= \frac{(1-m_1)}{m_1}\exp\bigg\{-\frac{x^2}{2\sigma_0^2} - \log\sigma_0 + \frac{x^2}{2\sigma_1^2} + \log\sigma_1\bigg\}\\ &\leqslant \frac{(1-m_1)}{m_1}\exp\bigg\{-\frac{a_n^2}{2\sigma_0^2} - \log\sigma_0 + \frac{E_n^2}{2\sigma_1^2} + \log\sigma_1\bigg\} \to 0.\end{aligned}$$

Note that the above convergence result holds since $E_n^2/\sigma_1^2 + \log\sigma_1 \asymp \log p_n$ and $\sigma_0 = O(a_n/\log p_n)$. Hence,

$$\frac{g_\lambda(x)}{m_1\phi(x;0,\sigma_1^2)} = 1 + \frac{1-m_1}{m_1}\frac{\sigma_1\exp\{-x^2/2\sigma_0^2\}}{\sigma_0\exp\{-x^2/2\sigma_1^2\}} \to 1.$$

Therefore, we have

$$\begin{aligned}l_n &= \max_{j\in\xi^*} \sup_{\substack{x_1,x_2\in\beta_j^*/\sigma^*\pm c_0\epsilon_n\\ |x_1|,|x_2|\geqslant a_n}} \frac{g_\lambda(x_1)}{g_\lambda(x_2)}\\ &\asymp \max_{j\in\xi^*} \sup_{\substack{x_1,x_2\in\beta_j^*/\sigma^*\pm c_0\epsilon_n\\ |x_1|,|x_2|\geqslant a_n}} \phi(x_1/\sigma_1)/\phi(x_2/\sigma_1)\\ &\leqslant \max_{j\in\xi^*} \sup_{x_1,x_2\in\beta_j^*/\sigma^*\pm c'\epsilon_n} \exp\{(x_1^2-x_2^2)/2\sigma_1^2\}\\ &= \max_{j\in\xi^*}\exp\{2(\beta_j^*+c'\epsilon_n)c'\epsilon_n/\sigma_1^2\},\end{aligned}$$

which implies $s\log l_n \leqslant O(sE_n\epsilon_n)/\sigma_1^2$. The proof can be concluded by applying Theorems A.5, A.7 and A.8. □