

Supplementary Materials

A Proofs of Main Results

Proposition 2.1. Consider (4) with $T(\cdot)$ being ReLU and a hyper-prior on α_0 ,

$$\pi(\alpha_0) \propto \Phi(-\alpha_0)^{a_0-1} (1 - \Phi(-\alpha_0))^{b_0-1} \phi(\alpha_0),$$

where ϕ and Φ are the pdf and cdf of $N(0, 1)$, respectively. Then, the resulting prior distribution is identical to the form of (3) with the Beta prior (8) on η .

Proof of Proposition 2.1. It is clear that $P(\theta_j \neq 0 \mid \alpha_0) = P(\alpha_j > \alpha_0 \mid \alpha_0) = \Phi(-\alpha_0)$. The term $\Phi(-\alpha_0)$ in the neuronized prior controls the sparsity level, and it corresponds to the hyper-parameter η in (8) for the standard SpSL priors. Then, after applying a change of variable as $\Phi(-\alpha_0) = \eta$, where $\eta \sim \text{Beta}(a_0, b_0)$, we obtain the transformed density function of α_0 as $\Phi(-\alpha_0)^{a_0-1} (1 - \Phi(-\alpha_0))^{b_0-1} \phi(\alpha_0)$. \square

Lemma 2.2. With the activation function $T(t) = t$, the marginal density of θ resulting from the neuronized prior is proportional to $\int_0^\infty z^{-1} \exp\{-\theta^2/(2\tau_w^2 z^2) - z^2/2\} dz$.

Proof of Lemma 2.2. Let $\theta = \alpha w$ and $z = w$. With a change of variable, we obtain the Jacobian term is z^{-1} . A simple plug-in of $\alpha = \theta/z$ and $w = z$ completes the proof. \square

Proposition 2.3. Let π_L be the marginal density function of θ defined in (4) with $T(t) = t$ and $\alpha_0 = 0$. Then, $\forall \epsilon \in (0, 1)$, $\exists \theta_0$ and constants $c_1, c_2 > 0$, such that $c_1 \exp\{-(1 + \epsilon)^{1/2} |\theta|/\tau_w\} \leq \pi_L(\theta) \leq c_2 \exp\{-(1 - \epsilon)^{1/2} |\theta|/\tau_w\}$ when $\theta > \theta_0$.

Proof of Proposition 2.3. We first show that the lower bound holds. By the change of variable $u = z^2$, for any $0 < \epsilon < 1$, we have

$$\begin{aligned} \pi_L(\theta) &= \int_0^\infty z^{-1} \exp\{-\theta^2/(2\tau_w^2 z^2) - z^2/2\} dz \\ &= (2\tau_w^2)^{-1} \int_0^\infty u^{-1} \exp\{-\theta^2/(2\tau_w^2 u) - u/2\} du \\ &= (2\tau_w^2)^{-1} \int_0^\infty u^{-1/2} \exp\{\epsilon u/2\} u^{-1/2} \exp\{-\theta^2/(2\tau_w^2 u) - (1/2 + \epsilon/2)u\} du \\ &\geq (2\tau_w^2)^{-1} \epsilon^{1/2} \exp\{1\} \int_0^\infty u^{-1/2} \exp\{-\theta^2/(2\tau_w^2 u) - (1/2 + \epsilon/2)u\} du \\ &= (2\tau_w^2)^{-1} \epsilon^{1/2} \exp\{1\} (\pi/(1/2 + \epsilon/2))^{1/2} \exp\{-(1 + \epsilon)^{1/2} |\theta|/\tau_w\}. \end{aligned}$$

Second, we show that the upper bound holds.

$$\begin{aligned}\pi_L(\theta) &= \int_0^\infty z^{-1} \exp\{-\theta^2/(2\tau_w^2 z^2) - z^2/2\} dz \\ &\leq \int_0^\infty \exp\{-(1-\epsilon)\theta^2/(2\tau_w^2 z^2) - z^2/2\} dz \\ &\propto \exp\{-(1-\epsilon)^{1/2}|\theta|/\tau_w\}.\end{aligned}$$

□

Proposition 2.5. *Let π_E be the marginal density of θ defined in (4) with $T(t) = \exp(\lambda_1 \text{sign}(t)t^2)$ for $0 < \lambda_1 \leq 1/2$. Then, for any $\kappa > 0$, there exists θ_0 such that $c_1(\log|\theta|)^{-\frac{1}{2}}|\theta|^{(-1-\frac{1}{2\lambda_1})(1+\kappa)} \leq \pi_E(\theta) \leq c_2(\log|\theta|)^{-\frac{1}{2}}|\theta|^{(-1-\frac{1}{2\lambda_1})(1-\kappa)}$ if $\theta > \theta_0$, where c_1 and c_2 are some positive constants.*

Proof of Proposition 2.5. Without loss of generality, we assume $\alpha_0 = 0$ and $\tau_w^2 = 1$. Because the tail behavior of θ is governed by the positive region of α , we assume that $\alpha > 0$. Then, letting $\theta = T(\alpha)w$ and $z = w$, it follows that

$$\exp\left\{-\frac{\alpha^2}{2} - \frac{w^2}{2}\right\} d\alpha dw = J(\theta, z) \exp\left\{-\frac{\{T^{-1}(\theta/z)\}^2}{2} - \frac{z^2}{2}\right\} d\theta dz,$$

where $J(\theta, z)$ is the determinant of the Jacobian term, and one can show that $J(\theta, w) = \left[2\theta\lambda_1^{1/2}\{\log(\theta/z)\}^{1/2}\right]^{-1}$ when $T(t) = \exp\{\lambda_1 \text{sign}(t)t^2\}$. As a result, the marginal density of θ given z is proportional to

$$\pi_E(\theta) \propto \int_0^\infty \lambda_1^{-1/2} \{\log(\theta/z)\}^{-1/2} \theta^{-1/(2\lambda_1)-1} z^{1/(2\lambda_1)} \exp\left\{-\frac{z^2}{2}\right\} dz$$

By the dominated convergence theorem, the proof is completed. □

Proposition 4.1. *Let $r_j = \mathbf{y} - \sum_{k \neq j} X_k \theta_k$ and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)$, and let $N_{tr}(a, b; c, d)$ denote the truncated Gaussian with mean a and variance b on (c, d) . The conditional distribution $[\alpha_j \mid \boldsymbol{\alpha}_{-j}, \mathbf{w}, \mathbf{y}, \sigma^2]$ based on the posterior distribution (5) with the ReLU activation function is $\kappa N_{tr}(0, 1; -\infty, \alpha_0) + (1 - \kappa) N_{tr}(\tilde{\alpha}_j, \tilde{\sigma}_j^2; \alpha_0, \infty)$, where $\tilde{\alpha}_j = \frac{(r_j + X_j \alpha_0 w_j)^\top X_j w_j}{X_j^\top X_j w_j^2 + \sigma^2}$, $\tilde{\sigma}_j^2 = \sigma^2 (X_j^\top X_j w_j^2 + \sigma^2)^{-1}$, and*

$$\kappa = \frac{\Phi(\alpha_0) \exp\left\{-\frac{\|r_j\|_2^2}{2\sigma^2}\right\}}{\Phi(\alpha_0) \exp\left\{-\frac{\|r_j\|_2^2}{2\sigma^2}\right\} + \left\{1 - \Phi\left(\frac{\alpha_0 - \tilde{\alpha}_j}{\tilde{\sigma}_j}\right)\right\} \tilde{\sigma}_j \exp\left\{\frac{\tilde{\alpha}_j^2}{2\tilde{\sigma}_j^2} - \frac{\|r_j + X_j \alpha_0 w_j\|_2^2}{2\sigma^2}\right\}}.$$

Proof of Proposition 4.1. We note that the conditional distribution of α_j given the others is

$$\pi(\alpha_j \mid \boldsymbol{\alpha}_{-j}, \mathbf{w}, \sigma^2, \mathbf{y}) \propto (2\pi)^{-1/2} \exp\left\{-\frac{\|r_j - X_j T(\alpha_j - \alpha_0) w_j\|_2^2}{2\sigma^2} - \frac{\alpha_j^2}{2}\right\}.$$

Since the activation function is the ReLU function, it follows that

$$\pi(\alpha_j \mid \alpha_{(-j)}, w, \sigma^2, \mathbf{y}) \propto \begin{cases} (2\pi)^{-1/2} \exp \left\{ -\|r_j\|_2^2 / (2\sigma^2) - \alpha_j^2 / 2 \right\}, & \text{if } \alpha_j < \alpha_0 \\ (2\pi)^{-1/2} \exp \left\{ -\|\tilde{r}_j - X_j \alpha_j w_j\|_2^2 / (2\sigma^2) - \alpha_j^2 / 2 \right\}, & \text{if } \alpha_j \geq \alpha_0, \end{cases}$$

where $\tilde{r}_j = r_j + X_j \alpha_0 w_j$. By doing a simple calculation, we obtain that

$$\begin{aligned} & (2\pi)^{-1/2} \exp \left\{ -\|\tilde{r}_j - X_j \alpha_j w_j\|_2^2 / (2\sigma^2) - \alpha_j^2 / 2 \right\} \\ &= \tilde{\sigma}_j \exp \left\{ -\|\tilde{r}_j\|_2^2 / (2\sigma^2) + \tilde{\alpha}_j^2 / (2\tilde{\sigma}_j^2) \right\} \phi(\alpha_j; \tilde{\alpha}_j, \tilde{\sigma}_j^2), \end{aligned}$$

where $\phi(\cdot; u, z)$ is the Gaussian density function with mean u and variance z , and $\tilde{\alpha}_j$ and $\tilde{\sigma}_j^2$ are defined in the statement of the proposition. This completes the proof. \square

Theorem 5.1. *Assume that (A1) – (A4) hold and σ^2 is known. Suppose, for the neuronized prior defined in Definition 1.1 with T be the ReLU function, $(n \log p)^{-1}/16 \leq \tau_w^2 \leq n^{-1}p^2$ and α_0 follows the distribution in (9) with $(a_0, b_0) = (1, p^u)$ for some constant $u > 1$. Then, the posterior distribution based on this neuronized prior achieves the optimal posterior contraction rate ϵ_n , i.e.,*

$$\epsilon_n = \begin{cases} |\mathbf{t}| \sqrt{\log p / n}, & \text{under } l_1 \text{ norm,} \\ \sqrt{|\mathbf{t}| \log p / n}, & \text{under } l_2 \text{ norm.} \end{cases}$$

Proof of Theorem 5.1. Castillo et al. (2015) investigated asymptotic posterior behaviors for high-dimensional linear regression models. They suggested some sufficient conditions for a certain class of priors to achieve the model selection consistency and the optimal posterior contraction rate. We will show that the conditions on the neuronized SpSL prior satisfies the sufficient conditions proposed in Castillo et al. (2015) to achieve an optimal posterior contraction rate. The first condition is imposed on the model prior as

$$A_1 p^{-A_3} \pi(|\gamma| - 1) \leq \pi(|\gamma|) \leq A_2 p^{-A_4} \pi(|\gamma| + 1), \quad (15)$$

where $\gamma = \{\gamma_1, \dots, \gamma_p\}^T$ for some positive constants A_1, A_2, A_3 , and A_4 , and $|\gamma|$ indicates the number of non-zero γ_j 's. It was shown that the condition (15) is met when a beta prior, $Beta(1, p^u)$ for some $u > 1$, is imposed on η in (3). For the neuronized prior, this condition can be satisfied by imposing a hyper-prior of α_0 proposed in Proposition 2.1 with $a_0 = 1$ and $b_0 = p^u$.

The other condition they considered is on the Laplace slab prior as follows:

$$\pi_1(\theta_j) = 2^{-1} \lambda_n \exp\{-\lambda_n |\theta_j|\} \text{ with } \|X\|/p \leq \lambda_n \leq 4 \|X\| (\log p)^{1/2}, \quad (16)$$

where $\|X\| = \max_{1 \leq j \leq p} \|X_j\|_2$.

As shown in Proposition 2.3, the tail behavior of the neuronized BL prior is decaying at a rate of $\exp\{-t/\tau_w\}$ when t is large enough, so by plug-in $1/\tau_w$ in λ_n ,

its asymptotic property can be preserved by setting $(n \log p)^{-1}/16 \leq \tau_w^2 \leq n^{-1}p^2$ in the neuronized prior under (A2).

One important concept in Castillo et al. (2015) is the *compatibility condition* that is defined as below:

$$\phi(\mathbf{k}) = \inf_{\theta} \left\{ \frac{\|X\theta\|_2 |\mathbf{k}|^{1/2}}{\|X\| \|\theta_{\mathbf{k}}\|_1} : \|\theta_{\mathbf{k}^c}\|_1 \leq 7 \|\theta_{\mathbf{k}}\|_1 \right\}.$$

The other definitions used in Castillo et al. (2015) follow

$$\bar{\phi}(s) = \inf_{\theta_{\mathbf{k}, \mathbf{k}}} \left\{ \frac{\|X_{\mathbf{k}}\theta_{\mathbf{k}}\|_2}{\|X_{\mathbf{k}}\| \|\theta_{\mathbf{k}}\|_1} : 0 \neq |\mathbf{k}| \leq s \right\}, \quad \tilde{\phi}(s) = \inf_{\theta_{\mathbf{k}, \mathbf{k}}} \left\{ \frac{\|X_{\mathbf{k}}\theta_{\mathbf{k}}\|_2}{\|X_{\mathbf{k}}\| \|\theta_{\mathbf{k}}\|_2} : 0 \neq |\mathbf{k}| \leq s \right\} \quad (17)$$

The first equation in (17) is a stronger version of the compatibility condition, which uniformly controls the minimum eigenvalue of Gram matrices in a l_1 sense, and the second equation in (17) is a restricted eigenvalue condition that is similar with (A3). Under these notations, one can show that $\bar{\phi}(s) \geq C_2^{-1} C_3^{1/2} s^{-1/2}$ by using (A2) and (A3). Then, consider

$$\begin{aligned} \bar{\psi}(\mathbf{k}) &= \bar{\phi} \left(\left(2 + \frac{3}{A_4} + \frac{33\lambda_n}{2\phi(\mathbf{k})^2 \|X\| \sqrt{\log p}} \right) |\mathbf{k}| \right) \\ \tilde{\psi}(\mathbf{k}) &= \tilde{\phi} \left(\left(2 + \frac{3}{A_4} + \frac{33\lambda_n}{2\phi(\mathbf{k})^2 \|X\| \sqrt{\log p}} \right) |\mathbf{k}| \right), \end{aligned}$$

where A_4 is defined in (15) and λ_n appears in (16).

Theorem 1 in Castillo et al. (2015) states that $\sup_{\theta_0} \mathbb{E}_{\theta_0} \pi(|\mathbf{k}| > |\mathbf{t}| + M(1 + 32/\phi(\mathbf{t})^2)|\mathbf{t}|/A_4 \mid \mathbf{y}) \rightarrow 0$, and the condition (A3) (restricted eigen value condition) implies a compatibility condition, i.e. $\phi(\mathbf{k}) > 0$ for $|\mathbf{k}| \leq |\mathbf{t}| \log n$, as shown in van de Geer et al. (2009). It thus follows that $\sup_{\theta_0} \mathbb{E}_{\theta_0} \pi(|\mathbf{k}| > |\mathbf{t}| \log n \mid \mathbf{y}) \rightarrow 0$, since the term $M(1 + 32/\phi(\mathbf{t})^2)|\mathbf{t}|/A_4$ is bounded when $\phi(\mathbf{t}) > 0$. Now we can restrict our focus on models such that $\{\mathbf{k} : |\mathbf{k}| \leq |\mathbf{t}| \log n\}$.

Using the aforementioned results, Theorem 2 in Castillo et al. (2015) shows the following results:

$$\begin{aligned} \sup_{\theta_0} \mathbb{E}_{\theta_0} \pi \left(\|\theta - \theta_0\|_2 > \frac{M}{\tilde{\psi}(\mathbf{t})^2} \frac{\sqrt{|\mathbf{t}| \log p}}{\|X\| \phi(\mathbf{t})} \mid \mathbf{y} \right) &\rightarrow 0 \\ \sup_{\theta_0} \mathbb{E}_{\theta_0} \pi \left(\|\theta - \theta_0\|_1 > \frac{M}{\tilde{\psi}(\mathbf{t})^2} \frac{|\mathbf{t}| \sqrt{\log p}}{\|X\| \phi(\mathbf{t})^2} \mid \mathbf{y} \right) &\rightarrow 0, \end{aligned}$$

for a large enough constant $M > 0$. Since the restricted eigenvalue condition (A3) implies that $\phi(\mathbf{k}) > 0$ for $|\mathbf{k}| < |\mathbf{t}| \log n$, by using condition (A2) and (A3)), it follows that

$$\begin{aligned} \sup_{\theta_0} \mathbb{E}_{\theta_0} \pi \left(\|\theta - \theta_0\|_2 > M' C_2^2 C_3^{-1} \sqrt{|\mathbf{t}| \log p/n} \mid \mathbf{y} \right) &\rightarrow 0 \\ \sup_{\theta_0} \mathbb{E}_{\theta_0} \pi \left(\|\theta - \theta_0\|_1 > M'' C_2^2 C_3^{-1} |\mathbf{t}| \sqrt{\log p/n} \mid \mathbf{y} \right) &\rightarrow 0, \end{aligned}$$

for some constant M' and M'' that are larger than M . \square

Theorem 5.2. *Assume that (A1) – (A4) hold and σ^2 is known. Suppose that $T(t) = \exp\{t^2/\{2(r-1)\}\}$ for $r \geq 2$, and let $\tau_w \preceq p^{-(u+1)/(r-1)}|\mathbf{t}|\log p/n$ and $-\log \tau_w = O(\log p)$ for some $u > 0$, and $\alpha_0 = 0$. Then, the posterior distribution of θ based on the corresponding neuronized prior achieves the optimal contraction rate in (14).*

Proof of Theorem 5.2. We will show that our proposed conditions on the continuous neuronized prior satisfy the sufficient conditions introduced in Song and Liang (2017), and as a result, the optimal contraction rate for the standard shrinkage prior also can be applied to its neuronized counterpart.

We first list the regularity conditions in Song and Liang (2017) as follows:

$B_1(1)$: All covariates are uniformly bounded.

$B_1(2)$: The dimensionality is high $p \succeq n$.

$B_1(3)$: There exist some integer \bar{p} and fixed constant λ_0 such that

$$\bar{p} \succ |\mathbf{t}|, \text{ and } \inf_{\mathbf{k}: |\mathbf{k}| < \bar{p}} \lambda_{\min}(X_{\mathbf{k}}^T X_{\mathbf{k}}) \geq n\lambda_0.$$

$B_2(1)$: $|\mathbf{t}|\log p \prec n$.

$B_2(2)$: $\max_{1 \leq j \leq p} |\theta_{0,j}/\sigma_0^2| \leq \gamma_3 E_n$ for some fixed $\gamma \in (0, 1)$ and E_n is a non-decreasing sequence.

It is clear that our condition (A2) guarantees $B_1(1)$, and our (A1) and (A3) imply $B_1(1)$, and $B_2(1)$. We further assume that $\bar{p} = |\mathbf{t}|\log n$ to assure that (A3) leads to $B_1(3)$. Also, (A4) leads to $B_2(2)$. Thus, our conditions (A1) – (A4) satisfy these regularity conditions.

In Corollary 3.1 in Song and Liang (2017), under B_1 and B_2 , they proposed some conditions on the shrinkage prior to achieve the optimal posterior contraction rate for standard continuous shrinkage priors. Consider a continuous prior with r degree of polynomial tails, e.g. a Cauchy attains $r = 2$, and the prior has a scale parameter λ_n . Then, their conditions on the global shrinkage parameter follows:

$$\tau_w \leq a_n p^{-(u+1)/(r-1)+1}, \quad -\log \tau_w = O(\log p),$$

for some $u > 0$ and $a_n \asymp (|\mathbf{t}|\log p/n)^{1/2}/p$.

By Proposition 2.5, setting $T(t) = \exp\{t^2/\{2(r-1)\}\}$ guarantees that the resulting marginal density of the coefficient decays at a polynomial rate with $r \geq 2$. Also, we set $-\log \tau_w = O(\log p)$ and $\tau_w = O(p^{-(u+1)/(r-1)}\sqrt{|\mathbf{t}|\log p/n})$ for some $u > 0$. This completes the proof. \square

Theorem 5.4. *Consider the case with X being orthogonal, σ^2 known, and α_0 fixed. Suppose the activation function T for a neuronized prior has stable tails. Then, Algorithm 1 is geometrically ergodic.*

Proof of Theorem 5.4. Without loss of generality, we assume that $\sigma^2 = 1$ and $\alpha_0 = 0$. Since $n^{-1/2}X$ is orthogonal, it follows that

$$\begin{aligned}\pi(\alpha_j \mid \mathbf{y}) &= \int \pi(\alpha_j, w_j \mid \mathbf{y}) dw_j \\ &\propto \{nT^2(\alpha_j) + 1/\tau_w^2\}^{-\frac{1}{2}} \exp[(X_j^T \mathbf{y})^2 / \{2(nT^2(\alpha_j) + 1/\tau_w^2)\} - \alpha_j^2/2]\end{aligned}$$

and $\pi(\boldsymbol{\alpha} \mid \mathbf{y}) = \prod_{j=1}^p \pi(\alpha_j \mid \mathbf{y})$. Then, it follows that

$$\|P^t(\boldsymbol{\alpha}^{(0)}, \cdot) - \pi_{\mathbf{y}}(\cdot)\|_{TV} \leq \max_{1 \leq j \leq p} \|P_j^t(\alpha_j^{(0)}, \cdot) - \pi_{\mathbf{y},j}(\cdot)\|_{TV}, \quad (18)$$

where $\pi_{\mathbf{y},j}(\alpha_j) = \pi(\alpha_j \mid \mathbf{y})$, $\pi_{\mathbf{y}}(\boldsymbol{\alpha}) = \pi(\boldsymbol{\alpha} \mid \mathbf{y}) = \prod_{j=1}^p \pi(\alpha_j \mid \mathbf{y})$, and P_j^t is a Markov transition kernel of the Metropolis algorithm for α_j at iteration t . Since the conditional posterior distribution of \mathbf{w} given $\boldsymbol{\alpha}$ is explicitly represented, which is a product of independent Gaussians with mean $X_j^T \mathbf{y} / (nT^2(\alpha_j) + 1/\tau_w^2)$ and variance $(nT^2(\alpha_j) + 1/\tau_w^2)^{-1}$, the convergence behavior of Algorithm 1 is solely determined by the convergence rate of $\max_{1 \leq j \leq p} \|P_j^t(\alpha_j^{(0)}, \cdot) - \pi_{\mathbf{y},j}(\cdot)\|_{TV}$, so it is sufficient to show that P_j^t results in a geometrical ergodicity for any $j \in \{1, \dots, p\}$.

To simplify the description, we first introduce some concepts regarding a distribution. We consider a distribution with a density function π , and define

$$V = \limsup_{|x| \rightarrow \infty} \frac{x}{|x|} \nabla \log \pi(x). \quad (19)$$

The distribution is called *super-exponentially light* if $V = -\infty$ in (19); *exponentially light* if V is a negative constant; and *sub-exponentially light* if $V = 0$ (Johnson and Geyer, 2012; Mengersen and Tweedie, 1996; Roberts and Tweedie, 1996). Using these definitions, Theorem 4.3 in Jarner and Hansen (2000) considers a Metropolis transition kernel induced by a proposal density that contains strictly positive amount of density around zero. Since we are using a Gaussian kernel in Algorithm 1, our case satisfies this condition. Then, their theorem implies that the resulting random-walk Metropolis algorithm targeting π is geometrically ergodic, if π is super-exponentially light and satisfies

$$\limsup_{|x| \rightarrow \infty} \frac{x}{|x|} \frac{\nabla \pi(x)}{|\nabla \pi(x)|} < 0. \quad (20)$$

However, in one-dimensional cases, equation (19) implies (20). Thus, the proof will be completed if we show that $\pi_{\mathbf{y},j}$ is super-exponentially light.

Note that

$$\frac{x}{|x|} \nabla \log \pi_{\mathbf{y},j}(x) = \text{sgn}(x) \left\{ -\frac{nT(x)T'(x)}{nT^2(x) + 1/\tau_w^2} - \frac{nT(x)T'(x)(X_j^T \mathbf{y})^2}{(nT^2(x) + 1/\tau_w^2)^2} - x \right\}, \quad (21)$$

where sgn is a sign function. Since the activation function T has stable tails, i.e., $\exists C_1, C_2, C_3 > 0$ such that (a) when $x < -C_3$, either $|T'(x)| \leq C_1$ or $|T'(x)| \geq C_2$ and the sign of $T'(x)$ does not change; and (b) when $x > C_3$, either $|T'(x)| \leq C_1$ or

$|T'(x)| \geq C_2$ and the sign of $T'(x)$ does not change. It is clear that for either tail, if $|T'(x)|$ is bounded from above, then the RHS of (21) is dominated by $-|x|$ and hence diverges to $-\infty$ as either $x \rightarrow \infty$ or $x \rightarrow -\infty$. If $|T'(x)|$ is bounded from below and $T'(x)$ does not change sign after $x > C_3$, then, as $x \rightarrow \infty$, either $T'(x) \geq C_2$, which implies that $T(x)$ will become positive eventually and thus $\lim_{x \rightarrow \infty} T(t)T'(t) \geq 0$; or $T'(x) \leq -C_2$, which means that $T(x)$ will become negative eventually and also $\lim_{x \rightarrow \infty} T(t)T'(t) \geq 0$. Thus, all the three terms inside the parenthesis of the RHS of (21) are of the same sign and, hence, the RHS diverges to $-\infty$. As $x \rightarrow -\infty$, we see by the same argument as above that, if $|T'(x)| \geq C_2$ and $T'(x)$ does not change sign after $x < -C_3$, $\lim_{x \rightarrow -\infty} T(x)T'(x) < 0$. Thus, all terms inside the parenthesis of the RHS of (21) are of the same sign and hence (21) diverges to $-\infty$.

As a result, there exist C_j and $\rho_j \in (0, 1)$ such that

$$\|P_j^t(\alpha_j^{(0)}, \cdot) - \pi_{\mathbf{y},j}(\cdot)\|_{TV} \leq C_j(\alpha_j)\rho_j^t,$$

for $j = 1, \dots, p$. By plugging this to (18), it follows that

$$\|P^t(\boldsymbol{\alpha}^{(0)}, \cdot) - \pi_{\mathbf{y}}(\cdot)\|_{TV} \leq \max_{1 \leq j \leq p} \{C_j(\alpha_j)\} \max_{1 \leq j \leq p} \{\rho_j\}^t.$$

□

Theorem 5.5. *Under the standard Bayesian linear regression setting, suppose we employ a standard continuous shrinkage prior as in (2) with a heavy-tailed distribution π_τ such that $\pi_\tau(x) \succeq \exp\{-cx^\kappa\}$, $x > 0$, for some constants $c > 0$ and $0 < \kappa < 1$. Then, the corresponding MCMC algorithm cannot achieve geometric ergodicity if one updates τ_j conditional on other variables by a RWMH algorithm.*

Proof of Theorem 5.5. We first note that when there exists no moment generating function of a target density of the Metropolis-Hastings algorithm, the resulting MH algorithm cannot achieve the geometric ergodicity (Mengersen and Tweedie, 1996). Moreover, it is well-known that if any single conditional density in a Metropolis-Hastings-within-Gibbs sampler is not geometrically ergodic, neither the full MCMC is (Diaconis et al., 2008; Robert, 1995; Roberts et al., 2001). So, it is sufficient to show that the moment generating function of $\pi(\tau_j | \beta_j)$ does not exist regardless of the value of β_j .

Consider the following conditional posterior density of τ_j for some $j \in \{1, \dots, p\}$:

$$\log \pi(\tau_j | \beta_j) = -(1/2) \log(\tau_j^2) - \beta_j^2/(2\tau_j^2) - c\tau_j^\kappa + C,$$

where C is some constant. Because $0 < \kappa < 1$, it is clear that for any $t > 0$ and $\beta_j \in \mathbb{R}$, $\tau_j t + \log \pi(\tau_j | \beta_j)$ diverges to infinity as τ_j increases, which concludes that this conditional posterior density cannot have a proper moment generating function. □

B Updating Matrix Inversion and Determinant

In this section, under a discrete SpSL Gaussian-conjugate prior, we provide an instruction on how to efficiently evaluate some linear algebra calculations that are

required to implement the fully-collapsed Gibbs sampler for the Bayesian linear model selection. When implementing the collapsed Gibbs sampler, one needs to compute the inversion and determinant of a modified sample covariance matrix at each iteration. To improve computational efficiencies, we can use the following linear algebra techniques.

Let A be a $m \times m$ symmetric matrix and $B = \begin{pmatrix} A & b \\ b^T & c \end{pmatrix}$, where b is an $m \times 1$ vector. Then,

$$B^{-1} \equiv \begin{pmatrix} Q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix} = \begin{pmatrix} A^{-1} + \frac{1}{k} A^{-1} b b^T A^{-1} & -\frac{1}{k} A^{-1} b \\ -\frac{1}{k} b^T A^{-1} & \frac{1}{k} \end{pmatrix}, \quad (22)$$

where $k = c - b^T A^{-1} b$, and

$$\det(B) \equiv \det \begin{pmatrix} A & b \\ b^T & c \end{pmatrix} = \det(A) \times (c - b^T A^{-1} b). \quad (23)$$

Conversely, if we want to update from B to A , we have

$$A^{-1} = Q_{11} - q_{12} \times q_{21} / q_{22},$$

and

$$\det(A) = \det(B) / (c - b^T A^{-1} b).$$

To apply the above updating formulas to the fully-collapsed Gibbs sampler, we let the current model be γ , randomly select one index $j \in \{1, \dots, p\}$. If $\gamma_j = 0$, we propose a candidate model by adding X_j to the current model, and the binary representation of the proposed model is $\gamma' = \{\gamma'_1, \dots, \gamma'_p\}$, where

$$\gamma'_h = \begin{cases} 1 & \text{if } \gamma_h = 1 \text{ or } h = j, \\ 0 & \text{otherwise} \end{cases},$$

for $h = 1, \dots, p$. Let $x = X_j$ and $A = X_\gamma^T X_\gamma + \frac{\sigma^2}{\tau_w^2} I$, and assume that for the current model, the inverse and the determinant of $X_\gamma^T X_\gamma + (\sigma^2 / \tau_w^2) I$ are known. We can obtain the inverse matrix and the determinant of $X_{\gamma'}^T X_{\gamma'} + (\sigma^2 / \tau_w^2) I$ economically using formulas (22) and (23):

$$\left(X_{\gamma'}^T X_{\gamma'} + \frac{\sigma^2}{\tau_w^2} I \right)^{-1} = \begin{pmatrix} A^{-1} + \frac{1}{k} A^{-1} X_\gamma^T x x^T X_\gamma A^{-1} & -\frac{1}{k} A^{-1} X_\gamma^T x \\ -\frac{1}{k} x^T X_\gamma A^{-1} & \frac{1}{k} \end{pmatrix},$$

where $k = x^T x + \sigma^2 / \tau - x^T X_\gamma A^{-1} X_\gamma^T x$, and

$$\det \left(X_{\gamma'}^T X_{\gamma'} + \frac{\sigma^2}{\tau_w^2} I \right) = \det(A) \times (x^T x + \sigma^2 / \tau_w^2 - x^T X_\gamma A^{-1} X_\gamma^T x).$$

If $\gamma_j = 1$, the candidate model is the same as the current model but with X_j excluded, i.e., γ' is

$$\gamma'_h = \begin{cases} 0 & \text{if } \gamma_h = 0 \text{ or } h = j, \\ 1 & \text{otherwise,} \end{cases}$$

for $h = 1, \dots, p$. Then, it follows that

$$\left(X_{\gamma'}^T X_{\gamma'} + \frac{\sigma^2}{\tau_w^2} I\right)^{-1} = Q_{11} - q_{12}q_{21}/q_{22}, \quad (24)$$

and

$$\det\left(X_{\gamma'}^T X_{\gamma'} + \frac{\sigma^2}{\tau_w^2} I\right) = \frac{\det(X_{\gamma'}^T X_{\gamma'} + \frac{\sigma^2}{\tau_w^2} I)}{c - b^T D^{-1} b},$$

where Q_{11} , q_{12} , q_{21} , and q_{22} are block components of

$$\left(X_{\gamma}^T X_{\gamma} + \frac{\sigma^2}{\tau_w^2} I\right)^{-1} = \begin{pmatrix} Q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix},$$

and the second block corresponds to X_j . Also, c , b , and D are block components of $X_{\gamma}^T X_{\gamma} + \frac{\sigma^2}{\tau_w^2} I$; i.e.,

$$X_{\gamma}^T X_{\gamma} + \frac{\sigma^2}{\tau_w^2} I = \begin{pmatrix} D & b \\ b^T & c \end{pmatrix} = \begin{pmatrix} X_{\gamma'}^T X_{\gamma'} + \frac{\sigma^2}{\tau_w^2} I & X_{\gamma'}^T X_j \\ X_j^T X_{\gamma'} & X_j^T X_j + (\sigma^2/\tau_w^2) \end{pmatrix},$$

where $\gamma \setminus j$ is the model where X_j is discarded from γ , and D^{-1} can be evaluated from (24).

Once these inverse matrix and determinant are evaluated, the Metropolis acceptance probability can be defined as $\min\left\{1, \frac{\pi(\gamma'|\mathbf{y}, \eta, \sigma^2)}{\pi(\gamma|\mathbf{y}, \eta, \sigma^2)}\right\}$, where

$$\pi(\gamma | \mathbf{y}, \eta, \sigma^2) \propto |X_{\gamma}^T X_{\gamma} + (\sigma^2/\tau_w^2) I|^{-1/2} \exp\left\{\mathbf{y}^T \tilde{P}_{\gamma} \mathbf{y} / 2\right\} \eta^{|\gamma|+a_0-1} (1-\eta)^{p-|\gamma|+b_0-1},$$

and $\tilde{P}_{\gamma} = X_{\gamma}(X_{\gamma}^T X_{\gamma} + \sigma^2/\tau_w^2 I)^{-1} X_{\gamma}^T$. We note that this posterior probability is based on a prior setting with $\theta_{\gamma} \sim N(0, \tau_w^2 I)$ and $\pi(\sigma^2) \propto 1/\sigma^2$.

The computational complexity of this linear algebra calculation, given the inverse matrix and determinant for the current model, is $O(|\gamma|n) + O(|\gamma|^2)$. This updating rule is more efficient than a naive evaluations without the guidance of the previous result, which requires $O(|\gamma|^2 n) + O(|\gamma|^3)$. However, the computational gain would be slightly diluted in overall, because after evaluating the inverse matrix and the determinant, evaluating the marginal likelihood takes an additional complexity $O(|\gamma|n)$ that is equally applied to both procedures.

In contrast, the half-collapsed Gibbs sampler and N-SpSL(Exact) do not require the evaluation of the determinant nor the inverse matrix, and their computational complexity for a single sampling γ_j is lower than that required for the fully-collapsed Gibbs, $O(n)$. The HCG and the neuronized SpSL procedure thus appear to be more efficient, in terms of ESS per second, than the FCG at least for our limited examples.

C Some Auxiliary Results

C.1 Additional optimization paths for CAAN

As a supplement of the synthetic example in Section 4.5, we examine a scenario where the true model size is five (the other settings are equivalent to the example

in the main text). Figure 8 show that the CAAN and the SSLasso procedures consistently chose the same model via EBIC across all ten random initial values, while the MM and the EMVS fail to achieve the consistency.

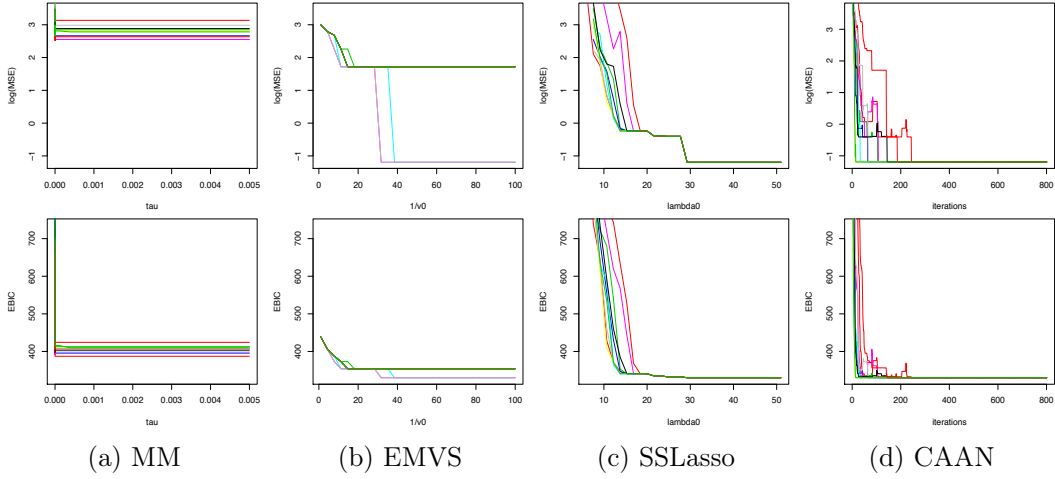


Figure 8: Trace plots of the log-MSE (top row) and EBIC (bottom row) paths from 10 different initial points for the four optimization algorithms, based on a synthetic data set generated from the Bardet-Biedl dataset ($n = 120$ and $p = 200$) with the true model size 5. The MM procedure used $\tau_3 = 10^{-2}$.

C.2 Comparisons Between Different MCMC Algorithms

In this section, we consider extra simulation studies. We first compare the ESS (per second) of “N-SpSL-L(Exact)” and “N-SpSL-L(RW)”, and the results are shown in Table 5. The column “Ind” and “Dep” indicates scenarios where the covariates are generated from iid standard Gaussian and from the Toeplitz design considered in Section 6, respectively. The other settings are exactly the same with these in the simulation studies in the main paper. The results show that “N-SpSL-L(Exact)” is at least two times more efficient in terms of ESS.

	Low-dimension							
Sample size	$(n = 200, p = 50)$				$(n = 400, p = 100)$			
Signal strength	Weak		Strong		Weak		Strong	
Covariate	Ind	Dep	Ind	Dep	Ind	Dep	Ind	Dep
N-SpSL-L(Exact)	7625.6	5949.1	8255.1	4551.7	2793.6	1123.9	3479.3	515.0
N-SpSL-L(RW)	2238.5	1666.8	2582.7	1397.9	1000.5	370.7	889.3	210.5
	High-dimension							
Sample size	$(n = 100, p = 300)$				$(n = 150, p = 1000)$			
Signal strength	Weak		Strong		Weak		Strong	
Covariate	Ind	Dep	Ind	Dep	Ind	Dep	Ind	Dep
N-SpSL-L(Exact)	919.7	874.5	1271.1	561.6	217.2	131.9	262.2	114.5
N-SpSL-L(RW)	221.9	203.6	294.7	136.6	69.4	43.7	80.3	36.6

Table 5: A comparison of ESS per second between different neuronized SpSL procedures.

Table 6 compares two different MCMC algorithms: the half-collapsed Gibbs sam-

pler used in the main manuscript (SpSL-G(HCG)) vs. the fully-collapsed Gibbs sampler (SpSL-G(FCG)). Briefly, by taking advantages of Gaussian conjugacy, SpSL-G(FCG) marginalizes out all the continuous coefficients to obtain the target distribution $\pi(\gamma \mid \mathbf{y})$ and considers as a proposal to flip a randomly selected indicator from γ_j to $1 - \gamma_j$. It is well-known that “SpSL-G(FCG)” is highly inefficient (Ji and Schmidler, 2013), and this finding is also confirmed again in Table 6. The ESS of “SpSL-G(FCG)” is significantly smaller than that from “SpSL-G(HCG)”. In particular, under high-dimensional settings, its ESS is less than 10, while “SpSL-G(HCG)” attains at least hundreds of ESS per second.

	Low-dimension							
Sample size	$(n = 200, p = 50)$				$(n = 400, p = 100)$			
Signal strength	Weak		Strong		Weak		Strong	
Covariate	Ind	Dep	Ind	Dep	Ind	Dep	Ind	Dep
SpSL-G(HCG)	15781.0	15422.8	20366.7	9752.9	6175.9	2521.9	8939.6	1205.3
SpSL-G(FCG)	82.2	48.9	487.5	54.0	184.6	21.6	540.7	38.26
	High-dimension							
Sample size	$(n = 100, p = 300)$				$(n = 150, p = 1000)$			
Signal strength	Weak		Strong		Weak		Strong	
Covariate	Ind	Dep	Ind	Dep	Ind	Dep	Ind	Dep
SpSL-G(HCG)	2773.3	3015.4	3960.2	1896.2	744.1	506.3	819.5	385.6
SpSL-G(FCG)	1.5	6.8	4.8	6.5	7.0	4.1	7.3	8.8

Table 6: A comparison of ESS per second between different SpSL procedures.

C.3 Additional simulation studies of sparse regression algorithms

We provide the results of more simulation studies for independent covariate cases with different signal strengths in Table 7 and 8, and Table 9 and 10 show simulations results for strong signals. The first five true regression coefficients are non-zero, and the non-zero coefficients of the low-dimensional and high-dimensional settings are set to be $\pm s$ and $s \times \{\pm 0.4, \pm 0.45, \pm 0.5, \pm 0.55, \pm 0.6\}$, respectively.

Method	Strong Signal ($s = 0.3$)									
	$(n = 200, p = 50)$					$(n = 400, p = 100)$				
	MSE	Cos(Angle)	MCC	FP	ESS	MSE	Cos(Angle)	MCC	FP	ESS
Oracle	0.025	0.979				0.028	0.987			
SpSL-G(HCG)	0.085	0.909	0.89	0.17	20877.1	0.036	0.981	0.99	0.14	6414.5
N-SpSL-L(Exact)	0.072	0.924	0.92	0.40	7957.3	0.042	0.978	0.98	0.47	4100.8
SpSL-C(HCG)	0.073	0.923	0.92	0.28	1608.1	0.038	0.981	0.98	0.27	795.7
N-SpSL-C(RW)	0.091	0.901	0.89	0.10	2423.0	0.036	0.981	0.99	0.13	1307.5
HS	0.087	0.906	0.90	0.15	815.6	0.054	0.972	0.99	0.14	718.5
N-HS(RW)	0.088	0.906	0.90	0.14	1754.4	0.052	0.973	0.99	0.13	619.5
BL	0.154	0.844	0.79	4.14	3374.9	0.134	0.918	0.92	1.69	846.0
N-BL(RW)	0.122	0.866	0.82	2.17	1822.3	0.114	0.936	0.97	0.67	575.5
SkG	0.074	0.922	0.91	0.26	9238.6	0.038	0.980	0.98	0.29	4712.8
SpSL(MM)	0.099	0.905	0.76	3.16		0.112	0.939	0.78	5.31	
N-SpSL-L(MAP)	0.078	0.928	0.88	1.07		0.058	0.970	0.93	1.49	
EMVS	0.238	0.718	0.71	0.03		0.089	0.954	0.96	0.00	
SSLasso	0.096	0.894	0.88	1.09		0.037	0.981	0.93	1.38	
Lasso(CV)	0.091	0.906	0.52	10.70		0.095	0.958	0.53	19.72	
SCAD(CV)	0.080	0.920	0.55	8.94		0.049	0.974	0.63	12.47	
Lasso(BIC)	0.216	0.852	0.90	0.94		0.339	0.904	0.95	1.10	
SCAD(BIC)	0.211	0.847	0.89	0.99		0.306	0.896	0.94	1.21	
N-BL(MAP)	0.107	0.881	0.78	2.94		0.106	0.942	0.94	1.34	

Table 7: Results for the low-dimensional setting with independent covariates. SpSL, HS, and BL indicate the procedure based on the discrete SpSL, the horseshoe, and Bayesian Lasso priors, respectively. The sign “N” stands for the neuronized version of the corresponding prior. The values of the best results are highlighted with bold.

Method	Strong Signal ($s = 1.5$)									
	$(n = 100, p = 300)$					$(n = 150, p = 1000)$				
	MSE	Cos(Angle)	MCC	FP	ESS	MSE	Cos(Angle)	MCC	FP	ESS
Oracle	0.055	0.992				0.037	0.995			
SpSL-G(HCG)	0.095	0.985	0.98	0.23	4409.6	0.054	0.992	0.99	0.15	1168.0
N-SpSL-L(Exact)	0.139	0.977	0.94	0.75	1317.1	0.084	0.987	0.96	0.49	389.2
SpSL-C(HCG)	0.120	0.981	0.96	0.48	157.9	0.068	0.989	0.97	0.34	55.7
N-SpSL-C(RW)	0.090	0.986	0.98	0.17	421.2	0.052	0.992	0.99	0.12	131.2
HS	0.164	0.973	0.86	1.98	56.6	0.191	0.968	0.68	6.20	7.3
N-HS(RW)	0.155	0.975	0.87	1.83	182.1	0.190	0.968	0.68	6.36	12.3
BL	1.015	0.808	0.41	24.20	42.1	1.512	0.699	0.65	5.32	12.6
N-BL(RW)	0.864	0.826	0.38	29.32	50.2	1.439	0.736	0.67	5.93	11.9
SkG	0.097	0.985	0.99	0.01	2827.0	0.054	0.992	1.00	0.00	949.7
SpSL(MM)	0.489	0.909	0.82	1.58		1.932	0.613	0.43	8.35	
N-SpSL-L(MAP)	0.109	0.982	0.98	0.03		0.041	0.994	1.00	0.03	
EMVS	0.483	0.910	0.88	0.01		1.215	0.743	0.71	0.00	
SSLasso	0.090	0.986	0.99	0.02		0.042	0.994	1.00	0.04	
Lasso(CV)	0.412	0.947	0.44	24.25		0.332	0.965	0.40	33.72	
SCAD(CV)	0.153	0.975	0.53	13.72		0.095	0.985	0.50	18.09	
Lasso(EBIC)	1.740	0.821	0.96	0.08		1.577	0.861	0.99	0.05	
SCAD(EBIC)	1.690	0.825	0.96	0.08		1.568	0.859	0.99	0.05	
N-BL(MAP)	0.394	0.941	0.36	30.40		0.332	0.955	0.30	48.70	

Table 8: Results for the high-dimensional setting with independent covariates.

Method	Strong Signal ($s = 0.3$)									
	$(n = 200, p = 50)$					$(n = 400, p = 100)$				
	MSE	Cos(Angle)	MCC	FP	ESS	MSE	Cos(Angle)	MCC	FP	ESS
Oracle	0.071	0.939				0.071	0.966			
SpSL-G(HCG)	0.305	0.645	0.56	0.05	12566.5	0.411	0.757	0.70	0.04	1243.8
N-SpSL-L(Exact)	0.269	0.683	0.59	0.14	4260.9	0.326	0.807	0.77	0.15	578.3
SpSL-C(HCG)	0.280	0.677	0.58	0.11	840.2	0.346	0.799	0.76	0.11	153.5
N-SpSL-C(RW)	0.311	0.637	0.56	0.05	1601.8	0.431	0.743	0.69	0.03	204.3
HS	0.278	0.669	0.61	0.09	584.0	0.362	0.782	0.76	0.02	101.1
N-HS(RW)	0.279	0.667	0.61	0.08	1224.2	0.369	0.778	0.76	0.02	168.7
BL	0.247	0.702	0.62	4.11	3124.1	0.265	0.814	0.79	6.41	548.0
N-BL(RW)	0.241	0.720	0.64	2.03	1396.5	0.285	0.834	0.81	5.18	302.3
SkG	0.289	0.662	0.57	0.10	5027.2	0.357	0.791	0.75	0.10	459.8
SpSL(MM)	0.328	0.573	0.49	1.23		0.464	0.711	0.61	2.66	
N-SpSL-L(MAP)	0.310	0.671	0.62	0.86		0.268	0.860	0.82	1.49	
EMVS	0.454	0.478	0.49	0.01		0.729	0.554	0.55	0.00	
SSLasso	0.363	0.572	0.53	0.81		0.499	0.714	0.68	1.19	
Lasso(CV)	0.244	0.686	0.46	6.96		0.316	0.815	0.47	20.19	
SCAD(CV)	0.357	0.614	0.41	4.89		0.367	0.800	0.50	12.74	
Lasso(BIC)	0.356	0.541	0.56	0.60		0.568	0.635	0.67	1.74	
SCAD(BIC)	0.385	0.524	0.52	0.89		0.628	0.602	0.59	2.97	
N-BL(MAP)	0.247	0.694	0.59	2.19		0.280	0.835	0.80	2.53	

Table 9: Results for the low-dimensional setting with dependent covariates. SpSL, HS, and BL indicate the procedure based on the discrete SpSL, the horseshoe, and Bayesian Lasso priors, respectively. The sign “N” stands for the neuronized version of the corresponding prior.

Method	Strong Signal ($s = 1.5$)									
	$(n = 100, p = 300)$					$(n = 150, p = 1000)$				
	MSE	Cos(Angle)	MCC	FP	ESS	MSE	Cos(Angle)	MCC	FP	ESS
Oracle	0.150	0.980				0.080	0.989			
SpSL-G(HCG)	0.980	0.823	0.76	0.11	2149.6	0.607	0.890	0.84	0.11	592.5
N-SpSL-L(Exact)	0.948	0.827	0.76	0.28	557.4	0.610	0.890	0.84	0.33	162.6
SpSL-C(HCG)	0.858	0.848	0.79	0.22	68.8	0.509	0.910	0.87	0.27	26.7
N-SpSL-C(RW)	1.046	0.809	0.75	0.11	175.9	0.628	0.886	0.84	0.10	58.1
HS	1.059	0.805	0.79	0.80	20.2	0.769	0.859	0.69	4.13	4.2
N-HS(RW)	1.019	0.814	0.79	0.84	95.2	0.719	0.868	0.71	4.03	7.9
BL	1.631	0.503	0.41	17.68	102.1	2.030	0.515	0.74	1.21	24.2
N-BL(RW)	1.494	0.682	0.62	7.53	130.0	1.715	0.622	0.74	0.96	34.0
SkG	1.402	0.736	0.65	0.02	1068.0	1.304	0.751	0.66	0.00	317.3
SpSL(MM)	2.106	0.577	0.50	1.36		2.143	0.544	0.44	2.24	
N-SpSL-L(MAP)	1.640	0.678	0.65	0.06		1.105	0.782	0.76	0.03	
EMVS	2.317	0.538	0.56	0.00		2.553	0.451	0.51	0.00	
SSLasso	0.993	0.821	0.81	0.07		0.491	0.910	0.90	0.10	
Lasso(CV)	1.240	0.759	0.46	14.56		1.258	0.762	0.38	26.23	
SCAD(CV)	1.484	0.731	0.39	12.40		1.338	0.752	0.36	21.14	
Lasso(EBIC)	2.422	0.544	0.62	0.00		2.311	0.562	0.65	0.08	
SCAD(EBIC)	2.467	0.537	0.60	0.07		2.383	0.547	0.61	0.14	
N-BL(MAP)	1.204	0.758	0.35	21.03		1.223	0.760	0.28	37.99	

Table 10: Results for the high-dimensional setting with dependent covariates.

C.4 Numerical Approximation Errors for Horseshoe Prior

In the simulation and real data studies examined in Section 6 and 7, it was shown that the horseshoe prior and its neuronized counterpart produced slightly different numerical results, even though they should have resulted in exactly the same posterior distribution for the coefficients. We here investigate a high-dimensional example with a much larger number of MCMC iterations and show that the observed differences are due to numerical approximation errors of MCMC.

We generate a synthesized data set by following the same high-dimensional setting used in Section 6, with a strong signal, $n = 150$ and $p = 1000$. We consider

100,000 iterations after 10,000 burn-in (20 thinning size). The resulting approximated posterior distributions for several coefficients are illustrated in the first two columns of Figure 9. A short chain with 10,000 iterations and 2,000 burn-in steps is also presented on the other columns.

Figure 9 shows that when the length of the chain is large enough, the standard horseshoe prior and its neuronized counterpart lead to nearly identical posterior distributions for θ_3 , θ_5 , and θ_9 . For short MCMC chains, the both standard and neuronized procedures successfully approximate the posterior distributions of θ_5 and θ_9 . However, the shorter chain did not provide a good mixing for the posterior distribution of θ_3 under the standard horseshoe prior (the left panel of (b)), with the chain stuck around the origin for a long time, leading to an over-estimation of the posterior probability around zero. Comparing with the result from the longer chain, we observe that the algorithm with the neuronized HS prior appears to have done a much better job mixing for the shorter chain.

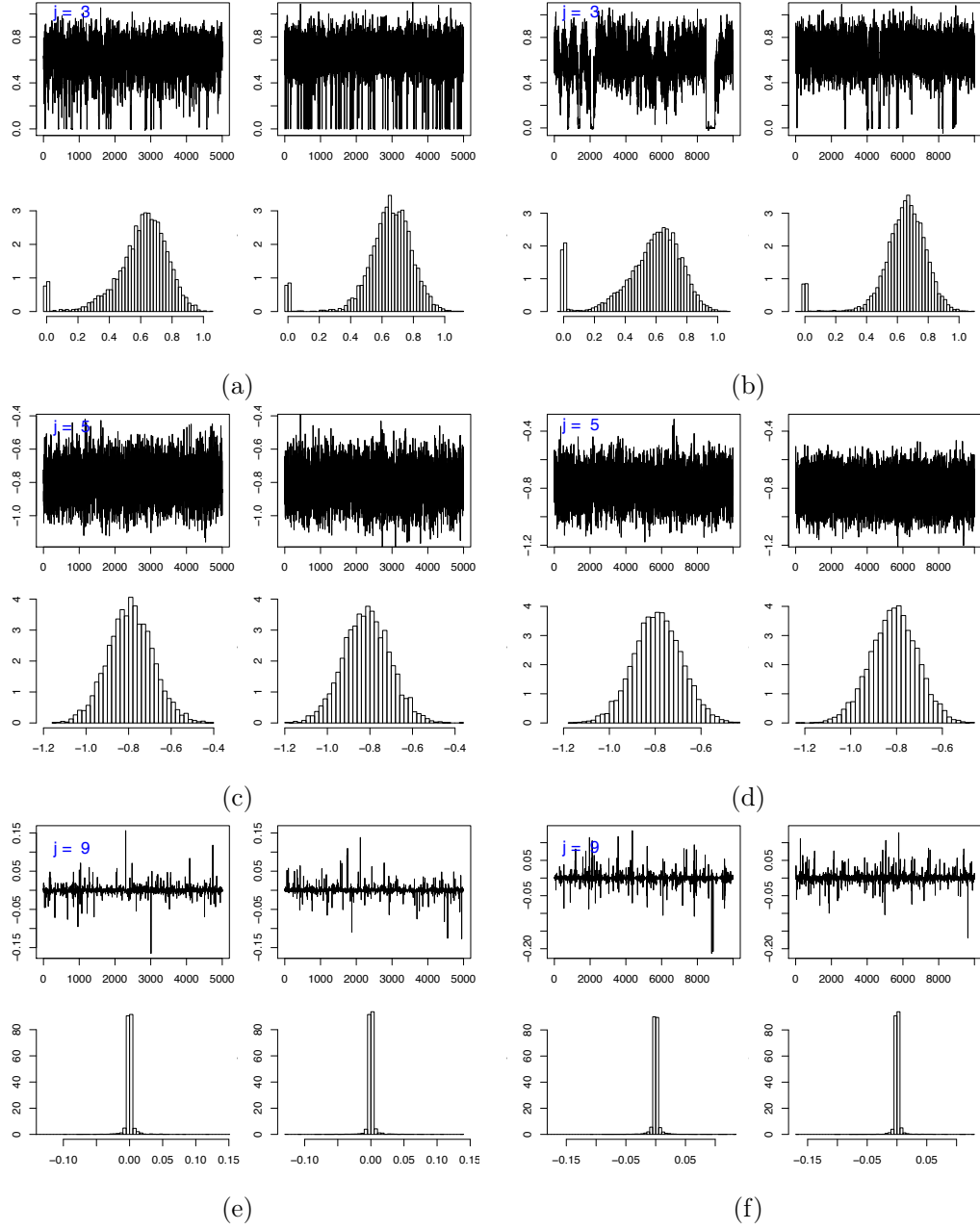


Figure 9: The first two columns indicate the cases with a long chain; the other columns show the results with a short chain. (a) and (b) illustrate the posterior distribution of θ_3 ; (c) and (d) are for the posterior distributions of θ_5 ; (e) and (f) are for θ_9 . The left and right panels within each sub-figure represent the standard horseshoe prior and the neuronized horseshoe prior, respectively.

References

- Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., et al. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.*, 43(5):1986–2018.
- Diaconis, P., Khare, K., Saloff-Coste, L., et al. (2008). Gibbs sampling, exponential families and orthogonal polynomials. *Statistical Science*, 23(2):151–178.
- Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of metropolis algorithms. *Stochastic processes and their applications*, 85(2):341–361.
- Ji, C. and Schmidler, S. C. (2013). Adaptive Markov chain Monte Carlo for bayesian variable selection. *Journal of Computational and Graphical Statistics*, 22(3):708–728.
- Johnson, L. T. and Geyer, C. J. (2012). Variable transformation to obtain geometric ergodicity in the random-walk Metropolis algorithm. *The Annals of Statistics*, 40(6):3050–3076.
- Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the hastings and metropolis algorithms. *The Annals of Statistics*, 24(1):101–121.
- Robert, C. P. (1995). Convergence control methods for Markov chain Monte Carlo algorithms. *Statistical Science*, pages 231–253.
- Roberts, G. O., Rosenthal, J. S., et al. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, 16(4):351–367.
- Roberts, G. O. and Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. *Biometrika*, 83(1):95–110.
- Song, Q. and Liang, F. (2017). Nearly optimal bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*.
- van de Geer, S. A., Bühlmann, P., et al. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.