Electronic Journal of Statistics
Vol. 14 (2020) 2714–2741
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1732

Bayesian shrinkage towards sharp minimaxity

Qifan Song

Department of Statistics, Purdue University,
250 University St, West Lafayette, Indiana, U.S.A. e-mail: qfsong@purdue.edu

Abstract: Shrinkage priors are becoming more and more popular in Bayesian modeling for high dimensional sparse problems due to its computational efficiency. Recent works show that a polynomially decaying prior leads to satisfactory posterior asymptotics under regression models. In the literature, statisticians have investigated how the global shrinkage parameter, i.e., the scale parameter, in a heavy tailed prior affects the posterior contraction. In this work, we explore how the shape of the prior, or more specifically, the polynomial order of the prior tail affects the posterior. We discover that, under the sparse normal means model, the polynomial order does affect the multiplicative constant of the posterior contraction rate. More importantly, if the polynomial order is sufficiently close to 1, it will induce the optimal Bayesian posterior convergence, in the sense that the Bayesian contraction rate is sharply minimax, i.e., not only the order, but also the multiplicative constant of the posterior contraction rate are optimal. The above Bayesian sharp minimaxity holds when the global shrinkage parameter follows a deterministic choice which depends on the unknown sparsity s. Therefore, a Beta-prior modeling is further proposed, such that our sharply minimax Bayesian procedure is adaptive to unknown s. Our theoretical discoveries are justified by simulation studies.

Keywords and phrases: Shrinkage prior, Bayesian sharp minimax, heavy-tailed prior, adaptive prior.

Received May 2019.

1. Introduction

In Bayesian inference for high dimensional sparse models, the prior distribution needs to incorporate certain *a priori* knowledge of the structural sparsity. The classical spike-and-slab modeling assigns a two-group prior to each entry of a sparse n-dimensional parameter vector $\theta^{(n)} = (\theta_1, \ldots, \theta_n)^T$, i.e., $\pi(\theta_i)$ is a mixture of two distributions which correspond to $\theta_i = 0$ and $\theta_i \neq 0$ respectively. This natural modeling however requires expensive posterior computation. An alternative one-group Bayesian modeling, or so-called shrinkage prior, is much more computational attractive, hence gains more and more popularity in the Bayesian community.

In this work, we consider the sparse normal means model, $y^{(n)} = \theta^{(n)} + \epsilon$, where $y^{(n)} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, $\epsilon \sim N(0, \sigma^2 I_n)$, and the parameter of interest is $\theta^{(n)} \in \mathbb{R}^n$. Throughout this work, the variance σ^2 of the error term is assumed to be known, and W.O.L.G, we let $\sigma^2 = 1$. Suppose that the true parameter

value θ^* is a sparse vector with $s^{(n)}$ nonzero entries, and asymptotically, we allow $s^{(n)}$ to increase simultaneously with n. For the sake of simplicity of the notation, we drop the superscript (n) from $y^{(n)}$, $\theta^{(n)}$ and $s^{(n)}$ in what follows. This normal means model can be viewed as the simplest regression problem under orthogonality. Theoretically, the study of the sparse normal means model can provide us important insights about Bayesian regression under a shrinkage prior. In practice, applications of the normal means model include multiple testing problems where y_i is the z-test statistics, and signal detection problems where y_i can be a noisy pixel in a functional magnetic resonance imaging (fMRI). In the literature, a variety of shrinkage priors have been proposed for Bayesian inferences of θ. Popular examples include Bayesian Lasso [23, 18], Horseshoe prior [8], Dirichlet-Laplace prior [6], Normal-Exponential-Gamma distribution [17], Generalized double Pareto distribution [2], Generalized Beta mixture of Gaussian distributions [1] and etc. A general form of shrinkage priors can be written as

$$\pi(\theta|\tau) = \prod_{i=1}^{n} \{(1/\tau)\pi_0(\theta_i/\tau)\}, \quad \tau \sim \pi(\tau), \tag{1.1}$$

where the scale parameter τ is called the global shrinkage parameter which controls the overall shrinkage effect. τ can either follow a prior distribution as in (1.1) or have a deterministic value. If furthermore, π_0 can be expressed as a scaled mixture of Gaussian distributions, (1.1) then leads to the so-called local-global shrinkage: $\theta_i \sim N(0, \lambda_i^2\tau^2)$, $\lambda_i^2 \sim \pi(\lambda_i^2), \tau \sim \pi(\tau)$ and λ_i's are called local shrinkage parameters. Note that the Dirichlet-Laplace prior is an exception and doesn't fit the general form (1.1).

Given a wide choice of shrinkage priors, certain criteria are necessary to evaluate and compare these different priors, e.g., computational efficiency and theoretical convergence. A benchmark for the theoretical performance is the posterior contraction rate, which characterizes how fast the posterior distribution (not just the Bayes estimator) converges to the true parameter. For example, the L_2 posterior contraction rate r_n satisfies:

$$\lim_{n\to\infty} E^*[\pi(\|\theta - \theta^*\| \geq r_n|y)] = 0, \text{ for any } \theta^*,$$

where E^* denotes the expectation with respect to the data generation measure of n dimensional data y under true parameter θ^*. It is well known that the frequentist minimax rate for normal means models is $\min_{\widehat{\theta}} \max_{\theta^*} \|\widehat{\theta} - \theta^*\| = \{(2+o(1))s\log(n/s)\}^{1/2}$ [11], and many Bayesian works show that the Bayesian contraction rates are comparable to this minimax rate. For example, Dirichlet-Laplace prior [6] (under certain conditions of θ^*) achieves $r_n \asymp \{s\log(n/s)\}^{1/2}$, and van der Pas, Kleijn and van der Vaart A. W. [27] showed that horseshoe prior achieves $r_n = M_n\{s\log(n/s)\}^{1/2}$ for any $M_n \to \infty$. Furthermore, recent works [e.g., 28, 16, 26] show that the tail behavior of π_0 plays an important role for posterior asymptotics, and suggest to choose a polynomial decaying π_0 in order to achieve (near-) optimal posterior contraction rate. It is worth noting that all the aforementioned shrinkage priors, except Dirichlet-Laplace prior,

have a polynomial tail. Thus we believe polynomially decaying shrinkage priors are, generally speaking, good choices for the high dimensional problem. As for the Dirichlet-Laplace prior, we notice that, although its posterior contraction is rate-minimax, its theory requires $\|\theta^*\| \leq s^{1/2}\log^2(n)$. We comment that this is actually a very strong condition. Under this condition, even the naive estimator $\widehat{\theta} = 0$ can achieve rate-$\{s^{1/2}\log^2(n)\}$ convergence, and minimax rate has merely logarithmic improvement.

Given that π_0 is polynomially decaying, existing literature focuses on how to (adaptively) choose the global shrinkage τ, i.e., the scale of the prior, such that the order of posterior contraction rate is (near-)optimal [e.g., 26, 16, 27, 29]. In this work, we will study another aspect of this story, that is how the shape of the prior distribution, i.e., the polynomial order of π_0, affects the posterior asymptotics. Our contribution of this work is two-fold. First, we show that if the polynomial order of π_0 is sufficiently close to 1, we can achieve Bayesian *sharp* minimax, i.e., $r_n/\{2s\log(n/s)\}^{1/2}$ is sufficiently close to 1. This sharp minimaxity holds for the L_1 norm as well. Our simulation study also demonstrates that it is necessary to choose a tiny polynomial order in order to obtain the optimal contraction. In Bayesian literature, the sharpness in term of multiplicative constant is barely investigated, and our work sharpens many existing results on Bayesian posterior convergence for the normal means problem. To attain such sharp minimaxity, the choice of τ will depend on true sparsity ratio (s/n) which in practice is unknown. Therefore, our second contribution is to propose a Beta modeling on τ. This leads to a Bayesian sharply minimax inference procedure that is adaptive to unknown sparsity. Simulations show that this adaptive Beta modeling has an excellent performance.

This paper is organized as follows. In Section 2, we study the relationship between the polynomial order of π_0 and the posterior contraction rate, and establish the Bayesian sharp minimax. In Section 3, we propose an adaptive modeling which doesn't depend on s. Two simulation studies and one real cancer data application are presented in Section 4. In the end, Section 5 provides more discussions and remarks. Technical proofs are provided in the Appendix.

Throughout this work, we use D_n to denote the observed data (i.e., $D_n = \{y_i\}_{i=1}^n$), and $\pi(\cdot|D_n)$ to denote the posterior based on data D_n. Given two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \succ b_n$ means $\lim(a_n/b_n) = \infty$, $a_n \asymp b_n$ means $0 < \liminf(a_n/b_n) \leq \limsup(a_n/b_n) < \infty$, and $a_n \sim b_n$ means $\lim(a_n/b_n) = 1$. $\|\cdot\|$ and $\|\cdot\|_1$ denote L_2 and L_1 norms of a vector respectively.

2. Sharp Bayesian minimaxity

As discussed in the Section 1, we consider Bayesian inferences for the sparse normal means model under a general prior specification (1.1), where π_0 has a polynomial tail; in other words, we assume the following conditions on the model sparsity and prior distribution π_0:

[C.1] The true model is sparse $s = o(n)$.

[C.2] The prior density $\pi_0(\cdot)$ is strictly decreasing on $(0, \infty)$ and increasing on $(-\infty, 0)$.

[C.3] The tail of $\pi_0(\cdot)$ is polynomially decaying with polynomial order $\alpha > 1$, i.e., there exist some positive constants M and $C_2 > C_1$ such that for any $|\theta| > M$, $C_1|\theta|^{-\alpha} \leq \pi_0(\theta) \leq C_2|\theta|^{-\alpha}$.

Note the condition C.3 (i.e., the polynomial decaying of π_0) implies the polynomial decaying of $\pi(\theta_i|\tau)$ as well: if $|\theta_i| \gg \tau$, $\pi(\theta_i|\tau) \asymp |\theta_i|^{-\alpha}\tau^{\alpha-1}$.

For the simplicity of analysis, in this section we only investigate the posterior asymptotics when the global shrinkage parameter τ is a deterministic value, under which the posteriors of θ_i's are mutually independent.

Let's first intuitively understand the posterior properties induced by a polynomial decaying prior. The posterior distributions of all θ_i's independently follow

$$\pi(\theta_i|y_i) = C\exp\{-(y_i-\theta_i)^2/2\}\pi(\theta_i),$$

for some normalizing constant C. Since both functions $\exp\{-(y_i-\theta_i)^2/2\}$ and prior $\pi(\theta_i)$ are unimodal, heuristically, the posterior $\pi(\theta_i|y_i)$ will have two major modes, around 0 and y_i respectively. The posterior mass of the mode around 0 is approximately $\pi\{\theta_i \in (-\delta,\delta)|y_i\} \approx C\exp\{-y_i^2/2\}\pi(\theta_i \in (-\delta,\delta)) \approx C\exp\{-y_i^2/2\}$ for some δ satisfying $\delta \succ \tau$; The posterior mass of the mode around y_i is approximately $\pi(\theta_i \in y_i \pm \{\delta'\log(n/s)\}^{1/2}|y_i) \approx C\int_{y_i\pm\sqrt{\delta'\log(n/s)}}\exp\{-(y_i-\theta_i)^2/2\}d\theta_i\pi(y_i) \approx C\sqrt{2\pi}\pi(y_i)$ for any small constant δ'. Therefore, if $\exp\{-y_i^2/2\} \succ \pi(y_i)$, the dominating posterior mode is the one at 0; if $\exp\{-y_i^2/2\} \prec \pi(y_i)$, the dominating posterior mode is the one at y_i. Note that in the above comparison, we consider different neighbor radiuses around 0 and y_i, this is due to the fact that the landscape of $\pi(\theta_i|y_i)$ has two unequally wide modes (refer to Figure 3 for an illustration). This heuristic comparison demonstrates a hard thresholding phenomenon for the Bayesian posterior center: when $|y_i| \ll t$, the posterior of $\pi(\theta_i|y_i)$ shrinks to 0; when $|y_i| \gg t$, the posterior will be concentrated around y_i, where the threshold value t satisfies $\exp\{-t^2/2\} \asymp \pi(t) \asymp t^{-\alpha}\tau^{\alpha-1}$. This is analogous to the hard thresholding estimator $\hat{\theta}_i = y_i 1(|y_i| \geq t)$ (refer to Figure 4 for more details). It is known that the hard thresholding estimator achieves asymptotic sharp minimaxity when $t = \{2\log(n/s)\}^{1/2}$, thus we conjecture that such sharp minimaxity carries over to the posterior mean of Bayesian hard thresholding if the same threshold value is used, i.e., the prior satisfies $\exp\{-\log(n/s)\} \asymp \tau^{\alpha-1}\{2\log(n/s)\}^{-\alpha/2}$. In addition, to derive minimax posterior contraction beyond minimax posterior mean, we also need to control the posterior variation, especially the posterior variation for these zero θ_i's. Hence, we need to impose a sufficiently strong shrinkage effect such that the posteriors of these zero θ_i's contract inside a small neighborhood around zero. Equivalently, this requires a sufficiently small global shrinkage parameter τ. Rigorously, we establish the following theorem whose proof is presented in Appendix A.1.

Theorem 2.1. *Given a positive constant ω, if $\tau^{\alpha-1} \geq (s/n)^c\{\log(n/s)\}^{1/2}$ for some $c \in (0, 1+\omega/2)$, and $\tau^{\alpha-1} \prec \{(s/n)\log(n/s)\}^{\alpha}$, then*

$$\lim E^*(\pi[\|\theta-\theta^*\| \geq C_1(\omega)\{s\log(n/s)\}^{1/2}|D_n]) = 0, \tag{2.1}$$

where $C_1(\omega) = \sqrt{2+\omega}+\sqrt{\omega}$ *and it satisfies* $\lim_{\omega\downarrow 0} C_1(\omega) = \sqrt{2}$. *If furthermore,* $\tau^{\alpha-1} \prec (s/n)^{\alpha}\{\log(n/s)\}^{(\alpha+1)/2}$, *then*

$$\lim E^*(\pi[\|\theta-\theta^*\|_1 \geq sC_2(\omega)\{\log(n/s)\}^{1/2}|D_n]) = 0, \tag{2.2}$$

where $C_2(\omega) = \sqrt{2+\omega}+\sqrt{w^2/5}+\sqrt{\omega/5}$, *and it satisfies* $\lim_{\omega\downarrow 0} C_2(\omega) = \sqrt{2}$.

We have several comments on this theorem. There are two conditions for the global shrinkage parameter τ. The first condition says that τ shall not be too small. An overly small τ may cause over-shrinkage for the true nonzero θ_i's. This condition also echoes our heuristic argument that $s/n \asymp \tau^{\alpha-1}(2\log(n/s))^{-\alpha/2}$ in the previous paragraph. The second condition says that τ shall not be too large, since an overly large τ fails to impose sufficient shrinkage effect on the zero θ_i's. To satisfy both conditions of τ, we need to choose $\alpha \in (1, 1+\omega/2)$. The theorem provides an L_1 contraction result as well. Convergence in L_1, comparing with L_2 convergence, requires stronger posterior variation control for the zero θ_i's. Note that L_1 contraction result can not be easily derived from the quantification of the posterior mean bias and posterior variance (which is the proof technique adopted by [27, 3]). Our proof technique, which directly studies the entry-wise posterior contraction for all the zero θ_i's, enables us to perform L_1 contraction analysis. There are several insights obtained from the theoretical results of this theorem. First, for any polynomially decaying prior, with proper choice of global shrinkage τ, the order of Bayesian contraction rate is exactly $O(\{s\log(n/s)\}^{1/2})$. Note that [27, 16, 28] only showed that the order of contraction is $O(M_n\{s\log(n/s)\}^{1/2})$ with $M_n \to \infty$, since their results are based on Markov's inequality. In contrast, our proof is based on constructing hypothesis testing that can test the true parameter versus balls of alternatives with exponentially small error probabilities. Secondly, the multiplicative constant of the Bayesian contraction rate is positively related to the polynomial order of $\pi_0(\cdot)$. Therefore, to obtain sharply L_2/L_1 minimax contraction,[1] i.e. the ω can be sufficiently small, a sufficient choice is to let the polynomial order of π_0 be very close to 1.

Since the logarithmic term $\log(n/s)$ is asymptotically dominated by $(n/s)^c$ for any small constant $c > 0$, the conditions of τ in Theorem 2.1 can be simplified (by replacing $\log(n/s)$ with arbitrarily small exponent of (n/s)), and we obtain the following corollary.

Corollary 2.1. *If* $\alpha \leq 1+\omega/2$ *and* $\tau^{\alpha-1} \asymp (s/n)^c$ *for some* $c \in [\alpha, 1+\omega/2)$, *then (2.1) and (2.2) hold.*

Theorem 2.1 and Corollary 2.1 suggest an optimal choice for the global shrinkage as $\tau \asymp (s/n)^{(\alpha+\delta)/(\alpha-1)}$ for some non-negative small value δ. In practice, when true sparsity is always unknown, this theoretical suggestion is not useful. Therefore in Section 3, we will discuss a full Bayesian approach which is adaptive to unknown sparsity. To end this section, we consider a possible alternative choice as $\tau \asymp (1/n)^{(\alpha+\delta)/(\alpha-1)}$, that is to substitute the unknown s with 1.

[1] The minimax L_1 contraction rate is $s\sqrt{(2+o(1))\log(n/s)}$ [10, 31].

Obviously, if we assume that the sparsity s is a fixed quantity, this choice of τ is of the same order of the optimal suggestion. But if s is increasing with n, this leads to suboptimal upper bound for the contraction rate. The details are provided in the next theorem.

Theorem 2.2. *If* $\tau^{\alpha-1} \geq (1/n)^c\{\log(n/s)\}^{1/2}$ *for some* $c \in (0, 1+\omega/2)$, *and* $\tau^{\alpha-1} \prec (s/n)^\alpha\{\log(n/s)\}^{(\alpha+1)/2}$, *then*

$$\begin{aligned} &\lim E^*(\pi[\|\theta-\theta^*\| \geq C_1(\omega)\{s\log(n)\}^{1/2}|D_n]) = 0, \\ &\lim E^*(\pi[\|\theta-\theta^*\|_1 \geq sC_2(\omega)\{\log(n)\}^{1/2}|D_n]) = 0, \end{aligned} \tag{2.3}$$

for the same functions $C_1(\omega)$ *and* $C_2(\omega)$ *used in Theorem 2.1.*

The proof of this theorem is similar to the proof of Theorem 2.1, hence is omitted in this manuscript. This result allows one to choose τ to be of order $(1/n)^{(\alpha+\delta)/(\alpha-1)}$ for any constant $\delta > 0$, and the resultant L_2 contraction rate is of order $\{s\log(n)\}^{1/2}$. The rate $\{s\log(n)\}^{1/2}$ is considered to be suboptimal in the literature. When comparing $C_1(\omega)\{s\log(n)\}^{1/2}$ versus $C_1(\omega)\{s\log(n/s)\}^{1/2}$, we have that (1) the two rates are asymptotically the same when $\log s \prec \log n$ (i.e., the sparsity s increases slower than polynomial rate); (2) the former one is of the same order with the latter, but has a large multiplicative constant, when $s \asymp n^c$ for some $c \in (0,1)$; (3) the former one is of strictly greater order, when $\log s \sim \log n$ (e.g., $s = n/\log n$). Note here we only compare the *upper bound* of posterior contraction rates obtained by Theorems 2.1 and 2.2, thus it is not rigorous to claim that the prior specification in Theorem 2.2 leads to suboptimal posterior convergence.

3. Adaptive Bayesian inference

In this section, we now consider that the information of s is not available, hence adaptive ways to determine the global shrinkage τ are necessary. In the Bayesian paradigm, there are at least two popular approaches to handle the hyperparameters: one is the empirical Bayes, i.e., to maximize the marginal likelihood of data, and the other one is the full Bayesian approach, i.e., to assign a prior distribution on the hyperparameters. For example, van der Pas, Szabo and van der Vaart [29] studied both approaches for the global shrinkage parameter in the horseshoe modeling, and established an adaptive horseshoe Bayesian inference with a suboptimal contraction rate of order $O(\{s\log n\}^{1/2})$. In this work, we are particularly interested in constructing an appropriate prior for τ.

Theorem 2.1 suggests that τ decreases to zero if τ is deterministic. This motivates us to design the prior $\pi(\tau)$ to be stochastically decreasing as n increases. More specifically, the distribution of prior $\pi(\tau)$ shrinks toward 0 under proper rate. In the meantime, this prior must not shrink too fast, such that it still assigns minimal prior density around the optimal choice $\tau \asymp (s/n)^{(\alpha+\delta)/(\alpha-1)}$. Our next theorem provides a sufficient condition for the prior $\pi(\tau)$, such that the Bayesian sharp minimaxity still holds.

Theorem 3.1. *If* $\alpha \leq 1+\omega/2$, $\pi(\tau)$ *satisfies that* $-\log \pi\{(s/n)^{(1+\omega/2)/(\alpha-1)} \leq \tau \leq (s/n)^{\alpha/(\alpha-1)}\} \prec s\log(n/s)$ *and* $-\log\pi\{\tau \geq (s/n)^{\alpha/(\alpha-1)}\} \succ s\log(n/s)$, *and* $\max_j |\theta_j^*| \leq (n/s)^{\omega/(5\alpha)}$, *then (2.1) and (2.2) still hold.*

The above theorem claims that under proper choice of $\pi(\tau)$, sharp minimaxity is still attainable when we choose α to be close to 1. Let us first discuss the two conditions on the hyper-prior $\pi(\tau)$. The first condition requires that $\pi(\tau)$ maintains a minimal prior probability on the optimal range $[(s/n)^{(1+\omega/2)/(\alpha-1)}, (s/n)^{\alpha/(\alpha-1)}]$. The second condition requires that $\pi\{\tau > (s/n)^{\alpha/(\alpha-1)}\}$ rapidly decays to 0, i.e., the distribution $\pi(\tau)$ becomes more and more concentrated around 0. A particular choice that satisfies these two conditions is that

$$\tau = \tau_0^c, \quad \tau_0 \sim \text{Beta}(1, n) \tag{3.1}$$

for some $c \in (\alpha/(\alpha-1), (1+\omega/2)/(\alpha-1))$. One can easily verify that

$$\pi\{\tau \geq (s/n)^{\alpha/(\alpha-1)}\} = (1-(s/n)^{\alpha/(\alpha-1)/c})^n \sim \exp\{-s(n/s)^{\delta}\}, \text{ and}$$
$$\pi\{(s/n)^{(1+\omega/2)/(\alpha-1)} \leq \tau \leq (s/n)^{\alpha/(\alpha-1)}\} \sim \exp\{-s(n/s)^{-\delta'}\}$$

where $\delta = 1-\alpha/(\alpha-1)/c$ and $\delta' = (1+\omega/2)/(\alpha-1)/c-1$ are two small positive constants. It is worth mentioning that the Beta prior is widely used as a hyperprior in spike-and-slab Bayesian modeling [9, 24]. For example, a common spike-and-slab modeling assigns a prior probability p for each θ_i being selected into the model, i.e. $\pi(\theta_i \neq 0) = p$, and Castillo and van der Vaart [9] suggested a hyperprior $p \sim \text{Beta}(1, 4n+1)$. In the literature, van der Pas, Szabo and van der Vaart [29] proposed a hyper truncated half Cauchy prior for the global shrinkage parameter in the horseshoe modeling, that is $\pi(\tau) \propto 1(\tau \in [1/n, 1])/(1+\tau^2)$. Both Beta modeling (3.1) and truncated half Cauchy prior have a compact support within $[0, 1]$, but a big difference between these two is that the Beta prior distribution converges to a Dirac measure at 0 as n goes to infinity, but the truncated half Cauchy prior converges to a non-degenerated distribution which is the half Cauchy distribution truncated within [0,1].

In the above theorem, we also impose an additional technical condition on the magnitude of the true nonzero θ_j such that $\log(\max|\theta_j^*|)/(\log(n/s)) \leq \omega/5\alpha$. Note that $\log(n/s) \to \infty$, hence this condition still allows that the true signal strength to grow. If the true signal grows sub-polynomially fast, i.e., $\log(\max|\theta_j^*|) = o(\log(n/s))$, then ω can be arbitrarily small and we obtain sharply minimax contraction; If the true signal grows polynomially fast, i.e., $\max|\theta_j^*| \asymp (n/s)^a$ for some constant $a > 0$, then Theorem 3.1 still ensures that the rate of contraction is $O(\{s\log(n/s)\}^{1/2})$, despite a larger multiplicative constant. This constraint on $|\theta_j^*|$ essentially is equivalent to that the prior density on true parameter (i.e., $\pi(\theta^*)$) is bounded away from 0. Similar conditions, which require that the prior is "thick" around the true parameter value θ^*, are regularly used in Bayesian theoretical literature [e.g., 19, 20, 14, 15]. As mentioned in Section 1, the Dirichlet-Laplace prior [6] also imposes an upper bound constraint on the magnitude of $\|\theta^*\|$, but our condition is much weaker.

It is worth to mention that this additional condition is mainly due to the fact that the posterior distributions among θ_i's are no longer independent when τ is subject to a prior distribution. This condition is only sufficient, and is not appealing to theoreticians. Theorem 3.7 in [29] showed that the Bayesian horseshoe with truncated half Cauchy prior on τ is capable to achieve order-$(s\log n)^{1/2}$ contraction rate without any assumption on $|\theta_j^*|$. A similar result can be derived here if one is only interested in a suboptimal upper bound for the posterior contraction rate:

Theorem 3.2. *If $\alpha \le 1+\omega/2$, the prior of τ has support on $[(1/n)^{c/(\alpha-1)}, \infty)$ for some $c<1+\omega/2$, and the $\pi(\tau)$ also satisfies that $-\log\pi\{(s/n)^{(1+\omega/2)/(\alpha-1)} \le \tau \le (s/n)^{\alpha/(\alpha-1)}\} \prec s\log(n/s)$ and $-\log\pi\{\tau \ge (s/n)^{\alpha/(\alpha-1)}\} \succ s\log(n/s)$, then (2.3) holds.*

The proof of this theorem is a combination of the proof of Theorems 2.1 and 3.1, hence is omitted. To construct a $\pi(\tau)$ that satisfies the conditions in this theorem, we can simply modify the above Beta modeling as: $\tau = \tau_0^c$ with $c \in (\alpha/(\alpha-1), (1+\omega/2)/(\alpha-1))$ and $\pi(\tau_0) \propto B(\tau_0; 1, n)1(\tau_0 \in [1/n, 1])$, where $B(\cdot; a, b)$ is the density of a Beta distribution.

4. Simulation and data anlaysis

This section, we will demonstrate two simulation studies to justify our theoretical discoveries, as well as a cancer data application. In the first simulation, we assume that the sparsity s is known in advance, and we empirically compare difference choices of the polynomial order of the prior tail. The theorems presented in Section 2 assert that it is sufficient to assign a very small α; and we would like to use numerical studies to evaluate how good is such a choice, and how necessary is this small-α condition. In the second simulation, we study the performance of the adaptive prior proposed by Theorem 3.1 when s is unknown, and compare it to the adaptive horseshoe prior proposed by [29]. We will also present a prostate cancer real data Bayesian analysis.

Simulation I: Comparison between different choices of polynomial order

In this simulation, to study the asymptotic behavior, we let the data dimension increases as $n = 50, 100, 500, 1000$ and the sparsity s equals to the rounded value of $n^{1/2}$. The nonzero coefficients are chosen to be $\theta_i^* = \{t\log(n/s)\}^{1/2}$ for $1 \le i \le s$, where $t = 1.2, 2.2, 4.2$ and 6.2. These different choices of t represent a range of levels of signal strength. To implement a class of polynomial priors with difference tail orders, we let π_0 be the t distribution with degree of freedom $\alpha - 1$, i.e.

$$\theta_i \sim \mathrm{N}(0, \lambda_i^2\tau^2);\ \lambda_i^2 \sim \mathrm{IG}((\alpha-1)/2, (\alpha-1)/2), \tag{4.1}$$

and τ follows a deterministic choice $\tau = (s/n)^c$. This leads to a simple Gibbs update

$$\lambda_j^2 \sim \text{IG}(\frac{\alpha}{2}, \frac{\alpha-1}{2} + \frac{\theta_i^2}{2\tau^2}),\ \theta_i \sim \text{N}((1+\frac{1}{\lambda_i^2\tau^2})^{-1}y_i, (1+\frac{1}{\lambda_i^2\tau^2})^{-1}). \tag{4.2}$$

We consider six different choices of prior modeling: (1) $\alpha = 1.1$, $c = (\alpha + 0.05)/(\alpha - 1)$; (2) $\alpha = 2.1$, $c = (\alpha + 0.05)/(\alpha - 1)$; (3) $\alpha = 3.1$, $c = (\alpha + 0.05)/(\alpha - 1)$; (4) $\alpha = 2.1$, $c = 1/(\alpha - 1)$; (5) $\alpha = 3.1$, $c = 1/(\alpha - 1)$; (6) horseshoe prior with global shrinkage $\tau = (s/n)\{\log(n/s)\}^{1/2}$. In the above five t-prior specifications, there are two choices for global shrinkage τ. One is $(s/n)^{(\alpha+0.05)/(\alpha-1)}$, which satisfies the upper bound condition on τ in Theorem 2.1. By our heuristic arguments in Section 2, under such prior specification, $\theta_i^* \approx \{2(\alpha+0.05)\log(n/s)\}^{1/2}$ posts the most difficult problem. The other choice is $(s/n)^{1/(\alpha-1)}$ which satisfies the lower bound condition on τ in Theorem 2.1. Note that Ghosh and Chakrabarti [16] claimed that an optimal choice for τ is $[(s/n)\{\log(n/s)\}^{1/2}]^{1/(\alpha-1)}$. The difference between $(s/n)^{1/(\alpha-1)}$ and $[(s/n)\{\log(n/s)\}^{1/2}]^{1/(\alpha-1)}$ is only a logarithmic term. The horseshoe prior has a polynomial tail with order $\alpha = 2$, and the choice of its τ follows the suggestion of [27]. And asymptotically, this horseshoe prior is almost the same to t-distribution with $\alpha = 2.1$ and $c = 1/(\alpha - 1)$. All simulation results are based on the average of 100 replications.

In Figure 1, we compare the posterior contraction among these 6 priors. We estimate their posterior probability $\pi(\|\theta-\theta^*\|^2 \geq 2.2s\log(n/s)|D_n)$ by posterior samples, and plot this probability with respect to the different n and t values. For a minimax Bayesian procedure, this probability converges to 0 as n increases, regardless of the magnitude of θ^*. The figure clearly indicates that $\alpha = 1.1$ has the best performance. The posterior probability always decreases toward 0 for all different t's. For the rest 5 priors, their posteriors don't contract into the $\{2.2s\log(n/s)\}^{1/2}$-neighborhood for some value of t. It is worth to mention that for the t prior with $\alpha = 1.1$, the case $t = 2.2$ leads to the slowest convergence, as the red curve is decreasing very slowly. This is because, as we discussed, $t = 2.2$ corresponds to the most difficult scenario for $\alpha = 1.1$. Besides, the plots of horseshoe and t-distribution with $\alpha = 2.1$ and $c = 1/(\alpha-1)$ have very similar patterns, since these two prior specifications are almost equivalent in terms of their tail behaviors.

In Figure 2, we present the some comparisons of the posterior mean of squared L_2 error $E(\|\theta - \theta^*\|^2|D_n)$ and posterior mean of L_1 error $E(\|\theta - \theta^*\|_1|D_n)$. As a reference for the comparison, we also plot the curves corresponding the the minimax squared L_2 and L_1 errors, namely, $2s\log(n/s)$ and $s\{2\log(n/s)\}^{1/2}$ respectively. Note that when $s = n^{1/2}$, the suboptimal contraction rate $2s\log(n) = 2(2s\log(n/s))$ and $s\{2\log n\}^{1/2} = \{2\}^{1/2}s\{2\log(n/s)\}^{1/2}$. When $t = 1.2$, i.e., signals are weak, the L_2 errors of all 4 priors don't exceed the minimax rate. This is not surprising, because under weak signals, any method that imposes enough shrinkage effect, including the naive estimator $\hat{\theta} = 0$, will induce an L_2 error that is smaller than the minimax error. It also shows that the t-prior with

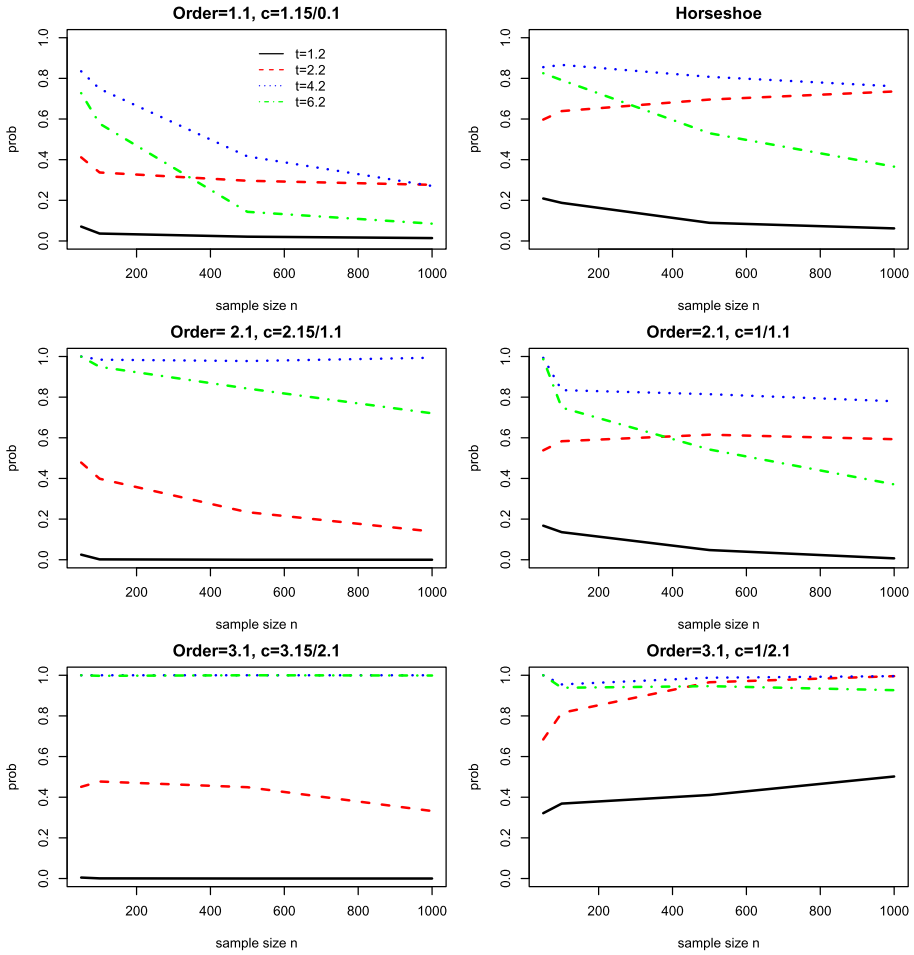

FIG 1. *Posterior contraction of the 6 prior specifications.*

$\alpha = 2.1$, $c = (\alpha + 0.05)/(\alpha - 1)$ does have a better L_2 error than t-prior with $\alpha = 1.1$ under weak signals. When $t = 2.2$, i.e., the signal strength is in the boundary case, horseshoe and t-prior with $c = 1/(\alpha - 1)$ begin to exceed the minimax L_2 error, while the two t-priors with $c = (\alpha + 0.05)/(\alpha - 1)$ have L_2 errors that is almost the same as, but slightly higher than, the minimax error. When $t = 4.2$, i.e., signals are strong, t-prior with $\alpha = 1.1$ is the only one that achieves asymptotic sharp minimaxity. When t is even larger (which is not presented in the Figure 2), the L_2 error of t-prior with $\alpha = 1.1$ will be much smaller than the minimax rate, but the L_2 errors for the rest three are much larger than the minimax rate. In summary, a small polynomial order universally ensures that the L_2 estimation error is asymptotically bounded by the minimax rate. As for the error rates under L_1 norm, it is much more sensitive to small varia-

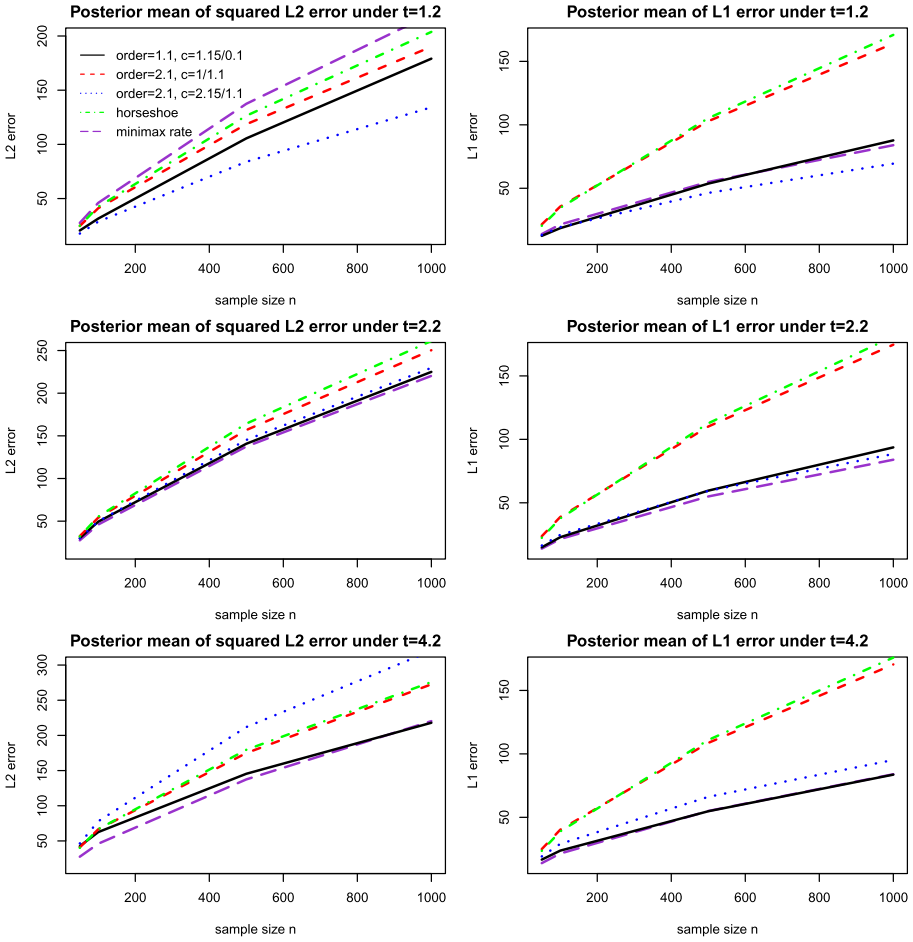

FIG 2. *Posterior mean sqaured L_2 and L_1 errors of the 4 selective prior specifications.*

tions in the coordinates than the L_2 error. The choice $\tau = (s/n)^{(\alpha)+0.05/(\alpha-1)}$ satisfies the upper bound condition on τ in Theorem 2.1, and as we discussed, guarantees sufficient posterior shrinkage for the zero θ_i's, hence it leads to small L_1 error. This argument is consistent to our simulation results: the two priors with choice $c = (\alpha+0.05)/(\alpha-1)$ have much smaller L_1 errors. The same holds for the comparison of L_2 errors, only the small polynomial order prior ($\alpha = 1.1$) ensures sharp minimaxity. Similar to Figure 1, we see that the t-distribution with $\alpha = 2.1$ and $c = 1/(\alpha - 1)$ has almost identical performance with the horseshoe prior.

The above simulation results successfully demonstrate the sharpness of the Bayesian minimax contraction when the polynomial order α of the prior is close to 1. However, there are still some discrepancies between the displayed finite-

sample behaviors and the Bayesian hard thresholding phenomenon described in Section 2. According to the Bayesian hard thresholding phenomenon, when $|y_i|$ is greater than the thresholding value $\sqrt{2\alpha \log(n/s)}$, its posterior will have one dominating mode which is approximately $\mathrm{N}(y_i, 1)$. This implies that when $t = 4.2 > 2\alpha = 2.2$, the posterior mean squared L_2 error or L_1 error of those nonzero θ_i's is asymptotically of order $O(s)$, which thus shall lead to a much smaller estimation error comparing with the boundary case of $t = 2.2$. But in Figure 2, there is no noticeable difference for posterior mean errors between $t = 2.2$ and $t = 4.2$ under the t prior with $\alpha = 1.1$. This is because our asymptotic theory relies on the sparsity assumption such that $\log(n/s) \to \infty$, while in our simulation experiments, the ratio of n/s is not large enough to reflect the asymptotic behavior. To illustrate it, Figure 3 plots the histograms of posterior $\pi(\theta_i | y_i = [3.2 \log(n/s)]^{1/2})$ using t-prior with $\alpha = 1.1$ and $\tau = (s/n)^{1.15/0.1}$, for different values of n/s. As showed, the posterior does have two modes, but the mode around y_i isn't the dominating one until n/s is as huge as 100,000. In contrast, other choices of prior with different polynomial order of tail decaying, for example, the horseshoe prior requires a much smaller value of n/s to have the mode around y_i be dominating. This implies that in a small-sample real application, prior with polynomial order close to 1 is less powerful in terms of detecting signals, i.e., yields a sparser model selection result, comparing with the horseshoe prior. But on the other hand, horseshoe prior specification fails to induce sufficient shrinkage effect for the zero θ_i's, therefore its estimation performance for the whole vector θ is still inferior. For t-prior with $\alpha = 1.1$, we also plot its posterior mean shrinkage coefficient $E(\theta_i|y_i)/y_i$ with respect to values of $c = y_i^2/\log(n/s)$ and n/s in Figure 4. As n/s tends to infinity, the posterior shrinkage does behave more and more similarly to the hard thresholding.

In additional, several simulation experiments are presented in the Supplementary Material, where we explore (1) the posterior contraction of shrinkage prior under varying nonzero θ's scenario, (2) the posterior convergence performances for nonzero θ's and zero θ's respectively, and (3) the uncertainty quantification and Bayesian model selection of shrinkage priors.

Simulation II: Adaptive Bayesian modeling and comparisons

In the second simulation, we no longer assume that s is known, and now the global shrinkage τ is chosen in an adaptive Bayesian manner. We compare the following two adaptive Bayesian procedures. The first prior is constructed based on Theorem 3.1. We consider a t-prior with $\alpha = 1.1$, and τ follows the Beta modeling (3.1) with $c = (\alpha + 0.05)/(\alpha - 1)$. As for the posterior sampling of this adaptive t-prior, in addition to (4.2), the marginal condition distribution of τ_0 is

$$\pi(\tau_0|\text{rest}) \propto \frac{1}{\tau_0^{cn}} \exp\left\{ -\frac{1}{\tau_0^{2c}} \sum_i \frac{\theta_i^2}{2\lambda_i^2} \right\} \frac{(1-\tau_0)^{n-1}}{n+1} 1(\tau_0 \in (0,1)).$$

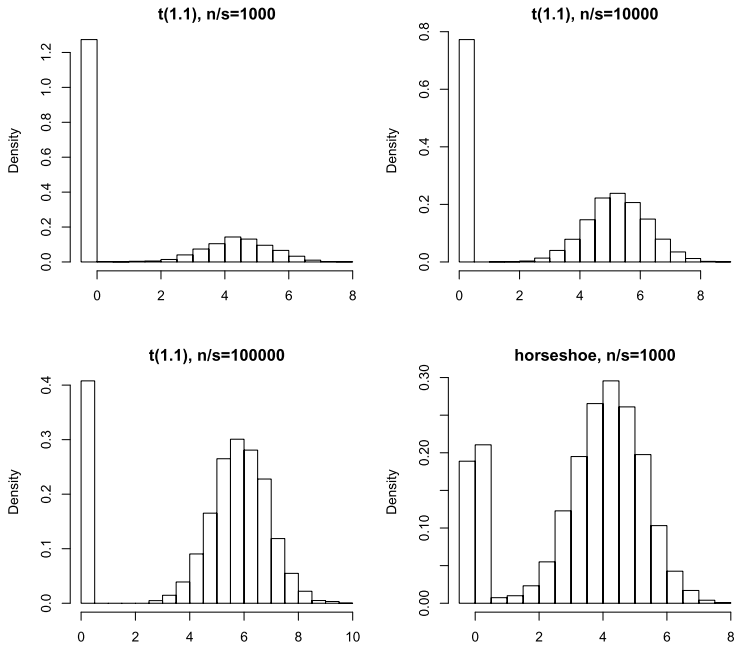

FIG 3. *Histograms for the posterior* $\pi(\theta_i|y_i = (3.2\log(n/s))^{1/2})$.

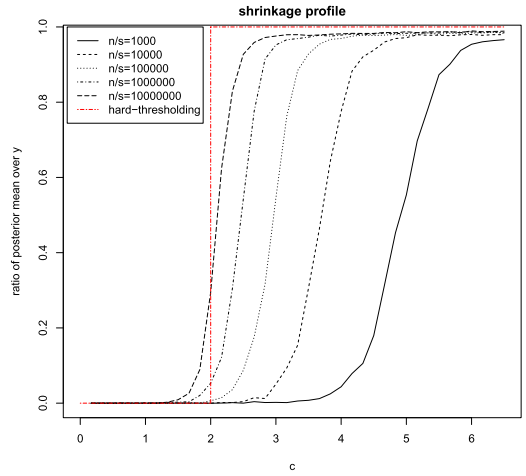

FIG 4. *Posterior shrinkage profile* $E(\theta_i|y_i)/y_i$ *under t-prior with* $\alpha = 1.1$ *and* $\tau = (s/n)^{1.15/0.1}$.

Since this conditional posterior has a compact support, it can be sampled via the inverse cumulative-distribution sampling. Another prior is the horseshoe prior with τ following a half-Cauchy distribution truncated on $[1/n, 1]$ [29]. The

simulation settings for data dimension and signal strength are exactly the same as the first simulation study. Obviously, one shall expect that the performances of adaptive Bayesian approaches are worse than the case that s is known.

Figure 5 demonstrates a comprehensive comparison between adaptive t-prior and adaptive horseshoe prior, in terms of posterior contraction, posterior mean square L_2 error and posterior mean L_1 error. It shows that the adaptive t-prior has a better performance in almost every aspect. The posterior contraction plots of t-prior always have a decreasing trend towards 0 under different signal strengths, and its convergence pattern is very similar to Figure 1. This implies that the Beta modeling of τ is a good substitute for the optimal choice $\tau = (s/n)^{(\alpha+0.05)/(\alpha-1)}$. The posterior contraction of the adaptive horseshoe, on the other side, doesn't converge at all. The plot pattern is also quite different from the posterior contraction plot in Figure 1, especially for the case $t = 6.2$. This somehow indicates that the truncated half-Cauchy prior doesn't adapt well to large signals. For the posterior mean L_2 error of the adaptive t-prior, when the signal is weak or strong, its error is well bounded by minimax rate if n is large. When the signal strength is moderate, its error slightly exceeds the minimax rate. However, there is no trend showing that the L_2 error will increase faster than the minimax rate, thus we believe that if n continues growing and α is closer to 1, the error of adaptive t-prior should be asymptotically bounded by the minimax rate. But the adaptive horseshoe prior induces a much larger error than the minimax rate, except for the weak signal situation. Similarly, in term of the L_1 norm, the adaptive t-prior attains the minimax rate, and it clearly outperforms the horseshoe prior regardless of the signal strength.

As a conclusion, the presented two simulation studies demonstrate the necessity of choosing α to be sufficiently close to 1. A prior specification as simple as t-distribution with a tiny degree of freedom ensures supreme Bayesian contraction and estimation. The proposed adaptive Beta modeling on the global shrinkage τ leads to a very stable result and significantly outperforms the adaptive horseshoe prior.

Real data set analysis

We consider a popular prostate cancer dataset [12, 25] from a microarray experiment which consists of expression levels for $n = 6033$ genes from 50 normal control subjects and 52 cancer patients.[2] Two-sample tests are performed to compare the expression level of each gene between control and patient groups. The corresponding p-values are thereafter converted into z-statistics, i.e., $z_i = \Phi^{-1}(p_i/2)$ for $i = 1, \ldots, n$. Hence, it is appropriate to model these z-statistics as a normal means model $z_i = \theta_i + \epsilon_i$, where $\theta_i = 0$ if the mean expression levels for the ith gene are the same between control and patient population. We use Bayesian shrinkage to make inference on the parameter θ.

We implement the adaptive t prior with $\alpha = 1.1$ and τ following the Beta modeling (3.1), and the adaptive horseshoe prior with τ following truncated

[2]The data set is available in the book [13].

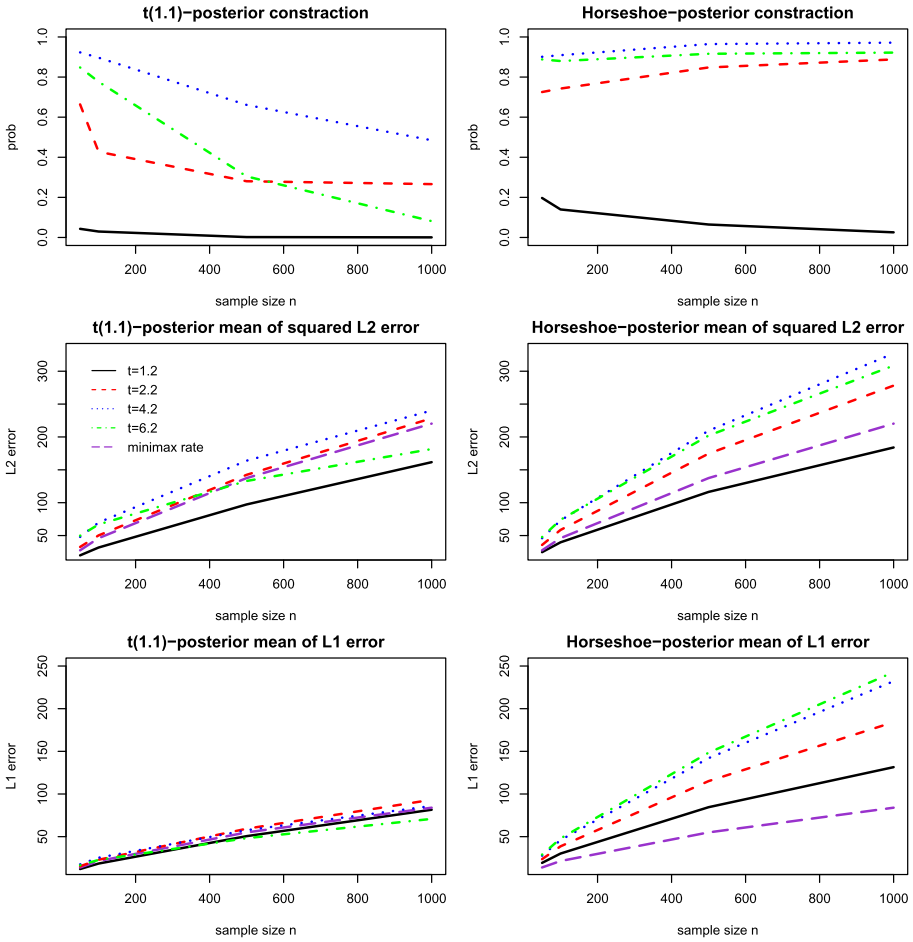

FIG 5. *Comparison between adaptive t-prior and adaptive horseshoe prior.*

half Cauchy prior. Note that horseshoe prior has already been used to analyze this prostate cancer data in the literature [6, 3, 5], but all these applications choose τ to follow the non-truncated half Cauchy prior. As illustrated by [29], the empirical performance between truncated half Cauchy hyper-prior and non-truncated half Cauchy hyper-prior are quite different, hence the horseshoe posterior summary presented in this section is not comparable to the results in the literature.

In Figure 6, we plot the posterior means $|E(\theta_i|D_n)|$ against the observations $|y_i|$. For larger $|y_i|$'s, the posterior means between adaptive t-prior and adaptive horseshoe prior are close. For smaller $|y_i|$'s, the adaptive t-prior apparently induces stronger shrinkage effect. This observation is consistent to our simulations results that horseshoe prior imposes insufficient posterior shrinkage

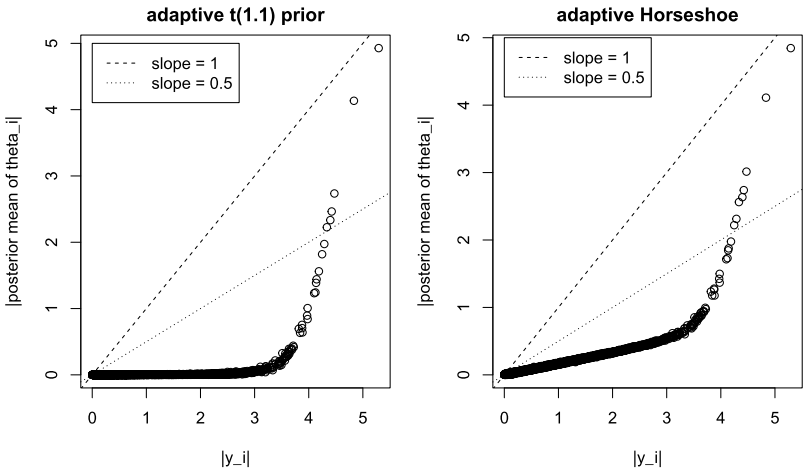

FIG 6. *Comparison between adaptive t-prior and adaptive horseshoe prior.*

for the zero θ_i's. If we perform a variable selection by selecting genes for which $|E(\theta_i|D_n)|/|y_i| > 1/2$, then horseshoe selects the top 8 genes and t-prior selects the top 6 genes. This is also consistent to our previous arguments that the t-prior with polynomial order close to 1 is less powerful than horseshoe prior. Note that the posterior of $\pi(\theta_i|D_n)$ has two major modes around 0 and y_i, hence the selection rule $|E(\theta_i|D_n)|/|y_i| > 1/2$ heuristically means that the posterior mass for the mode around y_i is greater than half.

5. Final remarks

In this work, we study the Bayesian inference on high dimensional sparse normal sequence model with a polynomially decaying prior distribution. Our main result Theorem 2.1 reveals the connection between the upper bound of the posterior contraction and the polynomial order α. This provides a sufficient condition to induce sharp posterior minimaxity. That is, choosing a sufficiently tiny $\alpha - 1$, the ratio between Bayesian posterior contraction rate and minimax will be sufficiently close to 1. We conjecture necessity holds for Theorem 2.1 as well, such that the smaller the $\alpha - 1$, the better the Bayesian contraction in terms of the multiplicative constant. Empirical studies also show great improvement for the accuracy of the Bayesian shrinkage procedure using a t-prior with a tiny degree of freedom. Our study considers α to be a fixed, sufficiently small hyperparameter. Alternatively, one can investigate the choice of letting the polynomial order α decrease to 1 as n increases, i.e., $\lim \alpha_n = 1$ under proper rate. Another related question will be: is it a good choice to use an improper prior with exactly $\alpha = 1$, e.g. $\pi_0(\theta) \propto \theta^{-1}$. Our theoretical results break down when $\alpha = 1$ (the term $\alpha - 1$ appears in the denominator), and our technical tool which follows the arguments of Le Cam-Birgé testing theory [7, 4, 22] only works for proper prior specifications.

The primary research interest of this paper is on the L_2 and L_1 posterior contraction rates. Another important research objective is sparsity recovery, i.e., to identify the set $\{j : \theta_j^* \neq 0\}$. Given a continuous posterior distribution induced by a shrinkage prior, one easy way to perform model selection is to do a threshold truncation, that is, a variable is selected if its posterior summary such as posterior mean is greater than some thresholding value. This simply approach has been widely used in the literature, however, it usually leads to over-selection with the number of false positives being of order of $O(s)$ [e.g., Theorem 3.4 of 6]. Another different model selection approach is to select θ_i's whose marginal credible intervals exclude 0, and the consistency of this Bayesian selection method is investigated by [30] for the horseshoe prior.

This work focus on the normal sequence model, thus it would be of substantial interest to conduct similar investigation for a more general regression model. Our results heavily rely on the independence among y_i's, and it is not trivial to extend these results to regression model with correlated design matrix. Song and Liang [26] studies the posterior asymptotics for the general linear regression model, including order-$(s \log p/n)^{1/2}$ L_2 contraction and model selection consistency, when a polynomially decaying prior is used. We believe that the choice of polynomial order also plays a role for the multiplicative constant of the posterior contraction rate under regression model, and we conjecture that the optimal choice of α will depend on the eigenstructure of the design matrix. If the design matrix X is nearly orthogonal, e.g., all entries of X follow independent Gaussian distribution, we conjecture that the same results as Theorem 2.1 will still hold, and one needs to choose $\alpha \approx 1$ in order to obtain optimal Bayesian contraction.

Acknowledgements

Dr. Song's research activities are partially supported by National Science Foundation DMS-1811812.

Appendix A: Technical proofs

A.1. Proof of Theorem 2.1

The proof consists of two parts. Since the posteriors of θ_i's are independent, in Part I, we study the posterior contraction for the nonzero θ_i's and in Part II, we study the posterior contraction for the zero θ_i's. First, let us state some useful lemmas.

Lemma A.1 (Lemma 1 of [21])**.** *Let $\chi_d^2(\kappa)$ be a chi-square random variable with degree of freedom d and noncentral parameter κ, then we have the following concentration inequality*

$$\Pr(\chi_d^2(\kappa) < d + \kappa - \{(4d + 8\kappa)x\}^{1/2}) \leq \exp(-x),$$

for any $x > 0$.

Lemma A.2 (Theorem 1 of [32])**.** *Let X be a Binomial random variable $X \sim B(n, v)$. For any $1 < k < n - 1$*

$$\Pr(X \geq k+1) \leq 1 - \Phi(sign(k-nv)\{2nH(v, k/n)\}^{1/2}),$$

where Φ is the cumulative distribution function of standard Gaussian distribution and $H(v, k/n) = (k/n)\log(k/nv) + (1-k/n)\log[(1-k/n)/(1-v)]$.

The next lemma is a refined result of Lemma 6 in [4]:

Lemma A.3. *Let f^* be the true probability density of data generation, f_θ be the likelihood function with parameter $\theta \in \Theta$, and E^*, E_θ denote the corresponding expectation respectively. Let B_n and C_n be two subsets of the parameter space Θ, and ϕ_n be some testing function satisfying $\phi_n(D_n) \in [0,1]$ for any realization D_n of the data generation. If $\pi(B_n) \leq b_n$, $E^*\phi(D_n) \leq b'_n$, $\sup_{\theta \in C_n} E_\theta(1-\phi(D_n)) \leq c_n$, and*

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

where $m(D_n) = \int_\Theta \pi(\theta) f_\theta(D_n) d\theta$ is the margin density of D_n, then,

$$E^*\left(\pi(C_n \cup B_n)|D_n\right)\right) \leq \frac{b_n + c_n}{a_n} + a'_n + b'_n.$$

Proof. Define Ω_n be the event of $(m(D_n))/(f^*(D_n)) \geq a_n$, and $m(D_n, C_n \cup B_n) = \int_{C_n \cup B_n} \pi(\theta) f_\theta(D_n) d\theta$. Then

$$\begin{aligned}
&E^*\pi(C_n \cup B_n)|D_n) = E^*\pi(C_n \cup B_n)|D_n)(1-\phi(D_n))1_{\Omega_n} \\
+&E^*\pi(C_n \cup B_n)|D_n)(1-\phi(D_n))(1-1_{\Omega_n}) + E^*\pi(C_n \cup B_n)|D_n)\phi(D_n) \\
\leq&E^*\pi(C_n \cup B_n)|D_n)(1-\phi(D_n))1_{\Omega_n} + E^*(1-1_{\Omega_n}) + E^*\phi(D_n) \\
\leq&E^*\pi(C_n \cup B_n)|D_n)(1-\phi(D_n))1_{\Omega_n} + b'_n + a'_n \\
\leq&E^*\{m(D_n, C_n \cup B_n)/a_n f^*(D_n)\}(1-\phi(D_n)) + b'_n + a'_n.
\end{aligned}$$

By Fubini theorem,

$$\begin{aligned}
&E^*(1-\phi(D_n))m(D_n, C_n \cup B_n)/f^*(D_n) \\
=&\int_{C_n \cup B_n} \int_{\mathcal{X}} [1-\phi(D_n)] f_\theta(D_n) dD_n \pi(\theta) d\theta \\
\leq&\int_{C_n} E_\theta(1-\phi(D_n))\pi(\theta)d\theta + \int_{B_n}\int_{\mathcal{X}} f_\theta(D_n) dD_n \pi(\theta) d\theta \leq b_n + c_n.
\end{aligned}$$

Combining the above inequalities leads to the conclusion. □

Part I: posterior contraction rate of nonzero θ_i's

It is equivalent to consider the situation that $\tilde{y} = \vartheta_1 + \epsilon$, where $\tilde{y}, \vartheta_1 \in \mathbb{R}^s$, $\epsilon \sim N(0, I_s)$. The parameter ϑ_1 is subject to prior $\prod_{j=1}^s \pi(\vartheta_{1,j})$. The parameter

ϑ_1 hence correspeonds to the subvector of θ: $\vartheta_1 = (\theta_j)_{j\in\{\theta_j^*\neq 0\}}$, and its true value is $\vartheta_1^* = (\theta_j^*)_{j\in\{\theta_j^*\neq 0\}}$. We want to show that

$$E^*\pi(\|\vartheta_1 - \vartheta_1^*\| \geq \{(2+\omega)s\log(n/s)\}^{1/2}|\tilde{y}) \leq \exp\{-c_0 s\log(n/s)\}, \tag{A.1}$$

for some positive constant c_0.

Let's consider the testing function $\phi(\tilde{y}) = 1(\|\tilde{y} - \vartheta_1^*\| \geq \{\delta s\log(n/s)\}^{1/2})$ where δ is a positive but tiny constant. This testing function satisfies

$$E_{\vartheta_1^*}\phi(\tilde{y}) = \Pr(\chi_s^2 \geq \delta s\log(n/s)) \leq \exp\left\{-\frac{\delta s\log(n/s)}{2+\delta_0}\right\} \tag{A.2}$$

where E_{ϑ_1} denotes the expectation over $\tilde{y}$ with respect to true parameter being ϑ_1, and the last inequality holds for any fixed $\delta_0 > 0$ when n is sufficiently large due to Lemma A.1. And for any $\vartheta_1 \in C_n = \{\vartheta_1 \in \mathbb{R}^s : \|\vartheta_1 - \vartheta_1^*\| \geq \{(2+\omega)s\log(n/s)\}^{1/2}\}$,

$$\begin{aligned}
&E_{\vartheta_1}[1-\phi(\tilde{y})] = \Pr(\|\tilde{y} - \vartheta_1^*\| \leq \{\delta s\log(n/s)\}^{1/2}|\vartheta_1)\\
\leq&\Pr(\|\tilde{y} - \vartheta_1\| \geq \|\vartheta_1 - \vartheta_1^*\| - \{\delta s\log(n/s)\}^{1/2}|\vartheta_1)\\
=&\Pr(\chi_s^2 \geq [\|\vartheta_1 - \vartheta_1^*\| - \{\delta s\log(n/s)\}^{1/2}]^2)\\
\leq&\exp\left\{-\frac{[\|\vartheta_1 - \vartheta_1^*\| - \{\delta s\log(n/s)\}^{1/2}]^2}{2+\delta_0}\right\} \leq \exp\left\{-\frac{\|\vartheta_1 - \vartheta_1^*\|^2}{2+2\delta_0}\right\},
\end{aligned} \tag{A.3}$$

where the last inequality holds as δ is sufficiently small. Denote $\Delta\vartheta_1 = \vartheta_1 - \vartheta_1^*$. With probability at least $1 - \exp\{-cs\log(n/s)\}$ for some positive c, $\|\epsilon\|^2 \leq \eta s\log(n/s)/2$ and

$$\begin{aligned}
\frac{m(\tilde{y})}{f^*(\tilde{y})} &= \int \exp\left(-\frac{-\|\epsilon\|^2 + \|\Delta\vartheta_1 + \epsilon\|^2}{2}\right)\pi(\vartheta_1)d\vartheta_1\\
&\geq \exp\left\{-\frac{\eta s\log(n/s)}{2}\right\}\pi(\|\Delta\vartheta_1\| \leq \{\eta s\log(n/s)/2\}^{1/2})
\end{aligned} \tag{A.4}$$

for any fixed small constant $\eta > 0$, where $m(\tilde{y}) = \int f(\tilde{y};\vartheta_1)\pi(d\vartheta_1)$ is the marginal likelihood of data $\tilde{y}$.

Therefore, let Ω_n be the event that (A.4) holds, by (A.2)–(A.4), we have

$$\begin{aligned}
&E^*\pi(C_n|\tilde{y})\\
=&E^*\pi(C_n|\tilde{y})(1-\phi(\tilde{y}))1_{\Omega_n} + E^*\pi(C_n|\tilde{y})(1-\phi(\tilde{y}))(1-1_{\Omega_n}) + E^*\pi(C_n|\tilde{y})\phi(\tilde{y})\\
\leq&E^*\pi(C_n|\tilde{y})(1-\phi(\tilde{y}))1_{\Omega_n} + E^*(1-1_{\Omega_n}) + E^*\phi(\tilde{y})\\
\leq&E^*\pi(C_n|\tilde{y})(1-\phi(\tilde{y}))1_{\Omega_n} + \exp\{-cs\log(n/s)\} + \exp\{-\delta s\log(n/s)/(2+\delta_0)\}\\
\leq&\frac{E^*\int_{C_n} f(\tilde{y};\vartheta_1)(1-\phi(\tilde{y}))/f^*(\tilde{y})\pi(d\vartheta_1)}{\exp\left\{-\frac{\eta s\log(n/s)}{2}\right\}\pi(\|\Delta\vartheta_1\| \leq \{\eta s\log(n/s)/2\}^{1/2})}\\
&+\exp\{-cs\log(n/s)\} + \exp\{-\delta s\log(n/s)/(2+\delta_0)\}.
\end{aligned}$$

And the first term in the above equation satisfies

$$\begin{aligned}
&\frac{E^* \int_{C_n} f(\tilde{y};\vartheta_1)(1-\phi(\tilde{y}))/f^*(\tilde{y})\pi(d\vartheta_1)}{\exp\left\{-\frac{\eta s\log(n/s)}{2}\right\}\pi(\|\Delta\vartheta_1\| \le \{\eta s\log(n/s)/2\}^{1/2})} \\
=&\frac{\int\int_{C_n} f(\tilde{y};\vartheta_1)(1-\phi(\tilde{y}))\pi(d\vartheta_1)d\tilde{y}}{\exp\left\{-\frac{\eta s\log(n/s)}{2}\right\}\pi(\|\Delta\vartheta_1\| \le \{\eta s\log(n/s)/2\}^{1/2})} \\
=&\frac{\int_{C_n} E_{\vartheta_1}(1-\phi(\tilde{y}))\pi(d\vartheta_1)}{\exp\left\{-\frac{\eta s\log(n/s)}{2}\right\}\pi(\|\Delta\vartheta_1\| \le \{\eta s\log(n/s)/2\}^{1/2})} \\
\le&\frac{\int_{C_n}\exp\{-\|\Delta\vartheta_1\|^2/(2+2\delta_0)\}\pi(d\vartheta_1)}{\pi(\|\Delta\vartheta_1\| \le \{\eta s\log(n/s)/2\}^{1/2})}\exp\left\{\frac{\eta s\log(n/s)}{2}\right\}.
\end{aligned}$$

Let us now study the quantity $\int_{C_n}\exp\{-\|\Delta\vartheta_1\|^2/(2+2\delta_0)\}\pi(d\vartheta_1)/\pi(\|\Delta\vartheta_1\| \le \{\eta s\log(n/s)/2\}^{1/2})$. Let $T_1 = \{j; |\vartheta^*_{1,j}| \ge 1\}$ and $T_2 = \{j; |\vartheta^*_{1,j}| < 1\}$, T_3 be the generic notation for a subset of T_1, $T_4 = T_1 \backslash T_3$ and $t_i = |T_i|$. ϑ_{1,T_i} denotes the subvector of ϑ_1 corresponding to T_i. Decompose $C_n = \cup_{T_3\subset T_1} C_n^{T_3} := \cup_{T_3\subset T_1}(C_n \cap \{\vartheta_1 : \{j : |\vartheta_{1,j}| < 1\} = T_3\})$. Then, we have

$$\begin{aligned}
&\frac{\int_{C_n}\exp\{-\|\Delta\vartheta_1\|^2/(2+2\delta_0)\}\pi(d\vartheta_1)}{\pi(\|\Delta\vartheta_1\| \le \{\eta s\log(n/s)/2\}^{1/2})} \\
\le&\frac{\int_{C_n}\exp\{-\|\Delta\vartheta_1\|^2/(2+2\delta_0)\}\pi(d\vartheta_1)}{\pi(\|\Delta\vartheta_{1,j}\| \le \{\eta\log(n/s)/2\}^{1/2} \text{ for all } j=1,\dots,s)} \\
\le&\frac{\int_{C_n}\exp\{-\|\Delta\vartheta_1\|^2/(2+2\delta_0)\}\prod_{j\in T_1}\pi(\vartheta_{1,j})\prod_{j\in T_2}\pi(\vartheta_{1,j})d\vartheta_1}{\prod_{j\in T_1}\pi(\vartheta^*_{1,j})\prod_{j\in T_2}\pi(\vartheta_{1,j}\in[-1,1])} \\
=&\sum_{T_3\subset T_1}\frac{\int_{C_n^{T_3}}\exp\{-\frac{\|\Delta\vartheta_1\|}{2+2\delta_0}\}\prod_{j\in T_4}\frac{\pi(\vartheta_{1,j})}{\pi(\vartheta^*_{1,j})}\prod_{j\in T_3}\frac{1}{\pi(\vartheta^*_{1,j})}\prod_{j\in T_2\cup T_3}\pi(\vartheta_{1,j})d\vartheta_1}{\prod_{j\in T_2}\pi(\vartheta_{1,j}\in[-1,1])},
\end{aligned} \tag{A.5}$$

where the second inequality holds asymptotically, because $\log(n/s)$ is sufficiently large and $\pi(|\cdot|)$ is a decreasing function.

By C.3, if $|\vartheta_{1,j}|$ and $|\vartheta^*_{1,j}|$ are both larger than 1, then $|\log(\pi(\vartheta_{1,j})/\pi(\vartheta^*_{1,j}))| \le |\log(C_2/C_1)| + \alpha|\log|\vartheta_{1,j}| - \log|\vartheta^*_{1,j}|| \le |\log(C_2/C_1)| + \alpha|\vartheta_{1,j} - \vartheta^*_{1,j}|$. If $|\vartheta^*_{1,j}| \ge 1$, then we have that $\log[1/\pi(\vartheta^*_{1,j})] \le \alpha\log|\vartheta^*_{1,j}| + (\alpha-1)\log(1/\tau) - \log C_1 \le \alpha|\vartheta^*_{1,j}| + (\alpha-1)\log(1/\tau) - \log C_1$. Using these facts, for any $T_3 \subset T_1$,

$$\begin{aligned}
&\frac{\int_{C_n^{T_3}}\exp\{-\frac{\|\Delta\vartheta_1\|}{2+2\delta_0}\}\prod_{j\in T_4}\frac{\pi(\vartheta_{1,j})}{\pi(\vartheta^*_{1,j})}\prod_{j\in T_3}\frac{1}{\pi(\vartheta^*_{1,j})}\prod_{j\in T_2\cup T_3}\pi(\vartheta_{1,j})d\vartheta_1}{\prod_{j\in T_2}\pi(\vartheta_{1,j}\in[-1,1])} \\
\le&\frac{\int_{C_n^{T_3}}\exp\{-\frac{\|\Delta\vartheta_1\|^2}{2+2\delta_0}+\alpha\|\Delta\vartheta_{1,T_4}\|_1+t_4\log\frac{C_2}{C_1}+\alpha\|\vartheta^*_{1,T_3}\|_1\}\prod_{j\in T_3\cup T_2}\pi(\vartheta_{1,j})d\vartheta_1}{C_1^{t_3}[\tau^{\alpha-1}]^{t_3}[1-2C_2(1/\tau)^{-(\alpha-1)}]^{t_2}}.
\end{aligned} \tag{A.6}$$

Since for any $j \in T_1$, $\Delta\vartheta_{1,j}^2 - (2+2\delta_0)\alpha|\Delta\vartheta_{1,j}| = (|\Delta\vartheta_{1,j}| - (1+\delta_0)\alpha)^2 - (1+\delta_0)^2\alpha^2$, we have that

$$\begin{aligned}
&\frac{\exp\{-\frac{\|\Delta\vartheta_1\|^2}{2+2\delta_0} + \alpha\|\Delta\vartheta_{1,T_4}\|_1 + t_4\log\frac{C_2}{C_1} + \alpha\|\vartheta^*_{1,T_3}\|_1\}}{C_1^{t_3}[\tau^{\alpha-1}]^{t_3}[1-2C_2(1/\tau)^{-(\alpha-1)}]^{t_2}} \\
\le&\frac{\exp\{-\frac{[\|\Delta\vartheta_1\| - s^{1/2}(1+\delta_0)\alpha]^2 - s(1+\delta_0)^2\alpha^2}{2+2\delta_0} + t_4\log\frac{C_2}{C_1} + \alpha t_3\}}{C_1^{t_3}[\tau^{\alpha-1}]^{t_3}[1-2C_2(1/\tau)^{-(\alpha-1)}]^{t_2}} \\
\le&\frac{\exp\{-\frac{[\{(2+\omega)s\log(n/s)\}^{1/2} - s^{1/2}(1+\delta_0)\alpha]^2 - s(1+\delta_0)^2\alpha^2}{2+2\delta_0} + t_4\log\frac{C_2}{C_1} + \alpha t_3\}}{C_1^{t_3}[\tau^{\alpha-1}]^{t_3}[1-2C_2(1/\tau)^{-(\alpha-1)}]^{t_2}} \\
\le&\frac{\exp\{-\kappa s\log(n/s)\}}{[\tau^{\alpha-1}]^{t_3}}
\end{aligned} \tag{A.7}$$

for any $0 < \kappa < (2+\omega)/(2+2\delta_0)$, where the last inequality holds since t_4, $t_3 < s$, $\tau \to 0$ and $n/s \to \infty$. Combining inequalities (A.5)–(A.7), (A.5) can be bounded by

$$\begin{aligned}
&\frac{\int_{C_n}\exp\{-\|\Delta\vartheta_1\|^2/(2+2\delta_0)\}\pi(d\vartheta_1)}{\pi(\|\Delta\vartheta_1\| \le \{\eta s\log(n/s)/2\}^{1/2})} \\
\le&\sum_{T_3\subset T_1}\frac{\int_{C_n^{T_3}}\exp\{-\kappa s\log(n/s)\}\prod_{j\in T_2\cup T_3}\pi(\vartheta_{1,j})d\vartheta_1}{[\tau^{\alpha-1}]^{t_3}} \\
\le&\sum_{T_3\subset T_1}\int_{C_n^{T_3}}\left\{\frac{\exp[-\kappa\log(n/s)]}{[\tau^{\alpha-1}]}\right\}^{t_3}\prod_{j\in T_3}\pi(\vartheta_{1,j})\prod_{j\in T_2}\pi(\vartheta_{1,j})d\vartheta_1 \\
\le&\int_{\|\Delta\vartheta_{1,T_1}\|\le\{(2+\omega)s\log(n/s)\}^{1/2}}\prod_{j\in T_1}\tilde\pi(\vartheta_{1,j})d\vartheta_{1,T_1} \\
\le&\frac{\exp\{-\kappa s\log(n/s)\}}{[\tau^{\alpha-1}]^s}[1+V_s(\{(2+\omega)s\log(n/s)\}^{1/2})],
\end{aligned} \tag{A.8}$$

where $\tilde\pi(\vartheta) = \exp[-\kappa\log(n/s)]\pi(\vartheta)/\tau^{\alpha-1}$ if $|\vartheta| < 1$, $\tilde\pi(\vartheta) = 1$ if $|\vartheta| \ge 1$; and $V_n(R)$ is the volume of n-dimensional ball with radius R.

Combining all the above calculus results, if $\tau^{\alpha-1} \ge (s/n)^c\{\log(n/s)\}^{1/2}$ for $0 < c < \kappa - \eta/2$ (which is guaranteed by the condition of the theorem, as long as δ_0 and η are sufficiently small), then

$$\frac{\int_{C_n}\exp\{-\|\Delta\vartheta_1\|^2/(2+2\delta_0)\}\pi(d\vartheta_1)}{\pi(\|\Delta\vartheta_1\| \le \{\eta s\log(n/s)/2\}^{1/2})}\exp\{\frac{\eta}{2}s\log(n/s)\} \le \exp\{-c's\log(n/s)\}$$

for some $0 < c < \kappa - \eta/2 - c$. And this concludes (A.1).

Part II: posterior contraction rate of zero θ_i's

It is equivalent to consider the situation that $\tilde y = \vartheta_2 + \epsilon$ where $\tilde y, \vartheta_2 \in \mathbb{R}^{n-s}$, $\epsilon \sim N(0, I_{n-s})$. The parameter ϑ_2 is subject to prior $\prod_{j=1}^{n-s}\pi(\vartheta_{2,j})$. The true

parameter $\vartheta_2^* = 0$, and we want to show that

$$E^*\pi(\|\vartheta_2\| \geq \sqrt{\omega s\log(n/s)}, \text{at most } c\delta s \text{ entries of } \vartheta_2 \\ \text{is larger than } \sqrt{\delta s\log(n/s)/n}|\tilde{y}) \leq \exp\{-c_0 s\log(n/s)\}, \tag{A.9}$$

if $\tau^{\alpha-1} \prec [(s/n)\log(n/s)]^{\alpha}$; and

$$E^*\pi(\|\vartheta_2\| \geq \sqrt{\omega s\log(n/s)}, \text{at most } c\delta s \text{ entries of } \vartheta_2 \\ \text{is larger than } s\sqrt{\delta\log(n/s)/n}|\tilde{y}) \leq \exp\{-c_0 s\log(n/s)\}, \tag{A.10}$$

if $\tau^{\alpha-1} \prec (s/n)^{\alpha}\{\log(n/s)\}^{(\alpha+1)/2}$, for $\delta = \omega/5$ and some constants $c < 1/2$ and $c_0 > 0$. We will apply Lemma A.3 to prove (A.9) and (A.10).

To proof (A.9), we consider the testing function $\phi(\tilde{y}) = \max_{|\xi|\leq c\delta s} 1(\|\tilde{y}_\xi\| \geq \{\delta s\log(n/s)\}^{1/2})$, where $\tilde{y}_\xi$ denotes the subvector of $\tilde{y}$ corresponding to model ξ. First, for any fixed $\delta_0 > 0$, by Lemma A.1 and Sterling's approximation, we have

$$\begin{aligned} E_{\vartheta_2^*}\phi(\tilde{y}) &= \lceil c\delta s\rceil \binom{n}{\lceil c\delta s\rceil} \Pr(\chi^2_{\lceil c\delta s\rceil} \geq \delta s\log(n/s)) \\ &\leq \lceil c\delta s\rceil \binom{n}{\lceil c\delta s\rceil} \exp\left\{-\frac{\delta s\log(n/s)}{2+\delta_0}\right\} \leq \exp\{-c' s\log(n/s)\} \end{aligned} \tag{A.11}$$

for some $0 < c' < \delta(1/(2+\delta_0) - c)$, when n is sufficiently large and we choose $c < 1/(2+\delta_0)$.

We define two sets in $\mathbb{R}^{n-s}$ as: $B_n = \{\text{more than } c\delta s \text{ entries of } |\vartheta_2| \text{ are bigger than } \sqrt{\delta s\log(n/s)/n}\}$, and $C_n = \{\|\vartheta_2\| \geq \sqrt{5\delta s\log(n/s)} \text{ and at most } c\delta s \text{ entries of } |\vartheta_2| \text{ are bigger than } \sqrt{\delta s\log(n/s)/n}\}$. For any $\vartheta_2 \in C_n$, let $\hat{\xi} = \hat{\xi}(\vartheta_2) = \{j : |\vartheta_{2,j}| \geq \{\delta s\log(n/s)/n\}^{1/2}\}\}$, thus we always have that $|\hat{\xi}| \leq c\delta s$, $\|\vartheta_{2,\hat{\xi}^c}\| \leq \{\delta s\log(n/s)\}^{1/2}$ and $\|\vartheta_{2,\hat{\xi}}\| \geq 2\{\delta s\log(n/s)\}^{1/2}$. Then we can derive that

$$\begin{aligned} &\sup_{\vartheta_2\in C_n} E_{\vartheta_2}[1-\phi(\tilde{y})] \leq \Pr(\|\tilde{y}_{\hat{\xi}}\| \leq \{\delta s\log(n/s)\}^{1/2}|\vartheta_2) \\ \leq &\Pr(\|\tilde{y}_{\hat{\xi}} - \vartheta_{2,\hat{\xi}}\| \geq \|\vartheta_{2,\hat{\xi}} - \vartheta^*_{2,\hat{\xi}}\| - \{\delta s\log(n/s)\}^{1/2}|\vartheta_2) \\ \leq &\Pr(\chi^2_{\lceil c\delta s\rceil} \geq [\|\vartheta_{2,\hat{\xi}} - \vartheta^*_{2,\hat{\xi}}\| - \{\delta s\log(n/s)\}^{1/2}]^2) \\ \leq &\exp\left\{-\frac{\delta s\log(n/s)}{2+\delta_0}\right\}. \end{aligned} \tag{A.12}$$

Note that with dominating probability, $\|y\| \leq (1+c')n^{1/2}$ for any $c' > 0$ and

$$\begin{aligned} \frac{m(\tilde{y})}{f^*(\tilde{y})} &\geq \int_{\|\vartheta_2\|_\infty \leq s\log(n/s)/n} \frac{f(\vartheta_2;\tilde{y})}{f(0;\tilde{y})}\pi(\vartheta_2) \\ &= \int_{\|\vartheta_2\|_\infty \leq \eta s\log(n/s)/n} \exp\left\{-\frac{1}{2}(\|\tilde{y}-\vartheta_2\|^2 - \|\tilde{y}\|^2)\right\}\pi(\vartheta_2) \\ &\geq \exp\{-3\eta s\log(n/s)\}\pi(\|\vartheta_2\|_\infty \leq \eta s\log(n/s)/n), \end{aligned} \tag{A.13}$$

for any positive η. By the condition of τ,

$$\begin{aligned}&\pi(\|\vartheta_2\|_\infty \le \eta s\log(n/s)/n) \ge (1 - C_2(\frac{\eta\frac{s}{n}\log\frac{n}{s}}{\tau})^{-(\alpha-1)})^{n-s}\\ &\ge(1-\eta'\frac{s}{n}\log\frac{n}{s})^{n-s} \to \exp\{-\eta' s\log(n/s)\}\end{aligned} \tag{A.14}$$

for any positive η'.

Similar, we have the $\pi(|\vartheta_{2,i}| \ge \{\delta s\log(n/s)/n\}^{1/2}) = o([(s/n)\log(n/s)]^{(\alpha+1)/2})$. Thus by Lemma A.2, it is easy to verify that the prior of B_n satisfies

$$-\log\pi(B_n) \gtrsim c\delta[(\alpha-1)/2]s\log(n/s) \tag{A.15}$$

Combining results (A.11)–(A.15), by Lemma A.3, one can see that (A.9) holds as long as we choose sufficiently small η and η'.

Similar arguments can be use to proof (A.10). The only difference is that we now define the set B_n = {more than $c\delta s$ entries of $|\vartheta_2|$ are bigger than $s\sqrt{\delta\log(n/s)/n}$} and set $C_n = \{\|\vartheta_2\| \ge \sqrt{5\delta s\log(n/s)}$ and at most $c\delta s$ entries of$|\vartheta_2|$ are bigger than $s\sqrt{\delta\log(n/s)/n}\}$. The details of proving (A.10) is left to the readers.

Due to the fact that posterior of ϑ_1 and ϑ_2 are independent, (A.1) and (A.9) imply

$$\begin{aligned}&E^*\pi(\|\theta-\theta^*\| \ge \{(2+\omega)s\log(n/s)\}^{1/2} + \{(\omega)s\log(n/s)\}^{1/2}|D_n)\\ &\le \exp\{-c_0' s\log(n/s)\}\end{aligned}$$

for some c_0'. And (A.1) and (A.10) imply that

$$\begin{aligned}E^*\pi(\|\theta-\theta^*\|_1 \ge s\{(2+\omega)\log(n/s)\}^{1/2} + s\{\omega\delta\log(n/s)\}^{1/2}\\ + s\{\delta\log(n/s)\}^{1/2}|D_n) \le \exp\{-c_0' s\log(n/s)\},\end{aligned}$$

for some c_0'. These conclude our results in Theorem 2.1.

A.2. Proof of Theorem 3.1

Consider the following testing function

$$\phi(D_n) = \max_{\xi\supset\xi^*, |\xi\setminus\xi^*|\le c\delta s} 1(\|y_\xi - \theta^*_\xi\| \ge \{\delta s\log(n/s)\}^{1/2}), \tag{A.16}$$

for some $c \le 1/2$, and define two sets in Θ: $B_n = \{\theta : |\{j : j \notin \xi^*, |\theta_j| \ge s\{\delta\log(n/s)\}^{1/2}/n\}| \ge c\delta s\}$, and $C_n = \{\theta : \|\theta-\theta^*\| \ge \sqrt{(2+2\omega)s\log(n/s)}\}\setminus B_n$, where $\xi^* = \{j : |\theta^*_j| \ne 0\}$. The δ is a small quantity depending on ω which we will determine later.

By the same arguments used in the proof of Theorem 2.1 and Lemma A.1, we have that, for any fixed small δ_0 satisfying $1/(2+\delta_0) > c$,

$$E_{\theta^*}\phi(D_n) \le \exp\left\{-[\frac{1}{2+\delta_0} - c]\delta s\log(n/s)\right\},$$

if n is sufficiently large, and

$$\sup_{\theta\in C_n} E_\theta(1-\phi(D_n)) \le \exp\{-[\{(2+2\omega-\delta)^{1/2}-\delta^{1/2}\}^2/(2+\delta_0)]s\log(n/s)\},$$

and we choose the values of δ and δ_0 to be very small, such that $\{\sqrt{2+2\omega-\delta}-\sqrt{\delta}\}^2/(2+\delta_0) > 1+3\omega/4$.

To derive the upper bound for $E^*\pi(C_n\cup B_n|D_n)$, let us study $E^*\pi(C_n|D_n)$ and $E^*\pi(B_n|D_n)$ separately.

Use the same notation and arguments in the proof of Theorem 2.1, let $m(D_n)$ be the marginal density of data, and $f^*(D_n)$ be the true likelihood. Let ϑ_1 and ϑ_2 be the subvectors of θ corresponding to ξ^* and ξ^{*c}. Thus with probability $1-\exp\{-cs\log(n/s)\}$,

$$\begin{aligned}\frac{m(D_n)}{f^*(D_n)} \ge& \exp\{-4\eta s\log(n/s)\}\\ &\times\pi(\|\vartheta_1-\vartheta_1^*\|\le\{\eta s\log(n/s)/2\}^{1/2},\|\vartheta_2\|_\infty\le\eta s\log(n/s)/n)\end{aligned}$$

for any positive η.

By Lemma A.3 and testing function (A.16), we have that

$$\begin{aligned}E^*\pi(C_n|D_n) \le& \exp\{-cs\log(n/s)\}+\exp\left\{-[\frac{1}{2+\delta_0}-c]\delta s\log(n/s)\right\}\\ &+\frac{\exp\{-[\{(2+2\omega-\delta)^{1/2}-\delta^{1/2}\}^2/(2+\delta_0)]s\log(n/s)\}}{\exp\{-4\eta s\log(n/s)\}\pi(\|\vartheta_1-\vartheta_1^*\|\le\sqrt{\eta s\log(n/s)/2},\|\vartheta_2\|_\infty\le\eta s\log(n/s)/n)}.\end{aligned}\tag{A.17}$$

When $\tau\in[(s/n)^{(1+\omega/2)/(\alpha-1)},(s/n)^{\alpha/(\alpha-1)}]$, as showed in the proof of Theorem 2.1, we have that $\pi(\|\vartheta_2\|_\infty\le\eta s\log(n/s)/n)|\tau)\ge\exp\{-\eta' s\log(n/s)\}$ for any positive η', and

$$\begin{aligned}&\pi(\|\vartheta_1-\vartheta_1^*\|\le\{\eta s\log(n/s)/2\}^{1/2}|\tau)\\ \ge&\{\eta\log(n/s)/2\}^{s/2}\left(\min_{|\theta|\le\{\eta\log(n/s)/2\}^{1/2}+\max|\theta_j^*|}\pi(\theta|\tau)\right)^s\\ \ge&\exp\{-(1+\omega/2+\omega/5+\eta'')s\log(n/s)\},\end{aligned}$$

for any positive η'', where the last inequality is due to the upper bound condition of $\max|\theta_j^*|$. Therefore,

$$\begin{aligned}&\pi\{\|\vartheta_1-\vartheta_1^*\|\le\sqrt{\eta s\log(n/s)/2},\|\vartheta_2\|_\infty\le\eta s\log(n/s)/n)\}\\ \ge&\pi\{(s/n)^{(1+\omega/2)/(\alpha-1)}\le\tau\le(s/n)^{\alpha/(\alpha-1)}\}\\ &\times\exp\{-(\eta'+1+\omega/2+\omega/5+\eta'')s\log(n/s)\}.\end{aligned}$$

Combining the above results, with the condition on the prior $\pi(\tau)$, we have that (A.17) $\le\exp\{-c's\log(n/s)\}$ for some positive c', given η' and η'' are sufficiently small.

Now we study the posterior $E^*\pi(B_n|D_n)$. The marginal distribution can be written as $m(D_n) = \int f(y^{(1)};\vartheta_1)f(y^{(2)};\vartheta_2)\pi(\theta)d\theta$, where $y^{(1)} = y_{\xi^*}$, $y^{(2)} = y_{\xi^{*c}}$, $f(y^{(1)};\vartheta_1)$ and $f(y^{(2)};\vartheta_2)$ are the likelihood functions for $y^{(1)}$ and $y^{(2)}$, and $\theta = (\vartheta_1, \vartheta_2)$. By the same arguments used in the proof of Theorem 2.1, with probability $1 - \exp\{-cs\log(n/s)\}$,

$$m(D_n) \geq f(y^{(2)};\vartheta_2^*)\exp\{-3\eta s\log(n/s)\}\int_{\|\vartheta_2\|_\infty \leq \eta s\log(n/s)/n} f(y^{(1)};\vartheta_1)\pi(\theta)d\theta,$$
$$\text{and } \|y^{(1)} - \vartheta_1^*\|^2 \leq \eta''' s\log(n/s), \tag{A.18}$$

for some positive c and η'''. Let Ω_n be the event that (A.18) holds.

Therefore, by the same argument in the Lemma A.3,

$$\begin{aligned}
&E^*\pi(B_n|D_n)\\
\leq&(1 - P^*(\Omega_n))\\
&+ E_{y^{(1)}}\frac{E_{y^{(2)}}\int_{B_n} f(y^{(1)};\vartheta_1)f(y^{(2)};\vartheta_2)\pi(\theta)d\theta / f(y^{(2)};\vartheta_2^*)}{\exp\{-3\eta s\log(n/s)\}\int_{\|\vartheta_2\|_\infty\leq\eta s\log(n/s)/n} f(y^{(1)};\vartheta_1)\pi(\theta)d\theta}1_{\Omega_n}\\
=&(1 - P^*(\Omega_n))\\
&+ E_{y^{(1)}}\frac{\int_{B_n} f(y^{(1)};\vartheta_1)\pi(\theta)d\theta}{\exp\{-3\eta s\log(n/s)\}\int_{\|\vartheta_2\|_\infty\leq\eta s\log(n/s)/n} f(y^{(1)};\vartheta_1)\pi(\theta)d\theta}1_{\Omega_n}.
\end{aligned} \tag{A.19}$$

Let's study the last term in the right handed side of (A.19). Define two sets in $\mathbb{R}^{n-s}$: $B_n^1 = \{\vartheta_2 : |\{j : |\vartheta_{2,j}| \geq s\{\delta\log(n/s)\}^{1/2}/n\}| \geq c\delta s\}$, $B_n^2 = \{\vartheta_2 : \|\vartheta_2\|_\infty \leq \eta s\log(n/s)/n\}$, hence

$$\begin{aligned}
&E_{y^{(1)}}\frac{\int_{B_n} f(y^{(1)};\vartheta_1)\pi(\theta)d\theta}{\int_{\|\vartheta_2\|_\infty\leq\eta s\log(n/s)/n} f(y^{(1)};\vartheta_1)\pi(\theta)d\theta}1_{\Omega_n}\\
\leq&E_{y^{(1)}}\frac{(\int_{\tau\leq\tau_0} + \int_{\tau>\tau_0})\int_{B_n^1}\int_{\mathbb{R}^s} f(y^{(1)};\vartheta_1)\pi(\vartheta_1|\tau)\pi(\vartheta_2|\tau)\pi(\tau)d\vartheta_1 d\vartheta_2 d\tau}{\int_{\tau\leq\tau_0}\int_{B_n^2}\int_{\mathbb{R}^s} f(y^{(1)};\vartheta_1)\pi(\vartheta_1|\tau)\pi(\vartheta_2|\tau)\pi(\tau)d\vartheta_1 d\vartheta_2 d\tau}1_{\Omega_n},
\end{aligned} \tag{A.20}$$

where $\tau_0 = (s/n)^{\alpha/(\alpha-1)}$. When $\tau \leq \tau_0$, by the same arguments used in the Part II of the proof of theorem 2.1, we have

$$\begin{aligned}
&\frac{\int_{B_n^1}\int_{\mathbb{R}^s} f(y^{(1)};\vartheta_1)\pi(\vartheta_1|\tau)\pi(\vartheta_2|\tau)d\vartheta_1 d\vartheta_2}{\int_{B_n^2}\int_{\mathbb{R}^s} f(y^{(1)};\vartheta_1)\pi(\vartheta_1|\tau)\pi(\vartheta_2|\tau)d\vartheta_1 d\vartheta_2} = \frac{\int_{B_n^1}\pi(\vartheta_2|\tau)d\vartheta_2}{\int_{B_n^2}\pi(\vartheta_2|\tau)d\vartheta_2}\\
\leq&\frac{\exp\{-c\delta(\alpha-1)s\log(n/s)/2\}}{\exp\{-\eta' s\log(n/s)\}}
\end{aligned} \tag{A.21}$$

for any fixed small $\eta' > 0$. And on event Ω_n,

$$\begin{aligned}
&\int_{\mathbb{R}^s} \exp\{-\|y^{(1)} - \vartheta_1\|^2/2\}\pi(\vartheta_1|\tau)d\vartheta_1 \\
&= \int_{\mathbb{R}^s} \exp\{-\|y^{(1)} - \vartheta_1^* + \vartheta_1^* - \vartheta_1\|^2/2\}\pi(\vartheta_1|\tau)d\vartheta_1 \\
&\geq \int_{\|\vartheta_1 - \vartheta_1^*\|_\infty \leq \{\log(n/s)\}^{1/2}} \exp\{-\|y^{(1)} - \vartheta_1\|^2/2\}\pi(\vartheta_1|\tau)d\vartheta_1 \\
&\geq \int_{\|\vartheta_1 - \vartheta_1^*\|_\infty \leq \{\log(n/s)\}^{1/2}} \exp[-\{\eta''' + 1 + 2(\eta''')^{1/2}\}s\log(n/s)/2]\pi(\vartheta_1|\tau)d\vartheta_1 \\
&= \exp\{-c's\log(n/s)\} \int_{\|\vartheta_1 - \vartheta_1^*\|_\infty \leq \{\log(n/s)\}^{1/2}} \pi(\vartheta_1|\tau)d\vartheta_1,
\end{aligned}$$

for some positive c'. Therefore, conditional on Ω_n, we have

$$\begin{aligned}
&\frac{\int_{\tau>\tau_0}\int_{B_n^1}\int_{\mathbb{R}^s} f(y^{(1)};\vartheta_1)\pi(\vartheta_1|\tau)\pi(\vartheta_2|\tau)\pi(\tau)d\vartheta_1 d\vartheta_2 d\tau}{\int_{\tau\leq\tau_0}\int_{B_n^2}\int_{\mathbb{R}^s} f(y^{(1)};\vartheta_1)\pi(\vartheta_1|\tau)\pi(\vartheta_2|\tau)\pi(\tau)d\vartheta_1 d\vartheta_2 d\tau} \\
&\leq \frac{\int_{\tau>\tau_0}\pi(\tau)d\tau}{\exp\{-c's\log(n/s)\}\int_{\tau_1}^{\tau_0}\int_{B_n^2}\int_{\|\vartheta_1-\vartheta_1^*\|_\infty\leq\sqrt{\log(n/s)}}\pi(\vartheta_1|\tau)\pi(\vartheta_2|\tau)\pi(\tau)d\vartheta_1 d\vartheta_2 d\tau} \\
&\leq \frac{\int_{\tau>\tau_0}\pi(\tau)d\tau}{\exp\{-c''s\log(n/s)\}\int_{\tau_1\leq\tau\leq\tau_0}\pi(\tau)d\tau} \\
&\leq \exp\{-c'''s\log(n/s)\} \qquad \text{(A.22)}
\end{aligned}$$

for any positive c'' and c''', where $\tau_1 = (s/n)^{(1+\omega/2)/(\alpha-1)}$, and the second inequality follows by the condition $\max|\vartheta_j^*| = \max|\theta_j^*| \leq (n/s)^{\omega/(5\alpha)}$.

Combining (A.19)–(A.22), we have that $E^*\pi(B_n|D_n) \leq \exp\{-c''''s\log(n/s)\}$ for some positive c'''' if η and η' are small enough.

These results conclude the theorem.

References

[1] Armagan, A., Clyde, M. and Dunson, D. B. (2011). Generalized beta mixtures of Gaussians. In *Advances in Neural Information Processing Systems* 523–531.

[2] Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U. and Strawn, N. (2013). Posterior consistency in linear models under shrinkage priors. *Biometrika* **100** 1011–1018. MR3142348

[3] Bai, R. and Ghosh, M. (2017). The inverse gamma-gamma prior for optimal posterior contraction and multiple hypothesis testing. *arXiv preprint* arXiv:1710.04369.

[4] Barron, A. (1998). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian Statistics 6* (J. M. Bernardo, J. Berger, A. Dawid and A. Smith, eds.) 27–52. MR1723492

[5] Bhadra, A., Datta, J., Polson, N. G., Willard, B. et al. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis* **12** 1105–1131. MR3724980

[6] Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110** 1479–1490. MR3449048

[7] Birgé, L. (1984). Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.* **3** 259–282. MR0764150

[8] Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751

[9] Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics* **40** 2069–2101. MR3059077

[10] Donoho, D. L. and Johnstone, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* 277–303. MR1278886

[11] Donoho, D. L., Johnstone, I. M., Hoch, J. C. and Stern, A. S. (1992). Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B (Methodological)* 41–81. MR1157714

[12] Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science* 1–22. MR2431866

[13] Efron, B. (2012). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* **1**. Cambridge University Press. MR2724758

[14] Ghosal, S., Ghosh, J. K. and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics* **28** 500–531. MR1790007

[15] Ghosal, S. and Van Der Vaart, A. W. (2007). Convergence rates of posterior distributions for noniid observations. *Annals of Statistics* **35** 192–223. MR2332274

[16] Ghosh, P. and Chakrabarti, A. (2014). Posterior concentration properties of a general class of shrinkage priors around nearly black vectors. *arXiv preprint* arXiv:1412.8161.

[17] Griffin, J. E. and Brown, P. J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics* **53** 423–442. MR2910027

[18] Hans, C. (2009). Bayesian lasso regression. *Biometrika* **96** 835–845. MR2564494

[19] Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: convergence rate of the fitted densities. *Annals of Statistics* **35** 1487–1511. MR2351094

[20] Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics* **34** 837–877. MR2283395

[21] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* 1302–1338. MR1805785

[22] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.*

Springer, New York. MR0856411
[23] Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103** 681–686. MR2524001
[24] Rocková, V. (2015). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. Submitted manuscript 1–34. MR3766957
[25] Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P. et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1** 203–209.
[26] Song, Q. and Liang, F. (2017). Nearly optimal Bayesian shrinkage for high dimensional regression. arXiv:1712.08964.
[27] van der Pas, S. L., Kleijn, B. J. K. and van der Vaart A. W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics* **2** 2585–2618. MR3285877
[28] van der Pas, S., Salomond, J.-B. and Schmidt-Hieber, J. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electronic Journal of Statistics* **10** 976–1000. MR3486423
[29] van der Pas, S. L., Szabo, B. and van der Vaart, A. (2017). Adaptive posterior contraction rates for the horseshoe. arXiv:1702.03698. MR3705450
[30] van der Pas, S., Szabó, B. and van der Vaart, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis* **12** 1221–1274. MR3724985
[31] Zhang, C.-H. (2012). Minimax L_q risk in L_p balls. In *Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman* 78–89. Institute of Mathematical Statistics. MR3202504
[32] Zubkov, A. and Serov, A. (2013). A complete proof of universal inequalities for the distribution function of the binomial law. *Theory of Probability & Its Applications* **57** 539–544. MR3196787