Conditions for posterior contraction in the sparse normal means problem

S van Der Pas, J.-B Salomond, J Schmidt-Hieber

Electronic Journal of Statistics
Vol. 10 (2016) 976–1000
ISSN: 1935-7524
DOI: 10.1214/16-EJS1130

Conditions for posterior contraction in the sparse normal means problem

S.L. van der Pas*

Leiden University, Mathematical Institute, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
e-mail: svdpas@math.leidenuniv.nl

J.-B. Salomond†

Université Paris Dauphine, Place du Maréchal DeLattre de Tassigny, 75016 Paris, France
e-mail: salomond@ceremade.dauphine.fr

and

J. Schmidt-Hieber

Leiden University, Mathematical Institute, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
e-mail: schmidthieberaj@math.leidenuniv.nl

Abstract: The first Bayesian results for the sparse normal means problem were proven for spike-and-slab priors. However, these priors are less convenient from a computational point of view. In the meanwhile, a large number of continuous shrinkage priors has been proposed. Many of these shrinkage priors can be written as a scale mixture of normals, which makes them particularly easy to implement. We propose general conditions on the prior on the local variance in scale mixtures of normals, such that posterior contraction at the minimax rate is assured. The conditions require tails at least as heavy as Laplace, but not too heavy, and a large amount of mass around zero relative to the tails, more so as the sparsity increases. These conditions give some general guidelines for choosing a shrinkage prior for estimation under a nearly black sparsity assumption. We verify these conditions for the class of priors considered in [12], which includes the horseshoe and the normal-exponential gamma priors, and for the horseshoe+, the inverse-Gaussian prior, the normal-gamma prior, and the spike-and-slab Lasso, and thus extend the number of shrinkage priors which are known to lead to posterior contraction at the minimax estimation rate.

MSC 2010 subject classifications: Primary 62F15; secondary 62G20.
Keywords and phrases: sparsity, nearly black vectors, normal means problem, horseshoe, horseshoe+, Bayesian inference, frequentist Bayes, posterior contraction, shrinkage priors.

Received October 2015.

*Research supported by Netherlands Organization for Scientific Research NWO.
†Research supported by NWO VICI project 'Safe Statistics'.

1. Introduction

In the sparse normal means problem, we wish to estimate a sparse vector θ based on a vector $X^n \in \mathbb{R}^n$, $X^n = (X_1, \ldots, X_n)$, generated according to the model

$$X_i = \theta_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the ε_i are independent standard normal variables. The vector of interest θ is sparse in the *nearly black* sense, that is, most of the parameters are zero. We wish to separate the signals (nonzero means) from the noise (zero means). Applications of this model include image reconstruction and nonparametric function estimation using wavelets [17].

The model is an important test case for the behaviour of sparsity methods, and has been well-studied. A great variety of frequentist and Bayesian estimators has been proposed, and the popular Lasso [25] is included in both categories. It is but one example of many approaches towards recovering θ; restricting ourselves to Bayesian methods, other approaches include shrinkage priors such as the spike-and-slab type priors studied by [17, 7] and [6], the normal-gamma prior [14], non-local priors [16], the Dirichlet-Laplace prior [3], the horseshoe [5], the horseshoe+ [2] and the spike-and-slab Lasso [24].

Our goal is twofold: *recovery* of the underlying mean vector, and *uncertainty quantification*. The benchmark for the former is estimation at the minimax rate. In a Bayesian setting, the typical choice for the estimator is some measure of center of the posterior distribution, such as the posterior mean, mode or median. For the purpose of uncertainty quantification, the natural object to use is a credible set. In order to obtain credible sets that are narrow enough to be informative, yet not so narrow that they neglect to cover the truth, the posterior distribution needs to contract to its center at the same rate at which the estimator approaches the truth.

For recovery, spike-and-slab type priors give optimal results ([17, 7, 6]). These priors assign independently to each component a mixture of a point mass at zero and a continuous prior. Due to the point mass, spike-and-slab priors shrink small coefficients to zero. The advantage is that the full posterior has optimal model selection properties but this comes at the price of, in general, too narrow credible sets. Another drawback of spike-and-slab methods is that they are computationally expensive although the complexity is much better than what has been previously believed ([27]).

Thus, we might ask whether there are priors which are smoother and shrink less than the spike-and-slab but still recover the signal with a (nearly) optimal rate. A naive choice would be to consider the Laplace prior $\propto e^{-\lambda\|\theta\|_1}$ with $\|\theta\|_1 = \sum_{i=1}^n |\theta_i|$, since in this case the maximum a posteriori (MAP) estimator coincides with the Lasso, which is known to achieve the optimal rates for sparse signals. In [6], Section 3, it was shown that although the MAP-estimator has good properties, the full posterior spreads a non-negligible amount of mass over large neighborhoods of the truth leading to recovery rates that are sub-optimal by a polynomial factor in n. This example shows that if the prior does not shrink enough, we loose the recovery property of the posterior.

Recently, shrinkage priors were found that are smoother than the spike-and-slab but still lead to (near) minimax recovery rates. Up to now, optimal recovery rates have been established for the horseshoe prior [26], horseshoe-type priors with slowly varying functions [12], the empirical Bayes procedure of [18], the spike-and-slab Lasso [24], and the Dirichlet-Laplace prior, although the latter result only holds under a restriction on the signal size [3]. Finding smooth shrinkage priors with theoretical guarantees remains an active area of research.

The question arises which features of the prior lead to posterior convergence at the minimax estimation rate. Qualitative discussion on this point is provided by [5]. Intuitively, a prior should place a large amount of mass near zero to account for the zero means, and have heavy tails to counteract the shrinkage effect for the nonzero means. In the present article, we make an attempt to quantify the relevant properties of a prior, by providing general conditions ensuring posterior concentration at the minimax rate, and showing that a large number of priors (including the ones listed above) meets these conditions.

We study scale mixtures of normals, as many shrinkage priors proposed in the literature are contained in this class and provide general conditions on the prior on the local variance such that posterior concentration at the minimax estimation rate is guaranteed. These conditions are general enough to recover the already known results for the horseshoe prior, the horseshoe-type priors with slowly varying functions and the spike-and-slab Lasso, and to demonstrate that the horsehoe+ [2], inverse-Gaussian prior [4] and the normal-gamma prior [4, 14] lead to posterior concentration at the correct rate as well. Our conditions in essence mean that a sparsity prior should have tails that are at least as heavy as Laplace, but not too heavy, and there should be a sizable amount of mass close to zero relative to the tails, especially when the underlying vector is very sparse.

This paper is organized as follows. We state our main result, providing conditions on sparsity priors such that the posterior contracts at the minimax rate in Section 2. We then show, in Section 3, that these conditions hold for the class of priors of [12], as well as for the horseshoe+, the inverse-Gaussian prior, the normal-gamma prior, and the spike-and-slab Lasso. A simulation study is performed in Section 4, and we conclude with a Discussion. All proofs are given in Appendix A.

Notation. Denote the class of nearly black vectors by $\ell_0[p_n] = \{\theta \in \mathbb{R}^n : \sum_{i=1}^n \mathbf{1}\{\theta_i \neq 0\} \leq p_n\}$. The minimum $\min\{a, b\}$ is given by $a \wedge b$. The standard normal density is denoted by ϕ, its cdf by Φ, and we set $\Phi^c(x) = 1 - \Phi(x)$. The norm $\|\cdot\|$ is the ℓ_2-norm.

2. Main results

Each coefficient θ_i receives a scale mixture of normals as a prior:

$$\theta_i \mid \sigma_i^2 \sim \mathcal{N}(0, \sigma_i^2), \qquad \sigma_i^2 \sim \pi(\sigma_i^2), \quad i = 1, \ldots, n, \tag{1}$$

where $\pi : [0, \infty) \to [0, \infty)$ is a density on the positive reals. While π might depend on further hyperparameters, no additional priors are placed on such

parameters, rendering the coefficients independent *a posteriori*. The goal is to obtain conditions on π such that posterior concentration at the minimax estimation rate is guaranteed.

We use the coordinatewise posterior mean to recover the underlying mean vector. By Tweedie's formula [23], the posterior mean for θ_i given an observation x_i is equal to $x_i + \frac{d}{dx}\log p(x_i)$, where $p(x_i)$ is the marginal distribution of x_i. The posterior mean for parameter θ_i is thus given by $\widehat{\theta}_i = X_i m_{X_i}$, where $m_x : \mathbb{R} \to [0,1]$ is

$$m_x := \frac{\int_0^1 z(1-z)^{-3/2} e^{\frac{x^2}{2}z} \pi\big(\frac{z}{1-z}\big)dz}{\int_0^1 (1-z)^{-3/2} e^{\frac{x^2}{2}z} \pi\big(\frac{z}{1-z}\big)dz} = \frac{\int_0^\infty u(1+u)^{-3/2} e^{\frac{x^2 u}{2+2u}} \pi(u)du}{\int_0^\infty (1+u)^{-1/2} e^{\frac{x^2 u}{2+2u}} \pi(u)du}. \tag{2}$$

We denote the estimate of the full vector θ by $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n) = (X_1 m_{X_1}, \ldots, X_n m_{X_n})$. An advantage of scale mixtures of normals as shrinkage priors over spike-and-slab-type priors, is that the posterior mean can be represented as the observation multiplied by (2). The ratio (2) can be computed via integral approximation methods such as a quadrature routine. See [21], [22] and [26] for more discussion on this point in the context of the horseshoe.

Our main theorem, Theorem 2.1, provides three conditions on π under which a prior of the form (1) leads to an upper bound on the posterior contraction rate of the order of the minimax rate. We first state and discuss the conditions. In addition, we present stronger conditions that are easier to verify. Condition 1 is required for our bounds on the posterior mean and variance for the nonzero means. The remaining two are used for the bounds for the zero means.

The first condition involves a class of regularly varying functions. Recall that a function ℓ is called *regular varying (at infinity)* if for any $a > 0$, the ratio $\ell(au)/\ell(u)$ converges to the same non-zero limit as $u \to \infty$. For our estimates, we need a slightly different notion, that will be introduced next. We say that a function L is *uniformly regular varying,* if there exist constants $R, u_0 \geq 1$, such that

$$\frac{1}{R} \leq \frac{L(au)}{L(u)} \leq R, \quad \text{for all } a \in [1,2], \text{ and all } u \geq u_0. \tag{3}$$

In particular, $L(u) = u^b$, and $L(u) = \log^b(u)$ with $b \in \mathbb{R}$ are uniformly regular varying (take for example $R = 2^{|b|}$ and $u_0 = 2$). An example of a function that is not uniformly regular varying is $L(u) = e^u$. From the definition, we can easily deduce the following properties of functions that are uniformly regular varying. Firstly, $u \mapsto L(u)$ is on $[u_0, \infty)$ either everywhere positive or everywhere negative. If L is uniformly regular varying then also $u \mapsto 1/L(u)$ and if L_1 and L_2 are uniformly regular varying, then also their product $L_1 L_2$.

We are now ready to present Condition 1, and the stronger Condition 1', which implies Condition 1, as shown in Lemma A.1.

Condition 1. *For some $b \geq 0$, we can write $u \mapsto \pi(u) = L_n(u)e^{-bu}$, where L_n is a function that satisfies (3) for some $R, u_0 \geq 1$ which do not depend on n.*

Suppose further that there are constants $C', b' > 0$, $K \geq 0$, *and* $u_* \geq 1$, *such that*

$$C'\pi(u) \geq \left(\frac{p_n}{n}\right)^K e^{-b'u} \quad \text{for all } u \geq u_*. \tag{4}$$

Condition 1'. *Consider a global-local scale mixture of normals:*

$$\theta_i \mid \sigma_i^2, \tau^2 \sim \mathcal{N}(0, \sigma_i^2\tau^2), \quad \sigma_i^2 \sim \widetilde{\pi}(\sigma_i^2), \quad i = 1, \ldots, n. \tag{5}$$

Assume that $\widetilde{\pi}$ *is a uniformly regular varying function which does not depend on* n, *and* $\tau = (p_n/n)^\alpha$ *for* $\alpha \geq 0$.

Condition 1 assures that the posterior recovers nonzero means with the optimal rate. Thus, the condition can be seen as a sufficient condition on the tail behavior of the density π for ℓ^2-recovery. The tail may decay exponentially fast, which is consistent with the conditions found on the 'slab' in the spike-and-slab priors discussed by [7]. In general, π will depend on n through a hyperparameter. Condition 1 requires that the n dependence behaves roughly as a power of p_n/n.

In the important special case where each θ_i is drawn independently from a global-local scale mixture, Condition 1 is satisfied whenever the density on the local variance is uniformly regular varying, as stated in Condition 1'. Below, we give the conditions on π that guarantee posterior shrinkage at the minimax rate for the zero coefficients. The first condition ensures that the prior π puts some finite mass on values between $[0, 1]$.

Condition 2. *Suppose that there is a constant* $c > 0$ *such that* $\int_0^1 \pi(u)du \geq c$.

We turn to Condition 3 which describes the decay of π away from a neighborhood of zero. To state the condition it will be convenient to write

$$s_n := \frac{p_n}{n}\log(n/p_n). \tag{6}$$

Condition 3. *Let* $b_n = \sqrt{\log(n/p_n)}$ *and assume that there is a constant* C, *such that*

$$\int_{s_n}^{\infty} \left(u \wedge \frac{b_n^3}{\sqrt{u}}\right)\pi(u)du + b_n \int_1^{b_n^2} \frac{\pi(u)}{\sqrt{u}}du \leq Cs_n.$$

In order to allow for many possible choices of π, the tail condition involves several terms. Observe that $u \wedge b_n^3/\sqrt{u} = u$ if and only if $u \leq b_n^2$ and therefore the first integral in Condition 3 can also be written as $\int_{s_n}^{b_n^2} u\pi(u)du + b_n^3 \int_{b_n^2}^{\infty} u^{-1/2}\pi(u)du$. It is surprising that some control of $\pi(u)$ on the interval $[s_n, 1]$ is needed. But this turns out to be sharp. Theorem 2.2 proves that if we would relax the condition to $\int_{s_n}^1 u\pi(u)du \lesssim t_n$ for an arbitrary rate $t_n \gg s_n$, then there is a prior that satisfies all the other conditions needed for the zero coefficients, but which does not concentrate at the minimax rate.

Below we state two stronger conditions, each of which obviously imply Condition 2 and Condition 3 for sparse signals, that is, $p_n = o(n)$.

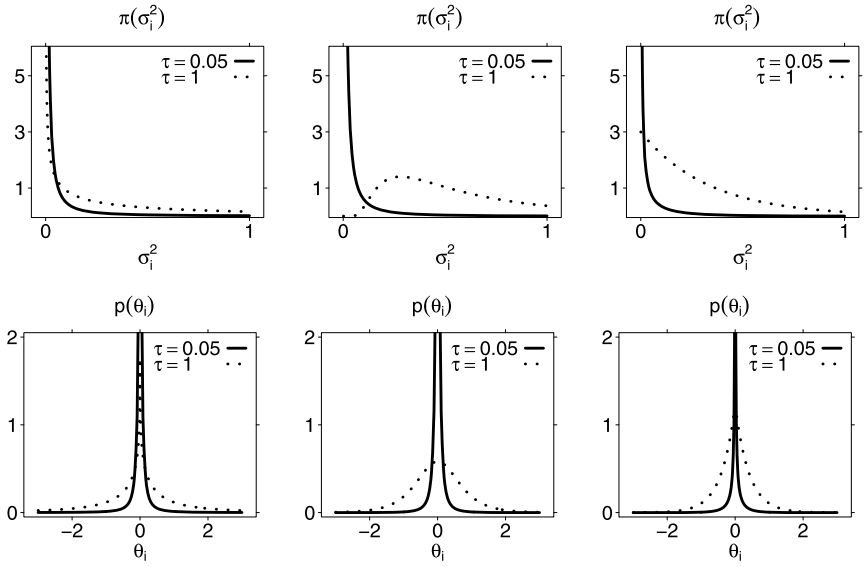

FIG 1. *Plots of priors on the local variance (first row) and the corresponding parameters (second row). From left to right: horseshoe, Inverse-Gaussian with* $a = 1/2, b = 1$, *and normal gamma with* $\beta = 3$. *The parameter* τ, *which in practice should be of the order* p_n/n, *is taken equal to 1 (dashed line) and 0.05 (solid line).*

Condition A. *Assume that there is a constant* C, *such that*

$$\pi(u) \leq \frac{C}{u^{3/2}} \frac{p_n}{n} \sqrt{\log(n/p_n)}, \quad \text{for all } u \geq s_n.$$

Condition B. *Assume that there is a constant* C, *such that*

$$\int_{s_n}^{\infty} \pi(u)du \leq \frac{Cp_n}{n}.$$

In this case, even a stronger version of Condition 2 holds in the sense that nearly all mass is concentrated in the shrinking interval $[0, s_n]$. Notice that Condition 3 does not imply Condition 2 in general. If, for example, the density π has support on $[n^2, 2n^2]$, then, Condition 3 holds but Condition 2 does not. Condition 1 and Condition 3 depend on the relative sparsity p_n/n. Indeed, Condition 1 becomes weaker if the signal is more sparse and at the same time Condition 3 becomes stronger. This matches intuition, as the prior should shrink more in this case and thus the assumptions that are responsible for the shrinkage effect should become stronger.

Figure 1 presents plots of the priors π on the local variance, and the corresponding priors on the parameters θ_i, for three priors for which the three conditions are verified in Section 3: the horseshoe, inverse-Gaussian, and normal-gamma. The parameter τ, in the notation of Section 3, should be thought of as the sparsity level p_n/n. Figure 1 shows that the priors start to resemble each

other when τ is decreased. If the setting is more sparse, corresponding to more zero means, the mass of the prior π on σ_i^2 concentrates around zero, leading to a higher peak at zero in the prior density on θ_i.

We now present our main result. The minimax estimation risk for this problem, under ℓ_2 risk, is given by $2p_n \log(n/p_n)$ [10]. We write $\theta_0 = (\theta_{0i})_{i=1,\ldots,n}$ and consider posterior concentration of the zero and non-zero coefficients separately. Asymptotics always refers to $n \to \infty$.

Theorem 2.1. *Work under model $X^n \sim \mathcal{N}(\theta_0, I_n)$ and assume that the prior is of the form (1). Suppose further that $p_n = o(n)$ and let M_n be an arbitrary positive sequence tending to $+\infty$. Let $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$ be the posterior mean. Under Condition 1,*

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi\big(\theta \ : \sum_{i:\theta_{0i}\neq 0} (\theta_i - \theta_{0i})^2 > M_n p_n \log(n/p_n) \ \big| \ X^n\big) \to 0$$

and

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \sum_{i:\theta_{0i}\neq 0} (\widehat{\theta}_i - \theta_{0i})^2 \lesssim p_n \log(n/p_n).$$

Under Condition 2 and Condition 3 (or either Condition A or B),

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi\big(\theta \ : \sum_{i:\theta_{0i}=0} \theta_i^2 > M_n p_n \log(n/p_n) \ \big| \ X^n\big) \to 0$$

and

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \sum_{i:\theta_{0i}=0} \widehat{\theta}_i^2 \lesssim p_n \log(n/p_n).$$

Thus, under Conditions 1–3 (or Condition 1 with either Condition A or B),

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi\big(\theta \ : \|\theta - \theta_0\|^2 > M_n p_n \log(n/p_n) \ \big| \ X^n\big) \to 0$$

and

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \big\|\widehat{\theta} - \theta_0\big\|_2^2 \lesssim p_n \log(n/p_n).$$

The statement is split into zero and non-zero coefficients of θ_0 in order to make the dependence on the conditions explicit. Indeed, posterior concentration of the non-zero coefficients follows from Condition 1 and posterior concentration for the zero-coefficients is a consequence of Conditions 2 and 3. In order to obtain posterior contraction, we need that $M_n \to \infty$. This is due to the use of Markov's inequality in the proof, simplifying the argument considerably. From the lower bound result [15], Theorem 2.1, one should expect that the result holds already for some sufficiently large constant M and that the speed at which the posterior mass of $\{\theta : \|\theta - \theta_0\|^2 > M p_n \log(n/p_n)\}$ converges to zero is $\exp(-C_1 p_n \log(n/p_n))$ for some positive constant C_1. It is well-known that posterior concentration at rate ϵ_n implies existence of a frequentist estimator with the same rate (cf. [11], Theorem 2.5 for a precise statement). Thus, the

rate of contraction around the true mean vector θ_0 must be sharp. This also means that credible sets computed from the posterior cannot be so large as to be uninformative, an effect that, as discussed in the introduction, occurs for the Laplace prior connected to the Lasso. If one wishes to use a credible set centered around the posterior mean, then its radius might still be too small to cover the truth. The first step towards guarantees on coverage is a lower bound on the posterior variance. Such a lower bound was obtained for the horseshoe in [26], and for priors very closely resembling the horseshoe in [12]. No such results have been obtained so far for priors on σ_i^2 that have a tail of a different order than $(\sigma_i^2)^{-3/2}$. This is a delicate technical issue that we will not pursue further here.

The results also indicates how to build adaptive procedures. We consider adaptivity to the number of nonzero means, without accounting for the possibly unknown variance of the ε_i, for which a prior of the type suggested for the horseshoe in [5] or an empirical Bayes procedure may be used. The method for adapting to the sparsity does not require explicit knowledge of p_n but in order to get minimax concentration rates, we need to find priors that satisfy the conditions of Theorem 2.1. Consider for example the prior defined as

$$\pi(u) := \frac{1}{u^{3/2}} \frac{\sqrt{\log n}}{n}, \quad \text{for all } u \geq \frac{\sqrt{\log n}}{n}$$

and the remaining mass is distributed arbitrarily on the interval $[0, \sqrt{\log n}/n)$. Thus Condition A holds for any $1 \leq p_n = o(n)$ and thus also Condition 2 and Condition 3. Whenever we impose an upper bound $p_n \leq n^{1-\delta}$ with $\delta > 0$, then also Condition 1 holds and thus Theorem 2.1 follows. This shows that in principle priors can be constructed that adapt over nearly the whole range of possible sparsity levels and lead to some theoretical guarantee. The trick is that a prior that works for an extremely sparse model with $p_n = 1$ also adapts to less sparse models. This requires, however, a lot of prior mass near zero. Such a prior shrinks small non-zero components more than if we first get a rough estimate of the relative sparsity p_n/n and then use a prior that lies on the "boundary" of the conditions in the sense that the both sides in the inequality of Condition 3 are of the same order. An empirical Bayes procedure that first estimates the sparsity was found to work well in [26], arguing along the lines of [17]. The sparsity level estimator counts the number of observations that are larger than the 'universal threshold' of $\sqrt{2 \log n}$. Similar results are likely to hold in our setting, as long as the posterior mean is monotone in the parameter that is taken to depend on p_n.

2.1. Necessary conditions

The imposed conditions are nearly sharp. To see this, consider the Laplace prior, where each θ_i is drawn independently from a Laplace distribution with parameter λ. It is well-known that the Laplace distribution with parameter λ can be represented as a scale mixture of normals where the mixing density is exponential with parameter λ^2 (cf. [1] or [19], Equation (4)). Thus, the Laplace

prior fits our framework (1) with $\pi(u) = \lambda^2 e^{-\lambda^2 u}$, for $u \geq 0$. As mentioned in the introduction, the MAP-estimator of this prior is the Lasso but the full posterior does not shrink at the minimax rate. Indeed, Theorem 7 in [6] shows that if the true vector is zero, then, the posterior concentration rate has the lower bound n/λ^2 for the squared ℓ^2-norm provided that $1 \leq \lambda = o(\sqrt{n})$. This should be compared to the optimal minimax rate $\log n$ (the rate for sparsity zero is the same as the rate for sparsity $p_n = 1$). Thus, the lower bound shows that the rate is sub-optimal as long as

$$\lambda \ll \sqrt{\frac{n}{\log n}}. \tag{7}$$

If $\lambda \gtrsim \sqrt{n/\log n}$, the lower bound is not sub-optimal anymore, but in this case, the non-zero components cannot be recovered with the optimal rate. The lower bound shows that the posterior does not shrink enough if λ is not taken to be huge and thus either Condition 2 or Condition 3 must be violated, as these are the two conditions that guarantee shrinkage of the zero mean coefficients.

Obviously, $\int_0^1 \pi(u)du \geq \int_0^1 e^{-u}du > 0$ for $1 \leq \lambda$ and thus Condition 2 holds. For Condition 3 notice that the integral can be split into the integral $\int_0^1 u\pi(u)du$ plus an integral over $[1, \infty)$ Now, if λ tends to infinity faster than a polynomial order in n then the integral over $[1, \infty)$ is exponentially small in n. Thus Condition 3 must fail because the integral over $\int_{s_n}^1 u\pi(u)du$ is of a larger order than $s_n = n^{-1}\log n$. To see this, observe that for $\lambda \leq \sqrt{n/\log n}$,

$$\int_{s_n}^1 u\lambda^2 e^{-\lambda^2 u}du = \frac{1}{\lambda^2}\int_{s_n\lambda^2}^{\lambda^2} ve^{-v}dv \geq \frac{1}{\lambda^2}\int_1^{\lambda^2} e^{-v}dv \gtrsim \frac{1}{\lambda^2}.$$

Now, we see that Condition 3 fails if and only if (7) holds. Indeed, if $\lambda \ll \sqrt{n/\log n}$, then the r.h.s. is of larger order than s_n and if $\lambda \asymp \sqrt{n/\log n}$, then, Condition 3 holds. This shows that this bound is sharp.

In order to state this as a formal result, let us introduce the following modification of Condition 3. Let κ_n denote an arbitrary positive sequence.

Condition 3(κ_n). *Let $b_n = \sqrt{\log(n/p_n)}$ and assume that there is a constant C, such that*

$$\kappa_n \int_{s_n}^1 u\pi(u)du + \int_1^\infty \Big(u \wedge \frac{b_n^3}{\sqrt{u}}\Big)\pi(u)du + b_n \int_1^{b_n^2} \frac{\pi(u)}{\sqrt{u}}du \leq Cs_n.$$

In particular, we recover Condition 3 for $\kappa_n = 1$.

Theorem 2.2. *Work under model $X^n \sim \mathcal{N}(\theta_0, I_n)$ and assume that the prior is of the form (1). For any positive sequence $(\kappa_n)_n$ tending to zero, there exists a prior π satisfying Condition 2 and Condition 3(κ_n) for $p_n = 1$ and a positive sequence $(M_n)_n$ tending to infinity, such that*

$$\mathbb{E}_{\theta_0=0}\Pi\big(\theta \ : \|\theta\|_2^2 \leq M_n \log(n) \ \big| \ X^n\big) \to 0, \quad \text{as} \quad n \to \infty. \tag{8}$$

This theorem shows that the posterior puts asymptotically all mass outside an ℓ^2-ball with radius $M_n \log(n) \gg \log(n)$ and is thus suboptimal. The proof can be found in the appendix.

3. Examples

In this section, Conditions 1–3 are verified for the horseshoe-type priors considered by [12] (which includes the horseshoe and the normal-exponential gamma), the horseshoe+, the inverse-Gaussian prior, the normal-gamma prior, and the spike-and-slab Lasso. There are, to the best of our knowledge, no existing results yet showing that the horseshoe+, the inverse-Gaussian and the normal-gamma priors lead to posterior contraction at the minimax estimation rate. Posterior concentration for the horseshoe and horseshoe-type priors were already established in [26] and [12], and for the spike-and-slab Lasso in [24] . Here, we obtain the same results but thanks to Theorem 2.1 the proofs become extremely short. In addition, we can show that a restriction on the class of priors considered by [12] can be removed.

3.1. Global-local scale mixtures of normals

In [12], the priors under consideration are normal priors with random variances of the form

$$\theta_i \mid \sigma_i^2, \tau^2 \sim \mathcal{N}(0, \sigma_i^2 \tau^2), \quad \sigma_i^2 \sim \pi'(\sigma_i^2), \quad i = 1, \ldots, n,$$

for priors π' with density given by

$$\pi'(\sigma_i^2) = K \frac{1}{(\sigma_i^2)^{a+1}} L(\sigma_i^2), \tag{9}$$

where $K > 0$ is a constant and $L : (0, \infty) \to (0, \infty)$ is a non-constant, *slowly varying* function, meaning that there exist $c_0, M \in (0, \infty)$ such that $L(t) > c_0$ for all $t \geq t_0$ and $\sup_{t \in (0,\infty)} L(t) \leq M$. [12] prove an equivalent of Theorem 2.1 for these priors, for $a \in [1/2, 1)$ and $\tau = (p_n/n)^\alpha$ with $\alpha \geq 1$.

The horseshoe prior, with $\pi(u) = (\pi\tau)^{-1} u^{-1/2}(1 + u/\tau^2)^{-1}$, is contained in this class of priors, by taking $a = 1/2$, $L(t) = t/(1 + t)$, and $K = 1/\pi$. This class also contains the normal-exponential-gamma priors of [13], for which $\pi(u) = \lambda/\gamma^2(1 + u/\gamma^2)^{-(\lambda+1)}$ with parameters $\lambda, \gamma > 0$. This class of priors is of the form (9) for the choice $\tau = \gamma$, $a = \lambda$ and $L(t) = (t/(1 + t))^{1+\lambda}$. In [12], it is stated that the three parameter beta normal mixtures, the generalized double Pareto, the inverse gamma and half-t priors are of the form (9) as well.

The global-local scale prior is of the form (1) with

$$\pi(u) = \frac{K\tau^{2a}}{u^{1+a}} L\Big(\frac{u}{\tau^2}\Big).$$

We assume that the polynomial decay in u is at least of order 3/2, that is $a \geq \frac{1}{2}$. In particular, the horseshoe lies directly at the boundary in this sense. Depending on a, we allow for different values of τ. If $\frac{1}{2} \leq a < 1$, we assume $\tau^{2a} \leq (p_n/n)\sqrt{\log(n/p_n)}$; if $a = 1$, we assume $\tau^2 \leq p_n/n$; and if $a > 1$, we assume $\tau^2 \leq (p_n/n)\log(n/p_n)$.

Below, we check Conditions 1–3.

Condition 1': It is enough to show that π' is a uniformly regular varying function. Notice that L is uniformly regular varying and satisfies (3) with $R = M/c_0$ and $z_0 = t_0$. If two functions are uniformly regular varying, then also their product, and thus π' is uniformly regular varying.

Condition 2: Because of $p_n = o(n)$, $\tau^2 \to 0$. Observe that $u \geq t_0\tau^2$ implies $L(u/\tau^2) \geq c_0$ and thus

$$\int_0^1 \pi(u)du \geq \int_{t_0\tau^2}^{(t_0+1)\tau^2} \pi(u)du \geq \int_{t_0\tau^2}^{(t_0+1)\tau^2} \frac{c_0 K\tau^{2a}}{u^{1+a}}du = \frac{c_0 K}{(t_0+1)^{1+a}}.$$

Condition 3: Since L is bounded in sup-norm by M, and $s_n \geq \tau^2$, we find that $\pi(u) \leq KM\tau^{2a}u^{-1-a}$, for all $u \geq s_n$. With this bound, it is straightforward to verify Condition 3.

Thus, we can apply Theorem 2.1. □

In particular, the posterior concentration theorem holds even more generally than shown by [12], as the restriction $a < 1$ can be removed. Thus, for example, we recover Theorem 3.3 of [26] and in addition, find that the normal-exponential-gamma prior of [13] contracts at at most the minimax rate for $\gamma = p_n/n$ and any $\lambda \geq 1/2$.

3.2. The inverse-Gaussian prior

Caron and Doucet [4] propose to use the inverse-Gaussian distribution as prior for σ^2. For positive constants b and τ the variance σ^2 is drawn from an inverse Gaussian distribution with mean $\sqrt{2}\tau$ and shape parameter $\sqrt{2b}$. Thus the prior on the components is of the form (1) with

$$\pi(u) = \frac{C_{b,\tau}\tau}{u^{3/2}}e^{-\frac{\tau^2}{u}-bu},$$

where $C_{b,\tau} = e^{2\sqrt{b}\tau}/\sqrt{\pi}$ is the normalization factor. (In the notation of [4], this corresponds to reparametrizing $\gamma = \sqrt{2b}$, $\alpha/n = \sqrt{2}\tau$, and $K = n$ is the dimension of the unknown mean vector.) As τ becomes small the distribution is concentrated near zero. [4] suggests to take τ proportional to $1/n$, and we find that optimal rates can be achieved if $(p_n/n)^K \lesssim \tau \leq (p_n/n)\sqrt{\log(n/p_n)}$ for some $K > 1$.

Below we verify Condition 1 and Condition A, which together imply Theorem 2.1. The inverse-Gaussian prior does not fit within the class considered by [12], because of the additional exponential factors.

Condition 1: For $u \geq 1$, $e^{-1} \leq e^{-\tau^2/u} \leq 1$. Thus, $u \mapsto e^{-\tau^2/u}$ is uniformly regular varying with constants $R = e$ and $z_0 = 1$. Since products of uniformly regular varying functions are again uniformly regular varying, we can write $\pi(u) = L_n(u)e^{-bu}$ with L_n uniformly regular varying.

For $u \geq 1$, $\pi(u) \geq \pi^{-1/2}e^{-1}\tau u^{-3/2}e^{-bu}$, using the explicit expression for the constant $C_{b,\tau}$. Thus, (4) holds with $b' > b$, $K = \alpha$, $z_* = 1$, and C' a sufficiently large constant.

Condition A: Observe that $\pi(u) \leq C_{b,1}\tau u^{-3/2}$.

Hence, the statement of Theorem 2.1 follows. □

3.3. The horseshoe+ prior

The horseshoe+ prior was introduced by [2]. It is an extension of the horseshoe including an additional latent variable. A Cauchy random variable with parameter λ that is conditioned to be positive is said to be half-Cauchy and we write $C^+(0, \lambda)$ for its distribution. The horseshoe+ prior can be defined via the hierarchical construction

$$\theta_i \mid \sigma_i \sim \mathcal{N}(0, \sigma_i^2), \quad \sigma_i \mid \eta_i, \tau \sim C^+(0, \tau\eta_i), \quad \eta_i \sim C^+(0, 1).$$

and should be compared to the horseshoe prior

$$\theta_i \mid \sigma_i \sim \mathcal{N}(0, \sigma_i^2), \quad \sigma_i \mid \tau \sim C^+(0, \tau).$$

The additional variable η_i allows for another level of shrinkage, a role which falls solely to τ in the horseshoe prior. In [2], the claim is made that the horseshoe+ is an improvement over the horseshoe in several senses, but no posterior concentration results are known so far. With Theorem 2.1, we can show that the horseshoe+ enjoys the same upper bound on the posterior contraction rate as the horseshoe, if $(p_n/n)^K \lesssim \tau \lesssim (p_n/n)(\log(n/p_n))^{-1/2}$, for some $K > 1$.

The horseshoe+ prior is of the form (1) with

$$\pi(u) = \frac{\tau}{\pi^2} \frac{\log(u/\tau^2)}{(u - \tau^2)u^{1/2}}.$$

Below, we verify Conditions 1–3.

Condition 1: Write $\pi(u) = L_n(u)$, that is, $b = 0$. Let us show that L_n is uniformly regular varying. For that define $u_0 := 2$. For $u > u_0$, and $\tau^2 \leq 1$ we have $u/2 \leq u - \tau^2 \leq u$, thus

$$\frac{1}{2}a^{-3/2}\frac{\log(u/\tau^2) + \log(a)}{\log(u/\tau^2)} \leq \frac{\pi(au)}{\pi(u)} \leq 2a^{-3/2}\frac{\log(u/\tau^2) + \log(a)}{\log(u/\tau^2)}.$$

Since

$$1 \leq \frac{\log(u/\tau^2) + \log(a)}{\log(u/\tau^2)} \leq 2,$$

L_n is regular varying. To check the second part of the assumption, observe that $\pi(u) \geq \pi^{-1}\tau u^{-3/2}\log(u/\tau^2)$. For any $K > \alpha$ and any $b' > 0$,

$$\pi(u)e^{b'u} \gtrsim \tau\log(1/\tau) \geq \Big(\frac{p_n}{n}\Big)^K, \quad \text{for all } u \geq u_0.$$

Thus, Condition 1 holds.

Condition 2: Observe that

$$\int_0^1 \pi(u)du \geq \frac{\tau}{\pi^2}\int_0^{\tau^2/2}\frac{\log(\tau^2/u)}{(\tau^2-u)u^{1/2}}du \geq \frac{\tau}{\pi^2}\frac{1}{(\tau^2/2)^{3/2}}\cdot\frac{\tau^2}{2}\log\tfrac{1}{2} \gtrsim 1.$$

Condition 3: For any $u \geq s_n$ we can use $(u-\tau^2) \geq u/2$. This shows that

$$\pi(u) \leq \frac{\tau\log(u)}{u^{3/2}} + \frac{\tau\log(1/\tau^2)}{u^{3/2}}, \quad \text{for all } u \geq s_n.$$

In particular, $\pi(u) \lesssim \tau\log(n/p_n)/u^{3/2}$ for $s_n \leq u \leq b_n^2$. For the integral on $[b_n^2,\infty)$, we use that $\frac{d}{du} - (\log(u)+1)/u = \log(u)/u^2$. Together, Condition 3 follows thanks to $\tau \lesssim (p_n/n)/\sqrt{\log(n/p_n)}$.

Thus, Theorem 2.1 can be applied. □

3.4. Normal-gamma prior

The normal-gamma prior, discussed by [4] and [14], takes the following form for shape parameter $\tau > 0$ and rate parameter $\beta > 0$:

$$\pi(u) = \frac{\beta^\tau}{\Gamma(\tau)}u^{\tau-1}e^{-\beta u} = \frac{\tau\beta^\tau}{\Gamma(\tau+1)}u^{\tau-1}e^{-\beta u}.$$

In [14], it is observed that decreasing τ leads to a distribution with a lot of mass near zero, while preserving heavy tails. This is also illustrated in the rightmost panels of Figure 1. The class of normal-gamma priors includes the double exponential prior as a special case, with $\tau = 1$. We now show that the normal-gamma prior satisfies the conditions of Theorem 2.1 for any fixed β, and for any $(p_n/n)^K \lesssim \tau \lesssim (p_n/n)\sqrt{\log(n/p_n)} \leq 1$ for some fixed K.

Below, we check Conditions 1–3.

Condition 1: We define $L_n(u) = \frac{\beta^\tau}{\Gamma(\tau)}u^{\tau-1}$, so $\pi(u) = L_n(u)e^{-bu}$ with $b = \beta$. Note that since $\tau \to 0$, we have that there exist a constant C such that $C^{-1} \leq \beta^\tau \leq C$. We now prove that L_n is regular varying. We have

$$\frac{L_n(au)}{L_n(u)} = a^{\tau-1}.$$

and thus for all $a \in [1,2]$, $a^{-1} \leq L_n(au)/L_n(u) \leq 1$. In addition for $u > u_* := 1$ we have, using $\Gamma(\tau+1) \geq \Gamma(1) = 1$,

$$L_n(u) = \frac{\tau\beta^\tau}{\Gamma(\tau+1)}u^{\tau-1} \geq \frac{(\beta\wedge 1)\tau}{\Gamma(2)u} \gtrsim \Big(\frac{p_n}{n}\Big)^K\frac{1}{u},$$

implying $\pi(u) = L_n(u)u^{-1}e^{-\beta u} \gtrsim (p_n/n)^K e^{-2\beta u}$. Thus Condition 1 is satisfied.

Condition 2:

$$\int_0^1 \pi(u)du \geq \frac{(\beta \wedge 1)e^{-bu}\tau}{\Gamma(2)} \int_0^1 u^{\tau-1}du = \frac{(\beta \wedge 1)e^{-bu}}{\Gamma(2)} \gtrsim 1.$$

Condition 3: Notice that $\pi(u) \leq (\beta \vee 1)\tau u^{\tau-1}$, for all $u \leq 1$. For $u \geq 1$, we find $\pi(u) \leq (\beta \vee 1)\tau e^{-\beta u}$. Since $e^{-\beta u}$ decays faster than any polynomial power of u, we see that Condition 3 holds thanks to $b_n\tau \lesssim s_n$.

Thus, we can apply Theorem 2.1.

In [14], it is discussed that the extra modelling flexibility afforded by generalizing the double exponential prior to include the parameter τ is essential, and indeed the double exponential ($\tau = 1$) does not allow a dependence on p_n and n such that our conditions are met.

3.5. Spike-and-slab Lasso prior

The spike-and-slab Lasso prior was introduced by [24]. It may be viewed as a continuous version of the usual spike-and-slab prior with a Laplace slab, as studied in [7, 6], where the spike component has been replaced by a very concentrated Laplace distribution. Recent theoretical results, including posterior concentration at the minimax rate, have been obtained in [24]. Here, we recover Corollary 6.1 of [24].

For a fixed constant $a > 0$ and a sequence $\tau \to 0$, we define the spike-and-slab Lasso as prior of the form (1) with hyperprior

$$\pi(u) = \omega a e^{-au} + (1-\omega)\frac{1}{\tau}e^{-\frac{u}{\tau}}, \quad u > 0 \tag{10}$$

on the variance. Recall that the Laplace distribution with parameter λ is a scale mixture of normals where the mixing density is exponential with parameter λ^2. Applied to model (1), the prior on θ_i is thus a mixture of two Laplace distributions with parameter $\sqrt{a}$ and $\tau^{-1/2}$ and mixing weights ω and $1-\omega$, respectively and this justifies the name.

We now prove that the prior satisfies the conditions of Theorem 2.1 for mixing weights satisfying $(p_n/n)^K \leq \omega \leq (p_n/n)\sqrt{\log(n/p_n)} \leq \frac{1}{2}$, for some $K > 1$ and $\tau = (p_n/n)^\alpha$ with $\alpha \geq 1$.

Condition 1: To prove that Condition 1 holds we rewrite the prior π as

$$\pi(u) = e^{-au}\left(a\omega + \frac{1-\omega}{\tau}e^{-u(\frac{1}{\tau}-a)}\right) =: e^{-au}L_n(u)$$

For n large enough, we have $1/\tau - a > 1/(2\tau)$. For all $u > 1$ and for $C > 0$ a constant depending only on K and α,

$$\frac{1-\omega}{\tau}e^{-u(\frac{1}{\tau}-a)} \leq \frac{1}{\tau}e^{-\frac{1}{2\tau}} \leq C\tau^{\frac{K}{\alpha}} \leq C\omega.$$

Hence, for sufficiently large n, $a\omega \leq L_n(u) \leq (a+C)\omega$ for all $u \geq 1$. Thus L_n is regular varying with $u_0 = 1$. Since also $\pi(u) \geq a\omega e^{-au}$ and $\omega \geq (p_n/n)^K$, Condition 1 holds.

Condition 2: $\int_0^1 \pi(u)du \geq (1-\omega)\int_0^\tau \frac{1}{\tau}e^{-\frac{u}{\tau}}du = (1-\omega)(1-e^{-1})$.

Condition 3: We might split the two mixing components in (10) and write $\pi =: \pi_1 + \pi_2$. To verify the condition for the first component π_1, we use that $e^{-au} \leq 1$ for $u \leq 1$ and that e^{-au} decays faster than any polynomial for $u > 1$. In order that Condition 3 is satisfied, we need thus $\omega \lesssim (p_n/n)\sqrt{\log(n/p_n)}$. For π_2, there exists a constant C such that $\pi_2(u) \leq C\tau/u^2$ for all $u \geq s_n$, due to $s_n \geq \tau$. Straightforward computations show that π_2 satisfies Condition 3 since $\tau \leq p_n/n$.

Thus, we can apply Theorem 2.1. □

4. Simulation results

To illustrate the point that our conditions are very sharp, we compute the average square loss for four priors that do not meet our conditions, and compare them with two of the examples from Section 3.

The two priors considered in this simulation study that do meet the conditions are the horseshoe and the normal-gamma priors, both with $\tau = p_n/n$. The four priors that do not meet the conditions are the Lasso (Laplace prior) with $\lambda = 1$ and $\lambda = 2n/\log n$ (see Section 3.4), and two prriors of the form (9) of Section 3.1 with $a = 0.1$ and $a = 0.4$, $L(u) = e^{-1/u}$ and density,

$$\pi(u) \propto u^{-(1+a)}e^{-\tau^2/u},$$

and we take $\tau = p_n/n$. This prior will be referred to as a $GC(a)$ prior hereafter. Note that π does not meet our conditions, as explained in Section 3.1.

For each of these priors, we sample from the posterior distribution using a Gibbs Sampling algorithm, following the one proposed for the horseshoe prior by [5]. To do so, we first compute the full conditional distributions

$$p(\beta|X,\sigma^2) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}}e^{-\frac{1}{2\hat{\sigma}^2}(\beta-\hat{\beta})^2}$$
$$p(\sigma^2|X,\beta) \propto (\sigma^2)^{-1/2}e^{-\frac{\beta^2}{2\sigma^2}}\pi(\sigma^2),$$

where $\hat{\sigma}^2 = \sigma^2/(1+\sigma^2)$ and $\hat{\beta} = X\sigma^2/(1+\sigma^2)$. The only difficulty is thus sampling from $p(\sigma^2|X,\beta)$. For the horseshoe prior we follow the approach proposed by [5]. We apply a similar method for the normal-gamma prior using the approach proposed by [8]. Sampling from the $GC(a)$ priors is even simpler given that in this case $p(\sigma|X,\beta)$ is an inverse gamma. We compute the mean integrated squared error (MISE) on 500 replicates of simulated data of size $n = 100, 250, 500, 1000$. The MISE is equal to $\mathbb{E}_{\theta_0}\sum_i[(\widehat{\theta}_i - \theta_{0i})^2 + \mathrm{Var}(\theta_i \mid X)]$. For each n, we fix the number of nonzero means at $p_n = 10$, and take the

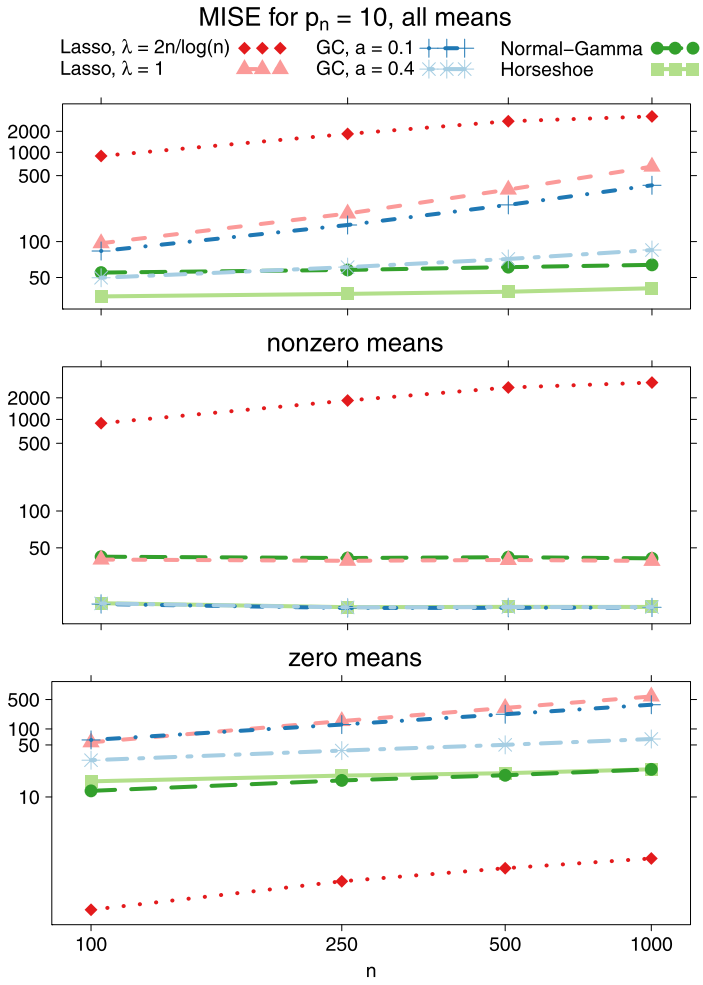

FIG 2. *The logarithm of the integrated square loss for the Lasso (Laplace) with* $\lambda = 2n/\log n$ *and* $\lambda = 1$*, the GC priors of [12] discussed in section 3.1 with* $a = 0.1$ *and* $a = 0.4$*, the normal-namma and horseshoe priors plotted against* $\log\log n$*, computed on 500 replicates of the data for each value of* n*. From top to bottom: MISE for all means, for only the* $p_n = 10$ *nonzero means, and for the* $(n - p_n)$ *zero means. The axis labels refer to the original, non-log-transformed scale.*

nonzero coefficients equal to $5\sqrt{2\log n}$. This value is well past the 'universal threshold' of $\sqrt{2\log n}$, and thus the signals should be relatively easy to detect. For each data set, we compute the posterior square loss using 5000 draws from the posterior with a burn-in of 20%.

The results are presented in Figure 2, for all means together and separately for the nonzero and zero means. Given that $p_n = 10$ is fixed, if the posterior contracts at the minimax rate, then the integrated square loss should be linear in $\log n$. However, we see that for both Laplace priors and the $GC(a = 0.1)$

priors, and less so for the $GC(a = 0.4)$ prior, the slope of the loss grows with n, when it remains steady for the other two considered priors. In addition, we see the expected trade-off for the two choices of the tuning parameter λ for the Lasso. A large value of λ results in strong shrinkage and thus low MISE on the zero means, but very high MISE on the nonzero means, while a small value of λ leads to barely any shrinkage, and we observe a relatively low MISE on the nonzero means but a high MISE on the zero means. The $GC(a)$ prior with $a = 0.1$ does not perform well, because it undershrinks. The same effect is visible for $a = 0.4$, but less so. The normal-gamma and horseshoe priors both have low MISE on the zero and nonzero means; the horseshoe outperforms the normal-gamma because it shrinks the nonzero means less.

These results suggest that the horseshoe and normal-gamma strike a better balance between shrinking the zero means without affecting the nonzero means than the four priors that do not meet our conditions, leading to lower risk and illustrating that our conditions are very sharp.

5. Discussion

Our main theorem, Theorem 2.1, expands the class of shrinkage priors with theoretical guarantees for the posterior contraction rate. Not only can it be used to obtain the optimal posterior contraction rate for the horseshoe+, the inverse-Gaussian and normal-gamma priors, but the conditions provide some characterization of properties of sparsity priors that lead to desirable behaviour. Essentially, the tails of the prior on the local variance should be at least as heavy as Laplace, but not too heavy, and there needs to be a sizable amount of mass around zero compared to the amount of mass in the tails, in particular when the underlying mean vector grows to be more sparse.

In [20] global-local scale mixtures of normals like (5) are discussed, with a prior on the parameter τ^2. Their guidelines are twofold: the prior on the local variance σ_i^2 should have heavy tails, while the prior on the global variance τ^2 should have substantial mass around zero. They argue that any prior on σ_i^2 with an exponential tail will force a tradeoff between shrinking the noise towards zero and leaving the large nonzero means unshrunk, while the shrinkage of large signals will go to zero when a prior with a polynomial tail is chosen. This matches the intuition behind our conditions, with the remark that exponential tails *are* possible, but they should not be lighter than Laplace.

Besides the three discussed goals of recovery, uncertainty quantification, and computational simplicity, we might have mentioned a fourth: performing *model selection* or *multiple testing*. Priors of the type studied in this paper are not directly applicable for this goal, as the posterior mean will, with probability one, not be exactly equal to zero. A model selection procedure can be constructed however, for example by thresholding using the observed values of m_{x_i}: if m_{x_i} is larger than some constant, we consider the underlying parameter to be a signal, and otherwise we declare it noise. Such a procedure was proposed for the horseshoe by [5], and was shown to enjoy good theoretical properties by

[9]. Similar results were found for the horseshoe+ [2]. The same thresholding procedure, and similar analysis methods, may prove to be fruitful for the more general prior (1).

Appendix A: Proofs

This section contains the proofs of Theorem 2.1 and Theorem 2.2, followed by the statement and proofs of the supporting Lemmas. The proof of Theorem 2.1 follows the same structure as that of Theorem 3.3 in [26], but requires more general methods to bound the integrals involved in the proof.

In the course of the proofs, we use the following two transformations of π,

$$g(z) = \frac{1}{z^2}\pi\left(\frac{1-z}{z}\right) \quad \text{and} \quad h(z) = \frac{1}{(1-z)^{3/2}}\pi\left(\frac{z}{1-z}\right). \tag{11}$$

The function g is a density on $[0,1]$, resulting from transforming the density π on σ_i^2 to a density for $z = (1+\sigma_i^2)^{-1}$. The function h is a rescaled version of π.

Lemma A.1. *Condition 1' implies Condition 1.*

Proof. Observe that $\pi(u) = \widetilde{\pi}(u/\tau^2)/\tau^2$. Since by assumption $\widetilde{\pi}$ is uniformly regular varying, (3) holds for some constants R and u_0 which do not depend on n. To check the first part of Condition 1, it is enough to see that $\widetilde{\pi}(\cdot/\tau^2)$ is uniformly regular varying as well and satisfies (3) with the same constants as $\widetilde{\pi}$.

It remains to prove a lower bound (4). Thanks to $\tau^2 \leq 1$ and Lemma A.3, for any $u \geq u_* := u_0$, $\widetilde{\pi}(u/\tau^2) \geq \widetilde{\pi}(u_0)(\tau^2 u_0/2u)^{\log_2 R}$. This implies the lower bound (4) with $K = 2\alpha \log_2 R$, $b' > 0$, and C' a sufficiently large constant. □

Proof of Theorem 2.1. Applying Lemma A.5 gives under Condition 1, $\sum_{i:\theta_i \neq 0} \mathbb{E}_{\theta_i}(\theta_i - \widehat{\theta}_i)^2 \lesssim p_n \log(n/p_n)$ and $\sum_{i:\theta_i \neq 0} \mathbb{E}_{\theta_i} \operatorname{Var}(\theta_i \mid X_i) \lesssim p_n \log(n/p_n)$. These inequalities combined with Markov's inequality prove the first two statements of the theorem. Similarly, under Condition 2 and Condition 3, we obtain from Lemma A.6 and Lemma A.7, $\mathbb{E}_\theta \sum_{i:\theta_i=0} \widehat{\theta}_i^2 \leq n\mathbb{E}_0(Xm_X)^2 \lesssim p_n \log(n/p_n)$ and $\sum_{i:\theta_i=0} \mathbb{E}_0 \operatorname{Var}(\theta_i \mid X_i) \lesssim p_n \log(n/p_n)$. Together with Markov's inequality, this proves the third and fourth statement of the theorem. □

Proof of Theorem 2.2. Without loss of generality, we can take κ_n such that $\kappa_n \geq n^{-1/4}$ for all n. Consider the prior, where θ_i is drawn from the Laplace density with parameter $\lambda = \sqrt{\kappa_n/s_n}$. This prior is of the form (1) with $\pi(u) = \lambda^2 e^{-\lambda^2 u}$ (cf. Section 2.1). Theorem 7 in [6] shows that (8) holds with $M_n = 1/\kappa_n \to \infty$. Thus it remains to prove that π satisfies Condition 2 and Condition 3(κ_n).

Condition 2 follows immediately. For Condition 3(κ_n) observe that due to $\kappa_n \geq n^{-1/4}$, $\lambda \geq n^{1/4}/\sqrt{\log n}$. Splitting the integral $\int_0^{\lambda^2} = \int_0^1 + \int_1^{\lambda^2}$, we find $\kappa_n \int_{s_n}^1 u\pi(u)du \leq \kappa_n \int_0^1 u\lambda^2 e^{-\lambda^2 u}du \leq \kappa_n \lambda^{-2} \int_0^{\lambda^2} ve^{-v}dv \lesssim \kappa_n \lambda^{-2} = s_n$. Also, $\int_1^{b_n^2} u\pi(u)du = \lambda^{-2} \int_{\lambda^2}^{b_n^2 \lambda^2} ve^{-v}dv \leq b_n^2 e^{-\lambda^2} = o(s_n)$ and $b_n^3 \int_1^\infty \pi(u)/\sqrt{u}du \leq b_n^3 \int_1^\infty \pi(u)du \leq b_n^3 e^{-\lambda^2} = o(s_n)$. Hence, Condition 3($\kappa_n$) holds and this completes the proof. □

Lemma A.2. *The posterior variance can be written as*

$$\mathrm{Var}(\theta \mid x) = m_x - (xm_x - x)^2 + x^2 \frac{\int_0^1 (1-z)^2 h(z) e^{\frac{x^2}{2}z} dz}{\int_0^1 h(z) e^{\frac{x^2}{2}z} dz} \tag{12}$$

and bounded by

$$\mathrm{Var}(\theta \mid x) \leq 1 + x^2 \frac{\int_0^1 (1-z)^2 h(z) e^{\frac{x^2}{2}z} dz}{\int_0^1 h(z) e^{\frac{x^2}{2}z} dz} \quad \text{and} \quad \mathrm{Var}(\theta \mid x) \leq m_x + x^2 m_x. \tag{13}$$

Proof. By Tweedie's formula [23], the posterior variance for θ_i given an observation x_i is equal to $1 + (d^2/dx^2) \log p(x)|_{x=x_i}$, where $p(x_i)$ is the marginal distribution of x_i. Computing

$$p(x) = \int_0^1 \frac{1}{\sqrt{2\pi}} (1-z)^{-3/2} e^{-\frac{x^2}{2}(1-z)} \pi\left(\frac{z}{1-z}\right) dz,$$

taking derivatives with respect to x, and substituting $h(z) = (1-z)^{-3/2}\pi(z/(1-z))$ gives

$$\begin{aligned}\mathrm{Var}(\theta \mid x) = 1 + x^2 \frac{\int_0^1 (1-z)^2 h(z) e^{\frac{x^2}{2}z} dz}{\int_0^1 h(z) e^{\frac{x^2}{2}z} dz} - \frac{\int_0^1 (1-z) h(z) e^{\frac{x^2}{2}z} dz}{\int_0^1 h(z) e^{\frac{x^2}{2}z} dz} \\ - x^2 \left(\frac{\int_0^1 (1-z) h(z) e^{\frac{x^2}{2}z} dz}{\int_0^1 h(z) e^{\frac{x^2}{2}z} dz} \right)^2.\end{aligned}$$

From that we can derive (12) noting that the third term on the r.h.s. is $1 - m_x$. The last display also implies the first inequality in (13). Representation (12) together with the trivial bound $(1-z)^2 \leq (1-z)$ for $z \in [0,1]$ yields

$$x^2 \frac{\int_0^1 (1-z)^2 h(z) e^{\frac{x^2}{2}z} dz}{\int_0^1 h(z) e^{\frac{x^2}{2}z} dz} \leq x^2 \frac{\int_0^1 (1-z) h(z) e^{\frac{x^2}{2}z} dz}{\int_0^1 h(z) e^{\frac{x^2}{2}z} dz} = x^2(1 - m_x).$$

Combined with (12), we find $\mathrm{Var}(\theta \mid x) \leq m_x - x^2 m_x^2 + x^2 m_x \leq m_x + x^2 m_x$. □

Lemma A.3. *Suppose that L is uniformly regular varying. If R and u_0 are chosen such that (3) holds, then, for any $a \geq 1$, and any $u \geq u_0$,*

$$L(u) \leq (2a)^{\log_2 R} L(au),$$

where $\log_2$ denotes the binary logarithm.

Proof. Write $a = 2^r b$ with r a non-negative integer and $1 \leq b < 2$. By assumption (3) holds for some R and u_0. We apply the upper bound (3) repeatedly and obtain for $a \geq 1$, $L(u) \leq RL(2u) \leq \ldots \leq R^r L(2^r u) \leq R^{r+1} L(au)$. Since $R^{r+1} = (2^{r+1})^{\log_2 R} \leq (2a)^{\log_2 R}$, the result follows. □

Lemma A.4. *Assume that L is uniformly regular varying and satisfies (3) with R and u_0. Then, the shifted function $L(\cdot - 1)$ is also uniformly regular varying with constants R^3 and $u_0 \vee 2$.*

Proof. Write

$$\frac{L(az-1)}{L(z-1)} = \frac{L(az-1)}{L(az)} \cdot \frac{L(az)}{L(z)} \cdot \frac{L(z)}{L(z-1)}.$$

For $z \geq z_0 \vee 2$ we apply (3) to each of the three fractions and this completes the proof. □

The following lemma states that if the density g can be decomposed as a product of a function that is uniformly varying and possibly n dependent, and a factor of the form $z \mapsto e^{-bz}$, then the posterior recovers the size of the non-zero components of θ with the minimax estimation rate, provided that the n dependence is of the right order.

Lemma A.5. *If Condition 1 holds, there exists a constant C, which is independent of n, such that*

$$\sum_{i:\theta_i \neq 0} \mathbb{E}_{\theta_i}(X_i m_{X_i} - \theta_i)^2 \leq C p_n \log(en/p_n), \tag{14}$$

and

$$\sum_{i:\theta_i \neq 0} \mathbb{E}_{\theta_i} \operatorname{Var}(\theta_i | X_i) \leq C p_n \log(en/p_n). \tag{15}$$

Proof. We prove the two statements separately. The main argument is a careful analysis of the integral representation

$$|x(m_x - 1)| = |x| \frac{\int_0^1 e^{-\frac{x^2}{2}z} z^{-1/2} \pi\big(\frac{1}{z} - 1\big) dz}{\int_0^1 e^{-\frac{x^2}{2}z} z^{-3/2} \pi\big(\frac{1}{z} - 1\big) dz} = |x| \frac{\int_0^1 e^{-\frac{x^2}{2}u} u^{3/2} g(u) du}{\int_0^1 e^{-\frac{x^2}{2}u} u^{1/2} g(u) du}$$

(cf. (2) and (11)). Throughout the remaining proof, let C_1 be a generic constant which is independent of n and which might change from line to line. Without loss of generality, we may assume that $u_0 \geq 2$ in Condition 1.

Proof of (14): It is enough to show $\sup_{x>0} |x(m_x - 1)| \lesssim 1 + \sqrt{\log(n/p_n)}$. It is thus enough to consider the sup over $|x| > T_0 := 2 + 2(u_0 \vee u_*) + \sqrt{8u_0 K \log(n/p_n)}$, since otherwise, we simply use $|x(m_x - 1)| \leq |x|$.

For $0 \leq a < b \leq 1$, write $I(a,b) = \int_a^b e^{-\frac{x^2}{2}u} u^{3/2} g(u) du / \int_0^1 e^{-\frac{x^2}{2}u} u^{1/2} g(u) du$ and for $b \leq a$, set $I(a,b) = 0$. We need to prove that

$$I(0,1) = I\big(0, \tfrac{2b+4}{|x|}\big) + I\big(\tfrac{2b+4}{|x|}, \tfrac{1}{u_0}\big) + I\big(\tfrac{1}{u_0}, 1\big) =: (I) + (II) + (III) \lesssim \frac{1}{|x|}.$$

Bound for (I): Obviously, $I(0,v) \leq v$ for all $v \in (0,1]$. Thus, $I\big(0, \frac{2b+4}{|x|}\big) \leq C_1/|x|$.

Bound for (II): We first derive a lower bound for the denominator. Recall that by Condition 1, $\pi(u) = L_n(u)e^{-bu}$. Define $\widetilde{L}_n = L_n(\cdot - 1)$ and observe that due to $|x| \geq 2u_0$ we can use Lemma A.4 and substitute $v = u|x|/2$ to obtain

$$\begin{aligned}\int_0^1 e^{-\frac{x^2}{2}u}u^{-3/2}\pi\big(\tfrac{1}{u}-1\big)du &\geq \int_{1/|x|}^{2/|x|} e^{-\frac{x^2}{2}u}u^{-3/2}\widetilde{L}_n\big(\tfrac{1}{u}\big)e^{-\frac{b}{u}+b}du \qquad (16)\\ &\geq \frac{1}{4}e^{b-(1+b)|x|}|x|^{3/2}\int_{1/|x|}^{2/|x|}\widetilde{L}_n\big(\tfrac{1}{u}\big)du\\ &= \frac{1}{4}e^{b-(1+b)|x|}|x|^{1/2}2\int_{1/2}^{1}\widetilde{L}_n\big(\tfrac{1}{v}\cdot\tfrac{|x|}{2}\big)dv\\ &\geq \frac{1}{4R^3}e^{b-(1+b)|x|}|x|^{1/2}\widetilde{L}_n\big(\tfrac{|x|}{2}\big). \qquad (17)\end{aligned}$$

For the numerator, using Lemma A.3 with $u = |x|/v$ and $a = v/2$,

$$\begin{aligned}&\int_{(2b+4)/|x|}^{u_0^{-1}} e^{-\frac{x^2}{2}u}u^{-1/2}\pi\big(\tfrac{1}{u}-1\big)du\\ &= \sum_{k=1}^{\infty}\int_{(2b+4+k-1)/|x|}^{(2b+4+k)/|x|} e^{-\frac{x^2}{2}u}u^{-1/2}\widetilde{L}_n\big(\tfrac{1}{u}\big)e^{b-\frac{b}{u}}\mathbf{1}(u\leq u_0^{-1})du\\ &\leq e^b\sum_{k=1}^{\infty}e^{-\frac{|x|}{2}(2b+4+k-1)}\Big(\tfrac{|x|}{2b+4+k-1}\Big)^{1/2}\int_{(2b+4+k-1)/|x|}^{(2b+4+k)/|x|}\widetilde{L}_n\big(\tfrac{1}{u}\big)\mathbf{1}(u\leq u_0^{-1})du\\ &\leq e^b\sum_{k=1}^{\infty}e^{-\frac{|x|}{2}(2b+2+k)}|x|^{-1/2}\int_{2b+4+k-1}^{2b+4+k}\widetilde{L}_n\big(\tfrac{|x|}{v}\big)\mathbf{1}\big(v\leq\tfrac{|x|}{u_0}\big)dv\\ &\leq e^{-|x|(b+1)}|x|^{-1/2}\widetilde{L}_n\big(\tfrac{|x|}{2}\big)e^b\sum_{k=1}^{\infty}e^{-\frac{|x|}{2}k}(2b+4+k)^{3\log_2 R}.\end{aligned}$$

The sum $\sum_{k=1}^{\infty}e^{-\frac{|x|}{2}k}(2b+4+k)^{3\log_2 R}$ is bounded for $|x| > T_0$. Since by assumption, R does not depend on n, we find $I\big(\frac{2b+4}{|x|},\frac{1}{u_0}\big) \leq C_1/|x|$.

Bound for (III): Since g is a density, we obtain

$$\int_{u_0^{-1}}^{1} e^{-\frac{x^2}{2}u}u^{3/2}g(u)du \leq e^{-x^2/(2u_0)}.$$

For the denominator, we find using (17), $|x| \geq 2 + 2u_*$, and Condition 1,

$$\begin{aligned}\int_0^1 e^{-\frac{x^2}{2}u}u^{-3/2}\pi\big(\tfrac{1}{u}-1\big)du &\geq \frac{1}{4R^3}e^{-(1+\frac{b}{2})|x|}|x|^{1/2}\pi\big(\tfrac{|x|}{2}-1\big)\\ &\geq \frac{1}{4R^3C'}\big(\tfrac{p_n}{n}\big)^K e^{-(1+b+b')|x|}|x|^{1/2}.\end{aligned}$$

Combining this with the upper bound and $(1+b+b')|x| \leq (1+b+b')^2u_0 + x^2/(4u_0)$ gives

$$I\big(\tfrac{1}{u_0},1\big) \leq 4C'R^3\big(\tfrac{n}{p_n}\big)^K|x|^{-1/2}e^{(1+b+b')^2u_0}e^{-x^2/(4u_0)}.$$

Using that $x \mapsto |x|^{1/2} e^{-x^2/(8u_0)}$ is bounded and $|x| > T_0$ yields $I\big(\frac{1}{u_0}, 1\big) \leq C_1/|x|$.

The result for (14) follows by combining the bounds (I)–(III).

Proof of (15): Recall that (13) uses $h(u) = (1-u)^{-3/2}\pi(u/(1-u))$. With (11), $h(1-u) = u^{-3/2}\pi((1-u)/u) = u^{1/2}g(u)$. Therefore, we find

$$\mathrm{Var}(\theta|x) \leq 1 + x^2 \frac{\int_0^1 e^{-\frac{x^2}{2}u} u^{5/2} g(u)du}{\int_0^1 e^{-\frac{x^2}{2}u} u^{1/2} g(u)du}.$$

Arguing as for (14) completes the proof. □

Next, we provide the technical lemmas establishing the rate for the zero coefficients. Recall that $s_n = (p_n/n)\log(n/p_n)$ and define

$$q_n := \frac{p_n}{n}\sqrt{\log(n/p_n)}. \tag{18}$$

Suppose that Condition 2 and Condition 3 hold with constants c and C, respectively. With (2),

$$\begin{aligned} m_x &:= \frac{\int_0^\infty \frac{u}{(1+u)^{3/2}} e^{\frac{x^2 u}{2+2u}} \pi(u)du}{\int_0^\infty \frac{1}{(1+u)^{1/2}} e^{\frac{x^2 u}{2+2u}} \pi(u)du} \\ &\leq s_n + \frac{\sqrt{2}}{c}\int_{s_n}^\infty \frac{u e^{\frac{x^2 u}{2+2u}}}{(1+u)^{3/2}}\pi(u)du \\ &\leq s_n\big(1 + \frac{\sqrt{2}C}{c} e^{\frac{x^2}{4}}\big) + \frac{\sqrt{2}}{c}\int_1^\infty \frac{u e^{\frac{x^2 u}{2+2u}}}{(1+u)^{3/2}}\pi(u)du \\ &\leq s_n\big(1 + \frac{\sqrt{2}C}{c} e^{\frac{x^2}{4}}\big) + \frac{\sqrt{8}C}{c} q_n e^{\frac{x^2}{2}}, \end{aligned} \tag{19}$$

where for the last inequality, we split the integral $\int_1^\infty = \int_1^{\log(n/p_n)} + \int_{\log(n/p_n)}^\infty$ and used Condition 3 twice. These inequality will be very useful for the proofs below. For the variance bound, the last bound is not sharp enough and we need to work with the upper bound induced by the second inequality.

Lemma A.6. *Work under Condition 2 and Condition 3. Then,*

$$\mathbb{E}_0(Xm_X)^2 \lesssim \frac{p_n}{n}\log(n/p_n).$$

Proof. Let q_n be as in (18) and set $a_n := \sqrt{2\log(1/q_n)}$. Decompose

$$\mathbb{E}_0(Xm_X)^2 = \mathbb{E}_0(Xm_X)^2 \mathbf{1}\{|X| \leq a_n\} + \mathbb{E}_0(Xm_X)^2\mathbf{1}\{|X| > a_n\} =: I_1 + I_2.$$

To bound the term I_1, (19) and $x^2 e^{x^2/2} \leq \frac{d}{dx}[x e^{x^2/2}]$ yield

$$I_1 \lesssim s_n^2 \int_{-a_n}^{a_n} x^2 dx + q_n^2 \int_{-a_n}^{a_n} x^2 e^{x^2/2} dx \lesssim s_n^2 a_n^3 + q_n^2 a_n e^{a_n^2/2}.$$

There is a constant only depending on K such that $x^2 \log^K(1/x) \leq C_K x$ for all $x \leq 1$. Thus, $I_1 \lesssim (p_n/n)\log(n/p_n)$.

In order to bound I_2, we use $m_x \leq 1$, $\frac{d}{dx}[-xe^{-x^2/2}] = -e^{-x^2/2} + x^2 e^{-x^2/2}$ and Mills' ratio,

$$\begin{aligned} I_2 &\leq \mathbb{E}_0 X^2 \mathbf{1}\{|X| > a_n\} = 2\int_{a_n}^{\infty} x^2 \phi(x) dx \\ &= 2[-x\phi(x)]_{a_n}^{\infty} + \int_{a_n}^{\infty} \phi(x) dx \leq e^{-a_n^2/2}(2a_n + 1). \end{aligned}$$

Plugging the expression for a_n into the r.h.s. shows that $I_2 \lesssim (p_n/n)\log(n/p_n)$ as well and this finally gives $\mathbb{E}_0(X m_X)^2 \lesssim (p_n/n)\log(n/p_n)$. □

Lemma A.7. *Work under Conditions 2 and 3. Then,*

$$\sum_{i:\theta_i=0}^{n} \mathbb{E}_0 \operatorname{Var}(\theta_i \mid X_i) \lesssim p_n \log(n/p_n).$$

Proof. Let $a_n = \sqrt{2\log(n/p_n)}$. It is enough to show that $\mathbb{E}_0 \operatorname{Var}(\theta \mid X) \lesssim p_n \log(n/p_n)/n$. To prove this, we need to treat the cases that $|X|$ is larger/smaller than a_n, separately. To bound the variance, we use (13), that is $\operatorname{Var}(\theta \mid X) \leq m_x + x^2 m_x \leq 1 + x^2$.

Case $|X| > a_n$: Using the identity $d/dx[x\phi(x)] = \phi(x) - x^2\phi(x)$,

$$\begin{aligned} \mathbb{E}_0 \operatorname{Var}(\theta \mid X)\mathbf{1}_{\{|X|>a_n\}} &\leq 2\int_{a_n}^{\infty} (1+x^2)\phi(x)dx = 2\Phi^c(a_n) + 2\int_{a_n}^{\infty} x^2\phi(x)dx \\ &= 4\Phi^c(a_n) + 2[-x\phi(x)]_{a_n}^{\infty} \leq 4\phi(a_n) + 2a_n\phi(a_n). \quad (20) \end{aligned}$$

Using the expression for a_n shows that this can be bounded by $(p_n/n)\sqrt{\log(n/p_n)}$.

Case $|X| \leq a_n$: Notice that the variance bound implies $\operatorname{Var}(\theta \mid X) \leq m_x \mathbf{1}\{|x| \leq 1\} + 2x^2 m_x$. Below, we estimate $\mathbb{E}_0 m_X \mathbf{1}\{|X| \leq 1\}$ and $\mathbb{E}_0 X^2 m_X \mathbf{1}\{|X| \leq a_n\}$. For the first term, using (19),

$$\mathbb{E}_0 m_X \mathbf{1}\{|X| \leq 1\} \lesssim \int_{-1}^{1} (s_n e^{x^2/4} + q_n e^{x^2/2})\phi(x)dx \leq 4s_n. \quad (21)$$

For the second term $\mathbb{E}_0 X^2 m_X \mathbf{1}\{|X| \leq a_n\}$, we use the second inequality in (19) and find

$$\begin{aligned} \mathbb{E}_0 X^2 m_X \mathbf{1}\{|X| \leq a_n\} &\lesssim s_n \int_{-a_n}^{a_n} x^2 e^{\frac{x^2}{4}} \phi(x)dx \\ &\quad + \int_{-a_n}^{a_n}\int_{1}^{\infty} \frac{u\pi(u)}{(1+u)^{3/2}} x^2 e^{-\frac{x^2}{2+2u}} du dx. \end{aligned}$$

The first integral is bounded by a constant and for the second integral, we use Fubini's theorem, substitute $y = x/\sqrt{1+u}$, and use Condition 3

$$\begin{aligned}\int_{-a_n}^{a_n}\int_1^\infty \frac{u\pi(u)}{(1+u)^{3/2}} x^2 e^{-\frac{x^2}{2+2u}}\,du\,dx &= \int_1^\infty u\pi(u)\int_{-a_n/\sqrt{1+u}}^{a_n/\sqrt{1+u}} y^2 e^{-\frac{y^2}{2}}\,dy\,du \\ &\le \int_1^\infty u\pi(u)\Big[\Big(\frac{a_n}{\sqrt{1+u}}\Big)^3 \wedge \sqrt{2\pi}\Big]\,du \\ &\le 2^{3/2} C s_n.\end{aligned}$$

Together with (21) this shows that $\mathbb{E}_0 \operatorname{Var}(\theta \mid X)\mathbf{1}\{|X| \le a_n\} \lesssim s_n$. Since in both cases the upper bound is of order $(p_n/n)\log(n/p_n)$ the result follows. □

References

[1] Andrews, D. F., and Mallows, C. L. Scale mixtures of normal distributions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* (1974), 99–102. MR0359122

[2] Bhadra, A., Datta, J., Polson, N. G., and Willard, B. The horseshoe+ estimator of ultra-sparse signals. arXiv:1502.00560v2, 2015.

[3] Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. Dirichlet-Laplace priors for optimal shrinkage. arXiv:1401.5398, 2014. MR3210997

[4] Caron, F., and Doucet, A. Sparse Bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine Learning* (New York, NY, USA, 2008), ICML '08, ACM, pp. 88–95.

[5] Carvalho, C. M., Polson, N. G., and Scott, J. G. The horseshoe estimator for sparse signals. *Biometrika 97*, 2 (2010), 465–480. MR2650751

[6] Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. Bayesian linear regression with sparse priors. *Ann. Statist. 43*, 5 (10 2015), 1986–2018. MR3375874

[7] Castillo, I., and Van der Vaart, A. W. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist. 40*, 4 (2012), 2069–2101. MR3059077

[8] Damien, P., Wakefield, J., and Walker, S. Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. R. Stat. Soc. Ser. B Stat. Methodol. 61*, 2 (1999), 331–344. MR1680334

[9] Datta, J., and Ghosh, J. K. Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis 8*, 1 (2013), 111–132. MR3036256

[10] Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. Maximum entropy and the nearly black object (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol. 54*, 1 (1992), 41–81. MR1157714

[11] Ghosal, S., Ghosh, J. K., and Van der Vaart, A. W. Convergence rates of posterior distributions. *Ann. Statist. 28*, 2 (2000), 500–531. MR1790007

[12] Ghosh, P., and Chakrabarti, A. Posterior concentration properties of a general class of shrinkage estimators around nearly black vectors. arXiv:1412.8161v2, 2015.

[13] Griffin, J. E., and Brown, P. J. Alternative prior distributions for variable selection with very many more variables than observations. *Technical Report, University of Warwick.* (2005).

[14] Griffin, J. E., and Brown, P. J. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis 5*, 1 (2010), 171–188. MR2596440

[15] Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. On adaptive posterior concentration rates. *Ann. Statist. 43*, 5 (10 2015), 2259–2295. MR3396985

[16] Johnson, V. E., and Rossell, D. On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B Stat. Methodol. 72*, 2 (2010), 143–170. MR2830762

[17] Johnstone, I. M., and Silverman, B. W. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist. 32*, 4 (2004), 1594–1649. MR2089135

[18] Martin, R., and Walker, S. G. Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat. 8*, 2 (2014), 2188–2206. MR3273623

[19] Park, T., and Casella, G. The Bayesian lasso. *J. Amer. Statist. Assoc. 103*, 482 (2008), 681–686. MR2524001

[20] Polson, N. G., and Scott, J. G. Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics 9* (2010), 501–538. MR3204017

[21] Polson, N. G., and Scott, J. G. Good, great or lucky? Screening for firms with sustained superior performance using heavy-tailed priors. *Ann. Appl. Stat. 6*, 1 (2012), 161–185. MR2951533

[22] Polson, N. G., and Scott, J. G. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis 7*, 4 (2012), 887–902. MR3000018

[23] Robbins, H. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (Berkeley, California, 1956), University of California Press, pp. 157–163. MR0084919

[24] Ročková, V. Bayesian estimation of sparse signals with a continuous spike-and-slab prior. submitted manuscript, available at `http://stat.wharton.upenn.edu/~vrockova/rockova2015.pdf`, 2015.

[25] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol. 58*, 1 (1996), 267–288. MR1379242

[26] van der Pas, S., Kleijn, B., and van der Vaart, A. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electron. J. Stat. 8* (2014), 2585–2618. MR3285877

[27] Yang, Y., Wainwright, M. J., and Jordan, M. I. On the computational complexity of high-dimensional Bayesian variable selection. arXiv:1505.07925, 2015.