*The Annals of Statistics*
2018, Vol. 46, No. 1, 401–437
https://doi.org/10.1214/17-AOS1554


# BAYESIAN ESTIMATION OF SPARSE SIGNALS WITH A CONTINUOUS SPIKE-AND-SLAB PRIOR[1]

BY VERONIKA ROČKOVÁ

*University of Chicago*

We introduce a new framework for estimation of sparse normal means, bridging the gap between popular frequentist strategies (LASSO) and popular Bayesian strategies (spike-and-slab). The main thrust of this paper is to introduce the family of Spike-and-Slab LASSO (SS-LASSO) priors, which form a continuum between the Laplace prior and the point-mass spike-and-slab prior. We establish several appealing frequentist properties of SS-LASSO priors, contrasting them with these two limiting cases. First, we adopt the penalized likelihood perspective on Bayesian modal estimation and introduce the framework of Bayesian penalty mixing with spike-and-slab priors. We show that the SS-LASSO global posterior mode is (near) minimax rate-optimal under squared error loss, similarly as the LASSO. Going further, we introduce an adaptive two-step estimator which can achieve provably sharper performance than the LASSO. Second, we show that the whole posterior keeps pace with the global mode and concentrates at the (near) minimax rate, a property that is known *not to hold* for the single Laplace prior. The minimax-rate optimality is obtained with a suitable class of independent product priors (for known levels of sparsity) as well as with dependent mixing priors (adapting to the unknown levels of sparsity). Up to now, the rate-optimal posterior concentration has been established only for spike-and-slab priors with a point mass at zero. Thus, the SS-LASSO priors, despite being continuous, possess similar optimality properties as the "theoretically ideal" point-mass mixtures. These results provide valuable theoretical justification for our proposed class of priors, underpinning their intuitive appeal and practical potential.

**1. Normal-means revisited.** Sparse estimation is fundamental to high-dimensional statistical learning. Existing methods include the plentiful variants of the LASSO (a popular frequentist approach) and of spike-and-slab selection (a popular Bayesian approach). Relevant references include [5, 11, 12, 17, 23, 26, 34, 35, 38, 39]. Here, we cross-fertilize the two paradigms into one unifying framework. To this end, we introduce the family of Spike-and-Slab LASSO (SS-LASSO) priors. We show that Spike-and-Slab LASSO priors can be optimal from both penalized likelihood and fully Bayes perspectives. We provide rigorous

Received May 2015; revised February 2017.

[1]Supported by NSF Grant DMS-14-06563, AHRQ Grant R21-HS021854 and the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business.

*MSC2010 subject classifications.* Primary 62J99; secondary 62F15.

*Key words and phrases.* Asymptotic minimaxity, LASSO, posterior concentration, spike-and-slab.

frequentist assessments of the behavior of the global posterior mode, an analog of the LASSO estimator, and of the entire posterior distribution. For our theoretical investigation, we confine attention to the traditional normal means model. Nevertheless, the ideas developed here reach far beyond this framework.

We focus on the canonical problem of estimating a high-dimensional mean vector from a single multivariate observation [10, 21]. The observed vector $\boldsymbol{y}^{(n)} = (y_1, \ldots, y_n)'$ arises from

$$Y_i = \beta_{0i} + \varepsilon_i \qquad \text{where } \varepsilon_i \sim \mathcal{N}(0,1), i = 1, \ldots, n, \tag{1.1}$$

and the goal is estimating $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0n})'$ under squared error loss. We approach this classic problem from both the penalized likelihood and fully Bayes perspectives, assuming that $\boldsymbol{\beta}_0$ is possibly sparse. The sparsity here is defined in the nearly-black sense, where $\boldsymbol{\beta}_0 \in l_0[p_n; n] = \{\boldsymbol{\beta} \in \mathbb{R}^n : \sum_{i=1}^n \mathbb{I}(|\beta_i| \neq 0) \leq p_n\}$ and $p_n = o(n)$ as $n \to \infty$. With only one observation for each parameter, estimation can be made more effectual under such sparsity assumptions. The quality of recovery here will be assessed relative to the benchmark minimax risk $2p_n \log(n/p_n)(1 + o(1))$ [14] and the near-minimax risk $2p_n \log n(1 + o(1))$. Namely, we adopt the view that a good point estimator should have a maximum risk that is always within a universal constant multiple of the (near) minimax risk. For instance, the near-minimax rate optimality is known to hold for the LASSO mode estimator with a penalty $\lambda = \sqrt{2 \log n}$ [4].

A traditional Bayesian approach to estimating sparse $\boldsymbol{\beta}_0$ begins with a spike-and-slab prior on each $\beta_i$ that naturally segregates important coefficients from the ignorable [6, 12, 17, 23, 26]. This separation is labeled by a vector of latent binary indicators $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$, one $\gamma_i \in \{0, 1\}$ for each coordinate. A particularly appealing spike-and-slab variant has been

$$\pi(\beta_i|\gamma_i, \lambda_1) = (1 - \gamma_i)\delta_0(\beta_i) + \gamma_i \psi(\beta_i|\lambda_1), \qquad \boldsymbol{\gamma} \sim \pi(\boldsymbol{\gamma}|\theta), \tag{1.2}$$

where $\delta_0(\cdot)$ is the spike distribution (atom at zero) and $\psi(\cdot|\lambda_1)$ is an absolutely continuous slab distribution, indexed by a hyper-parameter $\lambda_1$, and $\theta$ is a prior mixing proportion. Continuous relaxations of (1.2), with $\delta_0(\cdot)$ replaced by a peaked continuous density, have also become popular [17–19, 27, 30]. Despite the ubiquity of the spike-and-slab methodology throughout science, the underlying theory "has not kept pace with the applications" [20]. Here, we narrow the gap by providing new theoretical insights for a class of *continuous* spike-and-slab priors.

One of the earliest theoretical analyses of continuous spike-and-slab priors was carried out by Ishwaran and Rao [19]. The authors proposed and studied a class of continuous bimodal priors (two-point student-$t$ mixtures), established oracle-like misclassification performance of the posterior mean for variable selection [19] and multigroup classification [18] and coined the term selective shrinkage for the asymptotic behavior of the posterior mean. More recently, Ishwaran and Rao [20] went further and established the oracle property [15] of the posterior mean under

two-point Gaussian mixtures, assuming nonorthogonal low-dimensional designs. In another development, Narisetty and He [27] established model selection consistency of Bayes factors under Gaussian mixture priors in more general designs with a diverging number of covariates.

While the literature on theory for continuous spike-and-slab priors has been relatively sparse, there is a large body of theoretical evidence endorsing point-mass mixture priors [1, 10, 16, 21, 22]. More specifically, [21] analyzed an empirical Bayes variant of the point-mass mixture prior (1.2) in model (1.1). With a (restricted) marginal maximum likelihood estimate of $\theta$, the posterior mean and median are shown to attain the minimax rate $p_n \log[n/p_n]$. Going further, [10] provided profound theoretical results concerning the entire posterior measure under a class of priors (1.2). For suitably chosen $\theta$, or a suitable beta-prior $\pi(\theta)$, the entire posterior concentrates at the minimax rate. This remarkable feature reinforces the prominent position of point-mass mixture priors as benchmark methodological ideals [7]. The rate-optimal posterior convergence is typically (under convex losses) inherited by many posterior functionals, related to both location and spread. Thus, the optimally concentrating posteriors are conformable to valid recovery and uncertainty quantification. The notion of the speed of posterior concentration has become a valuable instrument for frequentist assessments of Bayesian procedures [3, 9, 25, 28, 36].

Despite their theoretical appeal, the point-mass mixture priors can be impractical for posterior simulation. Continuous spike-and-slab priors, on the other hand, are amenable to fast deterministic computation [30–32]. The Spike-and-Slab LASSO (SS-LASSO) priors introduced here can thus be viewed as an intermediate between "the theoretically ideal" [the prior (1.2)] and "the computationally ideal" (the Laplace prior). Recent methodological implementations of the Spike-and-Slab LASSO priors are discussed in [31, 32]. In this present work, we focus primarily on theoretical underpinnings.

Our proposed class of SS-LASSO priors forms a continuum between the point-mass mixture prior (1.2) and the LASSO (Laplace) prior. This raises several compelling questions. How does this prior fare compared to the two extreme cases? First, does the SS-LASSO posterior mode attain (near) minimax rates (just like the LASSO mode)? Second, does the SS-LASSO prior, despite being continuous, yield optimal posterior concentration [just like the prior (1.2)]? Here, we provide rigorous answers to these intriguing possibilities.

To answer the first question, we adopt the penalized likelihood perspective on Bayesian modal estimation [1, 16]. We introduce and develop a framework for Bayesian penalty mixing under *continuous* spike-and-slab priors. We flesh out several revealing connections between the SS-LASSO and LASSO thresholding operators (for similar comparisons see, e.g., [2, 15, 37, 38] and references therein). We are primarily interested in SS-LASSO priors that are en route to the ideal prior (1.2) in the limit as $n \to \infty$. These priors produce highly nonconcave penalties [2, 15] and, inherently, multimodal posteriors. To begin, we provide a nonasymptotic

upper risk bound for the global posterior mode and establish conditions under which the global mode attains the (near) minimax risk rate. Our analysis aligns with an earlier development of Antoniadis and Fan [2], who obtained oracle inequalities [13, 38] for a broad class of penalty functions (including ours). Here, however, we position our results in terms of asymptotic minimaxity rather than relative ideal risk performance [13]. We demonstrate that the selective shrinkage ability [19] is not unique to the posterior mean, but is manifested also in posterior modes. Our new insights justify the suitability of spike-and-slab posterior modes for sparsity recovery, complementary to results known to hold for posterior mean/median under priors (1.2) [10, 21]. Going further, we propose a data-adaptive two-step SS-LASSO thresholding rule that achieves a sharper asymptotic rate relative to nonadaptive estimators (including the LASSO), when the level of sparsity is unknown. With a suitable beta-min condition, this rate is minimax.

The rate-optimality of the mode (mean/median) and the entire posterior do not necessarily come together [8]. One surprising example was provided recently for the LASSO prior. Castillo and van der Vaart [9] show that the posterior distribution concentrates far slower than the LASSO mode, dampening its usefulness for uncertainty quantification. This can be attributed to the fact that with just one penalty parameter, the Laplace prior falls short of satisfying the two conflicting demands of shrinkage and unbiasedness. The SS-LASSO prior can be viewed as a two-penalty refinement of the single Laplace prior proposed here, to avoid this conflict.

To provide an affirmative answer to the second question, we study the posterior concentration of SS-LASSO priors, following the line of research pioneered by [10] and further developed by [3]. We show that under suitable SS-LASSO priors, the asymptotic rate-optimality is a global property of the whole posterior measure, not only of the global mode. We consider independent continuous product priors obtained with $\theta$ fixed, adapting the approach of [3], and further extend the results to the case of dependent continuous priors induced with a prior $\pi(\theta)$. Our results have distinct implications in terms of calibration of continuous spike-and-slab priors and provide valuable theoretical evidence supporting the intuition that with two parameters, the SS-LASSO prior is well suited for the two simultaneous goals of estimation and selection.

We begin with a formal introduction of SS-LASSO priors in Section 2 and develop the penalized likelihood perspective in Section 3. Section 4 is devoted to the discussion of the global mode and its risk properties. Section 5 concerns the fully Bayes optimality aspects of SS-LASSO priors. Section 6 presents a simulation study and Section 7 concludes with a discussion.

**2. The spike-and-slab LASSO prior.** We introduce the family of Spike-and-Slab LASSO priors for Bayesian inference about $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$ in sparse settings. Specified hierarchically with an intermediate vector of latent binary variables

$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$, $\gamma_i \in \{0, 1\}$, the SS-LASSO prior on $\boldsymbol{\beta}$ is defined as

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \prod_{i=1}^{n}\big[(1-\gamma_i)\psi(\beta_i|\lambda_0) + \gamma_i\psi(\beta_i|\lambda_1)\big], \qquad \boldsymbol{\gamma} \sim \pi(\boldsymbol{\gamma}|\theta), \tag{2.1}$$

where $\psi(\beta|\lambda) = \frac{\lambda}{2}\exp\{-\lambda|\beta|\}$ is a Laplace distribution with mean 0 and variance $2/\lambda^2$. With $\lambda_0 \gg \lambda_1$, the spike distribution $\psi(\beta_i|\lambda_0)$ will be concentrated around zero, while the slab distribution $\psi(\beta_i|\lambda_1)$ will be relatively diffuse. Although the choice of $\pi(\boldsymbol{\gamma})$ offers rich potential for modeling $\boldsymbol{\gamma}$, we shall focus primarily on the exchangeable case $\pi(\boldsymbol{\gamma}|\theta)$ where the entries in $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$ are i.i.d. Bernoulli with

$$\mathsf{P}(\gamma_i = 1|\theta) = \theta. \tag{2.2}$$

The point-mass mixture prior (1.2) is obtained as a limiting special case of (2.1) when $\lambda_0 \to \infty$, whereas the LASSO prior is obtained by setting $\lambda_0 = \lambda_1$.

As will be seen, this mixture prior (2.1) induces a variant of "selective shrinkage" [19] that adaptively segregates the active coefficients from the ignorable. By treating the coefficients differentially, this property is crucial to obtaining a fine balance between bias and shrinkage. This aspect plays an integral role in our theoretical analysis. Thereby it is worthwhile to briefly expand on the Bayesian mechanism underlying this property. The implications for frequentist penalized likelihood estimation will be drawn in the next section.

Conditionally on $\theta$, the prior (2.1) induces a posterior $\pi(\boldsymbol{\beta}|\boldsymbol{y}^{(n)}, \theta) = \prod_{i=1}^{n}\pi(\beta_i|y_i, \theta)$ under which $\beta_1, \ldots, \beta_n$ are independent and

$$\begin{aligned}\pi(\beta_i|y_i, \theta) &= \pi(\beta_i|y_i, \gamma_i = 1)\mathsf{P}(\gamma_i = 1|y_i, \theta)\\ &\quad + \pi(\beta_i|y_i, \gamma_i = 0)\mathsf{P}(\gamma_i = 0|y_i, \theta).\end{aligned} \tag{2.3}$$

The posterior (2.3) puts more weight on mixture components best supported by the data through the distribution $\pi(\boldsymbol{\gamma}|\boldsymbol{y}^{(n)}, \theta) = \prod_{i=1}^{n}\pi(\gamma_i|y_i, \theta)$. For example, (2.3) will be dominated by the spike posterior $\pi(\beta_i|y_i, \gamma_i = 0)$ when $\pi(\gamma_i = 0|y_i, \theta)$ is large, signaling that $\beta_i$ has a higher probability of being small. On the other hand, when $\pi(\gamma_i = 1|y_i, \theta)$ is large, (2.3) will be dominated by the slab posterior $\pi(\beta_i|y_i, \gamma_i = 1)$ in which case $\beta_i$ is allowed to take larger values. This mechanism underlies the selective shrinkage ability of the posterior, and its functionals that is typical for the spike-and-slab priors.

Further adaptivity can be obtained with a fully Bayes variant of (2.2) by assuming $\theta \sim \mathcal{B}(a, b)$, where $\mathcal{B}(a, b)$ denotes the beta distribution with shape parameters $a$ and $b$. This prior renders the elements in $\boldsymbol{\beta}$ a-priori (and a-posteriori) dependent. We will study the property of the full posterior measure under this beta-Bernoulli hierarchical construction in Section 5.3.

The following notation will be used throughout the paper. For sequences $a_n$ and $b_n$, $a_n \simeq b_n$ means $a_n/b_n \to c$ for some $c > 0$, $a_n \preceq b_n$ means $a_n = \mathcal{O}(b_n)$.

We will denote by $\psi_0(\beta)$ the Laplace spike density $\psi(\beta|\lambda_0)$ and by $\psi_1(\beta)$ the Laplace slab density $\psi(\beta|\lambda_1)$. By $\phi(\beta)$, we denote the density of the standard normal distribution. Denote by $\|\cdot\|_1$ the $l_1$ norm, by $\|\cdot\|$ the $l_2$ norm, by $\|\cdot\|_0$ the $l_0$ norm and by $\boldsymbol{\beta}_S$ the subvector of $\boldsymbol{\beta}$ containing entries in $S \subset \{1,\dots,n\}$.

**3. Spike-and-slab: The penalized likelihood perspective.** Before studying the posterior distribution in its entirety, we will examine one of its functionals, the global posterior mode. A key to our approach will be drawing upon connections between posterior modes and penalized likelihood maximizers, the LASSO in particular.

Our interest in posterior modes was motivated by the following practical considerations. There are many ways to use information from the posterior for variable selection. Rather than relying on the post-data selection uncertainty in $\pi(\boldsymbol{\gamma}|\boldsymbol{y}^{(n)},\theta)$ to select coefficients, a strategy computationally very involved, we can let posterior modes automatically threshold out the irrelevant coordinates. In contrast, existing continuous spike-and-slab priors [17, 18, 27, 30] yield nonsparse posterior modes that must be thresholded for variable selection.

The modes can be viewed as penalized likelihood estimators associated with a penalty $\log\pi(\boldsymbol{\beta}|\theta)$. A single Laplace prior yields the familiar LASSO penalty $\lambda\sum_{i=1}^n|\beta_i|$. For our SS-LASSO prior, the penalty is obtained from (2.1) by marginalizing $\boldsymbol{\gamma}$ out with respect to $\pi(\boldsymbol{\gamma}|\theta)$. We will assume throughout this section that the mixing proportion $\theta$ is fixed. This facilitates manipulations with the SS-LASSO penalty, because it is separable conditionally on $\theta$. Thus, we can write $\log\pi(\boldsymbol{\beta}|\theta)=\sum_{i=1}^n \mathrm{pen}(\beta_i)$ where

$$\mathrm{pen}(\beta_i) \equiv \log\big[(1-\theta)\psi(\beta_i|\lambda_0)+\theta\psi(\beta_i|\lambda_1)\big], \qquad i=1,\dots,n. \tag{3.1}$$

Unlike the LASSO penalty that is linear both in $|\beta_i|$ and $\lambda$, the SS-LASSO penalty (3.1) is a nonlinear functional of both $|\beta_i|$ and $(\lambda_1,\lambda_0,\theta)$. Despite the apparent differences, there is an interesting connection between the two penalties. This connection is unveiled after taking a derivative. The derivative corresponds to an implicit bias term and plays a crucial role in estimation [15]. For the SS-LASSO penalty (3.1), we have

$$\frac{\partial\,\mathrm{pen}(\beta_i)}{\partial|\beta_i|} = -\lambda_1 p^\star(\beta_i) - \lambda_0\big[1-p^\star(\beta_i)\big] \equiv -\lambda^\star(\beta_i), \tag{3.2}$$

where

$$p^\star(\beta_i) = \frac{\theta\psi_1(\beta_i)}{\theta\psi_1(\beta_i)+(1-\theta)\psi_0(\beta_i)}. \tag{3.3}$$

On the other hand, a single Laplace prior yields $\frac{\partial\log\psi(\beta_i|\lambda)}{\partial|\beta_i|}=-\lambda$. Thus, the SS-LASSO bias term (3.2) is a convex combination of two LASSO bias terms. Importantly, the combination is adaptive, because $p^\star(\beta_i)$ depends on $\beta_i$. This is a

unique feature of a spike-and-slab penalty, which weights the contributions from the spike and the slab individually for every coefficient. In sharp contrast, the LASSO penalty assigns the same amount of bias to every single coefficient, regardless its size. This is often a source of conflict between shrinkage and bias. The additional flexibility of $\lambda^\star(\beta_i)$ greatly alleviates this conflict.

The mixing proportion $p^\star(\beta_i)$ can be viewed as a conditional probability of inclusion, having seen the regression coefficient $\beta_i$. This interpretation comes directly from (3.3), which is $\mathsf{P}(\gamma_i = 1|\beta_i, \theta)$ by the Bayes theorem. It is clear that

$$p^\star(\beta_i) = \frac{1}{1 + \frac{(1-\theta)}{\theta}\frac{\lambda_0}{\lambda_1}\exp[-|\beta_i|(\lambda_0 - \lambda_1)]} \tag{3.4}$$

is exponentially increasing in $|\beta_i|$. This function has a sudden increase from near-zero to near-one. The transition occurs at the intersection point between the spike and slab densities. The intersection point will be introduced formally later and will play a fundamental role in quantifying the speed of posterior concentration.

In the following, we will describe the implications of the Bayesian penalty mixing for the global posterior mode estimator. Conditionally on $\theta$, the posterior factorizes into an independent product:

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}^{(n)}, \theta) = \prod_{i=1}^{n} \exp\big[L(\beta_i, y_i)\big], \tag{3.5}$$

where

$$L(\beta_i, y_i) = -\frac{1}{2}(y_i - \beta_i)^2 + \mathrm{pen}(\beta_i) \tag{3.6}$$

is the log-posterior contribution from a single observation $y_i$. Define by

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^n} \pi(\boldsymbol{\beta}|\boldsymbol{y}^{(n)}, \theta) \tag{3.7}$$

the global posterior mode, which will be further referred to as the SS-LASSO estimator. Due to the separability, $\widehat{\boldsymbol{\beta}} = (\widehat{\beta_1}, \ldots, \widehat{\beta_n})'$ can be obtained coordinate-wise, where each $\widehat{\beta_i}$ is the global mode of the univariate log-posterior $L(\beta_i, y_i)$ in (3.6).

As with the LASSO penalty [34], an important necessary characterization of the solution $\widehat{\boldsymbol{\beta}}$ can be derived from first-order conditions.

LEMMA 3.1. *The individual entries $\widehat{\beta_i}$ of the global mode $\widehat{\boldsymbol{\beta}} = (\widehat{\beta_1}, \ldots, \widehat{\beta_n})'$ satisfy*

$$\widehat{\beta_i} = \big(|y_i| - \lambda^\star(\widehat{\beta_i})\big)_+ \mathrm{sign}(y_i), \tag{3.8}$$

*where*

$$\lambda^\star(\beta_i) = \lambda_1 p^\star(\beta_i) + \lambda_0\big[1 - p^\star(\beta_i)\big]. \tag{3.9}$$

PROOF. From the subdifferential calculus, it is necessary that

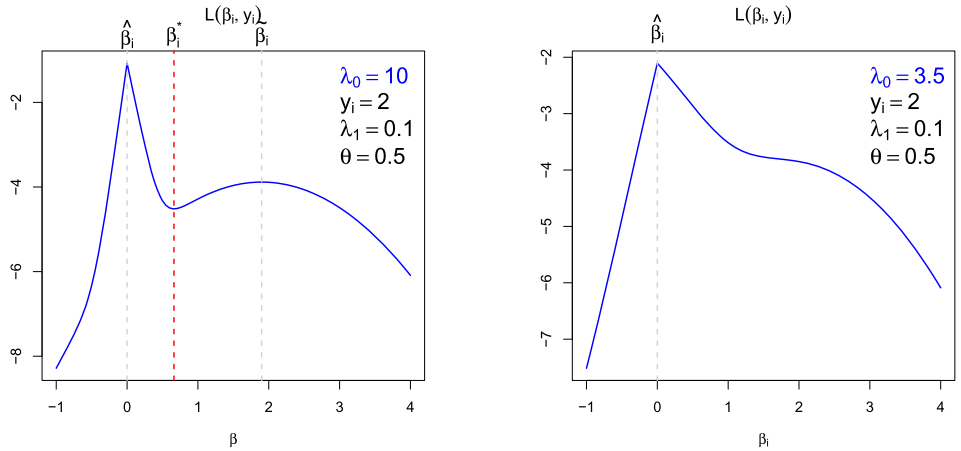
$$\frac{\partial(y_i-\beta_i)^2}{\partial\beta_i}\Big|_{\beta_i=\widehat{\beta}_i} = 2\lambda^\star(\widehat{\beta}_i)\,\mathrm{sign}(\widehat{\beta}_i) \qquad \text{if } \widehat{\beta}_i \neq 0,$$

$$\left|\frac{\partial(y_i-\beta_i)^2}{\partial\beta_i}\Big|_{\beta_i=\widehat{\beta}_i}\right| \leq 2\lambda^\star(\widehat{\beta}_i) \qquad \text{if } \widehat{\beta}_i = 0,$$

which completes the proof. □

Equation (3.8) resembles the necessary and sufficient characterization of the LASSO solution in orthogonal designs. There are, however, some fundamental differences. First, $\lambda^\star(\widehat{\beta}_i)$ is unique to each coefficient. This occurs also with the adaptive LASSO [38], which assigns *fixed* coefficient-specific penalties. The major difference here is that the penalties $\lambda^\star(\widehat{\beta}_i)$ are not fixed, but adaptive to the data through $\widehat{\beta}_i$. More precisely, $\lambda^\star(\widehat{\beta}_i)$ is a "self-adaptive" linear combination of spike and slab penalties, weighted by $p^\star(\widehat{\beta}_i)$. Promising coefficients have $p^\star(\widehat{\beta}_i)$ close to one and are shrunk less. This is because $\lambda^\star(\widehat{\beta}_i)$ is driven primarily by $\lambda_1$, which is set to be small to avoid overshrinkage. The opposite happens with small coefficients. Small $\widehat{\beta}_i$'s in the basin around zero have a small inclusion probability, where $\lambda^\star(\widehat{\beta}_i)$ is taken over by the large penalty $\lambda_0$. This is a manifestation of the selective shrinkage property behind the SS-LASSO estimator.

Conditionally on $\theta$, the coordinates are independent and the solution $\widehat{\beta}_i$ depends on $\boldsymbol{y}^{(n)}$ only through $y_i$. However, because $\lambda^\star(\widehat{\beta}_i)$ depends on $\widehat{\beta}_i$, obtaining $\widehat{\beta}_i$ from (3.8) is far from obvious. Ultimately, there are at most two local maxima of $L(\beta_i, y_i)$. As illustrated in Figure 1, a local maximum can occur at zero, elsewhere




(a) $y_i = 2, \lambda_0 = 10$ (b) $y_i = 2, \lambda_0 = 3.5$

FIG. 1. *Plots of $L(\beta_i, y_i)$ for different values of $\lambda_0$, assuming $\lambda_1 = 0.1$, $\theta = 0.5$ and $y_i = 2$.*

or both. If $|y_i| < \lambda_1$, the global mode $\widehat{\beta}_i$ occurs at zero. Similarly, if $|y_i| > \lambda_0$, then $\widehat{\beta}_i \neq 0$. From the characterization (3.8), the nonzero posterior mode $\widehat{\beta}_i \neq 0$, if it exists, satisfies the implicit relationship

$$|y_i| = |\widehat{\beta}_i| + \lambda^\star(\widehat{\beta}_i). \tag{3.10}$$

However, if $L(\beta_i, y_i)$ is bi-modal, (3.10) is also satisfied by a local maximum $\widetilde{\beta}_i$ as well as a local minimum separating the posterior modes [$\beta_i^\star$ in Figure 1(b)]. Thus, the global mode is not uniquely characterized by (3.8). This may occur when the SS-LASSO penalty is strongly nonconcave. The extent of the nonconcavity can be quantified by the maximal nonconcavity number $\kappa(\lambda_0, \lambda_1)$ [24], defined as $\kappa(\lambda_0, \lambda_1) = \max_\beta\{\frac{\partial^2 \operatorname{pen}(\beta)}{\partial|\beta|^2}\}$. The second derivative of the penalty function can be written as

$$\frac{\partial^2 \operatorname{pen}(\beta)}{\partial|\beta|^2} = p^\star(\beta)\big[1 - p^\star(\beta)\big](\lambda_0 - \lambda_1)^2, \tag{3.11}$$

and where $p^\star(\beta)$ is defined in (3.3). With $\kappa(\lambda_0, \lambda_1) < 1$, the posterior $L(\beta, y)$ will be concave and thereby unimodal. The second derivative is maximized when $p^\star(\beta) = 0.5$. This occurs at the two intersection points $\pm\delta(\lambda_0, \theta)$ between the spike and slab densities where

$$\delta(\lambda_0, \theta) = \frac{1}{\lambda_0 - \lambda_1} \log\left[\frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1}\right]. \tag{3.12}$$

The intersection point (3.12) is here expressed as a function of $(\lambda_0, \theta)$, the two sparsity parameters of the main focus, and will reoccur in Section 5 and Section 5.3, where it will be an important ingredient for characterizing a generalized notion of sparsity. To continue, the maximal nonconcavity then equals $\kappa(\lambda_0, \lambda_1) = \frac{1}{4}(\lambda_0 - \lambda_1)^2$. Thus, the posterior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{y}^{(n)}, \theta)$ in (3.5) has a unique mode, whenever $(\lambda_0 - \lambda_1)^2 < 4$.

The single Laplace prior, obtained as a special case when $\lambda_1 = \lambda_0$, is known to yield single-mode posteriors. Thus, it is not surprising that unimodal posteriors occur when $\lambda_0$ and $\lambda_1$ are not too different. However, here we focus primarily on penalties that are en route to the limiting ideal (1.2) when $\lambda_0 \to \infty$ as $n \to \infty$. Letting $\lambda_0$ and $\lambda_1$ grow apart as $n \to \infty$ will be essential for achieving optimal rates of convergence. The ambient penalties in this asymptotic regime are, however, very nonconcave and the posteriors, thus, may possess many local optima.

In order to investigate the estimation aspects of the global mode, we need a more refined characterization, which sets it apart from all the local solutions satisfying the first-order condition (3.8). The following function plays a principal role in this characterization:

$$g(x) = \big[\lambda^\star(x) - \lambda_1\big]^2 + 2\log p^\star(x). \tag{3.13}$$

We also need the following notation. Denote by

$$c_+ = 0.5\big(1 + \sqrt{1 - 4/(\lambda_0 - \lambda_1)^2}\big)$$

and

$$\delta_{c+} = 1/(\lambda_0 - \lambda_1)\log\left[\frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1}\frac{c_+}{1-c_+}\right]. \tag{3.14}$$

Note that $p^\star(\delta_{c+}) = c_+ > 0.5$ and $\text{pen}''(\delta_{c+}) = 1$. Thus, $\delta_{c+}$ is an inflection point of $L(\beta; y)$. Because $c_+ > 0.5$, $\delta_{c+}$ is greater than the intersection point $\delta(\lambda_0, \theta)$. The significance of $\delta_{c+}$ is in the fact that the curvature of $L(\beta; y)$ at $\delta_{c+}$ indicates the presence of multimodality and the degree of separation between posterior modes. An important quantity for us will be the value $g(\delta_{c+})$.

The following theorem uniquely characterizes the global mode and formalizes the intuition that, unlike the posterior mean, the mode is a strict thresholding rule.

THEOREM 3.1. *Denote by* $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_n)'$ *the global posterior mode* (3.7). *Then*

$$\widehat{\beta}_i = \begin{cases} 0 & \text{when } |y_i| \le \Delta, \\ \big[|y_i| - \lambda^\star(\widehat{\beta}_i)\big]_+ \operatorname{sign}(y_i) & \text{when } |y_i| > \Delta. \end{cases}$$

*Furthermore, when* $(\lambda_0 - \lambda_1) > 2$ *and* $g(0) > 0$ *we can write*

$$\Delta^L < \Delta \le \Delta^U,$$

*where*

$$\Delta^L = \sqrt{2\log[1/p^\star(0)] - d} + \lambda_1 \quad \text{and} \quad \Delta^U = \sqrt{2\log[1/p^\star(0)]} + \lambda_1 \tag{3.15}$$

*and* $0 < d = -g(\delta_{c+})$ *and* $\delta_{c+}$ *is as in* (3.14).

PROOF. See the supplementary materials ([29], Section 1.1.1). □

The global mode thresholds out values below $\Delta^L < \Delta \le \Delta^U$, where $\Delta^U$ depends on $(\lambda_1, \lambda_0, \theta)$ through $p^\star(0)$. Recall that $p^\star(0) = \frac{\theta\lambda_1}{\theta\lambda_1 + (1-\theta)\lambda_0}$ is the relative height of the slab density at zero. This quantity will be fundamental for controlling the risk of the global mode and posterior concentration properties. The assumption $g(0) > 0$ in Theorem 3.1 guarantees that $p^\star(0)$ is sufficiently far away from zero. This condition will be satisfied, for instance, when $\lambda_1 \le \mathrm{e}^{-2}$ and $\lambda_0 \ge 1/\theta + 3$, where $0 < \theta \le 0.5$.

Theorem 3.1 has important practical implications for calibrating the SS-LASSO priors. Suppose $\lambda_0 \to \infty$ and $\lambda_1 < 1$ is fixed. According to Lemma 1.2 (supplementary material [29]), we have $d < 2 - (\frac{1}{\lambda_0 - \lambda_1} - \sqrt{2})^2$ so that $d \to 0$ as $(\lambda_0 - \lambda_1) \to \infty$. Thus, $\Delta^L$ approaches the pseudo-threshold $\Delta^U$ as $(\lambda_0 - \lambda_1) \to \infty$. Moreover, (3.4) implies $\lim_{\lambda_0 \to \infty} p^\star(|x|) = 1$ for any $|x| > 0$. Consequently, $\lim_{\lambda_0 \to \infty} \lambda^\star(|x|) = \lambda_1$ for $|x| > 0$. Therefore, as $\lambda_0 \to \infty$, the global mode approaches the following estimator:

$$\bar{\beta} = \begin{cases} 0 & \text{when } |y| \le \Delta, \\ \big(|y| - \lambda_1\big)_+ \operatorname{sign}(y) & \text{when } |y| > \Delta. \end{cases} \tag{3.16}$$

This limiting estimator highlights the two distinct roles of the spike and slab penalty parameters $(\lambda_0, \lambda_1, \theta)$. The slab penalty $\lambda_1$ controls the bias of nonzero effects, whereas $(\theta, \lambda_0)$ control the size of the selected model (through $\Delta$). Interestingly, the estimator (3.16) relates to known thresholding operators [13] through the following connection: $|\widehat{\beta}_{\mathrm{ST}}| \leq |\bar{\beta}| \leq |\widehat{\beta}_{\mathrm{HT}}|$, where $\widehat{\beta}_{\mathrm{ST}} = (|y| - \Delta)_+ \operatorname{sign}(y)$ is the soft-thresholding operator and $\widehat{\beta}_{\mathrm{HT}} = |y|\mathbb{I}(|y| > \Delta)$ is the hard-thresholding operator. Both $\widehat{\beta}_{\mathrm{ST}}$ and $\widehat{\beta}_{\mathrm{HT}}$ satisfy the so called oracle inequality [13] when $\Delta$ equals (is sufficiently close to) $\sqrt{2\log n}$. The oracle inequality implies that the attained performance in terms of squared error loss differs from the ideal performance by at most a factor of $2\log n$. The estimator $\bar{\beta}$ ought to possess a similar optimality property when $\Delta \simeq \sqrt{2\log n}$ [i.e., $1/p^\star(0) \simeq n$] [2]. This asymptotic consideration provides useful insight into tuning $\lambda_0, \theta$ and $\lambda_1$, and conveys the intuition that controlling $1/p^\star(0)$ will be crucial for achieving good risk properties of $\widehat{\boldsymbol{\beta}}$.

**4. Risk properties of the global mode.** The similarities between the LASSO and SS-LASSO estimators apparent in (3.1) suggest that there be similarities also in terms of risk performance. Because the SS-LASSO estimator is a two-penalty refinement of the LASSO estimator, we would expect it to perform at least as well. In this section, we provide an affirmative answer. In Section 4.1, we show how improvements over the LASSO can be obtained with a two-step SS-LASSO estimator.

The following theorem presents a nonasymptotic upper risk bound for the SS-LASSO estimator $\widehat{\boldsymbol{\beta}}$ over $l_0[p_n; n]$ under the squared error loss. The bound is expressed in terms of $1/p^\star(0) = 1 + \frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1}$.

THEOREM 4.1. *Suppose* $(\lambda_0 - \lambda_1) > 2$ *and* $\lambda_1 < \mathrm{e}^{-2}$. *Assume the model* (1.1) *with* $\boldsymbol{\beta}_0 \in l_0[p_n; n]$. *Let* $d$ *be as in Theorem* 3.1 *and assume* $g(0) > 0$. *Then the risk of the global mode* $\widehat{\boldsymbol{\beta}}$ *satisfies*

$$\begin{aligned} \mathsf{E}_{\boldsymbol{\beta}_0}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 &< 8p_n\bigl(1 + \log\bigl[1/p^\star(0)\bigr]\bigr) \\ &\quad + \frac{4n}{\sqrt{\pi}}\mathrm{e}^d p^\star(0)\bigl(1 + \sqrt{\log\bigl[1/p^\star(0)\bigr]}\bigr). \end{aligned} \tag{4.1}$$

PROOF. See the supplementary materials ([29] Section 1.2.1). □

According to Theorem 4.1, the triplet $(\lambda_0, \lambda_1, \theta)$ affects the risk of $\widehat{\boldsymbol{\beta}}$ through the functional $1/p^\star(0)$. Here, we consider $\lambda_1$ fixed to a small constant. Thereby, the asymptotic behavior of (4.1) is affected by $(\lambda_0, \theta)$, the two parameters controlling the sparsity of the solution. The question remains, what values of $(\lambda_0, \theta)$ yield $\widehat{\boldsymbol{\beta}}$ with the minimax risk, possibly up to a multiplicative constant. The answer is provided in Corollary 4.1.

COROLLARY 4.1. *Assume model* (1.1) *with* $\boldsymbol{\beta}_0 \in l_0[p_n; n]$, *where* $p_n, n \to \infty$ *and* $p_n = o(n)$. *Suppose* $\lambda_1 < \mathrm{e}^{-2}$ *and*

$$\theta \simeq \left(\frac{p_n}{n}\right)^{\alpha} \quad \text{and} \quad \lambda_0 \simeq \left(\frac{n}{p_n}\right)^{\nu}, \tag{4.2}$$

*for* $\alpha \geq 0$, $\nu > 0$ *and* $\alpha + \nu \geq 1$. *If* $(\lambda_0 - \lambda_1) > 2$ *and* $g(0) > 0$, *then we have*

$$\sup_{\boldsymbol{\beta}_0 \in l_0[p_n; n]} \mathsf{E}_{\boldsymbol{\beta}_0} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 \simeq p_n \log(n/p_n).$$

PROOF. Applying Theorem 4.1, we find $p^\star(0) \preceq \lambda_1 (p_n/n)^{\alpha+\nu}$ and $\lambda_1 \mathrm{e}^d < 1$. If $\alpha + \nu \geq 1$, both summands in (4.1) are dominated by the term $p_n \log[1/p^\star(0)] \simeq (\alpha + \nu) p_n \log(n/p_n)$. □

Corollary 4.1 provides valuable insights into the delicate interplay between $\theta$ and $\lambda_0$. The mixing weight $\theta$ in (4.2) relates to the true proportion of the nonzero elements $p_n/n$. This is no surprise, since $\theta$ is often regarded as a proxy for the proportion of active coefficients. A more surprising finding is that $\lambda_0$ should increase ideally at a rate $(n/p_n)^\nu$. Any faster increase would have to be compensated by a slower decay of $\theta$ in order to maintain the balance. Thus, the parameters $\lambda_0$ and $\theta$ are ultimately tied with each other [through $p^\star(0)$] and have to cooperate in order to achieve optimal performance. This conveys an important conclusion that the rate at which the SS-LASSO prior approaches the limiting ideal (1.2) should not be arbitrary.

Interestingly, $\theta$ actually *does not* have to be adaptive as long as $\lambda_0$ is. Indeed, with $\alpha = 0$ and $\lambda_0 \sim (n/p_n)^\nu$ for $\nu \geq 1$ we can still obtain the minimax performance. This shows that the traditional interpretation of $\theta$ as the sparsity level does not necessarily pertain to the continuous spike-and-slab priors. The use of adaptive $\theta$ is thus not crucial for the posterior mode. However, it will be crucial for the posterior distribution, as will be seen in Section 5 and Section 5.3.

Recommendations for the choice of hyper-parameters $(\theta, \lambda_0)$ can be obtained by matching the dominant term $8p_n \log[1/p^\star(0)]$ in (4.1) to the minimax risk $2p_n \log[n/p_n]$. For instance, with $\lambda_0 = n/p_n$ and $(1-\theta)/\theta = \lambda_1$, the leading term becomes $8p_n \log[1 + n/p_n]$, inflating the minimax rate only by a factor of 4. Another possibility is setting $\lambda_0 = (1-\theta)/(\theta\lambda_1) = (n/p_n)^{1/2}$ which yields the same upper bound. Generally, larger values $\alpha + \nu > 1$ will inflate the multiplication constant in front of the minimax rate. We explore the practical implications of these hyper-parameter choices in Section 6.

Corollary 4.1 provides asymptotic minimax optimality when $p_n$ is known, an assumption rarely available. However, with the following automatic choice of $(\lambda_0, \theta)$, near-minimax performance can be achieved when $p_n$ is unknown.

COROLLARY 4.2. *Assume model* (1.1) *with* $\boldsymbol{\beta}_0 \in l_0[p_n; n]$, *where* $p_n, n \to \infty$ *and* $p_n = o(n)$. *Suppose* $\lambda_1 < \mathrm{e}^{-2}$ *and*

$$\theta \simeq 1/n^{\alpha} \quad \text{and} \quad \lambda_0 \simeq n^{\nu}, \tag{4.3}$$

*where* $\alpha \geq 0$, $\nu > 0$ *and* $\alpha + \nu \geq 1$. *If* $(\lambda_0 - \lambda_1) > 2$ *and* $g(0) > 0$, *then we have*

$$\sup_{\boldsymbol{\beta}_0 \in l_0[p_n;n]} \mathsf{E}_{\boldsymbol{\beta}_0} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 \preceq p_n \log(n).$$

PROOF. With (4.3), we obtain $1/p^{\star}(0) = 1 + n^{\alpha+\nu}/\lambda_1$, yielding a risk of order $p_n \log n(1 + o(1))$. □

REMARK 4.1 (Connection to the LASSO estimator). Recall that the LASSO estimator $\widehat{\boldsymbol{\beta}}_{\lambda}$ attains the near-minimax risk $2p_n \log n(1 + o(1))$ with the universal penalty $\lambda = \sqrt{2 \log n}$. As in Theorem 4.1, we can obtain an analogous upper risk bound for $\widehat{\boldsymbol{\beta}}_{\lambda}$: $\mathsf{E}_{\boldsymbol{\beta}_0} \|\widehat{\boldsymbol{\beta}}_{\lambda} - \boldsymbol{\beta}_0\|^2 \leq p_n(2 + 4\lambda^2) + (n - p_n)4\lambda\psi(\lambda)$. This bound, despite improvable, showcases the analogy between $\lambda$ (the LASSO complexity penalty) and $\sqrt{\log[1/p^{\star}(0)]}$ (the SS-LASSO complexity penalty). When $\lambda_0$ and $\theta$ are as in (4.3), $\sqrt{\log[1/p^{\star}(0)]}$ is of the same order as the universal penalty $\sqrt{2 \log n}$. Thus, $\widehat{\boldsymbol{\beta}}$ also attains the near-minimax risk up to a constant.

So far, our minimax-rate optimality result in Corollary 4.1 was obtained under the assumption that $p_n$ was known. Now, we show how the performance of the SS-LASSO posterior mode can be sharpened in the absence of knowledge of $p_n$.

4.1. *Adapting to unknown levels of sparsity.* We design adaptive penalized likelihood estimators that aspire to mimic the oracle performance obtained with $(\theta, \lambda_0)$ chosen so that $\Delta \simeq \sqrt{2 \log(n/p_n)}$ when $p_n$ is unknown. We pursue a two-step approach inspired by empirical Bayes considerations.

We begin by showing that under the assumptions of Corollary 4.2, the estimator $\widehat{p}_n \equiv \|\widehat{\boldsymbol{\beta}}\|_0$ overshoots $p_n$ by at most a constant factor (with large probability).

THEOREM 4.2. *Under the assumptions of Corollary* 4.2, *the following two statements hold with probability at least* $1 - \frac{2}{n}$. *The estimator* $\widehat{p}_n \equiv \|\widehat{\boldsymbol{\beta}}\|_0$ *satisfies*

$$\widehat{p}_n \leq p_n(1 + C),$$

*where* $0 < C < 2$ *whenever* $\alpha + \nu - 4(1 + \lambda_1) > c > 0$. *Moreover, if* $|\beta_{0i}| > b_0 > D\sqrt{p_n \log n}$ *for each* $\beta_{0i} \neq 0$ *and some suitable* $D > 0$, *then* $p_n \leq \widehat{p}_n$.

PROOF. See the supplementary materials ([29], Section 1.2.2). □

Now we introduce a two-step SS-LASSO approach.

DEFINITION 4.1. Denote by $\widehat{p}_n = \|\widehat{\boldsymbol{\beta}}\|_0$, where $\widehat{\boldsymbol{\beta}}$ is the SSL posterior mode from Theorem 4.2. The two-step SS-LASSO estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{TS}}$ is defined as the SS-LASSO posterior mode obtained with $0 < \lambda_1 < \mathrm{e}^{-2}$, $\theta \simeq (\frac{\widehat{p}_n+1}{n})^{\alpha}$ and $\lambda_0 \simeq (\frac{n}{\widehat{p}_n+1})^{\nu}$ for some $\alpha \geq 0$, $\nu > 0$ and $\alpha + \nu = 1$.

Note that $\widehat{\boldsymbol{\beta}}_{\mathrm{TS}}$ corresponds to the one-step SS-LASSO estimator from Corollary 4.2 when $\widehat{p}_n = 0$. The following theorem provides an upper bound for the maximal risk of $\widehat{\boldsymbol{\beta}}_{\mathrm{TS}}$.

THEOREM 4.3. *Let $\widehat{\boldsymbol{\beta}}_{\mathrm{TS}}$ be the two-step estimator from Definition 4.1, where $\widehat{p}_n > 0$. Then*

$$\sup_{\boldsymbol{\beta}_0 \in l_0[p_n;n]} \mathsf{E}_{\boldsymbol{\beta}_0} \|\widehat{\boldsymbol{\beta}}_{\mathrm{TS}} - \boldsymbol{\beta}_0\|^2 \preceq p_n \mathsf{E}_{\boldsymbol{\beta}_0} \log\Big(\frac{n}{\widehat{p}_n}\Big). \tag{4.4}$$

*Moreover, if $|\beta_{0i}| > b_0 > D\sqrt{p_n \log n}$ for each $\beta_{0i} \neq 0$ and some suitable $D > 0$, then $\widehat{\boldsymbol{\beta}}_{\mathrm{TS}}$ achieves the minimax rate $p_n \log(n/p_n)$.*

PROOF. With $(\theta, \lambda_0, \lambda_1)$ given in Definition 4.1, we have $\frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1} \simeq \frac{n}{\widehat{p}_n+1}$ and thereby $p^\star(0) \simeq \frac{\widehat{p}_n+1}{n}$. Theorem 4.1 yields

$$\mathsf{E}_{\boldsymbol{\beta}_0} \|\widehat{\boldsymbol{\beta}}_{\mathrm{TS}} - \boldsymbol{\beta}_0\|^2 \preceq p_n \mathsf{E}_{\boldsymbol{\beta}_0} \log[1/p^\star(0)] + n \mathsf{E}_{\boldsymbol{\beta}_0} p^\star(0) \sqrt{\log[1/p^\star(0)]}. \tag{4.5}$$

According to Theorem 4.2, we have $\mathsf{P}_{\boldsymbol{\beta}_0}[\widehat{p}_n < (1+C)p_n] > 1 - \frac{2}{n}$. Thereby, the second term in (4.5) can be upper-bounded by a constant multiple of

$$\mathsf{E}_{\boldsymbol{\beta}_0} \widehat{p}_n \sqrt{\log[1/p^\star(0)]} < \Big[(1+C)p_n\Big(1 - \frac{2}{n}\Big) + \frac{2}{n}n\Big]\sqrt{\log[1/p^\star(0)]}.$$

The first term in (4.5) thus dominates the second term, where $\log[1/p^\star(0)] \simeq \log(\frac{n}{\widehat{p}_n+1})$. This observation directly implies (4.4). The second statement follows again from Theorem 4.2 by noting $\mathsf{E}_{\boldsymbol{\beta}_0} \log(\frac{n}{\widehat{p}_n+1}) < (1 - \frac{2}{n}) \log(\frac{n}{p_n}) + \frac{2\log n}{n} \preceq \log(\frac{n}{p_n})$. □

Theorem 4.3 shows that when $\widehat{p}_n \to \infty$ (as $p_n, n \to \infty$), $\widehat{\boldsymbol{\beta}}_{\mathrm{TS}}$ improves on the nonadaptive estimators (such as the LASSO with $\lambda \sim \sqrt{2\log n}$ or the SS-LASSO estimator from Corollary 4.2) by achieving a sharper upper bound on the maximal risk. In conclusion, the rate of $\widehat{\boldsymbol{\beta}}_{\mathrm{TS}}$ is at least as good as the near-minimax rate and, under the beta-min condition of Theorem 4.2, it is minimax. Note that the beta-min condition, though a bit stronger than usual, it is not required to obtain improved performance (as long as $\widehat{p}_n \to \infty$). We will illustrate the performance of $\widehat{\boldsymbol{\beta}}_{\mathrm{TS}}$ in the simulation study in Section 6, showing that it mimics the performance of the parent oracle estimator.

The two-step approach has an empirical Bayes flavor in the sense that we are estimating the unknown coefficients $(\theta, \lambda_0)$ from the data. Alternative empirical Bayes strategies are discussed in van der Pas et al. [36], in the context of horseshoe priors, and Johnstone and Silverman [21], in the context of point-mass mixtures.

An alternative route toward adaptive estimators could be obtained, for instance, through fully Bayes considerations. This strategy is developed in Section 5.3, where we show minimax rates of posterior concentration under suitable hierarchical priors.

**5. Spike-and-slab: The Bayesian perspective.** For fully Bayesian inference about $\boldsymbol{\beta}_0$, the fundamental instrument is the *entire* posterior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{y}^{(n)})$. This random measure serves as a vehicle for both estimation and uncertainty quantification. For estimation, we studied the global mode and showed that it can be (near) minimax-rate optimal. In high-dimensional settings, however, the behavior of the posterior and its aspects (mode, mean or median) can be very different [8]. For instance, the LASSO posterior mode is known to be asymptotically near-minimax when $\lambda = \sqrt{2 \log n}$. At the same time, such LASSO prior induces asymptotically vanishing posterior mass on balls centered at $\boldsymbol{\beta}_0$ with a radius of much larger order than the near-minimax rate [10]. The posterior thus contracts a lot slower than the posterior mode dampening its usefulness for uncertainty quantification. This limitation is overcome with the SS-LASSO prior.

We show that asymptotic (near) minimaxity can be achieved simultaneously for the global mode and the entire posterior with the SS-LASSO prior. Thus, the posterior does not prevent from valid posterior inference, behaving similarly as the rate-optimal posteriors obtained with the limiting prior (1.2). In this analysis, we build on Castillo and van der Vaart [10], who pioneered posterior convergence rate results for variable selection priors, and on Bhattacharya et al. [3], who extended them to one-group shrinkage priors.

For now (until Section 5.3), we will assume that $(\lambda_0, \theta)$ are fixed, in which case the SS-LASSO prior constitutes an independent product. We will denote such prior by $\mathrm{SSL}(\lambda_0; \lambda_1; \theta)$. The results obtained under this simpler scenario will be a stepping stone for the developments in Section 5.3 with dependent coordinates induced by putting a prior on $(\lambda_0, \theta)$.

The SS-LASSO prior, despite being continuous, is also a two-group prior. However, it segregates the coefficients into negligible and nonnegligible groups rather than into zero and nonzero groups. This separation is captured by the latent $\gamma_i$ indicators. However, the $\gamma_i$'s now indicate the magnitude of the nonzero coefficients $\beta_i$. Naturally, coefficients $\beta_i$ such that $|\beta_i| > \delta(\lambda_0, \theta)$ are more likely affiliated with the slab (rather than the spike) because $\mathsf{P}(\gamma_i = 1|\beta_i, \theta) > 0.5$. Because the posterior puts no mass on exactly sparse vectors, we will work with a more general notion of sparsity. For this purpose, we introduce generalized binary indicators

$$\gamma(\beta_i) = \mathbb{I}\big[|\beta_i| > \delta(\lambda_0, \theta)\big] \tag{5.1}$$

to designate whether the coefficients are nonnegligible. These indicators are again i.i.d. with a Bernoulli distribution, where the success probability relates to $\theta$ in the following way.

LEMMA 5.1. *Assume* $\beta \sim \mathrm{SSL}(\lambda_0; \lambda_1; \theta)$ *with* $\theta \in (0, 1/2)$, $\lambda_1 \leq \mathrm{e}^{-2}$ *and* $\lambda_0 > 1$. *Then*

$$\mathsf{P}\big[\gamma(\beta) = 1\big] < \theta.$$

PROOF. Denote by $\delta_\theta = \delta(\lambda_0, \theta)$. Using the fact $\theta\psi_1(\delta_\theta) = (1-\theta)\psi_0(\delta_\theta)$, we have $\mathsf{P}(|\beta| > \delta_\theta) = \theta \exp[-\delta_\theta\lambda_1](1+\frac{\lambda_1}{\lambda_0})$. Because $\lambda_1 \leq \mathrm{e}^{-2}$ and $\lambda_0(1-\theta)/\theta > 1$, we have $\log[\frac{\lambda_0(1-\theta)}{\lambda_1\theta}] > 2$. Thereby

$$\exp(-\delta_\theta\lambda_1) = \exp\left\{-\frac{\lambda_1}{\lambda_0-\lambda_1}\log\left[\frac{\lambda_0(1-\theta)}{\lambda_1\theta}\right]\right\} \leq \exp\left(-\frac{2\lambda_1}{\lambda_0}\right) \leq \left(1-\frac{\lambda_1}{\lambda_0}\right).$$

Altogether, we obtain $\mathsf{P}(|\beta| > \delta_\theta) \leq \theta(1-\frac{\lambda_1^2}{\lambda_0^2}) < \theta$. □

We now define the notion of effective dimensionality under the SS-LASSO prior, an analogue to the actual dimensionality $|\boldsymbol{\gamma}| = \sum_{i=1}^n \gamma_i$ under the point-mass mixture prior.

DEFINITION 5.1. For the intersection point $\delta(\lambda_0, \theta)$, let us define $\boldsymbol{\gamma}(\boldsymbol{\beta}) = [\gamma(\beta_1), \ldots, \gamma(\beta_n)]'$ the vector of indicators (5.1) of active coefficients. Then by effective dimensionality we refer to

$$|\boldsymbol{\gamma}(\boldsymbol{\beta})| = \sum_{i=1}^n \gamma(\beta_i). \tag{5.2}$$

It is worthwhile to note that the intersection point (3.12) depends on $1/p^\star(0)$ through $\delta(\lambda_0, \theta) = \frac{1}{\lambda_0-\lambda_1}\log[\frac{1}{p^\star(0)} - 1]$. We will be interested in situations when $\lambda_0 \to \infty$ and $\theta \to 0$ as $n \to \infty$, in which case $\delta(\lambda_0, \theta) \to 0$ and $|\boldsymbol{\gamma}(\boldsymbol{\beta})|$ coincides with $|\boldsymbol{\gamma}|$ in the limit. Of particular interest to us will be the two scenarios in Corollaries 4.1 and 4.2, where $\delta(\lambda_0, \theta) \sim p_n/n \log(n/p_n)$ and $\delta(\lambda_0, \theta) \sim \log(n)/n$, respectively.

As seen in the previous section, $p^\star(0)$ controls the risk of the global posterior mode. Going further, we provide a result showing that the entire posterior distribution concentrates at a rate which depends on $p^\star(0)$. A crucial step toward obtaining the convergence rates is the study of the posterior effective dimensionality $|\boldsymbol{\gamma}(\boldsymbol{\beta})|$.

5.1. *Effective posterior dimension.* It is desirable that the posterior effective dimensionality $|\boldsymbol{\gamma}(\boldsymbol{\beta})|$ accumulates roughly around the true dimensionality $p_n$. With our notion of sparsity (5.1) and (5.2), this is akin to a requirement that the posterior squeezes most of its mass between $\pm\delta(\lambda_0, \theta)$ in roughly $n - p_n$ directions. This is where $\boldsymbol{\beta}_0$ is presumed to reside.

One important aspect leading to this desired property is having a prior on $|\boldsymbol{\gamma}(\boldsymbol{\beta})|$ that decays exponentially with $p_n$, that is, $\mathsf{P}(|\boldsymbol{\gamma}(\boldsymbol{\beta})| > k) < \mathrm{e}^{-Ck}$ for some $C > 0$ and $k \geq p_n$. With a suitably small $\theta$, the distribution on $|\boldsymbol{\gamma}|$ is exponentially decaying in $p_n$ from Chernoff's inequality. Due to Lemma 5.1, we obtain that the prior on $|\boldsymbol{\gamma}(\boldsymbol{\beta})|$ is also exponentially decaying in $p_n$.

LEMMA 5.2. *Assume* $\boldsymbol{\beta} \sim \mathrm{SSL}(\lambda_0; \lambda_1; \theta)$ *with* $\lambda_1 \leq \mathrm{e}^{-2}$ *and* $\lambda_0 > 1$. *Assume* $\theta < Ap_n/n < 1/2$ *for some* $A > 0$. *Then for any* $C > 2A\mathrm{e}$, *we have*

$$\mathsf{P}\big[|\boldsymbol{\gamma}(\boldsymbol{\beta})| > Cp_n\big] \leq \exp(-p_n C \log 2). \tag{5.3}$$

PROOF. The random variable $B_\theta \equiv |\boldsymbol{\gamma}(\boldsymbol{\beta})|$ is distributed according to Binomial$(n, \pi_\theta)$, where $\pi_\theta = \mathsf{P}(|\beta_1| > \delta_\theta)$. A version of Chernoff's inequality for the binomial distribution states that for $t > 2\mathrm{e}\mathsf{E}B_\theta$ we have $\mathsf{P}(B_\theta > t) \leq 2^{-t}$. From Lemma 5.1, we have $\pi_\theta \leq \theta \leq Ap_n/n$. The result then follows from Chernoff's inequality with $t = 2A\mathrm{e}p_n$. □

Note that the space of vectors $\boldsymbol{\beta} \in \mathbb{R}^n$ can be partitioned into $2^n$ subsets, identified by $\boldsymbol{\gamma}(\boldsymbol{\beta})$. For each of the $n$ coefficients, we have two possibilities: either $|\beta_i| \leq \delta(\lambda_0, \theta)$ or $|\beta_i| > \delta(\lambda_0, \theta)$. This results in a tessellation of $\mathbb{R}^n$ into $3^n$ boxes [splitting the set $\{\beta_i : |\beta_i| > \delta(\lambda_0, \theta)\}$ into positive and negative values], each having a positive prior probability. Lemma 5.2 shows that with a suitably chosen $\theta$, the SS-LASSO prior assigns small probability to boxes away from the origin in more than $Cp_n$ directions. Importantly, this property will be transmitted to the posterior, as shown in the following theorem.

THEOREM 5.1. *Consider the model* (1.1) *with* $\boldsymbol{\beta} \sim \mathrm{SSL}(\lambda_0; \lambda_1; \theta)$. *Assume* $\boldsymbol{\beta}_0 \in l_0[p_n; n]$ *with* $p_n, n \to \infty$ *and* $p_n = o(n)$. *Assume* $\theta \leq Ap_n/n < 1/2$ *with* $A > 1/2$, $\lambda_1 < \mathrm{e}^{-2}$ *and* $\lambda_0 \geq n/p_n$, $\forall n \in \mathbb{N}$. *Then*

$$\lim_{n\to\infty} \mathsf{E}_{\boldsymbol{\beta}_0} \mathsf{P}\big[|\boldsymbol{\gamma}(\boldsymbol{\beta})| > Cp_n | \boldsymbol{y}^{(n)}\big] = 0$$

*for any constant* $C > 2A\mathrm{e}$.

PROOF. See Appendix A.1.1. □

Theorem 5.1 was obtained under suitable conditions on $(\theta, \lambda_0)$. The mixing weight $\theta$ has to be reasonably small to guarantee the exponential decay of the prior on $|\boldsymbol{\gamma}(\boldsymbol{\beta})|$. In addition, $\lambda_0$ has to be reasonably large so that the tessellation yields boxes narrow enough to be informative about the sparsity of $\boldsymbol{\beta}_0$.

5.2. *Optimal posterior concentration.* Our goal is to show that the posterior $\pi(\boldsymbol{\beta}|\boldsymbol{y}^{(n)})$ concentrates asymptotically on $n$-balls centered at $\boldsymbol{\beta}_0$ with a square radius proportional to an optimal rate $r_n$. Here, $r_n$ will be either the minimax rate $p_n \log(n/p_n)$ or the near-minimax rate $p_n \log n$, depending on the context. More precisely, we want to show

$$\sup_{\boldsymbol{\beta}_0 \in l_0[p_n;n]} \mathsf{E}_{\boldsymbol{\beta}_0} \mathsf{P}\big(\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 > M r_n | \boldsymbol{y}^{(n)}\big) \to 0 \tag{5.4}$$

for a sufficiently large constant $M > 0$.

Castillo and van der Vaart [10] establish the minimax-optimal rate of posterior concentration for a class of point-mass mixture priors using a testing argument. This strategy relies on two main ingredients: (a) an exponentially decaying prior on $|\boldsymbol{\gamma}|$ and (b) heavy tailed slab densities $\pi(\beta_i|\gamma_i = 1)$ that are i.i.d. and proportional to $\mathrm{e}^h$, for $h : \mathbb{R} \to \mathbb{R}$ where $|h(x) - h(y)| \preceq 1 + |x - y|, \forall x, y \in \mathbb{R}$. With a continuous prior like SS-LASSO, the conditions (a) and (b) have to be suitably modified.

In the previous section, we formalized an analogue to the condition (a) using instead the effective-dimensionality $|\boldsymbol{\gamma}(\boldsymbol{\beta})|$. To modify (b), we impose a condition on the tails of the whole mixture rather than only the slab. Intuitively, the tails will be dominated by the slab density for sufficiently large $|\beta|$. Thus, the SS-LASSO marginal prior ought to satisfy a similar Lipschitz property. This is formalized in the following lemma.

LEMMA 5.3. *Denote by* $h(x) \equiv \mathrm{pen}(x)$ *the logarithm of the* $\mathrm{SSL}(\lambda_0; \lambda_1; \theta)$ *density* (3.1), *where* $\lambda_0 > \lambda_1$. *Then*

$$\big|h(x) - h(y)\big| \le C(\lambda_0; \lambda_1; \theta) + \lambda_1 |x - y| \qquad \forall x, y \in \mathbb{R},$$

*where* $C(\lambda_0; \lambda_1; \theta) = \log[\frac{1}{p^\star(0)} - 1]$.

PROOF. See Appendix A.1.2. □

REMARK 5.1. Letting $\lambda_0 \to \infty$ and $\theta \to 0$ as $n \to \infty$, the constant $C(\lambda_0; \lambda_1; \theta)$ will grow to infinity at a rate controlled by the familiar functional $\log[1/p^\star(0)]$.

Now, we are ready to state the main theorem. We show that the contraction rate of $\pi(\boldsymbol{\beta}|\boldsymbol{y}^{(n)})$ under the $\mathrm{SSL}(\lambda_0; \lambda_1; \theta)$ prior depends on $p^\star(0)$, the quantity which was shown to control the risk of the global mode.

From Section 4, we know that $(\lambda_0, \theta)$ should work in tandem to control the rate of $p^\star(0)$. Again, we will need $\theta$ to decay at a rate no slower than $p_n/n$. To deploy Theorem 5.1, we will also need $\lambda_0 \ge n/p_n$.

THEOREM 5.2. *Consider the model* (1.1) *with* $\boldsymbol{\beta}_0 \in l_0[p_n; n]$, *where* $p_n, n \to \infty$ *and* $p_n = o(n)$ *with* $p_n/n < 1/2$. *Assume* $\boldsymbol{\beta} \sim \mathrm{SSL}(\lambda_0; \lambda_1; \theta)$ *where* $\theta \leq p_n/(p_n + n)$, $1/C_{\lambda_1} < \lambda_1 \leq \mathrm{e}^{-2}$ *and* $\lambda_0 \geq n/p_n$. *Denote by* $r_n^2 = p_n[\log(\frac{1}{p^\star(0)} - 1)]$. *Then*

$$\lim_{n\to\infty} \mathsf{E}_{\boldsymbol{\beta}_0} \mathsf{P}\big(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 > M r_n^2 | \boldsymbol{y}^{(n)}\big) = 0$$

*for some constant* $M > \sqrt{(1 + C_{\lambda_1})/3}$.

PROOF. See Appendix A.1.3. □

By Theorems 5.2 and 4.1, one can simultaneously achieve good reconstruction (using the global mode) and good posterior concentration by suitably controlling $p^\star(0)$. The following two immediate corollaries are companion results to Corollaries 4.1 and 4.2. They show that the posterior distribution under a properly tuned SS-LASSO prior contracts at the minimax rate when $p_n$ is known and at the near-minimax rate when $p_n$ is unknown.

COROLLARY 5.1. *Consider the model* (1.1) *with* $\boldsymbol{\beta}_0 \in l_0[p_n; n]$, *where* $p_n, n \to \infty$ *and* $p_n = o(n)$ *with* $p_n/n < 1/2$. *Assume* $\boldsymbol{\beta} \sim \mathrm{SSL}(\lambda_0; \lambda_1; \theta)$ *with* $1/C_{\lambda_1} < \lambda_1 \leq \mathrm{e}^{-2}$ *and*

$$\theta = p_n/(p_n + n) \quad \text{and} \quad \lambda_0 = n/p_n. \tag{5.5}$$

*Then*

$$\lim_{n\to\infty} \mathsf{E}_{\boldsymbol{\beta}_0} \mathsf{P}\big(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 > M p_n \log(n/p_n) | \boldsymbol{y}^{(n)}\big) = 0$$

*for some constant* $M > \sqrt{(1 + C_{\lambda_1})/3}$.

PROOF. The proof follows by noting $p_n[\log(\frac{1}{p^\star(0)} - 1)] \simeq p_n \log(n/p_n)$. □

COROLLARY 5.2. *Consider the model* (1.1) *with* $\boldsymbol{\beta}_0 \in l_0[p_n; n]$, *where* $p_n, n \to \infty$ *and* $p_n = o(n)$ *with* $p_n/n < 1/2$. *Assume* $\boldsymbol{\beta} \sim \mathrm{SSL}(\lambda_0; \lambda_1; \theta)$ *with* $1/C_{\lambda_1} < \lambda_1 \leq \mathrm{e}^{-2}$ *and*

$$\theta = 1/(1 + n) \quad \text{and} \quad \lambda_0 = n. \tag{5.6}$$

*Then*

$$\lim_{n\to\infty} \mathsf{E}_{\boldsymbol{\beta}_0} \mathsf{P}\big(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 > M p_n \log(n) | \boldsymbol{y}^{(n)}\big) = 0$$

*for some constant* $M > \sqrt{(1 + C_{\lambda_1})/3}$.

PROOF. The proof follows by noting $p_n[\log(\frac{1}{p^\star(0)} - 1)] \simeq p_n \log(n)$. $\square$

The concentration results in Corollaries 5.1 and 5.2 do not require any assumptions restricting the size of the true signal, such as limiting $\|\boldsymbol{\beta}_0\|^2$ to be small. Such assumptions would be needed if the priors were to over-shrink the large coefficients in magnitude. Such problems would occur, for instance, with Gaussian slab distributions if $\|\boldsymbol{\beta}_0\|^2 \gg p_n \log(n/p_n)$ [10]. Assuming $p_n$ is known, the SS-LASSO prior avoids even the weaker assumption $\|\boldsymbol{\beta}_0\|^2 < p_n \log^4 n$ of [3].

5.3. *Fully Bayes considerations*. Ideally, the prior inclusion probability $\theta$ and the reciprocal of the spike penalty $1/\lambda_0$ should be set close to $p_n/n$. When $p_n$ is unknown, it is natural to treat $(\lambda_0, \theta)$ also as unknown with a prior distribution. The hope is that with a suitable prior, the posterior can adapt to the unknown sparsity level and contract at the minimax rate. Remarkably, this happens with point-mass mixture priors, when combined with a suitable beta prior $\pi(\theta)$ [10].

Rather than treating $\lambda_0$ and $\theta$ as two independent parameters and assigning a prior to both, we induce a prior distribution on $\theta$ and set $\lambda_0$ deterministically to $(1-\theta)/\theta$. This functional is chosen here (and in Theorem 5.2) so that $1/p^\star(0) = 1 + (n/p_n)^2/\lambda_1$ for $\theta = p_n/(n + p_n)$. By tying $\lambda_0$ with $\theta$, putting a prior only on $\theta$ will be enough to obtain the desired adaptivity.

We consider the beta prior $\pi(\theta) \sim \mathcal{B}(a, b)$, where $a \ll b$, so that $\theta$ is small with high probability. In particular, we set $a = 1$ and $b = 4n$ in results about to follow. To achieve the rate-optimal concentration under the point-mass mixture priors, [10] recommend setting $a = 1$ and $b = \kappa n + 1$ for some $\kappa > 0$. With $\lambda_0 = (1 - \theta)/\theta$, this amounts to assigning $\lambda_0$ a beta-prime[2] distribution $\beta'(a, b)$. Although this prior assigns positive probability to $\{\theta : \lambda_0 = (1-\theta)/\theta < \lambda_1\}$, these values will not be supported by the data in the presence of sparsity.

The prior on $\theta$ (and inherently on $\lambda_0$) renders the coefficients $\boldsymbol{\beta}$ a priori (and hence a posteriori) dependent. This is in sharp contrast with earlier results in this section, where the coordinates were separable. Independent product priors were studied previously by Bhattacharya et al. [3], who pioneered posterior concentration results for continuous shrinkage priors. Going a step further, here we obtain posterior convergence rates for a dependent continuous prior. The dependence comes exclusively from the mixing over $\theta$. Other sources of dependence can be introduced through multivariate slab densities [10]. We do not pursue these alternatives here.

We will leverage results established earlier in Section 5. In the remainder of this section, we will be using the following notation. In light of Corollary 5.1, denote by

$$\theta^o = \frac{p_n}{p_n + n} \quad \text{and} \quad \lambda_0^o = \frac{n}{p_n} \tag{5.7}$$

[2]The beta-prime distribution $\beta'(a, b)$ has an expectation $b/(a-1)$ for $a > 1$.

the "oracle" choices of parameters which would yield the minimax concentration rate if $p_n$ was known. Similarly, let

$$\delta^o \equiv \delta(\lambda_0^o, \theta^o) = \frac{1}{(\frac{n}{p_n}) - \lambda_1} \log\left[\left(\frac{n}{p_n}\right)^2 \frac{1}{\lambda_1}\right] \tag{5.8}$$

be the "oracle" intersection point, which is again the ideal intersection point when $p_n$ is known. Furthermore, by $\gamma_o(\beta_i)$ we will denote the indicator function (5.1) with $\delta(\lambda_0, \theta) = \delta^o$. We will denote by M-SSL$(\lambda_1; a; b)$ the mixture of SSL$(\lambda_0; \lambda_1; \theta)$ priors with $\pi(\theta) \sim \mathcal{B}(a, b)$ and $\lambda_0 = (1-\theta)/\theta$.

To begin, we generalize Theorem 5.1 to M-SSL$(\lambda_1; a; b)$ priors. One side-effect of the dependence is that the random variable $|\boldsymbol{\gamma}_o(\boldsymbol{\beta})|$ is no longer distributed binomially, but rather beta-binomially. Although this precludes from using the Chernoff's inequality, the result can be obtained with suitable modifications.

THEOREM 5.3. *Consider the model* (1.1) *with* $\boldsymbol{\beta} \sim$ M-SSL$(\lambda_1; a; b)$, *where* $\lambda_1 < \mathrm{e}^{-2}$, $a = 1$ *and* $b = 4n$. *Assume* $\boldsymbol{\beta}_0 \in l_0[p_n; n]$ *with* $p_n, n \to \infty$ *and* $p_n = o(n)$. *Assume* $p_n/n < 1/2$. *Then*

$$\lim_{n\to\infty} \mathsf{E}_{\boldsymbol{\beta}_0} \mathsf{P}\big(|\boldsymbol{\gamma}_o(\boldsymbol{\beta})| > C p_n | \boldsymbol{y}^{(n)}\big) = 0$$

*for any constant* $C > 2\mathrm{e}$.

PROOF. See Appendix A.2.1. □

Theorem 5.3 is only one of the two pieces needed to carry out the concentration result. The second piece is the Lipschitz property of the logarithm of the marginal prior $\pi(\boldsymbol{\beta})$. We will show that such property is satisfied by the $|S|$-variate marginal M-SSL$(\lambda_1; a; b)$ prior, confined to coordinates in a set $S \subset \{1, \ldots, n\}$. We denote this marginal prior by $\pi_S(\boldsymbol{\beta})$. By Fubini's theorem, we have

$$\begin{aligned} \pi_S(\boldsymbol{\beta}) &= \int_0^1 \prod_{i\in S} \pi(\beta_i|\theta)\, \mathrm{d}\pi(\theta) \\ &= \left(\frac{\lambda_1}{2}\right)^{|S|} \mathrm{e}^{-\lambda_1|\boldsymbol{\beta}_S|} \int_0^1 \theta^{|S|} \prod_{i\in S} \frac{1}{p^\star_\theta(\beta_i)}\, \mathrm{d}\pi(\theta). \end{aligned} \tag{5.9}$$

Here, $p^\star_\theta(\beta)$ is as in (3.3), where the $\theta$ subscript is added to emphasize the dependence on $\theta$.

The following lemma generalizes Lemma 5.3 to the M-SSL$(\lambda_1; a; b)$ prior.

LEMMA 5.4. *Assume* $S \subset \{1, \ldots, n\}$ *and let* $\pi_S(\boldsymbol{\beta})$ *be as in* (5.9), *where* $\lambda_1 < \mathrm{e}^{-2}$. *Assume* $p_n/n < 1/2$ *and let* $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^n$ *be such that* $|\beta_i| > \delta^o$, $i \in S$. *Then*

$$\log \pi_S(\boldsymbol{\beta}) - \log \pi_S(\boldsymbol{\beta}') < |S| C(\lambda_1) + \lambda_1 \|\boldsymbol{\beta}_S - \boldsymbol{\beta}'_S\|_1,$$

*where* $C(\lambda_1) = \log[2 + \frac{1}{p^\star_{\theta^o}(0)}]$.

PROOF. See Appendix A.2.2. □

Note the similarities with Lemma 5.3 introduced earlier. There the additive constant $C(\lambda_0; \lambda_1; \theta)$ depends on $p^\star(0)$ that is evaluated at $\theta$, which may not be optimal. Here, the additive constant $C(\lambda_1)$ depends on $p^\star_\theta(0)$ evaluated at the oracle mixing proportion $\theta^o$, which is known to be the optimal one.

In Theorem 5.2, we established the posterior concentration rate in terms of $p^\star_\theta(0)$ with $\theta$ fixed. There, unless $\theta$ is set close to $p_n/n$, the posterior would not be minimax-rate optimal. Here, we cast the convergence rate in terms of $p^\star_{\theta^o}(0)$, which yields the minimax order $p_n \log[n/p_n]$. To this end, we will need one additional assumption regarding the nonzero elements of $\boldsymbol{\beta}_0$. For $\boldsymbol{\beta}_0 \in l_0[p_n; n]$, we will require that the nonzero entries satisfy $|\beta_{0i}| > \delta^o$. This mild requirement is needed to assure that the signal is strong enough to be recoverable. Due to the continuous spike, the M-SSL prior may otherwise over-shrink negligible, yet nonzero, coefficients. This is in line with the notion of "practical significance" that has been associated with continuous spike and slab priors [17]. The coefficients worth recovering should be of magnitude greater than the intersection point. Since $\delta^o \sim p_n/n \log[n/p_n] \to 0$ as $n \to \infty$, this condition vanishes asymptotically as the M-SSL prior approaches to the point-mass mixture prior.

THEOREM 5.4. *Consider the model* (1.1) *with* $\boldsymbol{\beta} \sim \text{M-SSL}(\lambda_1; a; b)$, *where* $\lambda_1 < \mathrm{e}^{-2}$, $a = 1$ *and* $b = 4n$. *Assume* $\boldsymbol{\beta}_0 \in l_0[p_n; n]$ *with* $p_n = o(n)$ *and* $p_n/n < 1/2$ *as* $p_n, n \to \infty$. *Denote by* $\delta^o$ *the oracle threshold* (5.8) *and assume that the nonzero entries of* $\boldsymbol{\beta}_0$ *satisfy* $|\beta_{0i}| > \delta^o$, *then*

$$\lim_{n\to\infty} \mathsf{E}_{\boldsymbol{\beta}_0} \mathsf{P}\big(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 > M p_n \log(n/p_n) | \boldsymbol{y}^{(n)}\big) = 0$$

*for some constant* $M > 0$.

PROOF. See Appendix A.2.3. □

Theorem 5.4 shows that the M-SSL prior mimics the performance of the ideal limiting prior (1.2) over $l_0[p_n; n]$ sparsity class, restricted by a mild "beta-min condition". With a prior on $\theta$, the M-SSL prior achieves the desired adaptivity, attaining the minimax rate without assuming $p_n$ is known.

**6. Simulation study.** We assess the performance of the SS-LASSO estimators relative to its two limiting cases: (a) the LASSO estimator and (b) estimators under point-mass mixture priors. We simulated observations from $Y_i \sim \mathcal{N}(\beta_{0i}, 1)$ for $i = 1, \ldots, n = 500$, assuming $\beta_{0i}$ are zero except for $p_n$ chosen positions, where $\beta_{0i} = b_0 \neq 0$. Following Castillo and van der Vaart [10], we consider various degrees of sparsity $p_n \in \{25, 50, 100\}$ and various degrees of the signal strength $b_0 \in \{3, 4, 5\}$.

We consider three classes of estimators. The first are the oracle estimators assuming that $p_n$ *is known*: (a) the oracle hard and LASSO thresholding rules (with the selection threshold $\sqrt{2\log n/p_n}$), (b) the oracle SS-LASSO estimators from Corollary 4.1 assuming two choices of $\alpha \geq 0$ and $\nu > 0$ as well as two different values $\lambda_1 < e^{-2}$. Regarding the choice of $(\alpha, \nu)$, we consider $\alpha = \nu = 1/2$ so that $1/p^\star(0) = 1 + n/p_n$. It is worth noting that with $\alpha = 0$ and $\nu = 1$ we obtained nearly identical results. The second category are *nonadaptive estimators*: (a) the universal hard and LASSO thresholding rules (with the selection threshold $\sqrt{2\log n}$) and (b) nonadaptive SS-LASSO estimators from Corollary 4.2 with $\alpha = \nu = 1/2$. The third class includes estimators *intended to be adaptive to* $p_n$: (a) empirical Bayes median estimator under point-mass mixture priors [21] with Cauchy and Laplace tails, (b) the SLOPE estimator [33] with $q = 0.1$ and (c) the two-step SS-LASSO estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{TS}}$ from Theorem 4.3 with $\alpha = \nu = 1/2$.

Table 1 reports the empirical average estimates of the mean squared error $\mathsf{E}_{\boldsymbol{\beta}_0}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ from 100 independently generated data vectors. Within each of

TABLE 1

*Average square errors computed on* 100 *data vectors of length* $n = 500$ *with* $p_n$ *of the signal values set equal to a nonzero value* $b_0$, *and the remainder zero. The estimators are HTO*: *oracle hard thresholding*; *HTU*: *universal hard thresholding*; *LASSO*: *soft thresholding*; *SSL*: *Spike-and-Slab LASSO*; *EB Laplace/EB Cauchy*: *empirical Bayes posterior median as in Johnstone and Silverman* [21] *with Laplace/Cauchy tails*; *SLOPE*: *slope estimator* [33] *with* $q = 0.1$; *SSL* 2*step*: *two-step SSL estimator as in Theorem* 4.3; *Top performance within each of the three blocks for each of the nine settings is in bold font*

| | | | | $p_n = 25$ | | | $p_n = 50$ | | | $p_n = 100$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Signal $b_0$** | $\lambda_1$ | $\lambda_0$ | $\frac{1-\theta}{\theta}$ | **3** | **4** | **5** | **3** | **4** | **5** | **3** | **4** | **5** |
| As if $p_n$ were known | | | | | | | | | | | | |
| HTO | | | | 131 | 99 | 78 | 207 | 158 | 143 | 316 | 257 | 244 |
| LASSO | | | | 144 | 172 | 175 | 247 | 276 | 282 | 397 | 428 | 437 |
| SSL | 0.1 | $\sqrt{\frac{n}{p_n}}$ | $\sqrt{\frac{n}{p_n}} \times \lambda_1$ | **124** | **87** | **64** | **195** | **139** | **117** | **288** | **218** | **200** |
| SSL | 0.01 | $\sqrt{\frac{n}{p_n}}$ | $\sqrt{\frac{n}{p_n}} \times \lambda_1$ | 130 | 96 | 74 | 203 | 152 | 134 | 303 | 237 | 221 |
| Does not adapt to $p_n$ | | | | | | | | | | | | |
| HTU | | | | **171** | **143** | **65** | **342** | **275** | **123** | **683** | **563** | **262** |
| LASSO | | | | 201 | 290 | 326 | 403 | 572 | 652 | 810 | 1157 | 1318 |
| SSL | 0.1 | $\sqrt{n}$ | $\sqrt{n} \times \lambda_1$ | 176 | 155 | 72 | 353 | 301 | 140 | 704 | 618 | 292 |
| SSL | 0.01 | $\sqrt{n}$ | $\sqrt{n} \times \lambda_1$ | 172 | 145 | 66 | 344 | 278 | 125 | 685 | 568 | 264 |
| Adapts to $p_n$ | | | | | | | | | | | | |
| EB Laplace | | | | 144 | 95 | **54** | 275 | 167 | **96** | 621 | 371 | 208 |
| EB Cauchy | | | | 156 | 110 | 58 | 307 | 197 | 106 | 696 | 461 | 250 |
| SLOPE | | | | 190 | 241 | 251 | 355 | 422 | 440 | 650 | 746 | 769 |
| SSL 2step | 0.1 | $\sqrt{\frac{n}{p+1}}$ | $\sqrt{\frac{n}{p+1}} \times \lambda_1$ | 139 | **90** | 64 | 236 | **134** | 114 | 373 | **206** | **195** |
| SSL 2step | 0.01 | $\sqrt{\frac{n}{p+1}}$ | $\sqrt{\frac{n}{p+1}} \times \lambda_1$ | **134** | 92 | 74 | **227** | 143 | 131 | **350** | 219 | 216 |

the three blocks of estimators, we highlight the top performance in bold font for each of the nine considered settings.

Among the oracle estimators, SS-LASSO with $\alpha = \nu = 1/2$ appears to perform better than oracle hard-thresholding (HTO), especially when $\lambda_1 = 0.1$. While the selection threshold for the LASSO is very similar when $\alpha = \nu = 1/2$, there are dramatic differences in terms of performance. This is clearly a consequence of the segregation ability of the SS-LASSO estimator. Among the nonadaptive estimators, the universal hard-thresholding rule (HTU) performs best. However, SS-LASSO with $\lambda_1 = 0.01$ is very similar. Again, we observe marked differences between SS-LASSO and LASSO, despite the similarity of their selection thresholds. Among the adaptive estimators, empirical Bayes (EB) median estimators perform very well, especially for larger signals ($b_0 = 5$). SS-LASSO seems to outperform EB estimates when the signal is small (for both $b_0 = 3$ and $b_0 = 4$). It is worthwhile to note that, for large enough signals, the two step SS-LASSO estimator essentially mimics the performance of the oracle SS-LASSO estimator (under similar hyperparameter choices). This finding is consistent with Theorem 4.3. While SLOPE controls very well for false discoveries, it induces bias by shrinking more the large effects. SS-LASSO, on the other hand, shrinks the large effects less, yielding a smaller recovery error.

We also compared our procedures in terms of their subset recovery ability. We report the average Hamming distance (false positives plus false negatives) between the true and estimated subsets of nonzero coefficients in Table 2. The oracle estimators seem to perform best overall when the signal is small whereas universal thresholding rules dominate when the signal is large. The two step estimator is again seen to perform similarly as the oracle estimators (improving upon them when $b_0 = 4$). For subset recovery, SLOPE is now competitive across all the scenarios, yielding improved performance as the signal gets stronger.

**7. Discussion.** In this paper, we have provided a unifying perspective on Bayesian estimation of sparse signals with continuous spike-and-slab priors, combining penalized likelihood perspectives and fully Bayes perspectives. For our proposed class of SS-LASSO priors, we provided rigorous frequentist analysis of the entire posterior distribution and its modes. These results provide valuable theoretical evidence supporting the intuitive appeal of SS-LASSO priors.

The SS-LASSO priors are especially appealing due to their implementation potential. As shown by Rockova and George [32], by deploying a sequence of SS-LASSO priors, one can dynamically explore the posterior in a manner analogous to the LASSO method. This type of deployment greatly enhances the practical value of this prior, freeing its implementation from the confinement to posterior simulation.

TABLE 2

*Average Hamming distance between true and estimated support computed on* 100 *data vectors of length* $n = 500$ *with* $p_n$ *of the signal values set equal to a nonzero value* $b_0$, *and the remainder zero. The estimators are HTO*: *oracle hard thresholding*; *HTU*: *universal hard thresholding*; *LASSO*: *soft thresholding*; *SSL*: *Spike-and-Slab LASSO*; *EB Laplace/EB Cauchy*: *empirical Bayes posterior median as in Johnstone and Silverman* [21] *with Laplace/Cauchy tails*; *SLOPE*: *slope estimator* [33] *with* $q = 0.1$; *SSL 2step*: *two-step SSL estimator as in Theorem 4.3*; *Top performance within each of the three blocks for each of the nine settings is in bold font*

| | | | | $p_n = 25$ | | | $p_n = 50$ | | | $p_n = 100$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Signal $b_0$** | $\lambda_1$ | $\lambda_0$ | $\frac{1-\theta}{\theta}$ | **3** | **4** | **5** | **3** | **4** | **5** | **3** | **4** | **5** |
| As if $p_n$ were known | | | | | | | | | | | | |
| HTO | | | | 14 | 8 | 7 | 24 | 16 | 14 | 42 | 31 | 28 |
| LASSO | | | | 14 | 8 | 7 | 24 | 16 | 14 | 42 | 31 | 28 |
| SSL | 0.1 | $\sqrt{\frac{n}{p_n}}$ | $\sqrt{\frac{n}{p_n}} \times \lambda_1$ | **13** | **7** | **5** | **22** | **12** | **10** | **37** | **24** | **22** |
| SSL | 0.01 | $\sqrt{\frac{n}{p_n}}$ | $\sqrt{\frac{n}{p_n}} \times \lambda_1$ | 14 | 8 | 6 | 23 | 14 | 13 | 39 | 28 | 25 |
| Does not adapt to $p_n$ | | | | | | | | | | | | |
| HTU | | | | **18** | **8** | **2** | **35** | **16** | **4** | **71** | **32** | **8** |
| LASSO | | | | **18** | **8** | **2** | **35** | **16** | **4** | **71** | **32** | **8** |
| SSL | 0.1 | $\sqrt{n}$ | $\sqrt{n} \times \lambda_1$ | **18** | 9 | **2** | 37 | 17 | **4** | 74 | 36 | 9 |
| SSL | 0.01 | $\sqrt{n}$ | $\sqrt{n} \times \lambda_1$ | **18** | **8** | **2** | **35** | **16** | **4** | **71** | **32** | **8** |
| Adapts to $p_n$ | | | | | | | | | | | | |
| EB Laplace | | | | 16 | **6** | 3 | 28 | 9 | 5 | 64 | 14 | **6** |
| EB Cauchy | | | | 17 | **6** | **2** | 32 | 9 | **4** | 72 | 16 | **6** |
| SLOPE | | | | 14 | **6** | 3 | **23** | **8** | 5 | 35 | **13** | 9 |
| SSL 2step | 0.1 | $\sqrt{\frac{n}{\hat{p}+1}}$ | $\sqrt{\frac{n}{\hat{p}+1}} \times \lambda_1$ | 14 | **6** | 5 | 24 | 9 | 9 | 37 | 15 | 19 |
| SSL 2step | 0.01 | $\sqrt{\frac{n}{\hat{p}+1}}$ | $\sqrt{\frac{n}{\hat{p}+1}} \times \lambda_1$ | **13** | 7 | 6 | **23** | 11 | 12 | **34** | 19 | 23 |

## APPENDIX

**A.1. Proofs of Section 5.** We use the technique of Castillo and van der Vaart [10] (CvdV12) and Bhattacharya et al. [3] (BPPD14). Throughout this section, we will be using the following notation. For a set $S \subset \{1, \ldots, n\}$, $\boldsymbol{\beta}_S$ denotes the $|S|$-dimensional subvector of $\boldsymbol{\beta} \in \mathbb{R}^n$ comprised of entries in $S$. By $\delta$, we will simply denote the intersection point $\delta(\lambda_0, \theta)$ in (3.12). Let $\pi_S(\boldsymbol{\beta}) = \prod_{i \in S} \pi(\beta_i|\theta)$, where $\pi(\beta|\theta)$ is the SSL$(\lambda_0; \lambda_1; \theta)$ prior density (3.1).

A.1.1. *Proof of Theorem 5.1.* We will need the following lemma.

LEMMA A.1. *Assume that* $\beta \sim \text{SSL}(\lambda_0; \lambda_1; \theta)$ *with a fixed* $\theta \in (0, 1)$. *For* $r > 1$, *we have*

$$\mathsf{P}\left(|\beta| \le \frac{r}{\sqrt{n}}\right) > (1-\theta)\left(1 - \frac{1}{1 + \lambda_0^2 r^2/n}\right).$$

PROOF. We have

$$\begin{aligned}
&\mathsf{P}\Big(|\beta| \le \frac{r}{\sqrt{n}}\Big) \\
&\quad = 1 - \Big[\theta \exp\Big(-\frac{r}{\sqrt{n}}\lambda_1\Big) + (1-\theta)\exp\Big(-\frac{r}{\sqrt{n}}\lambda_0\Big)\Big] \\
&\quad > (1-\theta)\big[1 - \exp(-\lambda_0 r/\sqrt{n})\big].
\end{aligned}$$

Using $\exp(-x) < \frac{1}{1+x^2}$ for $x > 0$, we obtain for $r > 1$

$$\mathsf{P}\Big(|\beta| \le \frac{r}{\sqrt{n}}\Big) > (1-\theta)\Big(1 - \frac{1}{1+\lambda_0^2 r^2/n}\Big). \qquad \square$$

For $\boldsymbol{\beta} \in \mathbb{R}^n$, let $f_{\boldsymbol{\beta}}(\cdot)$ denote the probability density function of a $\mathcal{N}(\boldsymbol{\beta}, \mathrm{I}_n)$ distribution and $f_{\beta_i}(\cdot)$ denote the univariate marginal $\mathcal{N}(\beta_i, 1)$. Let $S_0 = \{1 \le i \le n : \beta_{0i} \neq 0\}$ and $S_0^c$ be its complement. Since $|S_0| = p_n$, it suffices to show that

$$\lim_{n\to\infty} \mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}\Big(\sum_{i=1}^n \gamma(\beta_i)\mathbb{I}(i \in S_0^c) > C p_n \Big| \boldsymbol{y}^{(n)}\Big) \to 0.$$

Let $B_n = \{\sum_{i=1}^n \gamma(\beta_i)\mathbb{I}(i \in S_0^c) > C p_n\}$. Following CvdV12, we can write

$$\mathsf{P}\big(B_n|\boldsymbol{y}^{(n)}\big) = \frac{\int_{B_n} \prod_{i\in S_0^c} \frac{f_{\beta_i}(y_i)}{f_{\beta_0}(y_i)}\,\mathrm{d}\pi(\beta_i)}{\int \prod_{i\in S_0^c} \frac{f_{\beta_i}(y_i)}{f_{\beta_0}(y_i)}\,\mathrm{d}\pi(\beta_i)} \equiv \frac{N_n}{D_n}. \tag{A.1}$$

Note that

$$\mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}\big(B_n|\boldsymbol{y}^{(n)}\big) \le \mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}\big(B_n|\boldsymbol{y}^{(n)}\big)\mathbb{I}_{A_n} + \mathsf{P}_{\boldsymbol{\beta}_0}(A_n^c), \tag{A.2}$$

where $A_n = \{D_n \ge \mathrm{e}^{-r_n^2}\mathsf{P}(\|\boldsymbol{\beta}_{S_0^c}\| \le r_n)\}$, where $r_n$ is a sequence of positive real numbers. According to Lemma 5.2 of CvdV12, we have $\mathsf{P}_{\boldsymbol{\beta}_0}(A_n^c) \le \mathrm{e}^{-r_n^2}$. From (A.1), we obtain

$$\mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}\big(B_n|\boldsymbol{y}^{(n)}\big) \le \frac{\mathsf{P}(B_n)}{\mathrm{e}^{-r_n^2}\mathsf{P}(\|\boldsymbol{\beta}_{S_0^c}\| \le r_n)} + \mathrm{e}^{-r_n^2}.$$

Now, we have $\mathsf{P}(\|\boldsymbol{\beta}_{S_0^c}\| \le r_n) \ge \mathsf{P}(|\beta_1| < r_n/\sqrt{n})^{n-p_n}$. Choosing $r_n^2 = p_n$ and assuming $\lambda_0 > n/p_n$, Lemma A.1 yields

$$\mathsf{P}\big(\|\boldsymbol{\beta}_{S_0^c}\| \le r_n\big) > (1-\theta)^{n-p_n}\Big(1 - \frac{1}{1+\lambda_0^2 r_n^2/n}\Big)^{n-p_n} \tag{A.3}$$

$$> \Big(1 - \frac{Ap_n}{n}\Big)^n \Big(1 - \frac{1}{n}\Big)^n \ge \Big(\frac{1}{2\mathrm{e}}\Big)^{Ap_n+1}. \tag{A.4}$$

The last inequality follows from $(1-x)^{1/x} > 1/(2\mathrm{e})$ for $0 < x < 0.5$. From Lemma 5.2, we have $\mathsf{P}(B_n) \le \exp(-Cp_n)$ for $C > 2A\mathrm{e}$. With $r_n^2 = p_n$, we have

$$\mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}\big(B_n|\boldsymbol{y}^{(n)}\big) \le 2\mathrm{e}^{-p_n[C-1-A\log(2\mathrm{e})]+1} + \mathrm{e}^{-p_n}. \tag{A.5}$$

Because $C > 2A\mathrm{e} > 1 + A\log(2\mathrm{e})$ for $A > 0.5$, (A.5) yields $\mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}(B_n|\boldsymbol{y}^{(n)}) \to 0$.

A.1.2. *Proof of Lemma 5.3.*

PROOF. Let us assume w.l.o.g. $|x| > |y|$, then

$$\big|h(x) - h(y)\big| \le \left|\log\frac{\psi_1(x)}{\psi_1(y)}\right| + \left|\log\frac{(1-\theta)}{\theta}\frac{\psi_0(x)}{\psi_0(y)}\right|.$$

This yields

$$\big|h(x) - h(y)\big| \le \lambda_1|x-y| + \log\left[\frac{(1-\theta)\lambda_0}{\theta\lambda_1}\exp\big(-|x|\lambda_0 + |y|\lambda_1\big)\right] \tag{A.6}$$

$$\le \lambda_1|x-y| + \log\left[\frac{(1-\theta)\lambda_0}{\theta\lambda_1}\exp\big[-|x|(\lambda_0-\lambda_1)\big]\right] \tag{A.7}$$

$$\le \lambda_1|x-y| + \log\left[\frac{(1-\theta)\lambda_0}{\theta\lambda_1}\right]. \tag{A.8}$$

□

A.1.3. *Proof of Theorem 5.2.* We will need the following two lemmata. In the sequel, we denote $r_n^2 = p_n\log[1/p^\star(0) - 1]$.

LEMMA A.2. *Let $S, S_0 \subset \{1,\dots,n\}$ and $\boldsymbol{\beta}', \boldsymbol{\beta} \in \mathbb{R}^n$. Assume $\lambda_0 > 2$, $0 < \theta < 1/2$ and $0 < \lambda_1 < \mathrm{e}^{-2}$. Assume $\gamma(\beta_i') = 1$, $i \in S$. Then*

$$\log\left[\frac{\pi_S(\boldsymbol{\beta}')}{\pi_{S_0}(\boldsymbol{\beta})}\right] \le \lambda_1\big|\boldsymbol{\beta}_{S_0\cap S} - \boldsymbol{\beta}'_{S_0\cap S}\big|_1 + \lambda_1|\boldsymbol{\beta}_{S_0\setminus S}|_1 + |S_0|\left[2\log\left(\frac{1}{p^\star(0)} - 1\right) + \log\left(\frac{1}{2}\right)\right].$$

PROOF. Similarly as CvdV12, let us decompose

$$\frac{\pi_S(\boldsymbol{\beta}')}{\pi_{S_0}(\boldsymbol{\beta})} = \frac{\pi_S(\boldsymbol{\beta}')}{\pi_{S\cap S_0}(\boldsymbol{\beta}')}\frac{\pi_{S\cap S_0}(\boldsymbol{\beta}')}{\pi_{S\cap S_0}(\boldsymbol{\beta})}\frac{\pi_{S\cap S_0}(\boldsymbol{\beta})}{\pi_{S_0}(\boldsymbol{\beta})}.$$

Denote by $h(\beta) = \log\pi(\beta|\theta)$. From the assumption $|\beta_i'| \ge \delta$, $i \in S$, and using the fact $(1-\theta)\psi_0(\delta) = \theta\psi_1(\delta)$, we obtain $h(\beta_i') \le h(\delta) = \log[\theta\lambda_1\exp(-\lambda_1\delta)] < \log(\theta\lambda_1)$ for $i \in S$. This yields

$$\log\frac{\pi_S(\boldsymbol{\beta}')}{\pi_{S\cap S_0}(\boldsymbol{\beta}')} = \sum_{i\in S\setminus S_0}\log h\big(\beta_i'\big) < |S\setminus S_0|\log[\theta\lambda_1] < 0. \tag{A.9}$$

Denote by $C_n = \log[1/p^\star(0) - 1]$. Then, using Lemma 5.3, we obtain

$$\begin{aligned}\left|\log\frac{\pi_{S\cap S_0}(\boldsymbol{\beta}')}{\pi_{S\cap S_0}(\boldsymbol{\beta})}\right| &\le \sum_{i\in S\cap S_0}\left|h(\beta_i') - h(\beta_i)\right| \\ &< \sum_{i\in S\cap S_0}\left[\lambda_1|\beta_i' - \beta_i| + C_n\right] = \lambda_1|\boldsymbol{\beta}_{S\cap S_0} - \boldsymbol{\beta}'_{S\cap S_0}|_1 + |S\cap S_0|C_n.\end{aligned}$$

Finally, using the fact $|h(0)| < \log(\lambda_0/2)$ for $\lambda_0 > 2$ and $\lambda_1 \le \mathrm{e}^{-2}$, we obtain

$$\left|\log\frac{\pi_{S_0}(\boldsymbol{\beta})}{\pi_{S_0\cap S}(\boldsymbol{\beta})}\right| \le \sum_{i\in S_0\setminus S}\left|h(\beta_i) - h(0)\right| + |S_0\setminus S|\log(\lambda_0/2) \tag{A.10}$$

$$\le \lambda_1|\boldsymbol{\beta}_{S_0\setminus S}|_1 + |S_0\setminus S|\big[C_n + \log(\lambda_0/2)\big]. \tag{A.11}$$

Altogether, because $\log(\lambda_0/2) < C_n + \log(1/2)$

$$\log\frac{\pi_S(\boldsymbol{\beta}')}{\pi_{S_0}(\boldsymbol{\beta})} \le \lambda_1|\boldsymbol{\beta}_{S_0\cap S} - \boldsymbol{\beta}'_{S_0\cap S}|_1 + \lambda_1|\boldsymbol{\beta}_{S_0\setminus S}|_1 + |S_0|\big[2C_n + \log(1/2)\big]. \qquad \square$$

LEMMA A.3. *Let $S, S_0 \subset \{1,\dots,n\}$, $|S_0| = p_n$ and $j \ge 1$. Assume a prior $\boldsymbol{\beta} \sim \mathrm{SSL}(\lambda_0;\lambda_1;\theta)$ with $\lambda_0 \ge n/p_n$, $\lambda_1 < \mathrm{e}^{-2}$ and $\theta \le p_n/n$. Let $\widetilde{\boldsymbol{\beta}}^{S,j} \in \mathbb{R}^n$ satisfy: $\widetilde{\beta}_i^{S,j} = 0$ for $i \notin S$; $\gamma(\widetilde{\beta}_i^{S,j}) = 1$ for $i \in S$; and $\|\widetilde{\boldsymbol{\beta}}^{S,j} - \boldsymbol{\beta}_0\| < 2(j+1)r_n$. Denote by*

$$\pi\big(\widetilde{\boldsymbol{\beta}}^{S;j};S;j\big) = \frac{\mathsf{P}(\boldsymbol{\beta}\in\mathbb{R}^n : \gamma(\beta_i) = \mathbb{I}(i\in S), \|\boldsymbol{\beta}_S - \widetilde{\boldsymbol{\beta}}_S^{S,j}\| < 2jr_n)}{\mathrm{e}^{-4r_n^2}\mathsf{P}(\boldsymbol{\beta}\in\mathbb{R}^n : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < 2r_n)}. \tag{A.12}$$

*Then*

$$\log\pi\big(\widetilde{\boldsymbol{\beta}}^{S;j};S;j\big) \le 3 + |S|(3+\log j) + r_n^2\big[9 + \lambda_1(4j+3+\delta)\big].$$

PROOF. Denote by $N_n$ and $D_n$ the numerator and the denominator of (A.12). The numerator can be upper-bounded as follows:

$$N_n \le |v_S(2jr_n)|\mathsf{P}\big[\gamma(\beta_1) = 0\big]^{n-|S|}\sup_{\boldsymbol{\beta}\in\mathcal{A}}\pi_S(\boldsymbol{\beta}),$$

where $\mathcal{A} = \{\boldsymbol{\beta}\in\mathbb{R}^n : \gamma(\beta_i) = 1, i\in S; \|\boldsymbol{\beta}_S - \widetilde{\boldsymbol{\beta}}_S^{S,j}\| < 2jr_n\}$, $v_S(r)$ denotes the $|S|$-dimensional ball centered at zero with a radius $r$ and $|v_S(r)|$ is its volume. To bound the denominator $D_n$ of (A.12), we begin with the inequality

$$\mathsf{P}\big(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < 2r_n\big) > \mathsf{P}\big(\|\boldsymbol{\beta}_{S_0^c}\| < r_n\big)\mathsf{P}\big(\|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}\| < r_n\big). \tag{A.13}$$

As in the proof of Theorem 5.1, we have $\mathsf{P}(\|\boldsymbol{\beta}_{S_0^c}\| < r_n) > (\frac{1}{2\mathrm{e}})^{Ap_n+1}$ with $A = 1$. The denominator can be thus lower-bounded as follows:

$$D_n > \mathrm{e}^{-4r_n^2-(p_n+1)\log(2\mathrm{e})}\big|v_{S_0}(r_n)\big| \inf_{\boldsymbol{\beta}\in\mathcal{B}} \pi_{S_0}(\boldsymbol{\beta}),$$

where $\mathcal{B} = \{\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}\| < r_n\}$. Assume $\boldsymbol{\beta}' \in \mathcal{A}$ and $\boldsymbol{\beta} \in \mathcal{B}$, then invoking Lemma A.2 we obtain

$$\frac{\pi_S(\boldsymbol{\beta}')}{\pi_{S_0}(\boldsymbol{\beta})} \le \exp\{\lambda_1\sqrt{|S_0|}(\|\boldsymbol{\beta}'_{S_0\cap S} - \boldsymbol{\beta}_{S_0\cap S}\| + \|\boldsymbol{\beta}_{S_0\setminus S}\|) + |S_0|[2C_n + \log(1/2)]\},$$

where $C_n = \log[1/p^\star(0) - 1]$. Denote by $S_1 = S_0 \cap S$, then by expanding and splitting both norms we obtain

$$\begin{aligned}
&\|\boldsymbol{\beta}'_{S_1} - \boldsymbol{\beta}_{S_1}\| + \|\boldsymbol{\beta}_{S_0\setminus S}\| \\
&\quad\le \|\boldsymbol{\beta}'_{S_1} - \widetilde{\boldsymbol{\beta}}^{S,j}_{S_1}\| + \|\widetilde{\boldsymbol{\beta}}^{S,j}_{S_0} - \boldsymbol{\beta}_{0S_0}\| + \|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}\| + \|\widetilde{\boldsymbol{\beta}}^{S,j}_{S_0\setminus S}\| \\
&\quad\le 2jr_n + 2(j+1)r_n + r_n + \sqrt{|S_0|}\delta = 4jr_n + 3r_n + \sqrt{|S_0|}\delta.
\end{aligned}$$

Using the facts $p_n \le r_n^2$ and $p_n C_n = r_n^2$, we obtain

$$\pi(\widetilde{\boldsymbol{\beta}}^{S;j}; S; j) \le \frac{|v_S(2jr_n)|}{|v_{S_0}(r_n)|}\mathrm{e}^{r_n^2[7+\lambda_1(4j+3+\delta)]+\log(2\mathrm{e})}.$$

Let $v_S = v_S(1)$, then $|v_S(r)| = v_S r^{|S|}$. According to Lemma 5.4 of CvdV12, we have

$$\frac{r_n^{|S|}v_S}{r_n^{|S_0|}v_{S_0}} \le \mathrm{e}^{1/6}(2\pi\mathrm{e})^{(|S|-|S_0|)/2}\left(\frac{r_n}{\sqrt{|S|}}\right)^{|S|}\left(\frac{\sqrt{S_0}}{r_n}\right)^{|S_0|}\frac{\sqrt{|S_0|}}{\sqrt{|S|}}.$$

With $\sqrt{|S_0|} \le r_n$, the last display can be bounded from above by

$$\begin{aligned}
&\mathrm{e}^{1/6}(2\pi\mathrm{e})^{|S|/2}\left(\frac{r_n}{\sqrt{|S|}}\right)^{|S|} r_n \\
&\quad\le \exp\left[\frac{1}{6} + \frac{|S|}{2}\log(2\pi\mathrm{e}) + \log r_n + \frac{r_n^2}{2\mathrm{e}}\right] \\
&\quad\le \exp\big(1 + 2|S| + 2r_n^2\big).
\end{aligned}$$

Altogether, we obtain

$$\log\pi(\widetilde{\boldsymbol{\beta}}^{S;j}; S; j) \le 3 + |S|(3 + \log j) + r_n^2[9 + \lambda_1(4j + 3 + \delta)]. \qquad \square$$

Now we embark on the proof of Theorem 5.2. In view of Theorem 5.1, it suffices to work with

$$\mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}\big(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| > Mr_n, |\boldsymbol{\gamma}(\boldsymbol{\beta})| \le Cp_n|\boldsymbol{y}^{(n)}\big). \tag{A.14}$$

Let $\mathcal{S}$ be the collection of subsets $S \subset \{1, \ldots, n\}$ such that $|S| \leq Cp_n$. For each such $S$ and $j \in \mathcal{N}$, we denote by

$$\mathcal{B}_{S,j} = \{\boldsymbol{\beta} \in \mathbb{R}^n : \boldsymbol{\gamma}(\beta_i) = \mathbb{I}(i \in S); 2jr_n \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq 2(j+1)r_n\} \tag{A.15}$$

the $j$th shell of vectors with an "effective support" $S$. Let $\{\widetilde{\boldsymbol{\beta}}_k^{S,j} : k \in I_{S,j}\}$ be a $2jr_n$-isolated set of $\mathcal{B}_{S,j}$, constructed similarly as in BPPD14. First, we take a $jr_n$-separated net inside the set $\{\boldsymbol{\beta} \in \mathbb{R}^n : \beta_i = 0, i \notin S; \gamma(\beta_i) = 1, i \in S : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < 2(j+1)r_n\}$. This net can be chosen so that $|I_{S,j}| \leq \mathrm{e}^{D|S|}$, where $D > 1$. This set appears to be a $2jr_n$-separated net of $\mathcal{B}_{S,j}$ for $j \geq M$. This is because $\forall \boldsymbol{\beta} \in \mathcal{B}_{S,j}$ we have

$$\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}_k^{S,j}\|^2 < j^2 r_n^2 + (n - p_n)\delta^2.$$

Note that $\delta = 1/(\lambda_0 - \lambda_1)\log[1/p^\star(0) - 1] < 1 + 1/\lambda_1 < 1 + C_{\lambda_1}$. Because $r_n^2 = p_n \log[1/p^\star(0) - 1]$ and $\lambda_0 \geq n/p_n$ we have

$$\frac{n - p_n}{(\lambda_0 - \lambda_1)^2} \log^2\big[1/p^\star(0) - 1\big] < \delta r_n^2.$$

Assuming $M > \sqrt{(1 + C_{\lambda_1})/3}$, we have $\delta < 3j^2$ for $j \geq M$ and thereby $\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}_k^{S,j}\|^2 < 4j^2 r_n^2$. Therefore, $\{\widetilde{\boldsymbol{\beta}}_k^{S,j} : k \in I_{S,j}\}$ is a $2jr_n$-separated net of $\mathcal{B}_{S,j}$. Thus, each shell $\mathcal{B}_{S,j}$ can be covered with balls $B_k^{S,j}$ of radius $2jr_n$ centered at $\widetilde{\boldsymbol{\beta}}_k^{S,j}$.

To bound the implicit denominator in (A.14), we confine attention to sets $\mathcal{A}_n$ (as in Lemma 5.2 of CvdV12) on which

$$\int \prod_{i=1}^n \frac{f_{\beta_i}(y_i)}{f_{\beta_{0i}}(y_i)}\, \mathrm{d}\pi(\boldsymbol{\beta}) \geq \mathrm{e}^{-4r_n^2} \mathsf{P}\big(\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < 2r_n\big). \tag{A.16}$$

By Lemma 5.2 of CvdV12 $\mathsf{P}_{\boldsymbol{\beta}_0}(\mathcal{A}_n) \geq 1 - \exp(-r_n^2/2)$. Applying the test argument to each point-versus-ball (CvdV12, Proposition 5.1), we obtain

$$\begin{aligned} &\mathsf{E}_{\boldsymbol{\beta}_0} \mathsf{P}\big(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| > 2Mr_n, |\boldsymbol{\gamma}(\boldsymbol{\beta})| = S | \boldsymbol{y}^{(n)}\big) \mathbb{I}_{\mathcal{A}_n} \\ &\quad \leq \sum_{S \in \mathcal{S}} \sum_{j \geq M} \sum_{k \in I_{S,j}} 2\sqrt{\pi_k^{S,j}}\, \mathrm{e}^{-j^2 r_n^2/2}, \end{aligned} \tag{A.17}$$

where

$$\pi_k^{S,j} = \frac{\mathsf{P}(B_k^{S,j})}{\mathrm{e}^{-4r_n^2} \mathsf{P}(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < 2r_n)}.$$

Using Lemma A.3 and assuming $|S| < Cp_n$, we can write

$$\log \sqrt{\pi_k^{S,j}} \leq 3/2 + 0.5Cp_n(3 + \log j) + 0.5r_n^2\big[9 + \lambda_1(4j + 3 + \delta)\big] < j^2 r_n^2/4$$

for $j > M$, where $M$ is large enough. Then (A.17) can be (for $Cp_n/n < 1/2$) bounded from above by

$$\sum_{p=0}^{\lfloor Cp_n \rfloor} \binom{n}{p} \sum_{j \geq M} 2\mathrm{e}^{Dp - j^2 r_n^2/4} < 2Cp_n \mathrm{e}^{DCp_n - M^2 r_n^2/4 + \tilde{D} r_n^2}, \tag{A.18}$$

where we used the inequality $\binom{n}{p} \leq \mathrm{e}^{Cp_n \log[(n\mathrm{e})/(Cp_n)]} < \mathrm{e}^{\tilde{D} r_n^2}$ for $Cp_n/n < 1/2$ (an assumption satisfied for $n$ large enough). Thus, (A.18) goes to zero as $n \to \infty$, which completes the proof.

## A.2. Proofs of Section 5.3.

A.2.1. *Proof of Theorem* 5.3. We will need the following two lemmata.

LEMMA A.4. *Assume* $\beta \sim \mathrm{SSL}(\lambda_1; \lambda_0; \theta)$ *with* $\theta \in (0, 1)$, $\lambda_1 \leq \mathrm{e}^{-2}$ *and* $\lambda_0 = (1 - \theta)/\theta$. *Let* $\theta^o$ *and* $\delta^o$ *be as in* (5.7) *and* (5.8). *Assume* $\theta_o < 1/2$, *then*

$$\mathsf{P}\big[|\beta| > \delta^o | \theta\big] \leq \theta^o \qquad \forall \theta \leq \theta^o.$$

PROOF. The function $x \to \mathsf{P}(|\beta| > y|x)$ is increasing on $(0, 1)$ when $\lambda_0 > \lambda_1$, because $\mathsf{P}(|\beta| > y|x) = x \exp(-y\lambda_1) + (1 - x)\exp(-y\lambda_0)$ has a positive derivative $\forall y \in \mathbb{R}^+$. Invoking Lemma 5.1 we obtain

$$\mathsf{P}\big[|\beta| > \delta^o | \theta\big] \leq \mathsf{P}\big[|\beta| > \delta^o | \theta^o\big] \leq \theta^o. \qquad \square$$

LEMMA A.5. *Assume* $\boldsymbol{\beta} \sim \text{M-SSL}(\lambda_1, a, b)$ *with* $\lambda_1 \leq \mathrm{e}^{-2}$, $a = 1$ *and* $b = 4n$. *Denote by* $\delta^o$ *the oracle intersection point* (5.8). *Then for any* $C > 2\mathrm{e}$, *we have*

$$\mathsf{P}\big[|\boldsymbol{\gamma}_o(\boldsymbol{\beta})| > Cp_n\big] \leq 2\exp(-2p_n).$$

PROOF. Denote by $B \equiv |\boldsymbol{\gamma}_o(\boldsymbol{\beta})|$. Then

$$\begin{aligned} \mathsf{P}(B > Cp_n) &= \int_0^1 \mathsf{P}(B > Cp_n|\theta)\,\mathrm{d}\pi(\theta) \\ &\leq \int_0^{\theta^o} \mathsf{P}(B > Cp_n|\theta)\,\mathrm{d}\pi(\theta) + \mathsf{P}\big(\theta > \theta^o\big). \end{aligned} \tag{A.19}$$

Conditionally on $\theta$, $B$ has a binomial law $\mathrm{Bin}(n, \pi_\theta)$ with $\pi_\theta = \mathsf{P}(|\beta_1| > \delta^o|\theta)$. By Corollary A.4, we can write $\pi_\theta \leq \theta^o$ for $\theta \leq \theta^o$ and apply Chernoff's inequality to bound the integrand in (A.19). For $C > 2\mathrm{e}$, we get

$$\mathsf{P}(B > Cp_n|\theta) \leq \exp\big[-p_n C \log(2)\big], \qquad 0 \leq \theta \leq \theta^o.$$

Under the assumption $\theta \sim \mathcal{B}(a, b)$ with $a = 1$ and $b = 4n$, we can write

$$\mathsf{P}(\theta > \theta_n) = (1 - \theta_n)^n = \left(1 - \frac{p_n}{p_n + n}\right)^{4n} \leq \mathrm{e}^{-2p_n}.$$

Combining the last two displays, we obtain $\mathsf{P}(B > Cp_n) \leq 2\mathrm{e}^{-2p_n}$. $\square$

Similarly as in the proof of Theorem 5.1, let $B_n = \{\sum_{i=1}^n \gamma_o(\beta_i)\mathbb{I}(i \in S_0^c) > Cp_n\}$. We can write

$$\mathsf{P}\big(B_n|\boldsymbol{y}^{(n)}\big) = \frac{\int_{B_n}\prod_{i\in S_0^c}\frac{f_{\beta_i}(y_i)}{f_{\beta_{0i}}(y_i)}\,\mathrm{d}\pi(\boldsymbol{\beta})}{\int_0^1 D_{n,\theta}\,\mathrm{d}\pi(\theta)} \equiv \frac{N_n}{D_n}, \tag{A.20}$$

where $D_{n,\theta} = \int\prod_{i\in S_0^c}\frac{f_{\beta_i}(y_i)}{f_{\beta_{0i}}(y_i)}\,\mathrm{d}\pi(\beta_i|\theta)$. Let $r_n$ be a sequence of positive real numbers and denote by $A_n = \{\int_0^1[D_{n,\theta} - \mathrm{e}^{-r_n^2}\mathsf{P}(\|\boldsymbol{\beta}_{S_0^c}\| \leq r_n|\theta)]\,\mathrm{d}\pi(\theta) \geq 0\}$. Note that

$$\mathsf{P}_{\boldsymbol{\beta}_0}(A_n) \geq \mathsf{P}_{\boldsymbol{\beta}_0}(\widetilde{A}_n) \qquad \text{where } \widetilde{A}_n = \Big\{\inf_\theta\big[D_{n,\theta} - \mathrm{e}^{-r_n^2}\mathsf{P}\big(\|\boldsymbol{\beta}_{S_0^c}\| \leq r_n|\theta\big)\big] > 0\Big\}.$$

According to Lemma 5.2 of CvdV12, we have $\mathsf{P}_{\boldsymbol{\beta}_0}(\widetilde{A}_n) \geq 1 - \mathrm{e}^{-r_n^2}$ and thereby $\mathsf{P}_{\boldsymbol{\beta}_0}(A_n^c) \leq \mathrm{e}^{-r_n^2}$. Using (A.2), we obtain

$$\mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}\big(B_n|\boldsymbol{y}^{(n)}\big) \leq \frac{\mathsf{P}(B_n)}{\mathrm{e}^{-r_n^2}\int_0^1\mathsf{P}(\|\boldsymbol{\beta}_{S_0^c}\| \leq r_n|\theta)\,\mathrm{d}\pi(\theta)} + \mathrm{e}^{-r_n^2}.$$

For $r_n^2 = p_n \geq 1$, $n \geq 2$ and $\theta \leq 1/(1+n)$, we have $\mathsf{P}(\|\boldsymbol{\beta}_{S_0^c}\| \leq r_n|\theta) \geq (1 - \frac{1}{n})^{2(n-p_n)} > (\frac{1}{2\mathrm{e}})^2$ by Lemma A.1. Thereby

$$\int_0^1 \mathsf{P}\big(\|\boldsymbol{\beta}_{S_0^c}\| \leq r_n|\theta\big)\,\mathrm{d}\pi(\theta) > \mathrm{e}^{-2\log(2\mathrm{e})}\mathsf{P}(\theta \leq 1/n) > \mathrm{e}^{-4}\big(1 - \mathrm{e}^{-4}\big) > 0.$$

The last inequality follows from $\mathsf{P}(\theta \leq 1/n) = 1 - (1 - \frac{1}{n})^{4n} > 1 - \mathrm{e}^{-4}$. From Lemma A.5, we have $\mathsf{P}(B_n) \leq 2\exp(-2p_n)$. We have $\mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}(B_n|\boldsymbol{y}^{(n)}) \preceq \mathrm{e}^{-p_n}$ and, therefore, $\mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}(B_n|\boldsymbol{y}^{(n)}) \to 0$.

A.2.2. *Proof of Lemma 5.4.* To begin, we find an upper bound to (5.9). The mapping $\theta \to \delta(\lambda_0, \theta)$, after plugging $(1-\theta)/\theta$ in for $\lambda_0$, is monotone increasing on $\theta \in [0, 1/2]$ when $\lambda_1 < \mathrm{e}^{-2}$. Because $|\beta_i| > \delta^o$, $i \in S$, we have $p_\theta^\star(\beta_i) > 1/2$ when $\theta < \theta^o < 1/2$. For $\theta > \theta^o$, we can write $p_\theta^\star(\beta_i) > p_\theta^\star(0)$ to obtain

$$\int_0^1 \theta^{|S|}\prod_{i\in S}\frac{1}{p_\theta^\star(\beta_i)}\,\mathrm{d}\pi(\theta) < 2^{|S|}\int_0^{\theta^o}\theta^{|S|}\,\mathrm{d}\pi(\theta) + \int_{\theta^o}^1\theta^{|S|}\Big[\frac{1}{p_\theta^\star(0)}\Big]^{|S|}\mathrm{d}\pi(\theta).$$

Since $\theta \to (1-\theta)/\theta$ is monotone decreasing, so is the mapping $\theta \to 1/p_\theta^\star(0)$. Therefore, we can write

$$\int_0^1 \theta^{|S|}\prod_{i\in S}\frac{1}{p_\theta^\star(\beta_i)}\,\mathrm{d}\pi(\theta) < \Big[2^{|S|} + \Big(\frac{1}{p_{\theta^o}^\star(0)}\Big)^{|S|}\Big]\int_0^1\theta^{|S|}\,\mathrm{d}\pi(\theta). \tag{A.21}$$

Next, note that

$$\pi_S(\boldsymbol{\beta}') > \left(\frac{\lambda_1}{2}\right)^{|S|} \exp(-\lambda_1 |\boldsymbol{\beta}'_S|) \int_0^1 \theta^{|S|} \,\mathrm{d}\pi(\theta). \tag{A.22}$$

Using (5.9) and (A.22), together with $[2^{|S|} + (\frac{1}{p^\star_{\theta^o}(0)})^{|S|}] < (2 + \frac{1}{p^\star_{\theta^o}(0)})^{|S|}$, we obtain

$$\frac{\pi_S(\boldsymbol{\beta}')}{\pi_S(\boldsymbol{\beta}')} < \mathrm{e}^{\lambda_1 |\boldsymbol{\beta}_S - \boldsymbol{\beta}'_S|} \left(2 + \frac{1}{p^\star_{\theta^o}(0)}\right)^{|S|}.$$

A.2.3. *Proof of Theorem 5.4.* We denote $r_n^2 = p_n \log[1/p^\star_{\theta^o}(0) + 2]$. The beginning of the proof is the same as in Theorem 5.2. While keeping $\theta = \theta^o$, we again partition the set (A.14) into shells and find a net of points. The first major difference occurs in (A.16). Here, we instead confine attention to sets $\mathcal{A}_n$, on which

$$\int_0^1 \int \prod_{i=1}^n \frac{f_{\beta_i}(y_i)}{f_{\beta_{0i}}(y_i)} \,\mathrm{d}\pi(\boldsymbol{\beta}|\theta) \,\mathrm{d}\pi(\theta)$$

$$\geq \mathrm{e}^{-4r_n^2} \int_0^1 \mathsf{P}\big(\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < 2r_n|\theta\big) \,\mathrm{d}\pi(\theta).$$

As in the proof of Theorem 5.1, we have $\mathsf{P}_{\boldsymbol{\beta}_0}(\mathcal{A}_n^c) \leq \exp(-r_n^2/2)$. Confining attention to $\mathcal{A}_n$, we obtain the following analog to (A.17):

$$\pi_k^{S,j} = \frac{\mathsf{P}(B_k^{S,j})}{\mathrm{e}^{-4r_n^2} \int_0^1 \mathsf{P}(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < 2r_n) \,\mathrm{d}\pi(\theta)}. \tag{A.23}$$

To find an upper bound to (A.23), we need two additional lemmata.

LEMMA A.6. *Assume $S, S_0 \subset \{1, \ldots, n\}$ and $\boldsymbol{\beta}', \boldsymbol{\beta} \in \mathbb{R}^n$. Let $\pi_S(\boldsymbol{\beta})$ be the marginal* M-SSL *prior* (5.9) *with $\lambda_1 < \mathrm{e}^{-2}$, $a = 1$ and $b = 4n$. Assume $|\beta'_i| > \delta^o, i \in S$, and $|\beta_i| > \delta^o, i \in S_0$. Then*

$$\log\left[\frac{\pi_S(\boldsymbol{\beta}')}{\int_{\frac{p_n}{4n}}^{\frac{p_n}{n}} \pi_{S_0}(\boldsymbol{\beta}|\theta) \,\mathrm{d}\pi(\theta)}\right]$$

$$\leq \lambda_1 |\boldsymbol{\beta}_{S_0 \cap S} - \boldsymbol{\beta}'_{S_0 \cap S}| + \lambda_1 |\boldsymbol{\beta}_{S_0 \setminus S}| + (|S_0| + 1) \log\left(\frac{2}{\lambda_1}\right)$$

$$+ (|S| + 3|S_0|)\left[\log\left(2 + \frac{1}{p^\star_{\theta^o}(0)}\right)\right] + 4|S_0|.$$

PROOF. Similarly as in the proof of Lemma A.3, let us decompose

$$\frac{\pi_S(\boldsymbol{\beta}')}{\int_{\frac{p_n}{4n}}^{\frac{p_n}{n}} \pi_{S_0}(\boldsymbol{\beta}|\theta) \,\mathrm{d}\pi(\theta)} = \frac{\pi_S(\boldsymbol{\beta}')}{\pi_{S \cap S_0}(\boldsymbol{\beta}')} \frac{\pi_{S \cap S_0}(\boldsymbol{\beta}')}{\pi_{S \cap S_0}(\boldsymbol{\beta})} \frac{\pi_{S \cap S_0}(\boldsymbol{\beta})}{\int_{\frac{p_n}{4n}}^{\frac{p_n}{n}} \pi_{S_0}(\boldsymbol{\beta}|\theta) \,\mathrm{d}\pi(\theta)}. \tag{A.24}$$

By direct application of Lemma 5.4, we obtain

$$\frac{\pi_{S\cap S_0}(\boldsymbol{\beta}')}{\pi_{S\cap S_0}(\boldsymbol{\beta})} < \mathrm{e}^{\lambda_1|\boldsymbol{\beta}_{S\cap S_0}-\boldsymbol{\beta}'_{S\cap S_0}|}\left(2+\frac{1}{p^{\star}_{\theta^o}(0)}\right)^{|S\cap S_0|}. \tag{A.25}$$

Similarly, from (5.9) and (A.22) we obtain

$$\frac{\pi_S(\boldsymbol{\beta}')}{\pi_{S\cap S_0}(\boldsymbol{\beta}')} < \left(\frac{\lambda_1}{2}\right)^{|S\setminus S_0|}\left(2+\frac{1}{p^{\star}_{\theta^o}(0)}\right)^{|S|}. \tag{A.26}$$

To find an upper bound to the last term in (A.24), we begin with

$$\begin{aligned}\int_{\frac{p_n}{4n}}^{\frac{p_n}{n}} \pi_{S_0}(\boldsymbol{\beta}|\theta)\,\mathrm{d}\pi(\theta) &> \left(\frac{\lambda_1}{2}\right)^{|S_0|}\mathrm{e}^{-\lambda_1|\boldsymbol{\beta}_{S_0}|}\int_{\frac{p_n}{4n}}^{\frac{p_n}{n}}\theta^{|S_0|}\,\mathrm{d}\pi(\theta)\\ &> \left(\frac{\lambda_1}{2}\right)^{|S_0|}\mathrm{e}^{-\lambda_1|\boldsymbol{\beta}_{S_0}|}\left(\frac{p_n}{4n}\right)^{|S_0|}\mathsf{P}\left[\theta\in\left(\frac{p_n}{4n},\frac{p_n}{n}\right)\right].\end{aligned}$$

For $p_n \geq 1$, we have $\mathsf{P}[\theta \in (\frac{p_n}{4n}, \frac{p_n}{n})] = (1-\frac{p_n}{4n})^{4n} - (1-\frac{p_n}{n})^{4n} > \mathrm{e}^{-2p_n}(1-\mathrm{e}^{-2})$. Now, we use the following two facts: (a) $|\beta_i| > \delta_n, i \in S_0$; (b) $n/p_n < 2 + 1/p^{\star}_{\theta^o}(0)$. Using (A.21), we can write

$$\frac{\pi_{S\cap S_0}(\boldsymbol{\beta})}{\int_{\frac{p_n}{4n}}^{\frac{p_n}{n}}\pi_{S_0}(\boldsymbol{\beta}|\theta)\,\mathrm{d}\pi(\theta)} < 2\left(\frac{2}{\lambda_1}\right)^{|S_0\setminus S|}\mathrm{e}^{\lambda_1|\boldsymbol{\beta}_{S_0\setminus S}|}\left(2+\frac{1}{p^{\star}_{\theta^o}(0)}\right)^{2|S_0|}\mathrm{e}^{4|S_0|}. \tag{A.27}$$

Combining the three pieces (A.25), (A.26) and (A.27), we obtain

$$\begin{aligned}&\frac{\pi_S(\boldsymbol{\beta}')}{\int_{\frac{p_n}{4n}}^{\frac{p_n}{n}}\pi_{S_0}(\boldsymbol{\beta}|\theta)\,\mathrm{d}\pi(\theta)}\\ &\quad< \mathrm{e}^{\lambda_1|\boldsymbol{\beta}_{S_0\setminus S}|+\lambda_1|\boldsymbol{\beta}_{S\cap S_0}-\boldsymbol{\beta}'_{S\cap S_0}|}\left(\frac{2}{\lambda_1}\right)^{|S_0|+1}\left(2+\frac{1}{p^{\star}_{\theta^o}(0)}\right)^{|S|+3|S_0|}\mathrm{e}^{4|S_0|}. \qquad \square\end{aligned}$$

LEMMA A.7. *Let $S, S_0 \subset \{1,\dots,n\}$ where $|S_0| = p_n$, $|S| < C|S_0|$ and $j \geq 1$, $j \in \mathcal{N}$. Assume $\boldsymbol{\beta} \sim$ M-SSL$(\lambda_1; a; b)$ with $\lambda_1 < \mathrm{e}^{-2}$, $a = 1$ and $b = 4n$. Let $\widetilde{\boldsymbol{\beta}}^{S,j} \in \mathbb{R}^n$ satisfy: $\widetilde{\beta}_i^{S,j} = 0$ for $i \notin S$; $\gamma_{\theta^o}(\widetilde{\beta}_i^{S,j}) = 1$ for $i \in S$; and $\|\widetilde{\boldsymbol{\beta}}^{S,j} - \boldsymbol{\beta}_0\| < 2(j+1)r_n$. Assume $|\beta_{0i}| > \delta^o$, $i \in S_0$. Denote by*

$$\pi\big(\widetilde{\boldsymbol{\beta}}^{S;j}; S; j\big) = \frac{\mathsf{P}(\boldsymbol{\beta}\in\mathbb{R}^n : \gamma_{\delta^o}(\beta_i) = \mathbb{I}(i\in S), \|\boldsymbol{\beta}_S - \widetilde{\boldsymbol{\beta}}_S^{S,j}\| < 2jr_n)}{\mathrm{e}^{-4r_n^2}\int_0^1 \mathsf{P}(\boldsymbol{\beta}\in\mathbb{R}^n : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| < 2r_n)\,\mathrm{d}\pi(\theta)}. \tag{A.28}$$

*Then*

$$\begin{aligned}\log\pi\big(\widetilde{\boldsymbol{\beta}}^{S;j}; S; j\big) \leq\ & 3 + \log(2/\lambda_1) + r_n^2\big[9 + C + \lambda_1(4j+3+\delta^o)\big]\\ &+ p_n\big[7 + \log(2/\lambda_1) + 3C + C\log j\big].\end{aligned} \tag{A.29}$$

PROOF. The numerator of (A.28) can be upper-bounded by

$$\mathsf{P}\big(\|\boldsymbol{\beta}_S - \widetilde{\boldsymbol{\beta}}_S^{S,j}\| < 2jr_n; \gamma_{\delta^o}(\beta_i) = 1, i \in S\big) \le \big|v_S(2jr_n)\big| \sup_{\boldsymbol{\beta}\in\mathcal{A}} \pi_S(\boldsymbol{\beta}),$$

where $\mathcal{A} = \{\boldsymbol{\beta} \in \mathbb{R}^n : \gamma(\beta_i) = 1, i \in S; \|\boldsymbol{\beta}_S - \widetilde{\boldsymbol{\beta}}_S^{S,j}\| < 2jr_n\}$. As a lower bound to the denominator, we use (invoking Lemma A.1)

$$\int_{\frac{p_n}{4n}}^{\frac{p_n}{n}} \mathsf{P}\big(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < 2r_n|\theta\big)\,\mathrm{d}\pi(\theta)$$

$$> \left(\frac{1}{2\mathrm{e}}\right)^{p_n+1} \int_{\frac{p_n}{4n}}^{\frac{p_n}{n}} \mathsf{P}\big(\|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}\| < r_n|\theta\big)\,\mathrm{d}\pi(\theta).$$

Next, we use the following truncation $\mathsf{P}(\|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}\| < r_n|\theta) > \mathsf{P}(|\beta_i| > \delta_n, i \in S_0; \|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}\| < r_n|\theta)$. Because $|\beta_{0i}| > \delta^o, i \in S_0$, the volume of the truncated ball is at least as large as $|v_{S_0}(r_n)|/2^{|S_0|}$. The denominator can be then lower-bounded by

$$\left(\frac{1}{2}\right)^{p_n} \mathrm{e}^{-4r_n^2-(p_n+1)\log(2\mathrm{e})}\big|v_{S_0}(r_n)\big| \int_{\frac{p_n}{4n}}^{\frac{p_n}{n}} \inf_{\boldsymbol{\beta}\in\mathcal{B}} \pi_{S_0}(\boldsymbol{\beta}|\theta)\,\mathrm{d}\pi(\theta),$$

where $\mathcal{B} = \{\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0S_0}\| < r_n; |\beta_i| > \delta^o, i \in S_0\}$. Using Lemma A.6, splitting the norms as in the proof of Theorem 5.2, and using the fact $\sqrt{|S_0|} \le r_n$, we obtain (A.29). □

To continue with the proof of Theorem 7.2, we use Lemma A.7 to find an upper bound to (A.23). We have $p_n\delta^o < p_n \log[1/p^{\star}_{\theta^o}(0)] < r_n^2$ and $\delta^o < M$. Thus, for $p_n \ge 1$, (A.23) can be upper-bounded by $\mathrm{e}^{j^2r_n^2/2}$ for $M$ large enough. The rest of the proof is now analogous to the proof of Theorem 5.1.

**Acknowledgment.** The author would like to express gratitude to Edward George for valuable discussions and for his endless enthusiasm and encouragement during the course of this work.

## SUPPLEMENTARY MATERIAL

**Supplement to "Bayesian estimation of sparse signals with a continuous spike-and-slab prior"** (DOI: 10.1214/17-AOS1554SUPP; .pdf). Supplement contains proofs of Section 4.

## REFERENCES

[1] ABRAMOVICH, F., GRINSHTEIN, V. and PENSKY, M. (2007). On optimality of Bayesian testimation in the normal means problem. *Ann. Statist.* **35** 2261–2286. MR2363971

[2] ANTONIADIS, A. and FAN, J. (2001). Regularization of wavelet approximations. *J. Amer. Statist. Assoc.* **96** 939–967. MR1946364

[3] BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *J. Amer. Statist. Assoc.* **110** 1479–1490. MR3449048

[4] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[5] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

[6] CARLIN, B. P. and CHIB, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **57** 473–484.

[7] CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751

[8] CASTILLO, I. (2014). Bayesian nonparametrics, convergence and limiting shape of posterior distributions. Univ. Paris Diderot Paris 7.

[9] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. W. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874

[10] CASTILLO, I. and VAN DER VAART, A. W. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077

[11] CHIPMAN, H., GEORGE, E. I. and MCCULLOCH, R. E. (2001). The practical implementation of Bayesian model selection. In *Model Selection*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **38** 65–134. IMS, Beachwood, OH. MR2000752

[12] CLYDE, M., DESIMONE, H. and PARMIGIANI, G. (1994). Prediction via orthogonalized model mixing. *J. Amer. Statist. Assoc.* **91** 1197–1208.

[13] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

[14] DONOHO, D. L., JOHNSTONE, I. M., HOCH, J. C. and STERN, A. S. (1992). Maximum entropy and the nearly black object. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **54** 41–81. MR1157714

[15] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

[16] GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747.

[17] GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

[18] ISHWARAN, H. and RAO, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *J. Amer. Statist. Assoc.* **98** 438–455. MR1995720

[19] ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158

[20] ISHWARAN, H. and RAO, J. S. (2011). Consistency of spike and slab regression. *Statist. Probab. Lett.* **81** 1920–1928.

[21] JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649.

[22] JOHNSTONE, I. M. and SILVERMAN, B. W. (2005). Empirical Bayes estimates selection of wavelet thresholds. *Ann. Statist.* **33** 1700–1752.

[23] LEMPERS, F. B. (1971). *Posterior Probabilities of Alternative Linear Models*: *Some Theoretical Considerations and Empirical Experiments*. Rotterdam Univ. Press, Rotterdam.

[24] LV, J. and FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498–3528. MR2549567

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

[26] MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1032.

[27] NARISETTY, N. and HE, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.* **42** 789–817. MR3210987

[28] PATI, D., BHATTACHARYA, A., PILLAI, N. and DUNSON, D. (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Ann. Statist.* **42** 1102–1130. MR3210997

[29] ROČKOVÁ, V. (2018). Supplement to "Bayesian estimation of sparse signals with a continuous spike-and-slab prior." DOI:10.1214/17-AOS1554SUPP.

[30] ROČKOVÁ, V. and GEORGE, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *J. Amer. Statist. Assoc.* **109** 828–846. MR3223753

[31] ROCKOVA, V. and GEORGE, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *J. Amer. Statist. Assoc.* **111** 160–1622.

[32] ROCKOVA, V. and GEORGE, E. I. (2017). The Spike-and-Slab LASSO. *J. Amer. Statist. Assoc.* To appear.

[33] SU, W. and CANDÉS, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.* **44** 1038–1068. MR3485953

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

[35] TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused LASSO. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108.

[36] 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. *Electron. J. Stat.* **8** 2585–2618. MR3285877

[37] ZHANG, C. H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional regression. *Ann. Statist.* **36** 1567–1594. MR2435448

[38] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

[39] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

BOOTH SCHOOL OF BUSINESS
UNIVERSITY OF CHICAGO
369 CARLES M. HARPER CENTER
5807 S. WOODLAWN AVENUE
CHICAGO, ILLINOIS 60637
USA
E-MAIL: Veronika.Rockova@Chicagobooth.edu