Adaptive Bayesian Predictive Inference

Veronika Ročková*

September 13, 2023

Abstract

Bayesian predictive inference provides a coherent description of entire predictive uncertainty through predictive distributions. We examine several widely used sparsity priors from the predictive (as opposed to estimation) inference viewpoint. Our context is estimating a predictive distribution of a high-dimensional Gaussian observation with a known variance but an unknown sparse mean under the Kullback-Leibler loss. First, we show that LASSO (Laplace) priors are incapable of achieving rate-optimal performance. This new result contributes to the literature on negative findings about Bayesian LASSO posteriors. However, deploying the Laplace prior inside the Spike-and-Slab framework (for example with the Spike-and-Slab LASSO prior), rate-minimax performance can be attained with properly tuned parameters (depending on the sparsity level s_n). We highlight the discrepancy between prior calibration for the purpose of prediction and estimation. Going further, we investigate popular hierarchical priors which are known to attain *adaptive* rate-minimax performance for estimation. Whether or not they are rate-minimax also for predictive inference has, until now, been unclear. We answer affirmatively by showing that hierarchical Spike-and-Slab priors are adaptive and attain the minimax rate without the knowledge of s_n. This is the first rate-adaptive result in the literature on predictive density estimation in sparse setups. This finding celebrates benefits of a fully Bayesian inference.

Keywords: *Asymptotic Minimaxity, Kullback-Leibler Loss, Predictive Densities, Sparse Normal Means*

*Veronika Ročková is professor of Econometrics and Statistics at the Booth School of Business of the University of Chicago.

The author gratefully acknowledges support from the James S. Kemper Faculty Fund and the National Science Foundation (DMS:1944740). The author would like to thank Johannes Schmidt-Hieber and Edward George for useful conversations and encouraging remarks.

1 Introduction

A fundamental goal in predictive inference is using observed data to estimate an *entire* predictive distribution of a future observation. The Bayesian approach offers a complete description of predictive uncertainty through posterior predictive distributions. Distributional predictions, as opposed to point predictions, are far more preferred by practitioners.

This paper focuses on the problem of predicting a high-dimensional normal vector whose unknown mean is sparse. There is a wealth of literature on estimation theory for the sparse normal means model (under the ℓ_2 loss) starting with the seminal papers [14] and [7]. For example, sparsity inducing priors (point-mass Spike-and-Slab priors [7], Spike-and-Slab LASSO priors [26] or continuous shrinkage priors [3] including the horseshoe [5, 31]) yield rate-minimax posterior concentration in the ℓ_0-sparse Gaussian sequence model. It has also been established [6], that the Bayesian LASSO prior [24] which yields a rate-optimal posterior mode, at the same time, yields a rate-suboptimal posterior distribution. Rate-minimaxity has has become a useful criterion for prior calibration in order to acquire a frequentist certificate for Bayesian estimation procedures. This work focuses on rate-minimax certificates in the context of Bayesian predictive inference. While there are decision-theoretic parallels between predictive density estimation under the Kullback-Leibler (KL) loss and point estimation under the quadratic loss [9, 10, 21], the prediction problem is intrinsically different and may require different prior calibrations. This invites the question: "*How to calibrate priors in order to obtain rate-minimax optimal predictive density estimates?*". To answer this question, this paper examines popular shrinkage priors but from a predictive point of view. We build on [21] and [22] but our work is different in at least two fundamental ways. First, we focus on (a) popular priors that are widely used in practice (such as the Bayesian LASSO [24] and the Spike-and-Slab LASSO [26]), and (b) we study *adaptation* to sparsity levels through hierarchical priors (none of the proposed estimators in [21] and [22] are adaptive).

We consider the problem of predicting $\boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, r)$ from an independent observation $\boldsymbol{X} \sim N_n(\boldsymbol{\theta}, I)$ as $n \to \infty$, where the true underlying parameter $\boldsymbol{\theta}$ is sparse in the sense that $\boldsymbol{\theta} \in \Theta_n(s_n)$ where $\Theta_n(s_n) = \{\boldsymbol{\theta} \in \mathbb{R}^n : \|\boldsymbol{\theta}\|_0 \leq s_n\}$. We study the problem of obtaining a predictive density $\hat{p}(\boldsymbol{y} \mid \boldsymbol{x})$ for $\boldsymbol{Y}$ that is close to $\pi(\boldsymbol{y} \mid \boldsymbol{\theta})$ in terms of the Kullback-Leibler loss

$$L(\boldsymbol{\theta}, \hat{p}(\cdot \mid \boldsymbol{x})) = \int \pi(\boldsymbol{y} \mid \boldsymbol{\theta}) \log \frac{\pi(\boldsymbol{y} \mid \boldsymbol{\theta})}{\hat{p}(\boldsymbol{y} \mid \boldsymbol{x})} \mathrm{d}\boldsymbol{y}, \tag{1.1}$$

assuming that $r \in (0, \infty)$ is known. A more classical version of this problem (without sparsity) was examined in the foundational paper [9]. They studied predictive distributions and assessed the quality of the density estimator $\hat{p}(\cdot \mid \boldsymbol{x})$ by its risk

$$\rho(\boldsymbol{\theta}, \hat{p}) = \int \pi(\boldsymbol{x} \mid \boldsymbol{\theta}) L(\boldsymbol{\theta}, \hat{p}(\cdot \mid \boldsymbol{x})) \mathrm{d}\boldsymbol{x}.$$

For any prior distribution $\pi(\cdot)$, the average (Bayes) risk $r(\pi, \hat{p}) = \int \rho(\boldsymbol{\theta}, \hat{p}) \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$ is known to be minimized by the Bayes (posterior) predictive density

$$\hat{p}(\boldsymbol{y} \mid \boldsymbol{x}) = \int \pi(\boldsymbol{y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{x}) \mathrm{d}\boldsymbol{\theta}. \tag{1.2}$$

We review some perhaps known, yet interesting, facts about the subtleties of predictive inference. See [10] for a complete compendium on knowledge in low-dimensional (non-sparse) situations. A tempting, but deceiving, strategy is to use a plug-in an estimator (e.g the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{MLE}$) to obtain a predictive density estimate $\pi(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_{MLE})$. This misdemeanor was denounced by Aitchison [1] who showed that $\hat{p}_U(\boldsymbol{y} \mid \boldsymbol{x})$ defined as (1.2) under the uniform prior dominates the plug-in predictive density $\pi(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_{MLE})$. When $p = 1$, $\hat{p}_U(\boldsymbol{y} \mid \boldsymbol{x})$ is admissible under KL loss [17] but when $p \geq 3$, $\hat{p}_U(\boldsymbol{y} \mid \boldsymbol{x})$ is inadmissible and dominated by Bayesian predictive density under the harmonic prior [16]. George et al. [9] established general sufficient conditions under which a Bayes predictive density will be minimax and will dominate $\hat{p}_U(\boldsymbol{y} \mid \boldsymbol{x})$ and George and Xu [11] extended some of these results to a regression setup (known variance, fixed dimensionality). These developments testify that the Bayesian approach through integration (as opposed to a plug-in approach) is a far more suitable predictive framework. Our work focuses on *high-dimensional* scenarios where $n \to \infty$ and when $\boldsymbol{\theta}$ is sparse, focusing on *rate-minimaxity*.

Mukherjee and Johnstone [21] were first to study predictive density estimation for a multivariate normal vector with ℓ_0 sparse means and found several fascinating parallels between sparse normal means estimation and predictive inference under the KL loss. In particular, they quantified the minimax risk under the KL loss which equals the minimax risk of estimating sparse normal means up to a constant which depends on the ratio of variances of future and observed data, i.e. with $s_n/n \to 0$

$$\inf_{\hat{p}} \sup_{\boldsymbol{\theta}\in\Theta_n(s_n)} \rho(\boldsymbol{\theta}, \hat{p}) \sim \frac{1}{1+r} s_n \log(n/s_n)$$

where $r = r/1$ is the variance ratio and where the minimum is taken over all predictive density estimators. In addition, [21] constructed a predictive density estimator (inspired by hard thresholding) by pasting two predictive density estimators (under the uniform prior and under a discrete "cluster" prior) depending on the magnitude of the observed data. While minimax optimal (up to a constant), this estimator is not entirely Bayesian since it is not a posterior predictive distribution under any prior. This estimator also relies on discrete priors (which may not be as natural to implement in practice) and is not a smooth function of the data. Finally, these results are non-adaptive, i.e. the knowledge of sparsity level s_n is required to construct the optimal estimator. In a followup paper, [22] proposed proper Bayesian predictive densities under discrete Spike-and-Slab priors (i.e. sparse univariate symmetric priors with *discrete* slab distributions) and showed that they are minimax-rate optimal. Again, these results are unfortunately non-adaptive, where the knowledge of s_n is required to tune/construct the prior. Moreover, [22] again mainly focused on discrete slab priors which may not be as practical. Uniform continuous slab distributions were shown to be minimax-rate optimal if the parameter space is suitably constrained.

Our work provides new insights into predictive performance of shrinkage priors that are widely used in practice. Our goal is providing guidelines for calibrating popular priors in the context of prediction. Our contributions can be summarized as follows: (1) we study Bayesian LASSO priors and show that the predictive distributions are incapable of achieving rate-minimax performance, (2) we study Laplace-induced Spike-and-Slab priors (including the popular Spike-and-Slab LASSO prior [26]) which have a *continuous* slab (versus a discrete slab considered in [21]) and show that, if calibrated by an oracle, predictive densities are rate-minimax, (3) we investigate hierarchical variants of the Spike-and-Slab prior and show an *adaptive* rate-minimaxity. In conclusion, no knowledge of s_n is required for hierarchical Spike-and-Slab priors to be rate-minimax optimal. This self-adaptation property is new but is in line with previous findings in the context of estimation [7, 26].

The paper is structured as follows: Section 2 presents a lower-bound result showing that Bayesian LASSO is incapable of yielding posterior predictive densities with good properties in sparse setups. Section 3 focuses on Spike-and-Slab priors with a Dirac spike and a Laplace slab (a direct extension of the Bayesian LASSO prior) as well as the Spike-and-Slab LASSO prior [26]. Section 4 then shows adaptive rate-minimax performance of hierarchical Spike-and-Slab priors. We conclude with a discussion in Section 5 and the proof of Theorem 3 in Section 6.

Notation We define $a \lesssim b$ and $a \gtrsim b$ if, for some universal constant C, $a \leq Cb$ and $a \geq Cb$, respectively. We write $a \sim b$ when $a \lesssim b$ and $a \gtrsim b$. We denote with $\phi(\cdot)$ and $\Phi(\cdot)$ the density and cdf of a standard normal distribution. The Gaussian Mills ratio will be denoted by $R(x) = [1 - \Phi(x)]/\phi(x)$.

2 The Calibration Conflict of Bayesian LASSO

The LASSO method [30] is a staple in sparse signal recovery. The LASSO estimator is, in fact, Bayesian [24] as it corresponds to the posterior mode under the Laplace prior

$$\pi(\boldsymbol{\theta} \mid \lambda) = \prod_{i=1}^{n} \pi_1(\theta_i \mid \lambda) \quad \text{where} \quad \pi_1(\theta \mid \lambda) = \frac{\lambda}{2} \mathrm{e}^{-\lambda|\theta|} \quad \text{for some} \quad \lambda > 0. \tag{2.1}$$

In our sparse normal-means setting, the LASSO estimator is known to attain the (near) minimax rate $s_n \log n$ for the square Euclidean loss if the regularity parameter λ is chosen of the order $\sqrt{2 \log n}$. Since the LASSO estimator is a posterior mode, it is tempting to utilize the entire posterior distribution as an inferential object. This common misconception was soon rectified by Castillo et al. [6] who showed that the entire posterior distribution (known as the Bayesian LASSO posterior) for such λ puts no mass on balls centered around the sparse truth whose radius is of much larger order than the minimax rate. The discrepancy between the performance of a posterior mode and the entire posterior distribution is a revealing cautionary tale. Since the posterior predictive distribution is a functional of the posterior distribution, we should be skeptical about Bayesian LASSO in the context of prediction as well.

For the Bayesian LASSO independent product prior (2.1) [12, 24], the Bayesian predictive density has a product form

$$\hat{p}(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{i=1}^{n} \hat{p}(y_i \mid x_i)$$

and

$$L(\boldsymbol{\theta}, \hat{p}(\cdot \mid \boldsymbol{x})) = \sum_{i=1}^{n} \int \pi(y_i \mid \theta_i) \log \frac{\pi(y_i \mid \theta_i)}{\hat{p}(y_i \mid x_i)} \mathrm{d}y = \sum_{i=1}^{n} L(\theta_i, \hat{p}(\cdot \mid x_i)).$$

The predictive risk of a product rule over the $\Theta_n(s_n)$ is additive and satisfies [21]

$$(n - s_n)\rho(0, \hat{p}) < \rho(\boldsymbol{\theta}, \hat{p}) = \sum_{i=1}^{n} \rho(\theta_i, \hat{p}) \leq (n - s_n)\rho(0, \hat{p}) + s_n \sup_{\theta \in \mathbb{R}} \rho(\theta, \hat{p}), \tag{2.2}$$

where

$$\rho(\theta, \hat{p}) = \int \pi(x \mid \theta) \int \pi(y \mid \theta) \log[\pi(y \mid \theta)/\hat{p}(y \mid x)] \mathrm{d}y \mathrm{d}x.$$

The following Lemma precisely characterizes the univariate risk $\rho(\theta, \hat{p})$ under the prior (2.1). This Lemma, in fact, applies to *any* prior $\pi_1(\cdot)$.

Lemma 1. *The univariate prediction risk under the Bayesian LASSO prior (2.1) satisfies*

$$\rho(\theta, \hat{p}) = \theta^2/(2r) + E \log N_{\theta,1}^{LASSO}(Z) - E \log N_{\theta,v}^{LASSO}(Z) \tag{2.3}$$

where $v = 1/(1 + 1/r)$ *and* $N_{\theta,v}^{LASSO}(Z) = \int \exp\left\{\mu \left[Z/\sqrt{v} + \theta/v\right] - \frac{\mu^2}{2v}\right\} \pi_1(\mu \mid \lambda) \mathrm{d}\mu$ *and where the expectation is taken over* $Z \sim N(0, 1)$*.*

Proof. Section D.1.

The inability of the entire posterior to concentrate around the truth at an optimal rate [6] stems from a tuning conflict. For noise coordinates $\theta_i = 0$, λ needs to be large in order to push the coefficient to zero and for signals $\theta \neq 0$, λ needs to be small to avoid squashing large effects. We expect a similar conundrum for the prediction problem. The inequality (2.2) shows that, in order for Bayesian LASSO prediction distributions to be rate-minimax optimal, we would need

$$\sup_{\theta \in \mathbb{R}} \rho(\theta, \hat{p}) \lesssim \frac{\log(n/s_n)}{1 + r} \quad \text{and, at the same time,} \quad \rho(0, \hat{p}) \lesssim \frac{s_n \log(n/s_n)}{(n - s_n)(1 + r)}. \tag{2.4}$$

It will follow from Theorem 1 that these two goals are simultaneously unattainable with the same λ.

Theorem 1. *For $v = 1/(1+1/r)$ and the Laplace prior $\pi_1(\theta \mid \lambda)$ with $\lambda > 0$ we obtain*

$$\rho(0,\hat{p}) \le \log\left(1 + \frac{\sqrt{2}}{\lambda\sqrt{\pi v}}\right) + \frac{4}{\lambda^2 v} \quad \textit{and} \quad \sup_{\theta \neq 0} \rho(\theta,\hat{p}) \le \log(\lambda 4\sqrt{2\pi/v}) + \frac{\lambda^2}{2} + \lambda\sqrt{\frac{2}{\pi}} + \frac{4}{\lambda^2}$$

Proof. Section A

Lemma 1 and the inequality (2.4) suggest the following calibrations. For the signal-less scenarios with $\lambda \to \infty$, the dominant term in the risk is $1/(\lambda\sqrt{v})$ which means that $\lambda\sqrt{v}$ should increase to infinity at least as fast as $\frac{(n-s_n)(1+r)}{s_n \log(n/s_n)}$. For the signal scenarios, the calibration may need to depend on r. The dominant term is λ^2 which suggests that λ (a) should not increase faster than $[(1-v)\log(n/s_n)]^{1/2}$ (when $r > 1$) and $[v \log(n/s_n)]^{1/2}$ when $0 < r < 1$ and (b) should not decay slower than $[(1-v)\log(n/s_n)]^{-1/2}$. The calibration upper bound mirrors the oracle threshold for signal recovery (up to the multiplication factor which depends on r). Unfortunately, the two calibration goals for signal and noise are not attainable simultaneously. We can appreciate the calibration dilemma from a plot of the Bayesian LASSO univariate prediction risk $\rho(\theta,\hat{p})$ for various choices of λ when $r = 2$ (Figure 1 on the left). In order for the risk at zero to be small, we need large λ which will unfortunately inflate the risk for larger $|\theta|$. On the other hand, to verify that small λ will inflate the risk at zero, we have the following lower bound result implying that the Bayesian LASSO prediction risk is suboptimal for the calibration $\lambda \propto \sqrt{\log(n/s_n)}$ which mirrors optimal tuning for estimation.

Theorem 2. *Consider the Laplace prior (2.1) with $\lambda > 0$. For $v = 1/(1+1/r)$ the univariate Bayesian LASSO predictive distribution $\hat{p}$ satisfies for any $a > 0$*

$$\rho(0,\hat{p}) > [1-\Phi(a)]\log\left[1 + \left(\frac{1}{\sqrt{v}} - 1\right)\frac{a}{\lambda + a}\right] - \frac{4}{\lambda^2} - \mathrm{e}^{-\lambda^2 v/2}\left(\frac{1}{2} + \log(\lambda\sqrt{2\pi v}) + \frac{2}{\lambda\sqrt{2\pi v}}\right).$$

Proof. Section B

Theorem 2 implies that when $\lambda \to \infty$, the first term in the lower bound dominates the other two. The following Corollary shows that for the oracle tuning $\lambda_n \propto \sqrt{\log(n/s_n)}$ the prediction risk will increase to infinity at a faster than the minimax rate.

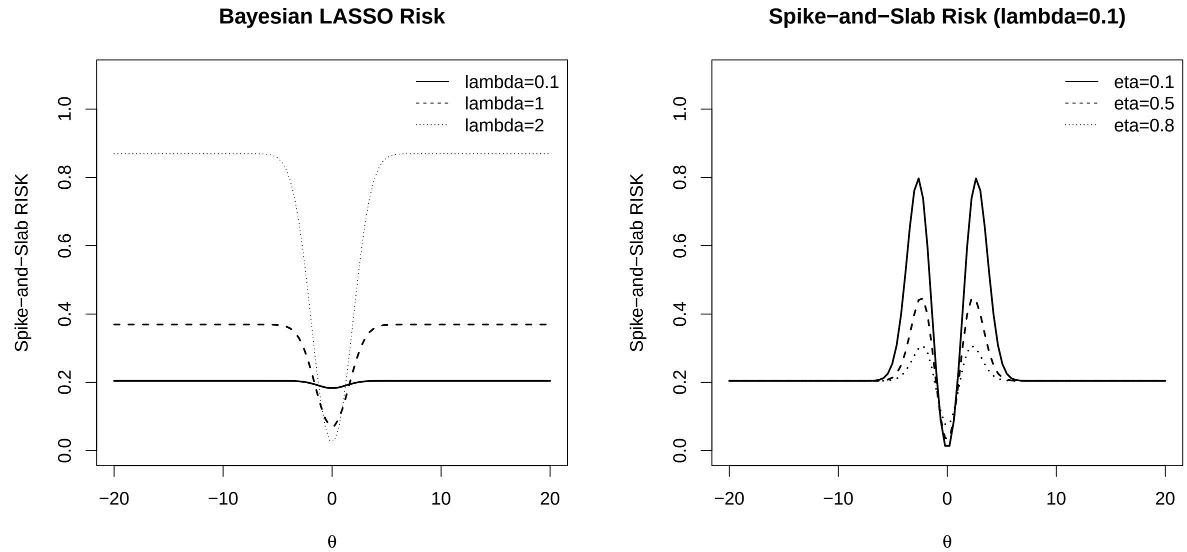

Figure 1: (Left) Bayesian LASSO prediction risk $\rho(\theta,\hat{p})$ for $\lambda\in\{0.1,1,2\}$; (Right) Spike-and-Slab risk (Dirac spike and Laplace slab with a penalty $\lambda=0.1$ with $\eta\in\{0.1,0.5,0.8\}$); Both plots correspond to $r=2$.

Corollary 1. *(Bayesian LASSO is Suboptimal) Consider the Bayesian LASSO prior in* (2.1) *with* $\lambda=\lambda_n\to\infty$ *as* $n\to\infty$. *For fixed* $v=1/(1+1/r)$ *and a suitable* $a>0$ *such that* $\left(\frac{1}{\sqrt{v}}-1\right)\frac{a}{\lambda_n+a}<1$ *we have*

$$\inf_{\boldsymbol{\theta}\in\Theta(s_n)}\rho(\boldsymbol{\theta},\hat{p})>(n-s_n)\left[(1-\Phi(a))\left(\frac{1}{\sqrt{v}}-1\right)\frac{a}{2(\lambda_n+a)}-\frac{1}{\lambda_n^2}\left(4+\frac{3}{v}+\frac{2}{\lambda_n\sqrt{v}}\right)\right].$$

Proof. It follows from Lemma 2 after noting that $\mathrm{e}^{-x^2/2}\log(x)\leq 1/x^2$ and that $\log(1+x)>x/2$ for $x\in(0,1)$.

The important takeaway message of Corollary 1 is that the lower bound on the Bayesian LASSO prediction risk increases to infinity at a *suboptimal rate* for calibrations λ_n that increase to infinity at a slower pace than $\frac{n-s_n}{s_n}\frac{1}{\log(n/s_n)}$.

3 Separable Spike-and-Slab Priors

Section 2 unveils the sad truth about the single Laplace prior that it cannot provide rate-optimal posterior predictive distributions in sparse situations. Precisely for the dual purpose of estimation and selection, the Spike-and-Slab prior framework [18] adaptively borrows from two prior components depending on the magnitude of observed data. Spike-and-Slab priors are known to have ideal adaptation properties in sparse signal recovery such as rate-minimax adaptation in a supremum-norm sense [13] (without any log factors!) or rate-minimax spatial adaptation [29]. This section highlights benefits of prior mixing for prediction and demonstrates why the spike distribution is an indispensable enhancement of the simple Laplace prior from Section 2.

We consider a separable (independent product) Spike-and-Slab prior for $\boldsymbol{\theta} \in \mathbb{R}^n$

$$\pi(\boldsymbol{\theta} \mid \lambda, \eta) = \prod_{i=1}^{n} \pi(\theta_i \mid \lambda, \eta), \quad \text{where} \quad \pi(\theta \mid \lambda, \eta) = \eta\pi_1(\theta) + (1-\eta)\pi_0(\theta) \quad \text{and} \quad \eta \in (0,1). \tag{3.1}$$

The spike density $\pi_0(\theta)$ serves as a model for the noise coefficients $\theta_i = 0$ while the slab density $\pi_1(\theta)$ serves as a model for the signals $\theta_i \neq 0$. The prior mixing weight $\eta \in (0,1)$ determines the amount of exchange between the two priors. We have, for now, silenced the dependence of π_1 on λ which will appear as a tuning parameter for the slab distribution in the next section. In order to appreciate the benefits of prior mixing for the purpose of prediction, we first point out that the predictive distribution is also a mixture. This simple fact is appreciated even more after noting that this mixture is conveniently mixed by a weight that depends on the Bayes factor between the slab/spike models. Notice that because (3.1) is an independent product prior, the predictive rule is again a product rule. Therefore we focus on the univariate predictive distributions below.

Lemma 2. *Denote by* $m_j(x) = \int \pi(x \mid \mu)\pi_j(\mu)\mathrm{d}\mu$ *for* $j = 0, 1$ *the marginal likelihoods under the spike/slab priors* π_0 *and* π_1. *For* $\eta \in (0,1)$, *we define a mixing weight*

$$\Delta_\eta(x) = \frac{\eta m_1(x)}{\eta m_1(x) + (1-\eta)m_0(x)}. \tag{3.2}$$

Then the Bayesian predictive density under the prior (3.1) *is a mixture, i.e.*

$$\hat{p}(y \mid x) = \Delta_\eta(x)\hat{p}_1(y \mid x) + [1 - \Delta_\eta(x)]\hat{p}_0(y \mid x) \tag{3.3}$$

where $\hat{p}_j(y \mid x) = \frac{\int \pi(y \mid \mu)\pi(x \mid \mu)\pi_j(\mu)\mathrm{d}\mu}{m_j(x)}$ *for* $j = 0, 1$ *are posterior predictive densities under the spike/slab priors (respectively).*

Proof. This follows readily from the definition $\hat{p}(y \mid x) = \frac{\int \pi(y \mid \mu)\pi(\mu \mid x)\mathrm{d}\mu}{\int \pi(\mu \mid x)\mathrm{d}\mu}$.

This elegant decomposition provides invaluable insights into the mechanics of the mixture prior. The predictive density will be dominated by the slab predictive density when $\Delta_\eta(x)$ is close to one, i.e. the observation x is more likely to have arisen from the marginal distribution m_1 as opposed to m_0. This is very intuitive. The opposite is true when the data x supports the spike model, in which case the predictive density is taken over by the spike predictive density. The amount of support for the spike/slab predictive densities is determined by the ratio of marginal likelihoods evaluated at the observed data x. In other words, we can rewrite the mixing weight as a functional of the Bayes factor

$$\Delta_\eta(x) = \left[1 + \frac{1-\eta}{\eta} BF(x; 0, 1)\right]^{-1},$$

where $BF(x; 0, 1) = \frac{m_0(x)}{m_1(x)}$ is the Bayes factor for the spike model versus the slab model. Being able to switch between two regimes has been exploited for the purpose of constructing predictive density estimators by [21]. These authors proposed a class of univariate predictive density estimators that are analogs of hard thresholding in that they paste two densities depending on the magnitude (signal detectability) of observed data $|x|$. These estimators, however, do not have a fully Bayesian motivation and are not smooth functions of data. The discussion above highlights that regime switching is achieved *automatically and smoothly* within the Spike-and-Slab framework. The predictive density decomposition in Lemma 2 invites the possibility of upper-bounding the risk in terms of separate spike/slab risks.

Lemma 3. *Denoting the average mixing weight* $\Lambda(\theta) = \int \Delta_\eta(x)\pi(x \mid \theta)\mathrm{d}x$ *and using the notation in Lemma 2, the prediction risk under the prior* (3.1) *satisfies*

$$\rho(\theta, \hat{p}) < \Lambda(\theta)\rho(\theta, \hat{p}_1) + (1 - \Lambda(\theta))\rho(\theta, \hat{p}_0). \tag{3.4}$$

Proof. It follows from a simple application of the Jensen's inequality $E \log X < \log EX$ which yields (with the expectation taken over a binary random variable $\gamma \in \{0, 1\}$ with

$P(\gamma = 1) = \Delta_\eta(x))$

$$
\begin{aligned}
\rho(\theta, \hat{p}) &= \int \pi(x \mid \theta) \left\{ \int \pi(y \mid \theta) \left[\log \pi(y \mid \theta) - \log E\left(\gamma \hat{p}_1(y \mid x) + (1-\gamma)\hat{p}_0(y \mid x) \right) \right] \mathrm{d}y \right\} \mathrm{d}x \\
&< \int \pi(x \mid \theta) \left\{ \int \pi(y \mid \theta) E\left[\log \frac{\pi(y \mid \theta)}{\gamma \hat{p}_1(y \mid x) + (1-\gamma)\hat{p}_0(y \mid x)} \right] \mathrm{d}y \right\} \mathrm{d}x \\
&= \int \pi(x \mid \theta) \left\{ \int \pi(y \mid \theta) \left[\Delta_\eta(x) \log \frac{\pi(y \mid \theta)}{\hat{p}_1(y \mid x)} + (1 - \Delta_\eta(x)) \log \frac{\pi(y \mid \theta)}{\hat{p}_0(y \mid x)} \right] \mathrm{d}y \right\} \mathrm{d}x \\
&= \Lambda(\theta)\rho(\theta, \hat{p}_1) + (1 - \Lambda(\theta))\rho(\theta, \hat{p}_0).
\end{aligned}
$$

This elegant upper bound shows how the risk is dominated either by the spike predictive density (for parameter values θ such that $\Lambda(\theta)$ is small) or the slab predictive density (for parameter values θ such that $\Lambda(\theta)$ is large). This bound is perhaps more intuitive than useful, however. Our analysis rests on a more precise characterization of the risk. In the Lemma 4 below, we obtain a risk decomposition for $\rho(\theta, \hat{p})$ but for *general* mixtures (3.1).

Lemma 4. *The univariate risk for the Spike-and-Slab prior* (3.1) *satisfies*

$$\rho(\theta, \hat{p}) = \rho(\theta, \hat{p}_1) + E \log N^{SS}_{\theta,1}(Z) - E \log N^{SS}_{\theta,v}(Z), \tag{3.5}$$

where

$$N^{SS}_{\theta,v}(z) = \left[1 + \frac{1-\eta}{\eta} \frac{\int \exp\left(\mu(\frac{\theta}{v} + \frac{z}{\sqrt{v}}) - \frac{\mu^2}{2v} \right) \pi_0(\mu) \mathrm{d}\mu}{\int \exp\left(\mu(\frac{\theta}{v} + \frac{z}{\sqrt{v}}) - \frac{\mu^2}{2v} \right) \pi_1(\mu) \mathrm{d}\mu} \right].$$

and $v = 1/(1 + 1/r)$ and where the expectation is taken over $Z \sim N(0, 1)$.

Proof. Section D.2

Lemma 4 is a simple generalization of Theorem 2.1 in [21], who showed risk decomposition for *point-mass* spike-and-slab priors. We present it here because it is useful for other Spike-and-Slab priors such as the Spike-and-Slab LASSO [26] studied in Section 3.2. In comparison with Lemma 1, we have a different expression $N^{SS}_{\theta,v}(z)$ which depends on the prior odds $(1-\eta)/\eta$ whose calibration will be crucial.

Remark 1. *Alternatively, we could write $\rho(\theta, \hat{p}) = \rho(\theta, \hat{p}_0) + E \log N_{\theta,1}(z) - E \log N_{\theta,v}(z)$ for $N_{\theta,v}(z) = \left[1 + \frac{\eta}{1-\eta} \frac{\int \exp\left(\mu(\frac{\theta}{v} + \frac{z}{\sqrt{v}}) - \frac{\mu^2}{2v} \right) \pi_1(\mu) \mathrm{d}\mu}{\int \exp\left(\mu(\frac{\theta}{v} + \frac{z}{\sqrt{v}}) - \frac{\mu^2}{2v} \right) \pi_0(\mu) \mathrm{d}\mu} \right]$. For the point-mass spike $\pi_0 = \delta_0$ we would then obtain the same expression as in Theorem 2.1 of [21].*

3.1 Dirac Spike and Laplace Slab

In Section 2, we convinced ourselves that a single Laplace prior will not be able to yield rate-optimal predictive distributions. We now show that using the Laplace slab $\pi_1(\theta \mid \lambda)$ *in addition* to a Dirac spike $\pi_0(\theta)$ inside the prior (3.1) does.

Theorem 3. *Assume the Spike-and-Slab prior (3.1) with a Dirac spike $\pi_0(\theta) = \delta_0(\theta)$ and a Laplace slab $\pi_1(\theta \mid \lambda)$ in (2.1). Denote $v = 1/(1+1/r)$ and set $(1-\eta)/\eta = n/s_n$. Choose $\lambda^2 = vC_r$ for $C_r > 2/[v(1/2+4)]$ when $0 < r < 1$ and $\lambda^2 = (1-v)C_r^*$ for $C_r^* > 2/[5(1-v)]$ when $r \geq 1$ then with $s_n/n \to 0$ we have for any fixed $r \in (0, \infty)$*

$$\sup_{\boldsymbol{\theta} \in \Theta(s_n)} \rho(\boldsymbol{\theta}, \hat{p}) \leq \frac{5}{1+r} s_n \log(n/s_n) + \widetilde{C}(r) \tag{3.6}$$

where $\widetilde{C}(r)$ is either (6.17) or (6.18).

Proof. Section 6.

Theorem 3 shows that the minimax-rate predictive performance is attainable by Spike-and-Slab priors with a simple Laplace slab distribution. This prior is widely used in practice [25]. Up until now, it has been perceived that this privilege is only reserved for discrete (grid) slab priors which may be simpler to treat analytically [22]. The only other continuous slab result in the literature so far is in [22] who analyzed *uniform* slabs over a finite parameter domain (which are perhaps not as widely used). Of course, our result in Theorem 3 is only rate-minimax, where the multiplication constant may not optimal.

Remark 2. *(Calibration) What is particularly noteworthy in Theorem 3 is the prior calibration. The oracle tuning for Spike-and-Slab priors in the context of estimation would be $\eta/(1-\eta) = s_n/n$. This is the oracle tuning for η in the prediction problem as well. However, in terms of calibrating λ, we distinguish between two regimes. When $r > 1$, the future observation is noisier than the observed data which makes the prediction problem simpler. In this case, we can choose λ to be a small fixed constant which does not need to depend on n. This corresponds to the usual Spike-and-Slab calibration and small λ regime discussed earlier in [26]. Here we tried to optimize λ as a function of r. For the more difficult case when $0 < r < 1$, i.e. when the observed data is noisier, we can also choose*

λ proportional to the usual oracle detection threshold $\sqrt{2\log(n/s_n)}$ suitably rescaled by a constant multiple of $\sqrt{v}$.

Remark 3. *Inside the proof of Theorem 3 we distinguish between two regimes: (1) when $|x| > \lambda + \sqrt{2v\log(n/s_n)}$, with i.e. the observed data is above the usual detection threshold rescaled by v and shifted by λ, and (2) when $|x| \leq \lambda + \sqrt{2v\log(n/s_n)}$. In the first case, the slab predictive density takes over and the reverse is true for case (2). [21] used a related regime switching idea to construct an estimator by pasting two predictive densities depending on the size of $|x|$. We leverage these two regimes only inside a proof, not for a construction of the estimator.*

It is interesting to compare the risk performance of the single Laplace prior and the Spike-and-Slab prior with a Laplace slab in Figure 1. The right plot corresponds to the calibration $\lambda = 0.1$, which benefits values $\theta \neq 0$. In order to diminish risk at zero, we need to decrease η. Compared with the left plot, Bayesian LASSO risk asymptotes towards the same value as $|\theta| \to \infty$, but has an elevated risk at zero. This confirms our intuition that the proper calibration for the Spike-and-Slab prior is a small (fixed constant) λ and a small η which depends on s_n/n. The difference between the risks for the Bayesian LASSO (left plot) with $\lambda = 0.1$ and Spike-and-Slab with a Laplace slab (right plot) with $\lambda = 0.1$ is striking. The mixture prior can suppress risk at zero by decreasing η, while keeping the risk small for larger $|\theta|$. Notably, there are certain values of θ for which the risk is inflated due to uncertainty whether or not the underlying parameter θ arrived more likely from the slab/spike.

3.2 The Spike-and-Slab LASSO

The Spike-and-Slab LASSO prior [26] has become popular for sparsity recovery due to its self-adaptive shrinkage properties. It generalizes the Laplace prior (2.1) by deplying a two-point Laplace mixture (3.1) where

$$\pi_0(\theta) = \lambda_0/2\mathrm{e}^{-\lambda_0|\theta|} \quad \text{and} \quad \pi_1(\theta) = \lambda_1/2\mathrm{e}^{-\lambda_1|\theta|} \quad \text{with } \lambda_1 < \lambda_0. \tag{3.7}$$

This prior has been successfully implemented in high-dimensional regression [28], graphical models [8], factor analysis [27], biclustering [20], among others [2]. The perception had been the Spike-and-Slab theory holds only for point-mass spikes $\pi_0(\cdot) = \delta_0(\cdot)$. Ročková [26], however, showed that asymptotic minimaxity is attainable also for these *continuous* spike distributions in the context of estimating sparse mean under the Euclidean loss. Here, we examine the ability of Spike-and-Slab LASSO priors to yield optimal predictive distributions. While the majority of implementations of the Spike-and-Slab LASSO focus on posterior mode detection, posterior simulation is available through traditional Gibbs sampling or Bayesian bootstrap techniques [23]. The ability to simulate from the posterior makes the posterior predictive distribution readily available. Here, we show that it is in fact rate-optimal when properly calibrated.

Theorem 4. *Assume the Spike-and-Slab LASSO prior* (3.1) *with* (3.7) *and* $(1-\eta)/\eta = c$ *for some fixed constant* $c > 0$. *Choose* $\lambda_0 = n/s_n$ *and* λ_1 *as in Theorem 3. When* $s_n/n \to 0$, *then we have*

$$\sup_{\boldsymbol{\theta}\in\Theta(s_n)} \rho(\boldsymbol{\theta}, \hat{p}) \sim \frac{s_n}{1+r}\log(n/s_n) \tag{3.8}$$

for any fixed $r \in (0,1)$. *The same conclusion holds for* $r \in [1,\infty)$ *for parameters* $\boldsymbol{\theta} \in \Theta_n(s_n) \cap \{\boldsymbol{\theta} \in \mathbb{R}^n : \min_{1\le i\le n} |\theta_i| > c\sqrt{\log(n/s_n)}\}$ *for* $c > 2\sqrt{v}$.

Proof. Section C.

Remark 4. *(Calibration) Ročková [26] concluded that* $\lambda_0/\lambda_1 \times (1-\eta)/\eta$ *should behave as* $(n/s_n)^c$ *for some* $c \ge 1$ *to yield rate-minimaxity in estimation. Figure 2 (middle plot) shows that the risk at zero will be small when* λ_0 *is large, not necessarily only when* η *is small (left plot). This observation is confirmed by the upper bound on* $\rho(0,\hat{p})$ *in* (**??**). *Interestingly, we* do not *require* $(1-\eta)/\eta$ *to be of the order* n/s_n *as long as* λ_0 *is increasing at a rate* n/s_n. *If we assumed* $(1-\eta)/\eta = n/s_n$, *the proof in Section C shows that the calibration* $\lambda_0 = n/s_n$ *would yield the optimal rate only for* $r \in (0,1)$. *For general* $r \in (0,\infty)$ *we need to fix* $(1-\eta)/\eta$ *and let* λ_0 *be of the order* n/s_n. *In terms of* λ_1, *we have the same assumption as in Theorem 3, where* λ_1 *is set proportional to either* $\sqrt{v}$ *or* $\sqrt{1-v}$. *The incentive is to keep it small so that the risk for large effects is small, but not too small so*

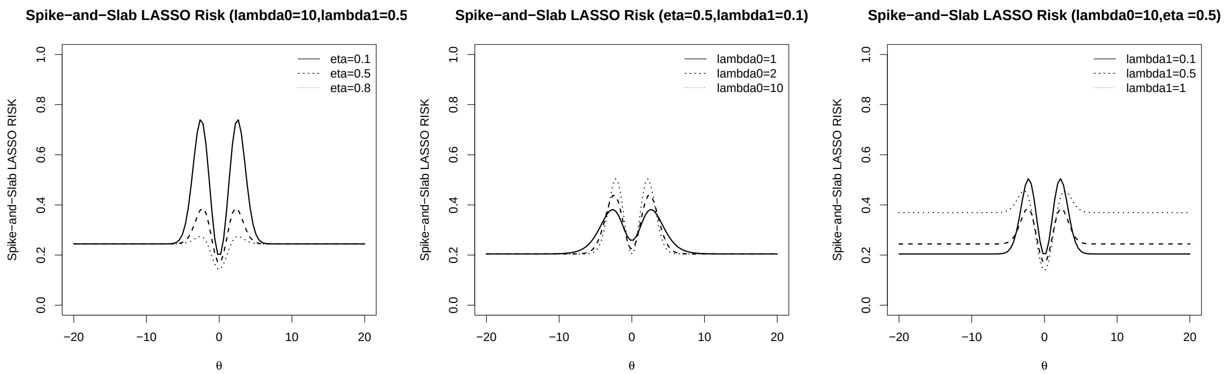

Figure 2: Spike-and-Slab LASSO prediction risk for various calibrations. (Left) Varying η for fixed $\lambda_0 = 10, \lambda_1 = 0.5$; (Middle) Varying λ_0 for fixed $\eta = 0.1, \lambda_1 = 0.1$; (Right) Varying λ_1 for fixed $\eta = 0.1, \lambda_0 = 10$

that the supremum over $\theta \neq 0$ is not too large. For $r > 1$, we require a signal-strength assumption that is similar to the one in [26].

4 Hierarchical Spike-and-Slab Priors

All results in [21, 22] are *non-adaptive*, i.e. the knowledge of s_n is required to construct or tune the estimator in order to obtain minimax performance. While we consider only rate-minimax performance, there are reasons to be hopeful that hierarchical priors (which do achieve automatic adaptation in parameter estimation) will yield an adaptive rate for predictive density estimation. One recent remarkable example is supremum-norm adaptation (without any log factors!) of hierarchical spike-and-slab priors in non-parametric wavelet regression [13]. We inquire into adaptability of hierarchical priors in the context of predictive inference.

We continue our investigation of the Dirac spike and Laplace slab prior from Section 3.1. However, now we assume a *hierarchical* version (not an independent product), i.e.

$$\pi(\boldsymbol{\theta}) = \int_\eta \prod_{i=1}^n [(1-\eta)\delta_0 + \eta\pi_1(\theta_i)]\pi(\eta)\mathrm{d}\eta \quad \text{and} \quad \pi(\eta) \sim Beta(a,b) \tag{4.1}$$

for some $a, b > 0$ In Section 3, we saw how independent product Spike-and-Slab mixtures yield within-coordinate mixing, adaptively borrowing from both the spike and the slab. The hierarchical prior (4.1) performs an additional across-coordinate mixing, borrowing strength and transmitting sparsity information. This point was highlighted in the estimation context under the squared error loss [26, 28]. The posterior predictive distribution under the hierarchical prior is no longer an independent product but a scale mixture of two-point mixtures. Indeed, using Lemma 2, we have

$$\hat{p}(\boldsymbol{y} \mid \boldsymbol{x}) = \int_{\eta} \prod_{i=1}^{n} \left[\Delta_{\eta}(x_i)\hat{p}_1(y_i \mid x_i) + (1-\Delta_{\eta}(x_i))\hat{p}_0(y_i \mid x_i)\right] \mathrm{d}\pi(\eta), \tag{4.2}$$

where $\hat{p}_1$ and $\hat{p}_0$ are the slab and spike univariate posterior predictive densities, $\Delta_{\eta}(x_i)$ is the mixing weight (3.2) and where m_1 and m_0 are the marginal likelihoods defined in Lemma 2. It is also useful to rewrite the predictive distribution as an average predictive density $\hat{p}(\boldsymbol{y} \mid \boldsymbol{x}, \eta)$ where the average is taken over the posterior $\pi(\eta \mid \boldsymbol{x})$, i.e.

$$\hat{p}(\boldsymbol{y} \mid \boldsymbol{x}) = E_{\eta \mid \boldsymbol{x}} \hat{p}(\boldsymbol{y} \mid \boldsymbol{x}, \eta). \tag{4.3}$$

This key property is utilized in the following lemma which allows us to bound the KL loss.

Lemma 5. *The Kullback-Leibler loss of the predictive distribution under the hierarchical prior* (4.1) *satisfies* $L(\boldsymbol{\theta}, \hat{p}(\cdot \mid \boldsymbol{x})) \leq E_{\eta \mid \boldsymbol{x}} L(\boldsymbol{\theta}, \hat{p}(\cdot \mid \boldsymbol{x}, \eta))$.

Proof. This follows from rewriting the posterior predictive distribution (4.2) as (4.3) and by applying the the Jensen's inequality $E \log X < \log EX$ to obtain

$$\begin{aligned} L(\boldsymbol{\theta}, \hat{p}(\cdot \mid \boldsymbol{x})) &= \int \pi(\boldsymbol{y} \mid \boldsymbol{\theta}) \log \pi(\boldsymbol{y} \mid \boldsymbol{\theta}) - \int \pi(\boldsymbol{y} \mid \boldsymbol{\theta}) \log E_{\eta \mid \boldsymbol{x}} \hat{p}(\boldsymbol{y} \mid \boldsymbol{x}, \eta) \mathrm{d}\boldsymbol{y} \\ &\leq E_{\eta \mid \boldsymbol{x}} L(\boldsymbol{\theta}, \hat{p}(\cdot \mid \boldsymbol{x}, \eta)). \end{aligned} \tag{4.4}$$

Lemma 5 shows that it is the conditional distribution $\pi(\eta \mid \boldsymbol{x})$ which drives the prediction performance. This posterior is expected to be concentrated around zero for sparse situations such as ours. In fact, we would expect that the posterior $\pi(\eta \mid \boldsymbol{x})$ will carry important information regarding the sparsity level s_n. To fortify this intuition, the following Lemma shows that the risk of the hierarchical prior is determined by the typical posterior mean of the log-odds $(1-\eta)/\eta$ of a spike versus a slab and its reciprocal.

Lemma 6. *The prediction risk under the hierarchical prior (4.1) satisfies for $\lambda > 2$*

$$\rho(\boldsymbol{\theta}, \hat{p}) \leq s_n \left\{ C(\lambda, v) + (1-v) \left[E_{\boldsymbol{x}\,|\,\boldsymbol{\theta}} E \log\left(\frac{1-\eta}{\eta}\right) \,|\, \boldsymbol{x} \right] \right\} + D(n - s_n) \sup_{i:\theta_i \neq 0} E_{\boldsymbol{x}_{\setminus i}\,|\,\boldsymbol{\theta}} E\left(\frac{\eta}{1-\eta} \,|\, \boldsymbol{x}_{\setminus i}\right).$$

for a suitable constant $C(\lambda, v) > 0$ defined in (S6) and $D = 1 + 2/(a-1)$, where $\boldsymbol{x}_{\setminus i}$ denotes the vector $\boldsymbol{x}$ without the i^{th} coordinate.

Proof. Section D.3

This Lemma shows how hierarchical priors achieve improved rates compared to the default tuning $(1-\eta)/\eta = n$ for when s_n is not known. Lemma 7 below characterizes the behavior of the posterior mean of prior model (spike/slab) odds.

Lemma 7. *Assume the hierarchical Spike-and-Slab prior (4.1) with $a, b > 0$. Then, under the Gaussian model $\boldsymbol{X} \sim N_n(\boldsymbol{\theta}, I)$, the posterior distribution $\pi(\eta \,|\, \boldsymbol{x})$ satisfies for any $\boldsymbol{\theta} \in \Theta(s_n)$ with $s_n(\boldsymbol{\mu}) = \|\boldsymbol{\mu}\|_0$*

$$E\left(\frac{\eta}{1-\eta} \,|\, \boldsymbol{x}\right) \leq \frac{a + E[s_n(\boldsymbol{\mu}) \,|\, \boldsymbol{x}] + 1}{b-1} \quad \text{and} \quad E\left(\frac{1-\eta}{\eta} \,|\, \boldsymbol{x}\right) \leq E\left(\frac{b+n}{s_n(\boldsymbol{\mu}) + a - 1} \,|\, \boldsymbol{x}\right).$$

Proof. Section D.4.

Remark 5. *This Lemma shows that the usual calibration [7] with $a = 1$ and $b = n+1$ implies $E\left(\frac{\eta}{1-\eta} \,|\, \boldsymbol{x}\right) \lesssim E[s_n(\boldsymbol{\mu})/n \,|\, \boldsymbol{x}]$ and $E\left(\frac{1-\eta}{\eta} \,|\, \boldsymbol{x}\right) \lesssim E[n/s_n(\boldsymbol{\mu}) \,|\, \boldsymbol{x}]$. While we focused on upper bounds in Lemma 7, one can easily show lower bounds as well implying that the order of these expectations is $s_n(\boldsymbol{\mu})/n$ and $n/s_n(\boldsymbol{\mu})$, respectively. In particular, for $\lambda > 2$ we obtain from (S9) that $E\left(\frac{\eta}{1-\eta} \,|\, \boldsymbol{x}\right) > \frac{a + E[s_n(\boldsymbol{\mu}) \,|\, \boldsymbol{x}]}{b-1}$.*

Going back to Lemma 6, it is crucial to understand the *typical* behavior of the posterior mean of the odds $(1-\eta)/\eta$ and $\eta/(1-\eta)$ under the model $\boldsymbol{X} \sim N_n(\boldsymbol{\theta}, I)$. The following Lemma utilizes the known property of Spike-and-Slab priors that the posterior does not *overshoot* the true dimensionality s_n by too much (Theorem 2.1 in [7]).

Lemma 8. *Assume $\boldsymbol{X} \sim N_n(\boldsymbol{\theta}, I)$ and the hierarchical Spike-and-Slab prior (4.1) with $a = 1$ and $b = n + 1$. Then for some suitable $M > 0$ we have*

$$\sup_{\boldsymbol{\theta} \in \Theta_n(s_n)} E_{\boldsymbol{x}\,|\,\boldsymbol{\theta}} E\left(\frac{\eta}{1-\eta} \,|\, \boldsymbol{x}\right) \leq M s_n/n + o(1) \quad \text{as } n \to \infty.$$

Proof. Section D.5

Lemma 8 takes care of the "noise" part of the risk bound in Lemma 6 where $(n - s_n) E_{\boldsymbol{x}_{\setminus i} \,|\, \boldsymbol{\theta}} E[\eta/(1-\eta) \,|\, \boldsymbol{x}_{\setminus i}] \lesssim s_n$. In order to bound the "signal" part of the risk bound, we need to show that $s_n E_{\boldsymbol{x} \,|\, \boldsymbol{\theta}} E[(1-\eta)/\eta \,|\, \boldsymbol{x}] \lesssim s_n \log(n/s_n)$. From Lemma 7, we need to make sure that the posterior $\pi(\eta \,|\, \boldsymbol{x})$ does not *undershoot* s_n by too much. In Lemma 9 below, we show that when the true underlying signal is strong enough, the posterior does not miss *any* signal. First, we define

$$\Theta_n(s_n, \widetilde{M}) = \Theta_n(s_n) \cap \left\{ \boldsymbol{\theta} \in \mathbb{R}^n : \min_{i:\theta_i \neq 0} |\theta_i| > \widetilde{M}\sqrt{\log n} \right\}. \tag{4.5}$$

A similar (but stronger) signal strength condition was used in [6] in the context of high-dimensional regression with Spike-and-Slab priors.

Lemma 9. *Assume $\boldsymbol{X} \sim N_n(\boldsymbol{\theta}, I)$ and the hierarchical Spike-and-Slab prior (4.1) with $a = 2$ and $b = n + 1$. Denote with S an index of all subsets of $\{1, \ldots, n\}$ and define $c = (\widetilde{M}^2 - 2)/4$. We have*

$$\sup_{\boldsymbol{\theta} \in \Theta_n(s_n, \widetilde{M})} P\Big(\exists j \;\; \textit{such that} \;\; \theta_j \neq 0 \quad \textit{and} \quad j \notin S \,|\, \boldsymbol{x}\Big) \leq \frac{C e^{\lambda^2/2} s_n}{n^{c-1}}$$

with probability at least $1 - 2/n$. Assume $\lambda > 0$ such that $\lambda^2 \leq 2d \log n$ for some $d > 0$. Then for $c > 2 + d$ we have

$$\sup_{\boldsymbol{\theta} \in \Theta_n(s_n, \widetilde{M})} E_{\boldsymbol{x} \,|\, \boldsymbol{\theta}} E\left(\frac{1-\eta}{\eta} \,|\, \boldsymbol{x}\right) \lesssim n/s_n.$$

Proof. Section D.6

Combining all the pieces, Theorem 5 below characterizes rate-minimax performance of the hierarchical Spike-and-Slab prior when the signals are large enough (i.e. over the parameter space (4.5)). The near-minimax performance is guaranteed over the entire parameter space $\Theta_n(s_n)$.

Theorem 5. *Assume the hierarchical prior (4.1) with a Laplace slab (2.1) and with $a = 2$ and $b = n+1$. Choose $\lambda^2 = vC_r$ for $C_r > 2/[v(1/2+4)]$ such that $\lambda > 2$ when $0 < r < 1$ and $\lambda^2 = (1-v)C_r^*$ for $C_r^* > 2/[5(1-v)]$ such that $\lambda > 2$ when $r \geq 1$. Denote $c = (\widetilde{M}^2 - 2)/4$*

where $\widetilde{M}$ *is the signal-strength constant in* (4.5) *then we have for* $c > 2$

$$\sup_{\boldsymbol{\theta} \in \Theta_n(s_n, \widetilde{M})} \rho(\boldsymbol{\theta}, \hat{p}) \lesssim \frac{s_n}{r+1} \log(n/s_n) \quad \textit{and} \quad \sup_{\boldsymbol{\theta} \in \Theta_n(s_n)} \rho(\boldsymbol{\theta}, \hat{p}) \lesssim \frac{s_n}{r+1} \log(n).$$

Proof. The proof follows directly from Lemma 6, 7, 8 and 9. The proposed calibrations ascertain that the term $C(\lambda, v)$ defined in (S6) is suitably small.

Theorem 5 establishes that rate-minimax predictive performance is achievable with hierarchical mixture priors that *do not* require the knowledge of s_n. There is no other rate-adaptive predictive density result in the literature so far. While we considered a fixed λ regime (as $n \to \infty$), for near-minimaxity (with a log factor $\log n$ instead of $\log(n/s_n)$)we could allow having λ increase at a rate $\sqrt{\log n}$ (rescaled by r). Another possible strategy to achieve adaptation would be via empirical Bayes or via a two-step procedure, plugging in an estimator of s_n/n in place of η. We have seen from Lemma 6 that hierarchical priors perform this plug-in automatically. Theorem 5 nicely complements adaptive rates of estimation results under the squared error loss for Spike-and-Slab priors obtained earlier by Castillo and van der Vaart [7] (Theorem 2.2) or Ročková [26].

5 Discussion

This paper investigates several widely used priors from a predictive inference point of view. We establish a negative result for the Bayesian LASSO, showing that posterior predictive densities under this prior cannot converge to the true density of future data at an optimal rate for the usual tuning that would yield an optimal posterior mode in estimation. Next, we study the popular Dirac Spike and Laplace Slab mixture prior and the Spike-and-Slab LASSO prior and show that proper calibrations (that depend on s_n and r) yield rate-minimaxity. Finally, by considering a hierarchical extension, we show that adaptation to s_n is possible with the usual beta-binomial prior on the sparsity pattern.

6 Proof of Theorem 3

We recall the risk decomposition in Remark 1 (or equivalently in Lemma 3 in [21])

$$\rho(\theta, \hat{p}) = Eg(Z, \theta, v), \quad \text{where} \quad g(z, \theta, v) = \theta^2/(2r) + \log N_{\theta,1}(z) - \log N_{\theta,v}(z), \tag{6.1}$$

where the expectation is taken over $Z \sim N(0,1)$ and where

$$N_{\theta,v}(z) = 1 + \frac{\eta}{1-\eta} \int \exp\left\{\mu\left[z/\sqrt{v} + \theta/v\right] - \frac{\mu^2}{2v}\right\} \pi_1(\mu \mid \lambda) \mathrm{d}\mu$$

with $\pi_1(\mu \mid \lambda) = \lambda/2\mathrm{e}^{-\lambda|\mu|}$ for $\lambda > 0$. We find that

$$\log N_{\theta,v}(z) = \log\left[1 + \frac{\eta}{1-\eta}\frac{\lambda}{2}(I_1^v + I_2^v)\right]$$

where

$$I_1^v = \sqrt{v}\frac{\Phi(\mu_1/\sqrt{v})}{\phi(\mu_1/\sqrt{v})} \quad \text{and} \quad I_2^v = \sqrt{v}\frac{\Phi(-\mu_2/\sqrt{v})}{\phi(-\mu_2/\sqrt{v})} \tag{6.2}$$

and

$$\mu_1 = z\sqrt{v} + \theta - v\lambda \quad \text{and} \quad \mu_2 = z\sqrt{v} + \theta + v\lambda. \tag{6.3}$$

The reversed risk characterization in Lemma 4 will appear later in the section below.

6.1 The case when $\theta \neq 0$.

Define $t_v = \lambda\sqrt{v} - |z + \theta/\sqrt{v}|$. When $z + \theta/\sqrt{v} > 0$ we have $I_1^v > I_2^v$ and the reverse is true when $z + \theta/\sqrt{v} \leq 0$. This observation yields the following bounds

$$\log N_{\theta,1}(z) \leq \log\left(1 + \frac{\eta\lambda}{1-\eta}R_1\right) \quad \text{and} \quad \log N_{\theta,v}(z) > \log\left(1 + \frac{\eta\lambda\sqrt{v}}{(1-\eta)2}R_v\right)$$

where (using the Mills ratio notation $R(x) = (1-\Phi(x))/\phi(x)$)

$$R_v = R(t_v) = \frac{1-\Phi(t_v)}{\phi(t_v)}.$$

We consider two scenarios for bounding the function $g(z, \theta, v)$ in (6.1) depending on the magnitude of $|z + \theta|$ (which has the same distribution as $|x|$). In Section 3 we emphasized that depending on $|x|$, the entire predictive density is dominated by either the slab or the spike predictive density. We exploit this idea but only for the proof, not for the construction of an actual estimator [21].

6.1.1 Case (i): $1/R_1 \leq [\eta/(1-\eta)]^v$

As will be seen later, this event is equivalent to $|z+\theta|$ being large enough. Since $z+\theta$ is a proxy for x, we would then expect the risk to be dominated by the slab risk. We can write

$$\log \frac{N_{\theta,1}(z)}{N_{\theta,v}(z)} \leq \log\left(\frac{R_1}{R_v}\right) + \log\left(\frac{\frac{1}{R_1} + \frac{\eta}{1-\eta}\lambda}{\frac{1}{R_v} + \frac{\eta}{1-\eta}\frac{\sqrt{v}}{2}\lambda}\right). \tag{6.4}$$

This expression, in fact, brings us back to the risk decomposition in Lemma 3.5 written in terms of the risk of the slab Bayesian LASSO predictive density $\hat{p}_1(\cdot)$ plus an extra component (the second summand in (6.4)). Indeed, $\theta^2/(2r) + \log(R_1/R_2)$ bounds the LASSO risk in Lemma 1 and the second summand in (6.4) corresponds to $E \log N_{\theta,1}^{SS}(z) - E \log N_{\theta,v}^{SS}(z)$ in Lemma 3.5. The goal is to show that the second summand in (6.4) is small. Indeed, we have

$$\log\left(\frac{1}{R_1} + \frac{\eta}{1-\eta}\lambda\right) - \log\left(\frac{1}{R_v} + \frac{\eta}{1-\eta}\frac{\sqrt{v}}{2}\lambda\right) \leq \log\left(\frac{2}{\sqrt{v}}\right) + \log\left[1 + \frac{1}{\lambda}\left(\frac{1-\eta}{\eta}\right)^{1-v}\right]. \tag{6.5}$$

Next, we have

$$\frac{R_1}{R_2} < \frac{1}{\Phi(-t_v)} \exp\left\{\frac{\lambda^2(1-v)}{2} + \lambda|z|(1-\sqrt{v}) + \theta^2\left(1-\frac{1}{v}\right) + 2z\theta\left(1-\frac{1}{\sqrt{v}}\right)\right\}. \tag{6.6}$$

Now we focus on the term $1/\Phi(-t_v)$ in (6.6). We have $-t_v > -\lambda\sqrt{v}$ and (because $\lambda\sqrt{v} > 0$ we can apply Lemma 13)

$$\frac{1}{\Phi(-t_v)} < \frac{1}{1-\Phi(\lambda\sqrt{v})} \leq \frac{\sqrt{\lambda^2 v + 4} + \lambda\sqrt{v}}{2\phi(\lambda\sqrt{v})} \leq \frac{\sqrt{\lambda^2 + 4} + \lambda}{2\phi(\lambda\sqrt{v})}.$$

This yields (using Lemma 12)

$$\log[1/\Phi(-t_v)] \leq \log\sqrt{2\pi} + \frac{\lambda^2 v}{2} + \frac{4}{\lambda^2} + \log(\lambda). \tag{6.7}$$

Now, we understand when the event $\{1/R_1 < [\eta\lambda/(1-\eta)]^v\}$ actually occurs. Because the Mills ratio function $R(x)$ is monotone decreasing and satisfies

$$0.5/\phi(x) \leq R(x) \leq 1/\phi(x) \quad \text{when} \quad x < 0$$

and because $(1-\eta)/\eta = n/s_n$ goes to infinity, we have

$$R(x_\eta) = \left[\frac{1-\eta}{\eta}\right]^v \quad \text{for some} \;\; h_\eta^L < x_\eta < h_\eta^U < 0 \tag{6.8}$$

where

$$h_\eta^U = -\sqrt{2v\log\left[(\pi 2)^{-v/2}(1-\eta)/\eta\right]} \quad \text{and} \quad h_\eta^L = -\sqrt{2v\log\left[(\pi/2)^{-v/2}(1-\eta)/\eta\right]}. \tag{6.9}$$

Then we have

$$\mathcal{A}_1 \equiv \left\{\frac{1}{R_1} \leq \left(\frac{\eta}{1-\eta}\right)^v\right\} = \{t_1 \leq x_\eta\} \subset \{t_1 \leq h_\eta^U\} = \{|z+\theta| \geq \lambda - h_\eta^U\}.$$

Interestingly, this corresponds to the regime when the magnitude of $z+\theta$ is above $-h_{\eta,\lambda}^U$, the usual detection threshold $\sqrt{2\log(n/s_n)}$ suitably rescaled by v. This implies that the slab is active when the size of the observed data $|x|$ is large, yielding an upper bound

$$g(z,\theta,v) < g_1(z,\theta,v)$$

where

$$g_1(z,\theta,v) = -\log\Phi(-t_v) + \frac{\lambda^2(1-v)}{2} + \lambda|z|(1-\sqrt{v}) + 2z\theta\left(1-\frac{1}{\sqrt{v}}\right) \tag{6.10}$$

$$+\log(4/\sqrt{v}) + (1-v)\log\left(\frac{1-\eta}{\eta}\right) + \log(1+1/\lambda). \tag{6.11}$$

6.1.2 Case (ii): $1/R_1 \geq [\eta/(1-\eta)]^v$

Using similar arguments as before in the Case (i), we have

$$\mathcal{A}_2 \equiv \left\{\frac{1}{R_1} > \left(\frac{\eta}{1-\eta}\right)^v\right\} = \{t_1 > x_\eta\} \subset \{t_1 > h_\eta^L\} = \{|z+\theta| < \lambda - h_\eta^L\}.$$

This regime mirrors the scenario when the observed data is below the detection threshold. In this case, the slab risk plays a minor role and we obtain a simplified expression

$$\log\frac{N_{\theta,1}(z)}{N_{\theta,v}(z)} < \log\left(\frac{1+\frac{\eta}{1-\eta}\lambda R_1}{1+\frac{\eta}{1-\eta}\frac{\sqrt{v}}{2}\lambda R_2}\right) < \log\left[1+\lambda\left(\frac{\eta}{1-\eta}\right)^{1-v}\right].$$

Next, we have (using $(a+b)^2 \leq 2a^2+2b^2$) and because $v/r = 1/(r+1) = (1-v)$

$$\frac{\theta^2}{2r} \leq \frac{|z+\theta|^2+|z|^2}{r} \leq \frac{2\lambda^2+|z|^2}{r} + 4(1-v)\log\left(\frac{1-\eta}{\eta}\right).$$

This implies an upper bound $g(z,\theta,v) \leq g_2(z,\theta,v)$ where

$$g_2(z,\theta,v) = \frac{2\lambda^2+|z|^2}{r} + 4(1-v)\log\left(\frac{1-\eta}{\eta}\right) + \log(1+\lambda). \tag{6.12}$$

6.1.3 Combining the Cases

We combine the two bounds in (6.11) and (6.12) to obtain

$$g(z,\theta,v) \leq g_1(z,\theta,v)\mathbb{I}(z \in \mathcal{A}_1) + g_2(z,\theta,v)\mathbb{I}(z \in \mathcal{A}_2). \tag{6.13}$$

Here, we are pasting two risk bounds depending on the two scenarios for $|z+\theta|$ as opposed to pasting two estimators [21]. Next, we define

$$f(\theta,z,v,\eta) = 2z\theta\left(1-\frac{1}{\sqrt{v}}\right)\mathbb{I}(z \in \mathcal{A}_1)$$

which occurs in the bound $g_1(z,\theta,v)$. It turns out that, on the event $\mathcal{A}_1$, this term averages out to a negative value, i.e. we have $E\,\mathbb{I}(z \in \mathcal{A}_1)f(\theta,z,v,\eta) < 0$. Indeed, with $c_\eta = \lambda - h_\eta^U$

$$E\,\mathbb{I}(z \in \mathcal{A}_1)(\theta z) = \int_{z>-\theta+c_\eta} \theta z\phi(z) + \int_{z<-\theta-c_\eta} \theta z\phi(z) = \theta[\phi(-\theta+c_\eta) - \phi(-\theta-c_\eta)] > 0$$

which yields $E\,\mathbb{I}(z \in \mathcal{A}_1)f(\theta,z,v,\eta) < 0$ because $1-1/\sqrt{v} < 0$.

Combining (6.11), (6.12) and (6.7) and using the fact that $|z|$ has a folded normal distribution with $E|z| = \sqrt{2/\pi}$ and $E|z|^2 = 1$

$$Eg(z,\theta,v) \leq \log\left(4\sqrt{2\pi/v}\right) + \frac{4}{\lambda^2} + \frac{\lambda^2}{2} + \frac{2\lambda^2+1}{r} + \lambda(1-\sqrt{v})\sqrt{2/\pi} \tag{6.14}$$

$$+ 5(1-v)\log\left(\frac{1-\eta}{\eta}\right) + \log(2\lambda + \lambda^2 + 1). \tag{6.15}$$

With fixed $r \in (0,\infty)$, fixed $\lambda > 0$ and $(1-\eta)/\eta = n/s_n$, we can see that the upper bound is of the order $(1-v)\log(n/s_n)$. What is concerning is the case when r is very small which make take a large n for the term $(1-v)\log(n/s_n)$ to dominate. It might be worthwhile to consider tuning λ according to r to make sure the the bound is valid also for smaller n.

We now consider two scenarios: (a) $r > 1$ when the variance of the future observation y is larger than that of x, and (b) $0 < r < 1$ when the training data is noisier. The case (a) $r > 1$ implies $v > 1/2$ and for calibrations $1/[(1-v)\log(n/s_n)] \lesssim \lambda^2 \lesssim (1-v)\log(n/s_n)$ (which includes setting λ equal to a fixed constant that does not depend on n) imply $\rho(\theta,\hat{p}) \lesssim (1-v)\log(n/s_n)$. For instance, we can choose $\lambda^2 = C_r^*(1-v)$ for $C_r^* > 0$ such that $C_r^* > 2/[5(1-v)]$. Then

$$\rho(\theta,\hat{p}) \leq 5(1-v)\log(n/s_n) + \widetilde{C}_r^*, \tag{6.16}$$

where

$$\widetilde{C}_r^* = \log(8\sqrt{\pi}) + 10 + 2C_r^*(1-v) + C_r^*\sqrt{2/\pi} + \log(2\sqrt{C_r^*(1-v)} + C_r^*(1-v)). \quad (6.17)$$

In the case (b) $0 < r < 1$ we have $0 < v < 1/2$. We can choose $\lambda^2 = C_r v$ for $C_r > \frac{2}{v(1/2+4)}$. This yields

$$\frac{4}{\lambda^2} + \frac{\lambda^2}{2} + \frac{2\lambda^2+1}{r} + \lambda(1-\sqrt{v})\sqrt{2/\pi} < 9 + \frac{\lambda^2(1/2+4)+2}{r} + \sqrt{C_r v}(1-\sqrt{v})\sqrt{2/\pi}$$
$$< 9 + 9\frac{C_r}{r+1} + \sqrt{C_r v}(1-\sqrt{v})\sqrt{2/\pi}$$

and

$$\log\left(4\sqrt{2\pi/v}\right) + \log(2\lambda + \lambda^2 + 1) < \log(8\sqrt{2\pi}) + \log(2\sqrt{C_r} + C_r)$$

and $\rho(\theta, \hat{p}) \leq 5(1-v)\log(n/s_n) + \widetilde{C}_r$, where

$$\widetilde{C}_r = \log(8\sqrt{2\pi}) + 9 + \frac{C_r}{r+1}(1/2+4) + \sqrt{C_r}\sqrt{1/(8\pi)} + \log(2\sqrt{C_r} + C_r). \quad (6.18)$$

6.1.4 The case when $\theta = 0$

This step is analogous to Section 2 in [21]. Because $N_{\theta,v}(z) > 1$ and from the Jensen's inequality and the fact that $E\mathrm{e}^{cZ} = \mathrm{e}^{-c^2/2}$

$$E\log N_{\theta,v}(Z) \leq \log EN_{\theta,v}(Z) = 1 + \frac{\eta}{1-\eta}\int \exp(\mu\theta/v)\pi_1(\mu \mid \lambda)\mathrm{d}\mu$$

we have for $\theta = 0$

$$\rho(0, \hat{p}) < \log\left(1 + \frac{\eta}{1-\eta}\right) < \frac{\eta}{1-\eta}. \quad (6.19)$$

The statement (3.8) of Theorem 3 follows from the fact that, for separable (independent product) priors,

$$\rho(\boldsymbol{\theta}, \hat{p}) = \sum_{i=1}^{n} \rho(\theta_i, \hat{p}) \leq (n - s_n)\rho(0, \hat{p}) + s_n \sup_{\theta\in\mathbb{R}} \rho(\theta, \hat{p}).$$

Plugging the inequalities (6.16) and (6.19) into the expression above, we obtain the desired statement.

References

[1] Aitchison, J. (1975). Goodness of prediction fit. *Biomerika 62*, 547–554.

[2] Bai, R., V. Ročková, and E. George (2020). Spike-and-Slab Meets LASSO: A Review of the Spike-and-Slab LASSO. *arXiv:2010.06451*, 1–30.

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

[4] Birnbaum (1942). An inequality for Mill's ratio. *Annals of Mathematical Statistics 13*, 245–246.

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

[6] Castillo, I., J. Schmidt-Hieber, and A. van der Vaart (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics 43*, 1986–2018.

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

[8] Deshpande, S., Ročková, and E. George (2019). Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *28*, 921–931.

[9] George, E., F. Liang, and X. Xu (2006). Improved minimax predictive densities under kullback-leibler loss. *The Annals of Statistics 34*, 78–91.

[10] George, E., F. Liang, and X. Xu (2012). From minimax shrinkage estimation to minimax shrinkage prediction. *Statistical Science 27*, 1102–1130.

[11] George, E. and X. Xu (2008). Predictive density estimation for multiple regression. *Econometric Theory 24*, 528–544.

[12] Hans, C. (2009). Bayesian LASSO regression. *Biometrika 96*, 835–845.

[13] Hoffmann, M. Rousseau, J. and J. Schmidt-Hieber (2015). On adaptive posterior concentration rates. *The Annals of Statistics 43*, 2259–2295.

[14] Johnstone, I. M. and B. W. Silverman (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics 32*, 1594–1649.

[15] Karp, D. and S. Sitnik (2009). Inequalities and monotonicity of ratios for generalized hypergeometric function. *Journal of Approximation Theory 161*, 337–352.

[16] Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika 88*, 859–864.

[17] Liang, F. and A. Barron (2004). Exact minimax strategies for predictive denstiy estimation, data compression, and model selection. *IEEE Transactions of Information Theory 50*, 2708–2723.

[18] Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*, 1023–1032.

[19] Mitrinovć, D., J. Pecarić, and M. Fink (1993). *Classical and new inequalities in Analysis*. Kluwer Academic Publishers.

[20] Moran, G., Ročková, and E. George (2021). Spike-and-Slab LASSO biclustering. *Annals of Applied Statistics 15*, 148–173.

[21] Mukherjee, G. and I. Johnstone (2015). Exact minimax estimation of the predictive density in sparse gaussian models. *The Annals of Statistics 43*, 81–106.

[22] Mukherjee, G. and I. Johnstone (2022). On minimax optimality of sparse Bayes predicitve density estimates. *The Annals of Statistics 50*, 81–106.

[23] Nie, L. and V. Ročková (2022). Bayesian bootstrap Spike-and-Slab LASSO. *Journal of the American Statistical Association (in press) 1*, 1–50.

[24] Park, T. and G. Casella (2008). The Bayesian LASSO. *Journal of the American Statistical Association 103*, 681–686.

[25] Ray, K. and B. Szabo (2022). Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association 117*, 1270–1281.

[26] Ročková (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics 46*, 401–437.

[27] Ročková, V. and E. George (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association 111*, 1608–1622.

[28] Ročková, V. and E. George (2018). The Spike-and-Slab LASSO. *113*, 431–444.

[29] Ročková, V. and J. Rousseau (2023). Ideal Bayesian spatial adaptation. *Journal of the American Statistical Association (In Press)*, 1–80.

[30] Tibshirani, R. (1994). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B 58*, 267–288.

[31] van der Pas, S., B. Kleijn, and A. van der Vaart (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics 8*, 2585–2618.

Appendix

Table of Contents

A Proof of Theorem 1

Similarly as in the proof of Theorem 3, we separate the cases when $\theta \neq 0$ and when $\theta = 0$.

A.1 The case when $\theta \neq 0$.

We find an upper bound on $\rho(\theta, \hat{p})$ for the case when $\theta \neq 0$ using Lemma 1. Using the prior (2.1), we have

$$\log N_{\theta,v}^{LASSO}(z) = \log \left[\frac{\lambda}{2}(I_1^v + I_2^v)\right]$$

where I_1^v and I_2^v were defined in (6.2). Recall also the definition of μ_1 and μ_2 in (6.3). To find and upper and lower bound to $\log N_{\theta,v}(z)$, we consider two plausible scenarios (a) $\mu_1 > -\mu_2$ which is equivalent to $I_1^v > I_2^v$ and to $\{z > -\theta/\sqrt{v}\}$, and (b) $\mu_1 \leq -\mu_2$ which is equivalent to $I_2^v \geq I_1^v$ and to $\{z \leq -\theta/\sqrt{v}\}$.

A.1.1 Upper bound on $E \log N_{\theta,v}^{LASSO}(Z)$

We consider the two cases (a) $I_1^v > I_2^v$ and (b) $I_1^v \leq I_2^v$ and write

$$E \log N_{\theta,v}^{LASSO}(Z) \leq \log(\lambda\sqrt{2\pi v}) + T_1^v(\theta) + T_2^v(\theta) \tag{S1}$$

where

$$T_1^v(\theta) = E\,\mathbb{I}(z > -\theta/\sqrt{v})\mu_1^2/(2v) \quad \text{and} \quad T_2^v(\theta) = E\,\mathbb{I}(z \leq -\theta/\sqrt{v})\mu_2^2/(2v). \tag{S2}$$

We find that

$$T_1^v(\theta) + T_2^v(\theta) = E\frac{(|z + \theta/\sqrt{v}| - \lambda\sqrt{v})^2}{2} = \frac{\theta^2}{2v} + \frac{1 + \lambda^2 v}{2} - \lambda E|z\sqrt{v} + \theta|. \tag{S3}$$

A.1.2 Lower bound on $E \log N_{\theta,v}^{LASSO}(Z)$

Considering again the two cases (a) $I_1^v > I_2^v$ and (b) $I_1^v \leq I_2^v$ we find that

$$E \log N_{\theta,v}^{LASSO}(z) > \log(\lambda\sqrt{\pi v/2}) + T_1^v(\theta) + T_2^v(\theta) + T_v^3(\theta) \tag{S4}$$

where $T_1^v(\theta)$ and $T_1^v(\theta)$ were defined earlier in (S2) and where (because in the case (a) $\mu_1 > -\lambda v$ and in the case (b) $-\mu_2 > -\lambda v$)

$$T_v^3(\theta) = [P(\mu_1 > 0) + P(\mu_2 < 0)] \log 1/2 + P\left(|z + \frac{\theta}{\sqrt{v}}| < \lambda\sqrt{v}\right) \log \Phi(-\lambda\sqrt{v}). \qquad \text{(S5)}$$

Recall the definition of the Gaussian Mills ratio $R(x) = (1 - \Phi(x))/\phi(x)$. Then

$$T_v^3(\theta) > \log 1/2 + P\left(|z + \frac{\theta}{\sqrt{v}}| < \lambda\sqrt{v}\right)\left(\log \phi(\lambda\sqrt{v}) + \log R(\lambda\sqrt{v})\right).$$

Next, using the lower bound on the Gaussian Mills ratio in Lemma 13 we have

$$T_v^3(\theta) > \log 1/2 + P\left(|z + \frac{\theta}{\sqrt{v}}| < \lambda\sqrt{v}\right)\left(-\log\sqrt{2\pi} - \frac{\lambda^2 v}{2} - \log\frac{\sqrt{\lambda^2 v + 4} + \lambda\sqrt{v}}{2}\right). \qquad \text{(S6)}$$

A.1.3 Combining the bounds

Combining the upper bound for $E \log N_{\theta,1}^{LASSO}(Z)$ and the lower bound for $E \log N_{\theta,v}^{LASSO}(z)$, Lemma 1 yields that for any $\theta \in \mathbb{R}$

$$\rho(\theta, \hat{p}) \leq \log(2/\sqrt{v}) + \frac{\lambda^2(1-v)}{2} + \lambda E|z\sqrt{v} + \theta| - \lambda E|z + \theta| - T_v^3(\theta).$$

Next, we use the fact that $|z\sqrt{v} + \theta| - |z + \theta| \leq |z|(1 - \sqrt{v})$ and that $|Z|$ has a folded normal distribution with a mean $\sqrt{2/\pi}$. From the bound

$$-T_v^3(\theta) < \log(2\sqrt{2\pi}) + \frac{\lambda^2 v}{2} + \log\frac{\sqrt{\lambda^2 v + 4} + \lambda\sqrt{v}}{2}$$

we obtain

$$\rho(\theta, \hat{p}) \leq \log(\lambda 4\sqrt{2\pi/v}) + \frac{\lambda^2}{2} + \lambda\sqrt{2/\pi} + \log\frac{\sqrt{1 + 4/(\lambda^2)} + 1}{2}.$$

Using Lemma 12 we obtain the desired upper bound on $\rho(\theta, \hat{p})$.

A.2 The case when $\theta = 0$.

By the Jensen's inequality $E \log X \leq \log EX$ we find from the Fubini's theorem and from the fact that $E\exp(\mu Z/\sqrt{v}) = \exp(\mu^2/(2v))$

$$E \log N_{0,1}^{LASSO}(Z) < \log E \int \exp\left\{\mu\, Z/\sqrt{v} - \frac{\mu^2}{2v}\right\} \pi_1(\mu \mid \lambda) \mathrm{d}\mu = 0.$$

This yields $\rho(0,\hat{p}) \leq -E\log N^{LASSO}_{0,v}(z)$. To find a lower bound for $E\log N^{LASSO}_{0,v}(z)$, we use the notation introduced I_1^v and I_2^v in (6.2) but now for the special case when $\theta = 0$. Similarly as in Section A.1 we consider two cases (a) when $z > 0$ we have $I_1^v > I_2^v$ and (b) when $z \leq 0$ we have $I_1^v \leq I_2^v$. Next, in the case (a) (since $\theta = 0$) we have $\mu_2 = z\sqrt{v} + \lambda v > 0$ and thereby we can use Lemma 13 which yields

$$\frac{2}{\sqrt{\mu_2^2/v+4}+\mu_2/\sqrt{v}} < I_2^v = \frac{1-\Phi(\mu_2/\sqrt{v})}{\phi(\mu_2/\sqrt{v})} < \frac{2}{\sqrt{\mu_2^2/v+2}+\mu_2/\sqrt{v}}. \tag{S7}$$

Similarly, in the case (b) we have $\mu_1 = z\sqrt{v} - \lambda v < 0$ and thereby

$$\frac{2}{\sqrt{\mu_1^2/v+4}-\mu_1/\sqrt{v}} < I_1^v = \frac{1-\Phi(-\mu_1/\sqrt{v})}{\phi(\mu_1/\sqrt{v})} < \frac{2}{\sqrt{\mu_1^2/v+2}-\mu_1/\sqrt{v}}. \tag{S8}$$

While here we only need the lower bounds, in the next Section B we utilize also the upper bounds. This yields

$$E\log\left[\frac{\lambda}{2}(I_1^v+I_2^v)\right] > \log(\lambda\sqrt{v}) - E\log(\lambda\sqrt{v}+|z|) - E\log\frac{\sqrt{\frac{4}{(\lambda\sqrt{v}+|z|)^2}+1}+1}{2} \tag{S9}$$

We use the Jensen's inequality $E\log(\lambda\sqrt{v}+|z|) \leq \log(\lambda\sqrt{v}+\sqrt{2/\pi})$ and Lemma 12 which yields

$$E\log\frac{\sqrt{\frac{4}{(\lambda\sqrt{v}+|z|)^2}+1}+1}{2} < \log\frac{\sqrt{\frac{4}{\lambda^2 v}+1}+1}{2} < \frac{4}{\lambda^2 v}.$$

Altogether, we find

$$\rho(0,\hat{p}) < -E\log\left[\frac{\lambda}{2}(I_1^v+I_2^v)\right] < \log\left(1+\frac{\sqrt{2}}{\lambda\sqrt{\pi v}}\right) + \frac{4}{\lambda^2 v}.$$

B Proof of Theorem 2

We know that for separable priors (such as the Laplace product prior (2.1)) we have

$$(n-s_n)\rho(0,\hat{p}) < \rho(\boldsymbol{\theta},\hat{p}) = \sum_{i=1}^{n}\rho(\theta_i,\hat{p}) = (n-s_n)\rho(0,\hat{p}) + \sum_{i:\theta_i\neq 0}\rho(\theta_i,\hat{p}).$$

We focus on the lower part of these inequalities and obtain a lower bound for

$$\rho(0,\hat{p}) = E\log N^{LASSO}_{1,0}(z) - E\log N^{LASSO}_{v,0}(z).$$

First, we recall the lower bound for $E \log N_{1,0}^{LASSO}(z)$ in (S9) obtained in Section A.2.

$$E \log N_{1,0}^{LASSO}(Z) > -E \log \left(1 + \frac{|z|}{\lambda}\right) + E \log \frac{2}{\sqrt{\frac{4}{(\lambda+|z|)^2} + 1} + 1}.$$

Now, we obtain an upper bound for $E \log N_{v,0}^{LASSO}(z) = E \log[\frac{\lambda}{2}(I_1^v + I_2^v)]$ using similar ideas as in Section A.2. Recall the notation I_1^v and I_2^v in (6.2) and μ_1 and μ_2 in (6.3). These quantities now tacitly assume $\theta = 0$. We again consider two cases (a) $z > 0$ and (b) $z \leq 0$ but we split them further depending on whether $|z| > \lambda\sqrt{v}$ or $|z| \leq \lambda\sqrt{v}$. In the case (a) the term I_1^v dominates I_2^v and when $\mu_1 \leq 0$ (i.e. $z \leq \lambda\sqrt{v}$) we can use the upper part in the Mills ratio bounds (S8). Similarly, in the case (b) when $\mu_2 < 0$ (i.e. $z > -\lambda\sqrt{v}$) we can use the upper part in the Mills ratio bound (S7). This yields

$$\begin{aligned}
&E\mathbb{I}(|z| \leq \lambda\sqrt{v}) \log\left[\frac{\lambda}{2}(I_1^v + I_2^v)\right] \\
&\quad \leq -E\,\mathbb{I}(|z| \leq \lambda\sqrt{v}) \log\left(1 + \frac{|z|}{\lambda\sqrt{v}}\right) + E\,\mathbb{I}(|z| \leq \lambda\sqrt{v}) \log \frac{2}{\sqrt{\frac{2}{(\lambda\sqrt{v}+|z|)^2} + 1} + 1} \\
&\quad \leq -E \log\left(1 + \frac{|z|}{\lambda\sqrt{v}}\right) + E\,\mathbb{I}(|z| > \lambda\sqrt{v}) \log\left(1 + \frac{|z|}{\lambda\sqrt{v}}\right) + \log \frac{2}{\sqrt{\frac{1}{2\lambda^2 v} + 1} + 1}.
\end{aligned}$$

Next we find that (using the fact that $|z|$ has a folded normal distribution with a density $2/\sqrt{2\pi}\mathrm{e}^{-z^2/2}$)

$$E\,\mathbb{I}(|z| > \lambda\sqrt{v}) \log\left(1 + \frac{|z|}{\lambda\sqrt{v}}\right) < \frac{2}{\sqrt{2\pi}} \int_{\lambda\sqrt{v}}^{\infty} \frac{z}{\lambda\sqrt{v}} \mathrm{e}^{-z^2/2} \mathrm{d}z = \frac{2}{\sqrt{2\pi}} \frac{\mathrm{e}^{-\lambda^2 v/2}}{\lambda\sqrt{v}}$$

and thereby

$$E\,\mathbb{I}(|z| \leq \lambda\sqrt{v}) \log N_{v,0}^{LASSO}(z) < -E \log\left(1 + \frac{|z|}{\lambda\sqrt{v}}\right) + \frac{2}{\sqrt{2\pi}} \frac{\mathrm{e}^{-\lambda^2 v/2}}{\lambda\sqrt{v}}.$$

For the remaining scenario when $|z| > \lambda\sqrt{v}$ we bound (using the Gaussian tail bound $(1 - \Phi(x)) \leq \mathrm{e}^{-x^2/2}$ for $x > 0$)

$$\begin{aligned}
E\,\mathbb{I}(|z| > \lambda\sqrt{v}) \log N_{0,v}^{LASSO}(z) \leq& \log(\lambda\sqrt{2\pi v})(1 - \Phi(\lambda\sqrt{v})) \\
&+ E[\,\mathbb{I}(z > \lambda\sqrt{v})\mu_1^2/(2v) + \mathbb{I}(z < -\lambda\sqrt{v})\mu_2^2/(2v)] \\
\leq& \log(\lambda\sqrt{2\pi v})\mathrm{e}^{-\lambda^2 v/2} + E(|z| - \lambda\sqrt{v})^2 \mathbb{I}(|z| \geq \lambda\sqrt{v})/2.
\end{aligned}$$

We note that

$$E(|z|-\lambda\sqrt{v})^2\mathbb{I}(|z|\geq\lambda\sqrt{v})=\frac{2}{\sqrt{2\pi}}\int_{\lambda\sqrt{v}}^{\infty}(z-\lambda\sqrt{v})^2\mathrm{e}^{-z^2/2}<\frac{2\mathrm{e}^{-\lambda^2 v/2}}{\sqrt{2\pi}}\int_0^\infty y^2\mathrm{e}^{-y^2/2}=\mathrm{e}^{-\lambda^2 v/2}.$$

Putting the bounds together, we have

$$\rho(0,\hat{p})>E\log\left(\frac{1+\frac{|z|}{\lambda\sqrt{v}}}{1+\frac{|z|}{\lambda}}\right)-\mathrm{e}^{-\lambda^2 v/2}\left(\frac{1}{2}+\log(\lambda\sqrt{2\pi v})+\frac{2}{\lambda\sqrt{2\pi v}}\right)+E\log\frac{2}{\sqrt{\frac{4}{(\lambda+|z|)^2}+1}+1}.$$

We use Lemma 12 to find that

$$E\log\frac{\sqrt{\frac{4}{(\lambda+|z|)^2}+1}+1}{2}<\log\frac{\sqrt{\frac{4}{\lambda^2}+1}+1}{2}<\frac{\sqrt{\frac{4}{\lambda^2}+1}-1}{2}<\frac{4}{\lambda^2}.$$

Next, for any $a>0$

$$E\log\left(1+\frac{\left(\frac{1}{\sqrt{v}}-1\right)\frac{|z|}{\lambda}}{1+\frac{|z|}{\lambda}}\right)>P(|z|\geq a)\log\left(1+\left(\frac{1}{\sqrt{v}}-1\right)\frac{a}{\lambda+a}\right).$$

Altogether, we have for any $a>0$

$$\rho(0,\hat{p})>P(|z|\geq a)\log\left[1+\left(\frac{1}{\sqrt{v}}-1\right)\frac{a}{\lambda+a}\right]-\frac{4}{\lambda^2}-\mathrm{e}^{-\lambda^2 v/2}\left(\frac{1}{2}+\log(\lambda\sqrt{2\pi v})+\frac{2}{\lambda\sqrt{2\pi v}}\right).$$

C Proof of Theorem 4

Recall the Spike-and-Slab LASSO prior

$$\pi(\theta)=\eta\pi_1(\theta)+(1-\eta)\pi_0(\theta)$$

where $\pi_0(\mu)=\lambda_0/2\mathrm{e}^{-\lambda_0|\mu|}$ and $\pi_1(\mu)=\lambda_1/2\mathrm{e}^{-\lambda_1|\mu|}$ for $\lambda_1<\lambda_0$. We also recall the definitions (6.2) and (6.3) but we now explicitly state their dependence on λ. We define

$$I_1^v(\lambda)=\sqrt{v}\frac{\Phi(\mu_1^v(\lambda)/\sqrt{v})}{\phi(\mu_1^v(\lambda)/\sqrt{v})}\quad\text{and}\quad I_2^v(\lambda)=\sqrt{v}\frac{\Phi(-\mu_2^v(\lambda)/\sqrt{v})}{\phi(-\mu_2^v(\lambda)/\sqrt{v})}$$

where

$$\mu_1^v(\lambda)=z\sqrt{v}+\theta-v\lambda\quad\text{and}\quad\mu_2^v(\lambda)=z\sqrt{v}+\theta+v\lambda.$$

We denote the rescaled ratio of two marginal likelihoods $\frac{\lambda_1}{\lambda_0}\frac{m_0(x)}{m_1(x)}$ for $x=z+\theta$ by

$$R_{\lambda_0,\lambda_1}(z)=\frac{I_1^1(\lambda_0)+I_2^1(\lambda_0)}{I_1^1(\lambda_1)+I_2^1(\lambda_1)}.$$

We can use Lemma 4 to decompose the prediction risk

$$\rho(\theta, \hat{p}) = \rho(\theta, \hat{p}_1) + E \log N^{SS}_{\theta,1}(z) - E \log N^{SS}_{\theta,v}(z) \tag{S1}$$

where

$$N^{SS}_{\theta,v}(z) = 1 + \frac{1-\eta}{\eta}\frac{\lambda_0}{\lambda_1} R_{\lambda_0,\lambda_1}(z).$$

Alternatively, we could write

$$\rho(\theta, \hat{p}) = \rho(\theta, \hat{p}_0) + E \log \widetilde{N}^{SS}_{\theta,1}(z) - E \log \widetilde{N}^{SS}_{\theta,v}(z) \tag{S2}$$

where

$$\widetilde{N}^{SS}_{\theta,v}(z) = 1 + \frac{\eta}{1-\eta}\frac{\lambda_1}{\lambda_0}\frac{1}{R_{\lambda_0,\lambda_1}(z)}.$$

We will utilize both of these characterizations. The decomposition (S1) is useful when the observed data x is large in magnitude, because we would expect the slab risk to be the dominant term (according to Lemma 3). The opposite is true when x is small. While the upper bound on the risk in Lemma 3 uses an average mixing weight $\Delta_\eta(x)$, we pause with averaging over the distribution $\pi(x \mid \theta)$ and bound the Kullback-Leibler loss in two regimes depending on the magnitude of x. Below, we show that $(n - s_n)\rho(0, \hat{p}) \lesssim s_n$ and $s_n \sup_{\theta\in\mathbb{R}} \rho(\theta, \hat{p}) \lesssim (1 - v)s_n \log(n/s_n)$ which will conclude the proof of the theorem.

C.1 The case when $\theta = 0$.

We utilize the expression (S2) and (using the risk expression at $\theta = 0$ for Bayesian LASSO from Section A.2) we find that

$$\rho(0, \hat{p}) \leq \frac{\sqrt{2}}{\lambda_0\sqrt{\pi v}} + \frac{4}{\lambda_0^2 v} + \log\left(1 + \frac{\eta}{1-\eta} E_{x\,|\,\theta=0}\frac{m_1(x)}{m_0(x)}\right) \tag{S3}$$

In order to bound the expectation $E_{x\,|\,\theta=0}\frac{m_1(x)}{m_0(x)}$, we consider 4 possible cases.

(1) When $x = z > \lambda_0 > 0$, we have $I_1^1(\lambda) > I_2^1(\lambda)$ and because $\mu_1^1(\lambda_0) = z - \lambda_0 > 0$ (and thereby $\Phi(\mu_1^1(\lambda_0)) > 1/2$) we have

$$\frac{m_1(x)}{m_0(x)} = 2\frac{\lambda_1}{\lambda_0} \times \frac{I_1^1(\lambda_1)}{I_1^1(\lambda_0)} < 4\frac{\lambda_1}{\lambda_0} \times \frac{\phi(\mu_1^1(\lambda_0))}{\phi(\mu_1^1(\lambda_1))} = 4\frac{\lambda_1}{\lambda_0} e^{x(\lambda_0-\lambda_1)-\lambda_0^2/2+\lambda_1^2/2}.$$

Then

$$\int_{\lambda_0}^{\infty} \phi(x)\frac{m_1(x)}{m_0(x)}\mathrm{d}x < \frac{4\lambda_1 \mathrm{e}^{-\lambda_0^2/2+\lambda_1^2/2+(\lambda_0-\lambda_1)^2/2}}{\lambda_0}\int_{\lambda_0}^{\infty}\frac{\mathrm{e}^{-[x-(\lambda_0-\lambda_1)]^2/2}}{\sqrt{2\pi}}\mathrm{d}x < \frac{4\lambda_1}{\lambda_0}\mathrm{e}^{-\lambda_0\lambda_1+\lambda_1^2}.$$

(2) When $\lambda_0 > x = z > 0$, we can use Lemma 13 (because $\mu_1^1(\lambda_0) = z - \lambda_0 < 0$) to find

$$\frac{m_1(x)}{m_0(x)} < 2\frac{\lambda_1}{\lambda_0} \times \frac{1}{\phi(\mu_1^1(\lambda_1))} \times \frac{\sqrt{(\lambda_0 - x)^2 + 4} + \lambda_0 - x}{2}.$$

This yields

$$\frac{m_1(x)}{m_0(x)} < 2\frac{\lambda_1}{\lambda_0} \times \sqrt{2\pi}\mathrm{e}^{(x-\lambda_1)^2/2} \times \frac{(1+\sqrt{2})\max\{2, (\lambda_0 - x)\}}{2}$$

and

$$\int_0^{\lambda_0} \phi(x)\frac{m_1(x)}{m_0(x)}\mathrm{d}x < \frac{(1+\sqrt{2})\max\{2, \lambda_0\}}{\lambda_0}\mathrm{e}^{\lambda_1^2/2}(1-\mathrm{e}^{-\lambda_0\lambda_1}).$$

The next two cases are mirror images of the previous too.

(3) When $-\lambda_0 < x = z \leq 0$ we have $I_1^1(\lambda) \leq I_2^1(\lambda)$ and (using Lemma 13)

$$\frac{m_1(x)}{m_0(x)} = 2\frac{\lambda_1}{\lambda_0} \times \frac{1}{\phi(\mu_2^1(\lambda_1))} \times \frac{\sqrt{(\lambda_0 + x)^2 + 4} + \lambda_0 + x}{2}.$$

This yields

$$\frac{m_1(x)}{m_0(x)} = 2\frac{\lambda_1}{\lambda_0} \times \sqrt{2\pi}\mathrm{e}^{(x+\lambda_1)^2/2} \times \frac{(1+\sqrt{2})\max\{2, (\lambda_0 + x)\}}{2}$$

and

$$\int_{-\lambda_0}^{0} \phi(x)\frac{m_1(x)}{m_0(x)}\mathrm{d}x \leq \frac{(1+\sqrt{2})\max\{2, \lambda_0\}}{\lambda_0}\mathrm{e}^{\lambda_1^2/2}(1-\mathrm{e}^{-\lambda_0\lambda_1}).$$

(4) When $x = z \leq -\lambda_0 < -\lambda_1$ we have $-\mu_2^1(\lambda_0) = -z - \lambda_0 > 0$ and thereby

$$\frac{m_1(x)}{m_0(x)} = 4\frac{\lambda_1}{\lambda_0} \times \frac{\phi(\mu_2^1(\lambda_0))}{\phi(\mu_2^1(\lambda_1))} = 4\frac{\lambda_1}{\lambda_0}\mathrm{e}^{-x(\lambda_0-\lambda_1)-\lambda_0^2/2+\lambda_1^2/2}.$$

and

$$\int_{-\infty}^{-\lambda_0} \phi(x)\frac{m_1(x)}{m_0(x)}\mathrm{d}x < \frac{4\lambda_1 \mathrm{e}^{-\lambda_0^2/2+\lambda_1^2/2+(\lambda_0-\lambda_1)^2/2}}{\lambda_0}\int_{-\infty}^{-\lambda_0}\frac{\mathrm{e}^{-[z+(\lambda_0-\lambda_1)]^2/2}}{\sqrt{2\pi}}\mathrm{d}z = \frac{4\lambda_1}{\lambda_0}\mathrm{e}^{-\lambda_0\lambda_1+\lambda_1^2}.$$

Putting all the pieces together, we find that keeping $\lambda_1 = \sqrt{v}C_v$ for some $C_v > 0$ we have for some $C_1 > 0$

$$E_{x\,|\,\theta=0}\frac{m_1(x)}{m_0(x)} < C_1.$$

With $\eta/(1-\eta) = n/s_n$ and $\lambda_0\sqrt{v} = n/s_n$ we find that $\rho(0,\hat{p}) \lesssim s_n/n$.

C.2 The case when $\theta \neq 0$

We consider two cases (similarly as in the proof of Theorem 3) for some $A > 0$:

Case i: On the event $A_\eta(\theta, A, d)^c \equiv \{z : R_{\lambda_0,\lambda_1}(z) > A(s_n/n)^d\}$ we have

$$\log \widetilde{N}_{\theta,v}^{SS}(z) \leq \log \left[1 + \left(\frac{\eta}{1-\eta}\right) \frac{\lambda_1}{\lambda_0} (n/s_n)^d / A\right].$$

Case ii: On the event $A_\eta(\theta, A, d) \equiv \{z : R_{\lambda_0,\lambda_1}(z) \leq A(s_n/n)^d\}$ we have

$$\log N_{\theta,v}^{SS}(z) \leq \log \left[1 + \left(\frac{1-\eta}{\eta}\right) \frac{\lambda_0}{\lambda_1} A (n/s_n)^{-d}\right].$$

Recall (from (S3)) that under the Laplace prior with a parameter λ, the KL loss equals (using the expression $x + y/r = z/\sqrt{v} + \theta/v$)

$$K_\lambda(\theta, z) = \theta^2/(2r) + \log[\lambda/2(I_1^1(\lambda) + I_2^1(\lambda))] - E_{y\,|\,\theta} \log[\lambda/2(I_1^v(\lambda) + I_2^v(\lambda))]. \tag{S4}$$

This means

$$\begin{aligned} \rho(\theta, \hat{p}) \leq & P[A_\eta(\theta, A, d)^c] \log \left[1 + \left(\frac{\eta}{1-\eta}\right) \frac{\lambda_1}{\lambda_0} (n/s_n)^d / A\right] + \int_{z \in A_\eta^c(\theta,A,d)} K_{\lambda_0}(\theta, z)\phi(z) \mathrm{d}z \\ & + P[A_\eta(\theta, A, d)] \log \left[1 + \left(\frac{1-\eta}{\eta}\right) \frac{\lambda_0}{\lambda_1} A (n/s_n)^{-d}\right] + \int_{z \in A_\eta(\theta,A,d)} K_{\lambda_1}(\theta, z)\phi(z) \mathrm{d}z. \end{aligned} \tag{S5}$$

Now we focus on the properties of the set $A_\eta^c(\theta, A, d)$.

Lemma 10. *For $\lambda_0/\lambda_1 = n/s_n$ and $\lambda_1 > 0$ is a fixed constant. When $0 < r < 1$, there exists $A > 0$ such that $A_\eta(\theta, A, 2v)^c = \emptyset$. When $r \in [1, \infty)$, we have*

$$A_\eta(\theta, A, 2v)^c \subset \{x : |x| < \widetilde{x}_c\},$$

where $\widetilde{x}_c = \lambda_1 + \sqrt{4v \log[n/s_n]}$ and $A = [\sqrt{2\pi}(1 + \sqrt{2})]^{-1}$.

Proof. The ratio $R_{\lambda_0,\lambda_1}(z)$ as a function of $x = z + \theta$ has a maximal value at $x = 0$ where (using the Mills ratio notation $R(x) = (1 - \Phi(x))/\phi(x)$ and Lemma 13)

$$R_{\lambda_0,\lambda_1}(z) \leq R_{\lambda_0,\lambda_1}(-\theta) = \frac{R(\lambda_0)}{R(\lambda_1)} < \frac{\lambda_1}{\lambda_0} \frac{\sqrt{1 + 4/\lambda_1^2} + 1}{\sqrt{1 + 2/\lambda_0^2} + 1} < \frac{s_n}{n} \frac{\sqrt{1 + 4/\lambda_1^2} + 1}{2}.$$

This means that for $A = \frac{\sqrt{1+4/\lambda_1^2}+1}{2}$ we have $R_{\lambda_0,\lambda_1}(z) < As_n/n$ which is strictly smaller than $A(s_n/n)^{2v}$ when $0 < v < 1/2$ (i.e. when $0 < r < 1$). Now we look into the case when $d = 2v \geq 1$. Because the ratio $m_0(x)/m_1(x)$ is symmetrical around zero and monotone decreasing on $(0, \infty)$, we have

$$A_\eta(\theta, A, d)^c = \left\{ x = z + \theta : \frac{m_0(x)}{m_1(x)} > \frac{\lambda_0}{\lambda_1} A(s_n/n)^d \right\} = \{x : |x| \leq x_c\},$$

where

$$\frac{m_0(x_c)}{m_1(x_c)} = \lambda_0/\lambda_1 A(s_n/n)^d.$$

We now find an upper bound to $\frac{m_0(x)}{m_1(x)}$. We first consider the case when $\lambda_1 < x < \lambda_0 - 2$. On this interval (similarly as in the case (2) in Section **??**) we obtain

$$\frac{m_0(x)}{m_1(x)} \leq \frac{\lambda_0}{\lambda_1} \frac{2\phi(\mu_1^1(\lambda_1))}{\sqrt{(\lambda_0 - x)^2 + 2} + \lambda_0 - x} \leq \frac{\lambda_0}{\lambda_1} \frac{2\phi(\mu_1^1(\lambda_1))}{\min\{2, \lambda_0 - x\}(1 + \sqrt{2})}.$$

Setting this upper bound equal to $A(s_n/n)^d$ yields

$$A_\eta(\theta, A, d)^c \subset \{x : |x| < \tilde{x}_c\},$$

where $\tilde{x}_c = \lambda_1 + \sqrt{2d \log[n/s_n]}$ and $A = [\sqrt{2\pi}(1 + \sqrt{2})]^{-1}$. When $0 < x < \lambda_1$, we have

$$\frac{m_0(x)}{m_1(x)} > \frac{\lambda_0}{2\lambda_1} \frac{\sqrt{(\lambda_1 - x)^2 + 2} + \lambda_1 - x}{\sqrt{(\lambda_0 - x)^2 + 4} + \lambda_0 - x}.$$

for $\lambda_0 = n/s_n$. For $d \geq 1$ and suitable $A > 0$ we will have $\{-\lambda_1, \lambda_1\} \subset A_\eta(\theta, A, d)^c$. Because $\tilde{x}_c < \lambda_0$ when $\lambda_0 = n/s_n$, we conclude that $A_\eta(\theta, A, d)^c \subset \{x : |x| < \tilde{x}_c\}$. □

Now we continue with the proof of Theorem 4.

C.3 When $r \in (0, 1)$

Going back to (S4), we find that for $A = \frac{\sqrt{1+4/\lambda_1^2}+1}{2}$ and $\lambda_1 = \sqrt{v}C_v$ for some $C_v > 0$ and $\lambda_0\sqrt{v} = n/s_n$ and $\eta/(1-\eta) = 1$ we have for $d = 2v$

$$\rho(\theta, \hat{p}) \leq \log\left[1 + A(n/s_n)^{2-2v}/C_v\right] + \rho(\theta, \hat{p}_1),$$

where $\hat{p}_1$ is the predictive distribution under the slab Laplace prior. From the proof of Theorem 1 in Section A.1.3, we know that

$$\rho(\theta,\hat{p}) \leq \log\left[1 + A(n/s_n)^{2-2v}/C_v\right] + \log(\lambda_1 4\sqrt{2\pi/v}) + \lambda_1^2/2 + \lambda_1\sqrt{2/\pi} + 4/\lambda_1^2.$$

Keeping $\lambda_1 = \sqrt{v}C_v$ for some $C_v > 0$, the first term is the dominant term and $\rho(\theta,\hat{p}) \lesssim (1-v)\log(n/s_n)$.

C.4 When $r \in [1,\infty)$

Using (S4), we need to make sure that the event $A_\eta(\theta, A, d)^c$ for $d = 2v$ is small enough when the signal (and $|x|$) is large so that the spike predictive distribution gets silenced. We have from Lemma 10 when $|\theta| > c\sqrt{\log(n/s_n)}$ for some $c > 2d$ (because $|z+\theta| > |\theta| - |z|$)

$$P(A_\eta(\theta,A,d)^c) \leq P(|z| > |\theta| - \widetilde{x}_c) \leq 2\mathrm{e}^{-[(c-\sqrt{2d})\sqrt{\log(n/s_n)}-\lambda_1]^2/2} \leq 2\mathrm{e}^{\lambda_1^2/4}\mathrm{e}^{-(c-\sqrt{2d})^2\log(n/s_n)}.$$

With $(c-\sqrt{2d})^2 \geq 2$ we have $P(A_\eta(\theta,A,d)^c) \lesssim (s_n/n)^2$. Using similar steps as in the proof of Theorem 1 in Section A.1.3 we find

$$\begin{aligned} K_\lambda(z) &\leq \frac{\theta^2}{2r} + \log(2/\sqrt{v}) + \frac{(|z+\theta|-\lambda)^2}{2} - \frac{(|z+\theta/\sqrt{v}|-\lambda\sqrt{v})^2}{2} - \log\Phi(-\lambda\sqrt{v}) \\ &\leq \frac{\theta^2}{2r} + \log(\lambda 4\sqrt{2\pi/v}) + \frac{(|z+\theta|-\lambda)^2}{2} - \frac{(|z+\theta/\sqrt{v}|-\lambda\sqrt{v})^2}{2} + \lambda^2 v/2 + \frac{1}{\lambda^2}. \end{aligned} \tag{S6}$$

On the set $A_\eta(\theta,A,d)^c$ we have $|z+\theta| \leq \widetilde{x}_c$ and

$$K_{\lambda_0}(z) \leq \frac{\theta^2}{2r} + \log(\lambda_0 4\sqrt{2\pi/v}) + \widetilde{x}_c^2 + (1+v)\lambda_0^2/2 + \frac{1}{\lambda_0^2}.$$

With $\lambda_0 = n/s_n$, the dominant term among the last three terms above is $\lambda_0^2(1+v)$. Above, we have shown that when the signal is strong enough we have $P(A_\eta(\theta,A,d)^c) \lesssim 1/\lambda_0^2 = (s_n/n)^2$. This means

$$\int_{A_\eta(\theta,A,d)^c} K_{\lambda_0}(z)\phi(z)\mathrm{d}z \lesssim \theta^2/(2r)P(A_\eta(\theta,A,d)^c) + O(1).$$

Combined with (S4) the term $\frac{\theta^2}{2r}P(A_\eta(\theta,A,d)^c$ can be combined with the term $\frac{\theta^2}{2r}P(A_\eta(\theta,A,d))$ contained inside the slab part. Altogether, we obtain an upper bound that is of the order $(1-v)\log(n/s_n)$.

D Proofs of Lemmata

D.1 Proof of Lemma 1

We have the following risk definition

$$\rho(\theta, \hat{p}) = \int \pi(x \mid \theta) \int \pi(y \mid \theta) \log[\pi(y \mid \theta)/\hat{p}(y \mid x)]\mathrm{d}y\mathrm{d}x. \tag{S1}$$

For the marginal likelihood $m(x) = \int \pi(x \mid \mu)\pi_1(\mu \mid \lambda)\mathrm{d}\mu$, we have

$$\hat{p}(y \mid x) = \frac{\int \pi(y \mid \mu)\pi(\mu \mid x)\mathrm{d}\mu}{m(x)} = \frac{\mathrm{e}^{-y^2/(2r)} \int \exp\{\mu(y/r + x) - \mu^2/2(1 + 1/v)\}\pi_1(\mu \mid \lambda)\mathrm{d}\mu}{\int \exp\{\mu x - \mu^2/2\}\pi_1(\mu \mid \lambda)\mathrm{d}\mu}$$

and

$$\frac{\pi(y \mid \theta)}{\hat{p}(y \mid x)} = \frac{\exp(y\theta/r - \theta^2/2r) \int \exp\{\mu x - \mu^2/2\}\pi_1(\mu \mid \lambda)\mathrm{d}\mu}{\int \exp\{\mu(y/r + x) - \mu^2/2(1 + 1/r)\}\pi_1(\mu \mid \lambda)\mathrm{d}\mu}. \tag{S2}$$

Then

$$\begin{aligned}\log \frac{\pi(y \mid \theta)}{\hat{p}(y \mid x)} =& y\theta/r - \theta^2/2r + \log \int \exp\{\mu x - \mu^2/2\}\pi_1(\mu \mid \lambda)\mathrm{d}\mu \\ & - \log \int \exp\{\mu(y/r + x) - \mu^2/2(1 + 1/r)\}\pi_1(\mu \mid \lambda)\mathrm{d}\mu. \end{aligned} \tag{S3}$$

The expectation of the first term with respect to $Y \sim N(\theta, r)$ is $\theta^2/(2r)$. Since $X \sim N(\theta, 1)$ and $Y \sim N(\theta, r)$, the expectation of the other two terms can be taken with respect to $X + Y/r \sim N(\theta/v, 1/v)$ where $v = 1/(1 + 1/r)$. This is the same as taking an expectation with respect to $Z/\sqrt{v} + \theta/v$. This concludes the proof.

D.2 Proof of Lemma 4

The Kullback-Leibler loss can be written as

$$\begin{aligned} L(\theta, \hat{p}) &= \int \pi(y \mid \theta) \log \frac{\pi(y \mid \theta)}{\Delta_\eta(x)\hat{p}_1(y \mid x) + (1 - \Delta_\eta(x))\hat{p}_0(y \mid x)}\mathrm{d}y \\ &= L(\theta, \hat{p}_1) - \log[\Delta_\eta(x)] - \int \pi(y \mid \theta) \log \left[1 + \frac{1 - \Delta_\eta(x)}{\Delta_\eta(x)} \frac{\hat{p}_0(y \mid x)}{\hat{p}_1(y \mid x)}\right] \mathrm{d}y. \end{aligned}$$

Next, note that

$$-\log[\Delta_\eta(x)] = \log\left(1 + \frac{(1 - \eta)}{\eta} \frac{m_0(x)}{m_1(x)}\right)$$

and

$$\frac{1-\Delta_\eta(x)}{\Delta_\eta(x)}\frac{\hat{p}_0(y\,|\,x)}{\hat{p}_1(y\,|\,x)} = \frac{(1-\eta)}{\eta}\frac{m_0(x)}{m_1(x)}\frac{\hat{p}_0(y\,|\,x)}{\hat{p}_1(y\,|\,x)}.$$

Next,

$$\frac{m_0(x)}{m_1(x)}\frac{\hat{p}_0(y\,|\,x)}{\hat{p}_1(y\,|\,x)} = \frac{\int \exp(\mu(x+y/r)-\mu^2/2(1+1/r))\pi_0(\mu)\mathrm{d}\mu}{\int \exp(\mu(x+y/r)-\mu^2/2(1+1/r))\pi_1(\mu)\mathrm{d}\mu}$$

To obtain $\rho(\theta,\hat{p})$ we take an expectation over $X \sim N(\theta,1)$ since $X+Y/r \sim N(\theta(1+1/r),1+1/r)$ which is the same as taking an expectation with respect to $\theta/v + Z/\sqrt{v}$ for $Z \sim N(0,1)$ and $v = 1/(1+1/r)$. Noting that $-E_{x\,|\,\theta}\log\Delta_\eta(x) = E_z \log N^{SS}_{\theta,1}(z)$ we obtain the desired expression.

D.3 Proof of Lemma 6

Because, given η, the Kullback-Leibler loss is separable, i.e.

$$L(\boldsymbol{\theta},\hat{p}(\cdot\,|\,\boldsymbol{x},\eta)) = \sum_{i=1}^n L(\theta_i,\hat{p}(\cdot\,|\,x_i,\eta)),$$

taking the expectation of both sides of (4.4) over the distribution $\pi(\boldsymbol{x}\,|\,\boldsymbol{\theta})$ yields

$$\rho(\boldsymbol{\theta},\hat{p}) \leq (n-s_n)E_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}E_{\eta\,|\,\boldsymbol{x}}L(0,\hat{p}(\cdot\,|\,x_i,\eta)) + E_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}E_{\eta\,|\,\boldsymbol{x}}\sum_{i:\theta_i\neq 0}L(\theta_i,\hat{p}(\cdot\,|\,x_i,\eta)). \tag{S4}$$

We use a similar expression as in the proof of Lemma 4 in Section D.2. We first focus on the case when $\theta_i \neq 0$. Define

$$\widetilde{g}(x,y,\theta,\eta) = \log\frac{\pi(y\,|\,\theta)}{\hat{p}(y\,|\,x)}$$

Then

$$\widetilde{g}(x,y,\theta,\eta) = \log\frac{\phi((y-\theta)/\sqrt{r})}{\phi(y/\sqrt{r})} - \log[1-\Delta_\eta(x)] - \log\left[1+\frac{\Delta_\eta(x)}{1-\Delta_\eta(x)}\frac{\hat{p}_1(y\,|\,x)}{\hat{p}_0(y\,|\,x)}\right]. \tag{S5}$$

which means

$$L(\theta_i,\hat{p}(\cdot\,|\,x_i,\eta)) = E_{y_i\,|\,\theta_i}\widetilde{g}(x_i,y_i,\theta_i,\eta).$$

We also denote

$$\frac{1}{1-\Delta_\eta(x)} = 1+\frac{\eta}{1-\eta}\frac{m_1(x)}{m_0(x)} = 1+\frac{\lambda}{2}\frac{\eta}{1-\eta}\widetilde{R}_1(x)$$

and

$$1+\frac{\Delta_\eta(x)}{1-\Delta_\eta(x)}\frac{\hat{p}_1(y\,|\,x)}{\hat{p}_0(y\,|\,x)}=1+\frac{\lambda}{2}\frac{\eta}{1-\eta}\widetilde{R}_v(x,y)$$

where $\widetilde{R}_1(x)=\widetilde{R}_{v=1}(x,y)$ where for $v=1/(1+1/r)$

$$\widetilde{R}_v(x,y)=\int\exp\left[\mu\left(x+y\frac{1-v}{v}\right)-\frac{\mu^2}{2v}-\lambda|\mu|\right]\mathrm{d}\mu.$$

Similarly as in the proof of Theorem 3, we consider two cases (i) and (ii) depending on the magnitude $|x|$. Note that unlike in the proof of Theorem 3, here we are explicitly using notation involving z and y as opposed to z. Section 6.1.1 and 6.1.2 show upper bounds on the risk which separate parameter η from z (and inherently also x and y). We can use the same ideas as in Section 6.1.1, 6.1.2 and 6.1.3 to find that

$$\log\frac{1+\frac{\lambda}{2}\frac{\eta}{1-\eta}\widetilde{R}_1(x)}{1+\frac{\lambda}{2}\frac{\eta}{1-\eta}\widetilde{R}_v(x,y)}<C(\lambda,v,x,y)+5(1-v)\log\left(\frac{1-\eta}{\eta}\right).$$

Then because the term $C(\lambda,v,x,y)$ does not depend on η, we can write

$$E_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}E_{\eta\,|\,\boldsymbol{x}}E_{y_i\,|\,\theta_i}\widetilde{g}(x_i,y_i,\theta_i,\eta)\leq E_z\widetilde{C}(\lambda,v,z)+5(1-v)E_{\eta\,|\,\boldsymbol{x}}\log\left(\frac{1-\eta}{\eta}\right)$$

where the term $\widetilde{C}(\lambda,v,z)$ contains aspects of the bounds (6.11) and (6.12) that do not involve η. After taking the expectation over $Z\sim N(0,1)$ we arrive at a version of the (6.15), only with $\log[(1-\eta)/\eta]$ replaced by its conditional expectation. It also follows from (6.15) that

$$C(\lambda,v)\equiv E_z\widetilde{C}(\lambda,v,z)=\log\left(4\sqrt{\frac{2\pi}{v}}\right)+\frac{1}{\lambda^2}+\frac{\lambda^2}{2}+\frac{2\lambda^2+1}{r}+\lambda(1-\sqrt{v})\sqrt{\frac{2}{\pi}}+\log(2\lambda+\lambda^2+1).. \tag{S6}$$

When $\theta_i=0$, we have from (S5)

$$\widetilde{g}(x,y,\theta,\eta)<-\log[1-\Delta_\eta(x)]$$

and using the Jensen's inequality $E\log X<\log EX$ we find that

$$E_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}E_{\eta\,|\,\boldsymbol{x}}\widetilde{g}(x,y,\theta,\eta)<\log\left[1+E_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}E\left(\frac{\eta}{1-\eta}\,|\,\boldsymbol{x}\right)\frac{m_1(x_i)}{\phi(x)}\right].$$

The conditional expectation $E\left(\frac{\eta}{1-\eta}\,|\,\boldsymbol{x}\right)$ will be (for large n) very similar to the conditional expectation $E\left(\frac{\eta}{1-\eta}\,|\,\boldsymbol{x}_{\backslash i}\right)$, where $\boldsymbol{x}_{\backslash i}$ denotes the sub-vector of $\boldsymbol{x}$ after excluding the i^{th} coordinate. For $\lambda > 2$, we can use the sandwich relation (S9) to find that

$$\frac{E\left(\frac{\eta}{1-\eta}\,|\,\boldsymbol{x}\right)}{E\left(\frac{\eta}{1-\eta}\,|\,\boldsymbol{x}_{\backslash i}\right)} \leq \frac{a+E[s_n(\boldsymbol{\mu})\,|\,\boldsymbol{x}]+1}{a+E[s_n(\boldsymbol{\mu}_{\backslash i})\,|\,\boldsymbol{x}_{\backslash i}]} \leq \frac{a+E[s_n(\boldsymbol{\mu})\,|\,\boldsymbol{x}]+1}{a+E[s_n(\boldsymbol{\mu})\,|\,\boldsymbol{x}]-1} < 1+\frac{2}{a-1}.$$

Then we can write

$$E_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}E\left(\frac{\eta}{1-\eta}\,|\,\boldsymbol{x}\right)\frac{m_1(x_i)}{\phi(x_i)} \leq E_{\boldsymbol{x}_{\backslash i}\,|\,\boldsymbol{\theta}}E\left(\frac{\eta}{1-\eta}\,|\,\boldsymbol{x}_{\backslash i}\right)\left(1+\frac{2}{a-1}\right)\int m_1(x_i)\mathrm{d}x_i.$$

The desired statement is obtained after noting that (because $E\mathrm{e}^{cx} = \mathrm{e}^{-c^2/2}$ for $X \sim N(0,1)$)

$$\int m_1(x_i)\mathrm{d}x_i = \int_\mu \mathrm{e}^{-\mu^2/2}E\mathrm{e}^{\mu x_i}\pi_1(\mu)\mathrm{d}\mu = 1.$$

D.4 Proof of Lemma 7

Since the parameter η is hierarchically separated from the data by $\boldsymbol{\mu}$, we can write

$$\pi(\eta\,|\,\boldsymbol{x}) = \int \pi(\eta\,|\,\boldsymbol{\mu})\pi(\boldsymbol{\mu}\,|\,\boldsymbol{x})\mathrm{d}\boldsymbol{\mu}.$$

Then

$$E\left(\frac{\eta}{1-\eta}\,|\,\boldsymbol{x}\right) = \int_\eta \frac{\eta}{1-\eta}\pi(\eta)\int_{\boldsymbol{\mu}}\frac{\pi(\boldsymbol{\mu}\,|\,\eta)\pi(\boldsymbol{\mu}\,|\,\boldsymbol{x})}{\pi(\boldsymbol{\mu})}\mathrm{d}\boldsymbol{\mu}\mathrm{d}\eta.$$

We will now work conditionally on $\boldsymbol{\mu}$. For $\boldsymbol{\mu} = (\mu_1,\dots,\mu_n)'$ and $\pi_1(\boldsymbol{\mu}) = \prod_{i=1}^n \pi_1(\mu_i) = (\lambda/2)^n \mathrm{e}^{-\lambda|\boldsymbol{\mu}|_1}$, we write

$$\pi(\boldsymbol{\mu}\,|\,\eta) = \eta^n\,\pi_1(\boldsymbol{\mu})\prod_{i=1}^n\left[1+\frac{1-\eta}{\eta}\frac{\pi_0(\mu_i)}{\pi_1(\mu_i)}\right] \text{ and } \pi(\boldsymbol{\mu}) = \int_\eta \pi(\boldsymbol{\mu}\,|\,\eta)\pi(\eta).$$

For $S_n(\boldsymbol{\mu}) = \{i : \mu_i \neq 0\}$ with size $s_n(\boldsymbol{\mu}) = |S_n(\boldsymbol{\mu})|$ we can write for the point-mass spike $\pi_0(\mu) = \delta_0(\mu)$

$$\pi(\boldsymbol{\mu}\,|\,\eta) = \eta^n\pi_1(\boldsymbol{\mu})\prod_{i=1}^n\left[1+\frac{1-\eta}{\eta}\frac{\pi_0(\mu_i)}{\pi_1(\mu_i)}\right] = \eta^n\pi_1(\boldsymbol{\mu})\left[1+\frac{1-\eta}{\eta}\frac{1}{\pi_1(0)}\right]^{n-s_n(\boldsymbol{\mu})}.$$

Then

$$E\left(\frac{\eta}{1-\eta}\,|\,\boldsymbol{x}\right) = E_{\boldsymbol{\theta}\,|\,\boldsymbol{x}}E\left(\frac{\eta}{1-\eta}\,|\,\boldsymbol{\theta}\right) = \int_{\boldsymbol{\mu}}\pi(\boldsymbol{\mu}\,|\,\boldsymbol{x})\frac{\int_\eta \frac{\eta^{n+1}}{1-\eta}\pi(\eta)\left[1+\frac{1-\eta}{\eta}\frac{2}{\lambda}\right]^{n-s_n(\boldsymbol{\mu})}}{\int_\eta \eta^n\pi(\eta)\left[1+\frac{1-\eta}{\eta}\frac{2}{\lambda}\right]^{n-s_n(\boldsymbol{\mu})}}. \quad \text{(S7)}$$

Next, we represent the posterior expectation above as a functional of a ratio of two Gauss hypergeometric functions. An Euler representation of the Gauss hypergeometric function writes as

$$F_2(a',b',c';z)=\frac{\Gamma(c')}{\Gamma(b')\Gamma(c'-b')}\int_0^1\eta^{b'-1}(1-\eta)^{c'-b'-1}(1-\eta z)^{-a'}\mathrm{d}\eta. \tag{S8}$$

We now bound the posterior expectation of the prior odds $\eta/(1-\eta)$, given $\boldsymbol{\mu}$. Using simple notation $s=s_n(\boldsymbol{\mu})$ and $z=\lambda/2-1$, we can write

$$\begin{aligned}E\left(\frac{\eta}{1-\eta}\mid\boldsymbol{\mu}\right)&=\frac{\int_\eta \eta^{a+s}(1-\eta)^{b-2}\left[1+\eta\left(\frac{\lambda}{2}-1\right)\right]^{n-s}}{\int_\eta \eta^{a-1+s}(1-\eta)^{b-1}\left[1+\eta\left(\frac{\lambda}{2}-1\right)\right]^{n-s}}\\&=\frac{\Gamma(b-1)}{\Gamma(b)}\frac{\Gamma(a+s+1)}{\Gamma(a+s)}\frac{F_2(s-n,a+s+1,b+a+s;-z)}{F_2(s-n,a+s,b+a+s;-z)}.\end{aligned}$$

In Lemma 11 we show that the ratio of two Gauss hypergeometric functions with a shifted second parameter is monotone increasing in z. This result is an extension of Theorem 1 in [S15] who considered a different integer shift on the second and third argument. We use the notation $f_\delta(a',b',c';z)$ in (S1) and use Lemma 11 to find that (when $z=\lambda/2-1<0$)

$$\begin{aligned}E\left(\frac{\eta}{1-\eta}\mid\boldsymbol{\mu}\right)&\le\frac{a+s+1}{b}f_1(s-n,a+s,b+a+s;0)\\&=\frac{B(a+s+1,b-1)}{B(a+s,b-1)}\frac{B(a+s,b-1)}{B(a+s,b)}=\frac{a+s}{b-1}<\frac{a+s+1}{b-1}.\end{aligned}$$

In the regime when $0<\lambda/2-1$ with $\lambda\to\infty$, we have (by a repeated application of the l'Hospital rule) $\lim_{z\to\infty}f_1(s-n,a+s,b+a+s;-z)\to 1$ and

$$\frac{a+s}{b-1}<E\left(\frac{\eta}{1-\eta}\mid\boldsymbol{\mu}\right)\le\frac{a+s+1}{b}<\frac{a+s+1}{b-1} \tag{S9}$$

Similarly, we have

$$\begin{aligned}E\left(\frac{1-\eta}{\eta}\mid\boldsymbol{\mu}\right)&=\frac{\int_\eta \eta^{a-2+s}(1-\eta)^{b}\left[1+\eta\left(\frac{\lambda}{2}-1\right)\right]^{n-s}}{\int_\eta \eta^{a-1+s}(1-\eta)^{b-1}\left[1+\eta\left(\frac{\lambda}{2}-1\right)\right]^{n-s}}\\&=\frac{\Gamma(a+s-1)\Gamma(b+1)}{\Gamma(a+s)\Gamma(b)}\frac{F_2(s-n,a+s-1,b+a+s;-z)}{F_2(s-n,a+s,b+a+s;-z)}.\end{aligned}$$

Applying Lemma 11 we find that the ratio above is monotone *decreasing* in $z = \lambda/2 - 1$ for $z > -1$. Thereby

$$E\left(\frac{1-\eta}{\eta}\,|\,\boldsymbol{\mu}\right) \leq \frac{\Gamma(a+s-1)\Gamma(b+1)}{\Gamma(a+s)\Gamma(b)} \frac{1}{f_1(s-n, a+s-1, b+a+s, -1)}$$
$$= \frac{B(a+s-1, b+n-s+1)}{B(a+s-1, b+n-s)} \frac{B(a+s-1, b+n-s)}{B(a+s, b+n-s)} = \frac{b+n-s}{a+s-1}.$$

□

D.5 Proof of Lemma 8

We start by noting that the prior $Beta(2, n+1)$ has a strict exponential decrease property (property (2.2) in [S7] $\pi(s_n(\boldsymbol{\mu}) = s) \leq \pi(s_n(\boldsymbol{\mu}) = s-1)C_\pi$ for some $0 < C_\pi < 1$ which can be shown as in their Example 2.2). We can thereby apply Theorem 2.1 in [S7]. Denoting $B_n(M) = \{\boldsymbol{\mu} \in \mathbb{R}^n : \|\boldsymbol{\mu}\|_0 \leq M\, s_n\}$, this theorem yields that for some suitable $M > 0$, the posterior concentrates on sparse vectors in the sense that $P_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}\Pi(B_n^c\,|\,\boldsymbol{x}) = o(1)$ for any $\boldsymbol{\theta} \in \Theta_n(s_n)$. Lemma 7 then yields that for $a = 2$ and $b = n+1$

$$\sup_{\boldsymbol{\theta}\in\Theta_n(s_n)} E_{\boldsymbol{x}\,|\,\boldsymbol{\theta}} E\left(\frac{\eta}{1-\eta}\,|\,\boldsymbol{x}\right) \leq \frac{3+M\,s_n}{n} + (3/n+1)P_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}\Pi(B_n^c\,|\,\boldsymbol{x}) = Ms_n/n + o(1).$$

D.6 Proof of Lemma 9

We rewrite the model $\boldsymbol{X} \sim N(\boldsymbol{\theta}, I)$ as

$$x_i = \theta_i + \epsilon_i \quad \text{for} \quad \epsilon_i \overset{\text{iid}}{\sim} N(0,1) \quad (1 \leq i \leq n)$$

and define an event

$$\mathcal{A}_n = \left\{\epsilon_i : \max_{1\leq i\leq n} |\epsilon_i| \leq 2\sqrt{\log n}\right\}.$$

This event has a large probability in the sense that $P[\mathcal{A}_n^c] \leq 2/n$ (see Lemma 4 in [S6]). Then for $\boldsymbol{\theta} \in \Theta_n(s_n, \widetilde{M})$, take any j such that $\theta_j \neq 0$. We will denote with S a variable indexing all possible subsets of $\{1, \ldots, n\}$. Denote with $\mathcal{S}_j = \{S \subseteq \{1, \ldots, n\} : j \notin S\}$ the set of subsets of $\{1, \ldots, n\}$ that do not include j. Take $S \in \mathcal{S}_j$ and denote with $S^+ = S \cup \{j\}$ an enlarged "model" obtained by including j. Then

$$\frac{\Pi(S\,|\,\boldsymbol{x})}{\Pi(S^+\,|\,\boldsymbol{x})} = \frac{\Pi(S)}{\Pi(S^+)} \frac{\pi(\boldsymbol{x}\,|\,S)}{\pi(\boldsymbol{x}\,|\,S^+)}.$$

The one-dimensional marginal likelihood under the Laplace prior satisfies

$$m_1(x) = \phi(x)\lambda/2\left[I_1(x) + I_2(x)\right],$$

where

$$I_1(x_i) = \frac{\Phi(x_i - \lambda)}{\phi(x_i - \lambda)} \quad \text{and} \quad I_2(x_i) = \frac{\Phi(-(x_i + \lambda))}{\phi(x_i + \lambda)}.$$

This yields that, given S, the marginal likelihood is an independent product

$$\pi(\boldsymbol{x} \mid S) = \prod_{i=1}^{n} \phi(x_i) \prod_{i \in S} \frac{\lambda}{2}\left[I_1(x_i) + I_2(x_i)\right].$$

and

$$\frac{\pi(\boldsymbol{x} \mid S)}{\pi(\boldsymbol{x} \mid S^+)} = \frac{2}{\lambda} \frac{1}{I_1(x_j) + I_2(x_j)}.$$

When $x_j > 0$ we have $I_1(x_j) > I_2(x_j)$ and $I_1(x_j) \leq I_2(x_j)$ when $x_j \leq 0$. This yields

$$\frac{\pi(\boldsymbol{x} \mid S)}{\pi(\boldsymbol{x} \mid S^+)} \leq \frac{2}{\lambda} \frac{\mathrm{e}^{-(|x_j| - \lambda)^2/2}}{\Phi(-\lambda)}.$$

Because of Lemma 13 and Lemma 12 we have for $\lambda > 1$

$$\frac{1}{\Phi(-\lambda)} < \frac{\left(\sqrt{1 + 4/\lambda^2} + 1\right)\phi(\lambda)\lambda}{2} < 2\phi(\lambda)\lambda.$$

On the event $\mathcal{A}_n$ we obtain (using the inequality $(a+b)^2 > a^2/2 - b^2$)

$$(|x_j| - \lambda)^2 = (|\theta_j + \epsilon_j| - \lambda)^2 > |\theta_j + \epsilon_j|^2/2 - \lambda^2 > \theta_j^2/4 - 2\log n - \lambda^2 > (\widetilde{M}^2/4 - 2)\log n - \lambda^2.$$

This yields for $c = \widetilde{M}^2/4 - 2$

$$\frac{\pi(\boldsymbol{x} \mid S)}{\pi(\boldsymbol{x} \mid S^+)} \leq 4\phi(\lambda)\mathrm{e}^{-c\log n + \lambda^2} = \frac{4\sqrt{2\pi}}{n^c}\mathrm{e}^{\lambda^2/2}.$$

Under the hierarchical prior $\eta \sim Beta(a, b)$, we have for $s = |S|$

$$\Pi(S) = \frac{B(a+s, n-s+b)}{B(a,b)} = \frac{\Gamma(s+a)\Gamma(n-s+b)}{\Gamma(a+b+n)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}.$$

The prior ratio for $b = n+1$ satisfies

$$\frac{\Pi(S)}{\Pi(S^+)} = \frac{\Gamma(s+a)\Gamma(n-s+b)}{\Gamma(s+a+1)\Gamma(n-s-1+b)} = \frac{n-s+b}{s+a+1} \leq 2n.$$

Because the mapping $S \to S^+$ is one-to-one, we have for $C = 8\sqrt{2\pi}$

$$\begin{aligned}\Pi(\mathcal{S}_j \mid \boldsymbol{x}) &= \sum_{S: j \notin S} \frac{\Pi(S \mid \boldsymbol{x})}{\Pi(S^+ \mid \boldsymbol{x})} \Pi(S^+ \mid \boldsymbol{x}) < \frac{C e^{\lambda^2/2}}{n^{c-1}} \sum_{S: j \notin S} \Pi(S^+ \mid \boldsymbol{x}) \\ &= \frac{C e^{\lambda^2/2}}{n^{c-1}} \sum_{S^+: \exists S \, s.t. \, S^+ = S \cup j} \Pi(S^+ \mid \boldsymbol{x}) \leq \frac{C e^{\lambda^2/2}}{n^{c-1}}.\end{aligned}$$

This means that the posterior probability of missing any signal satisfies for λ such that $\lambda^2 \leq 2d \log n$ for some $d > 0$ and for $c > 2 + d$

$$\Pi\left(\exists j : |\theta_j| \neq 0 \text{ and } j \notin S \mid \boldsymbol{x}\right) \leq \sum_{j: |\theta_j| \neq 0} \Pi(\mathcal{S}_j \mid \boldsymbol{x}) \leq \frac{s_n C e^{\lambda^2/2}}{n^{c-1}} = o(1).$$

This means that for any $\boldsymbol{\theta} \in \Theta_n(s_n, \widetilde{M})$, on the event $\mathcal{A}_n$, the posterior *does not undershoot* $s_n = \|\boldsymbol{\theta}\|_0$. In other words

$$\sup_{\boldsymbol{\theta} \in \Theta_n(s_n, \widetilde{M})} \Pi(s_n(\boldsymbol{\mu}) < s_n \mid \boldsymbol{x}) \leq \sup_{\boldsymbol{\theta} \in \Theta_n(s_n, \widetilde{M})} \Pi(s_n(\boldsymbol{\mu}) < s_n \mid \boldsymbol{x}) \mathbb{I}(\mathcal{A}_n) + o(1) = o(1)$$

and using Lemma 7 we have for $a = 2$ and $b = n + 1$

$$E_{\boldsymbol{x} \mid \boldsymbol{\theta}} E\left[\log\left(\frac{1-\eta}{\eta}\right) \mid \boldsymbol{x}\right] \leq P(\mathcal{A}_n) \log\left(\frac{2n}{s_n}\right) + \log(2n+1) P(\mathcal{A}_n^c) \lesssim \log(n/s_n).$$

E Auxiliary Lemmata

Lemma 11. *For the Gauss hypergeometric function $F_2(a', b', c'; z)$ defined in (S8), the ratio*

$$f_\delta(a', b', c'; z) = \frac{F_2(a', b' + \delta, c'; -z)}{F_2(a', b', c'; -z)} \tag{S1}$$

for $\delta > 0$ is monotone increasing when $a' < 0$ for any $z > -1$.

Proof. We will prove this analogously as in Theorem 1 of [S15]. We denote with $A_\delta = \frac{\Gamma(c')}{\Gamma(b'+\delta)\Gamma(c'-b'-\delta)}$ and write

$$f_\delta(a', b', c'; z) = \frac{A_\delta \int [\eta/(1-\eta)]^\delta q(\theta, z) d\theta}{A_0 \int q(\theta, z) d\theta}$$

where $q(\eta, z) = \eta^{b'-1}(1-\eta)^{c'-b'-1}(1+\eta z)^{-a'}$. By differentiating the ratio $f_\delta(z)$ with respect to z, we find that the function $f_\delta(z)$ is monotone increasing for $a' < 0$ if

$$\int_0^1 h(\eta) \times q(\eta, z) d\eta \int_0^1 f(\eta) \times q(\eta, z) d\eta < \int_0^1 h(\eta) \times f(\eta) \times q(\eta, z) d\eta \int_0^1 q(\eta, z) d\eta$$

where $h(\eta) = (\frac{\eta}{1-\eta})^\delta$ and $f(\eta) = \frac{\eta}{1+\eta z}$. The function $q(\eta, z)$ is positive, while the functions $h(\eta)$ and $f(\eta)$ are monotone increasing for fixed $z > -1$ and $0 < \eta < 1$. Hence, the above inequality is an instance of the Chebyshev inequality ([S19] Chapter IX, formula (1.1)). □

Lemma 12. *For any $x > 0$, we have for $0 < c$*

$$\log \frac{\sqrt{1 + c/x^2} + 1}{2} < \frac{\sqrt{1 + c/x^2} - 1}{2} < c/x^2.$$

Proof. This follows from the fact that $\log(1 + x) < x$ and $\sqrt{1 + x} - 1 < x$ for $x > 0$.

Lemma 13. *(Mills Ratio Bounds) For the Gaussian Mill's ratio $R(x) = (1 - \Phi(x))/\phi(x)$ we have for any $x > 0$*

$$\frac{2}{\sqrt{x^2 + 4} + x} < \frac{1 - \Phi(x)}{\phi(x)} < \frac{2}{\sqrt{x^2 + 2} + x}. \tag{S2}$$

Proof. This result shown is in [S4].

References

[S1] Aitchison, J. (1975). Goodness of prediction fit. *Biomerika 62*, 547–554.

[S2] Bai, R., V. Ročková, and E. George (2020). Spike-and-Slab Meets LASSO: A Review of the Spike-and-Slab LASSO. *arXiv:2010.06451*, 1–30.

[S3] Bhattacharya, A., D. Pati, N. Pillai, and D. Dunson (2015). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association 110*, 1479–1490.

[S4] Birnbaum (1942). An inequality for Mill's ratio. *Annals of Mathematical Statistics 13*, 245–246.

[S5] Carvalho, C. and N. Polson (2010). The horseshoe estimator for sparse signals. *Biometrika 97*, 465–480.

[S6] Castillo, I., J. Schmidt-Hieber, and A. van der Vaart (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics 43*, 1986–2018.

[S7] Castillo, I. and A. van der Vaart (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics 40*, 2069–2101.

[S8] Deshpande, S., Ročková, and E. George (2019). Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *28*, 921–931.

[S9] George, E., F. Liang, and X. Xu (2006). Improved minimax predictive densities under kullback-leibler loss. *The Annals of Statistics 34*, 78–91.

[S10] George, E., F. Liang, and X. Xu (2012). From minimax shrinkage estimation to minimax shrinkage prediction. *Statistical Science 27*, 1102–1130.

[S11] George, E. and X. Xu (2008). Predictive density estimation for multiple regression. *Econometric Theory 24*, 528–544.

[S12] Hans, C. (2009). Bayesian LASSO regression. *Biometrika 96*, 835–845.

[S13] Hoffmann, M. Rousseau, J. and J. Schmidt-Hieber (2015). On adaptive posterior concentration rates. *The Annals of Statistics 43*, 2259–2295.

[S14] Johnstone, I. M. and B. W. Silverman (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics 32*, 1594–1649.

[S15] Karp, D. and S. Sitnik (2009). Inequalities and monotonicity of ratios for generalized hypergeometric function. *Journal of Approximation Theory 161*, 337–352.

[S16] Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika 88*, 859–864.

[S17] Liang, F. and A. Barron (2004). Exact minimax strategies for predictive denstiy estimation, data compression, and model selection. *IEEE Transactions of Information Theory 50*, 2708–2723.

[S18] Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*, 1023–1032.

[S19] Mitrinovć, D., J. Pecarić, and M. Fink (1993). *Classical and new inequalities in Analysis*. Kluwer Academic Publishers.

[S20] Moran, G., Ročková, and E. George (2021). Spike-and-Slab LASSO biclustering. *Annals of Applied Statistics 15*, 148–173.

[S21] Mukherjee, G. and I. Johnstone (2015). Exact minimax estimation of the predictive density in sparse gaussian models. *The Annals of Statistics 43*, 81–106.

[S22] Mukherjee, G. and I. Johnstone (2022). On minimax optimality of sparse Bayes predicitve density estimates. *The Annals of Statistics 50*, 81–106.

[S23] Nie, L. and V. Ročková (2022). Bayesian bootstrap Spike-and-Slab LASSO. *Journal of the American Statistical Association (in press) 1*, 1–50.

[S24] Park, T. and G. Casella (2008). The Bayesian LASSO. *Journal of the American Statistical Association 103*, 681–686.

[S25] Ray, K. and B. Szabo (2022). Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association 117*, 1270–1281.

[S26] Ročková (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics 46*, 401–437.

[S27] Ročková, V. and E. George (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association 111*, 1608–1622.

[S28] Ročková, V. and E. George (2018). The Spike-and-Slab LASSO. *113*, 431–444.

[S29] Ročková, V. and J. Rousseau (2023). Ideal Bayesian spatial adaptation. *Journal of the American Statistical Association (In Press)*, 1–80.

[S30] Tibshirani, R. (1994). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B 58*, 267–288.

[S31] van der Pas, S., B. Kleijn, and A. van der Vaart (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics 8*, 2585–2618.