Adaptive Bayesian Predictive Inference for Sparse High-dimensional Regression

Veronika Ročková*

May 30, 2024

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. To start, we investigate predictive distributions in the context 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 predictive distributions. However, deploying the Laplace prior inside the Spike-and-Slab framework (e.g. 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. Building on the sparse normal-means model, we extend our adaptive rate findings to the case of sparse high-dimensional regression with Spike-and-Slab priors. All of these results underscore benefits of fully Bayesian predictive 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 mere point predictions, are a valuable decision-making device for practitioners.

This paper focuses on the problem of predicting a Gaussian vector whose mean depends on an unknown sparse high-dimensional parameter. First, we investigate predictive inference for the quintessential sparse normal-means model. Later, we extend our findings to high-dimensional sparse linear regression which is widely used in practice. There is a wealth of literature on *estimation theory* in terms of rate-minimax posterior concentration (under the ℓ_2 loss) for various priors, both for the ℓ_0-sparse Gaussian sequence model (point-mass Spike-and-Slab priors [7], Spike-and-Slab LASSO priors [25] or continuous shrinkage priors [3, 5, 30]) and sparse high-dimensional regression [6, 27]. Lower bound results have also been established [6], showing that the Laplace prior [23] that yields a rate-optimal posterior mode in sparse regression, at the same time, yields a rate-suboptimal posterior distribution. Rate-minimaxity has become a useful criterion for prior calibration in order to acquire a frequentist license for Bayesian estimation procedures. This work focuses on rate-minimax certification 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, 20], 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 [20] and [21] but our work is different in at least three fundamental ways. First, we focus on (a) popular priors that are widely used in practice (such as the Bayesian LASSO [23] and the Spike-and-Slab LASSO [25]), and (b) we study *adaptation* to sparsity levels through hierarchical priors (none of the proposed estimators in [20] and [21] are adaptive). Lastly, (c) in addition to the Gaussian sequence

model, we study the high-dimensional sparse regression model which is indispensable to practitioners.

Mukherjee and Johnstone [20] were first to study predictive density estimation for a multivariate normal vector with ℓ_0 sparse means (with up to s_n signals) and found several fascinating parallels between sparse normal means estimation and predictive inference under the KL loss. In addition, [20] constructed a predictive density estimator (inspired by hard thresholding) by pasting together 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 smooth with respect to the input 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, [21] 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, [21] 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. All these existing results have been obtained only under the Gaussian sequence model.

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 [25]) which have a *continuous* slab (versus a discrete slab considered in [20]) 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 *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, 25]. Results (1)-(3) are obtained under the sparse Gaussian sequence model. Our final contribution is obtaining adaptive rate results in the context of high-dimensional sparse regression under Spike-and-Slab priors. Minimax predictive inference in low-dimensional regression was previously studied by [11]. We investigate the high-dimensional regime where $p > n$ and where the sparsity level s_n is allowed to increase with the sample size n. Focusing on both the total variation distance as well as the (typical) KL divergence, we establish a rate of estimation of Spike-and-Slab predictive distributions which is adaptive in s_n and which mirrors the ℓ_2 estimation rate. Our proof relies on some techniques developed in [6] but has required several new steps including a novel upper bound on the marginal likelihood under a Laplace prior.

The paper has the following structure. Section 2 is dedicated to the sparse normal means model. Section 2.1 presents a lower-bound result showing that Bayesian LASSO is incapable of yielding posterior predictive densities with good properties in sparse setups. Section 2.2 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 [25]. Section 2.3 then shows adaptive rate-minimax performance of hierarchical Spike-and-Slab priors. Section 3 is dedicated to the sparse high-dimensional regression model. We conclude with a discussion in Section 4 and the proof of Theorem 3 in Section 5.

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 $R(x) = [1 - \Phi(x)]/\phi(x)$. We will denote with $\mathbb{E}$ expectation with respect to a data generating process and with E expectation over a latent variable.

2 Predictive Inference in the Sparse Normal Means Model

Within the context of the Gaussian sequence model, we aim to predict $\widetilde{Y} \sim N_n(\theta, r \times I)$ from an independent observation $Y \sim N_n(\theta, I)$ as $n \to \infty$, where the true underlying parameter θ is sparse in the sense that $\theta \in \Theta_n(s_n)$ where $\Theta_n(s_n) = \{\theta \in \mathbb{R}^n : \|\theta\|_0 \leq s_n\}$. We study the problem of obtaining an entire *predictive density* $\hat{p}(\widetilde{Y} \,|\, Y)$ for $\widetilde{Y}$ that is close to $\pi(\widetilde{Y} \,|\, \theta)$ in terms of the Kullback-Leibler loss

$$L(\theta, \hat{p}(\cdot \,|\, Y)) = \int \pi(\widetilde{Y} \,|\, \theta) \log \frac{\pi(\widetilde{Y} \,|\, \theta)}{\hat{p}(\widetilde{Y} \,|\, Y)} \mathrm{d}\widetilde{Y}, \tag{2.1}$$

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

$$\rho_n(\theta, \hat{p}) = \int \pi(Y \,|\, \theta) L(\theta, \hat{p}(\cdot \,|\, Y)) \mathrm{d}Y.$$

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

$$\hat{p}(\widetilde{Y} \,|\, Y) = \int \pi(\widetilde{Y} \,|\, \theta) \pi(\theta \,|\, Y) \mathrm{d}\theta. \tag{2.2}$$

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 estimator (e.g. the maximum likelihood estimator $\hat{\theta}_{MLE}$) to obtain a predictive density estimate $\pi(\widetilde{Y} \,|\, \hat{\theta}_{MLE})$. This malpractice was denounced by Aitchison [1] who showed that $\hat{p}_U(\widetilde{Y} \,|\, Y)$ defined as (2.2) under the uniform prior dominates the plug-in predictive density $\pi(\widetilde{Y} \,|\, \hat{\theta}_{MLE})$. When $p = 1$, $\hat{p}_U(\widetilde{Y} \,|\, Y)$ is admissible under KL loss [16] but when $p \geq 3$, $\hat{p}_U(\widetilde{Y} \,|\, Y)$ is inadmissible and dominated by Bayesian predictive density under the harmonic prior [15]. George et al. [9] established general sufficient conditions under which a Bayes predictive density will be minimax and will dominate $\hat{p}_U(\widetilde{Y} \,|\, Y)$. 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 θ is sparse, focusing on *rate-minimaxity*.

In their pathbreaking paper, Mukherjee and Johnstone [20] were the first to study predictive density estimation for a multivariate normal vector with ℓ_0 sparse means. These authors 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 $s_n/n \to 0$

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

where $r = r/1$ is the variance ratio (future over observed data) and where the minimum is taken over all predictive density estimators. We will be targeting this minimax rate using popular priors.

2.1 The Calibration Conflict of Bayesian LASSO

The LASSO method [29] is a staple in sparse signal recovery. The LASSO estimator is, in fact, Bayesian [23] as it corresponds to the posterior mode under the Laplace prior

$$\pi(\theta \,|\, \lambda) = \prod_{i=1}^{n} \pi_1(\theta_i \,|\, \lambda) \quad \text{where} \quad \pi_1(\theta \,|\, \lambda) = \frac{\lambda}{2} \mathrm{e}^{-\lambda|\theta|} \quad \text{for some} \quad \lambda > 0. \tag{2.3}$$

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 under the Laplace prior, it is tempting to utilize the entire posterior distribution as an inferential object. This inclination was soon corrected 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 inference as well.

For the Bayesian LASSO independent product prior (2.3) [12, 23], the Bayesian predictive density has a product form $\hat{p}(\widetilde{Y} \mid Y) = \prod_{i=1}^{n} \hat{p}(\widetilde{Y}_i \mid Y_i)$ and

$$L(\theta, \hat{p}(\cdot \mid Y)) = \sum_{i=1}^{n} \int \pi(\widetilde{Y}_i \mid \theta_i) \log \frac{\pi(\widetilde{Y}_i \mid \theta_i)}{\hat{p}(\widetilde{Y}_i \mid Y_i)} \mathrm{d}y = \sum_{i=1}^{n} L(\theta_i, \hat{p}(\cdot \mid Y_i)).$$

The predictive risk of a product rule over the $\Theta_n(s_n)$ is additive and satisfies [20]

$$(n - s_n)\rho(0, \hat{p}) < \rho_n(\theta, \hat{p}) = \sum_{i=1}^{n} \rho(\theta_i, \hat{p}) \leq (n - s_n)\rho(0, \hat{p}) + s_n \sup_{\theta \in \mathbb{R}} \rho(\theta, \hat{p}), \tag{2.4}$$

where (for any $1 \leq i \leq n$) and $\theta \in \mathbb{R}$

$$\rho(\theta, \hat{p}) = \int \pi(Y_i \mid \theta) \int \pi(\widetilde{Y}_i \mid \theta) \log[\pi(\widetilde{Y}_i \mid \theta)/\hat{p}(\widetilde{Y}_i \mid Y)] \mathrm{d}\widetilde{Y}_i \mathrm{d}Y_i$$

is the univariate risk. The following Lemma precisely characterizes the univariate risk $\rho(\theta, \hat{p})$ for $\theta \in \mathbb{R}$ under the prior (2.3). This Lemma, in fact, applies to *any* prior $\pi_1(\cdot)$.

Lemma 1. *The univariate prediction risk under the Bayesian LASSO prior* (2.3) *satisfies*

$$\rho(\theta, \hat{p}) = \theta^2/(2r) + E \log N_{\theta,1}^{LASSO}(Z) - E \log N_{\theta,v}^{LASSO}(Z) \tag{2.5}$$

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

Proof. Section 9.1.

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 $\theta_i \neq 0$, λ needs to be small to avoid squashing large effects. We expect a similar conundrum for the prediction problem. The inequality (2.4) shows that, in order for Bayesian LASSO prediction distributions to be rate-minimax optimal, we would need

$$\sup_{\theta \in \mathbb{R}} \rho(\theta, \hat{p}) \lesssim \frac{\log(n/s_n)}{1 + r} \quad \text{and, at the same time,} \quad \rho(0, \hat{p}) \lesssim \frac{s_n \log(n/s_n)}{(n - s_n)(1 + r)}. \tag{2.6}$$

It will follow from Theorem 1 and Theorem 2 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*

$$\rho(0,\hat{p}) \leq \log\left(1+\frac{\sqrt{2}}{\lambda\sqrt{\pi v}}\right)+\frac{4}{\lambda^2 v} \quad \textit{and} \quad \sup_{\theta\neq 0}\rho(\theta,\hat{p}) \leq \log\left(\sqrt{\frac{32\lambda^2\pi}{v}}\right)+\frac{\lambda^2}{2}+\lambda\sqrt{\frac{2}{\pi}}+\frac{4}{\lambda^2}.$$

Proof. Section 6

Theorem 1 and the inequality (2.6) 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 (a) λ 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) λ 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 ascknowledge 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. *(Bayesian LASSO is Suboptimal) Consider the Bayesian LASSO prior in (2.3) 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*

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

Proof. It follows from Lemma 13 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 Theorem 2 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 $\frac{n-s_n}{s_n}\frac{1}{\log(n/s_n)}$.

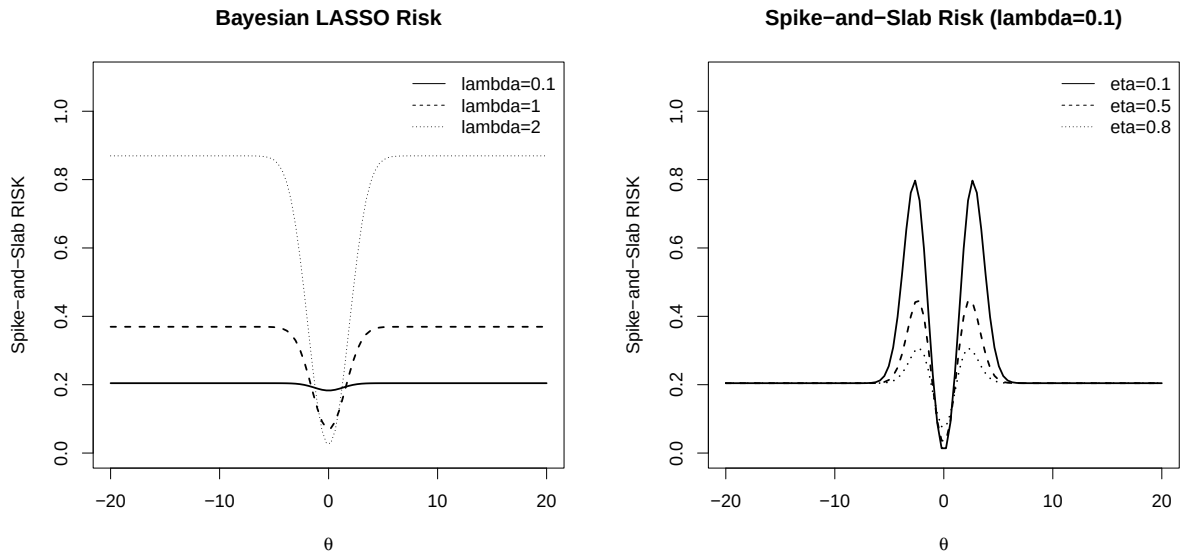

Figure 1: (Left) Bayesian LASSO prediction risk $\rho(\theta, \hat{p})$ for $\lambda \in \{0.1, 1, 2\}$; (Right) Spike-and-Slab risk (Dirac spike and Laplace slab with a penalty $\lambda = 0.1$ with $\eta \in \{0.1, 0.5, 0.8\}$); Both plots correspond to $r = 2$.

2.2 Separable Spike-and-Slab Priors

Section 2.1 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 [17] 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 the supremum-norm sense [13] (without any log factors) or rate-minimax spatial adaptation [28]. 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 2.1.

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

$$\pi(\theta\,|\lambda, \eta) = \prod_{i=1}^{n} \pi(\theta_i\,|\lambda, \eta), \quad \text{where } \pi(\theta_i\,|\lambda, \eta) = \eta\pi_1(\theta_i) + (1-\eta)\pi_0(\theta_i) \text{ and } \eta \in (0, 1). \quad (2.7)$$

The spike density $\pi_0(\theta_i)$ serves as a model for the noise coefficients $\theta_i = 0$ while the slab

density $\pi_1(\theta_i)$ 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 (2.7) 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* $m_j(Y_i) = \int \pi(Y_i \,|\, \mu)\pi_j(\mu)\mathrm{d}\mu$ *for* $j = 0, 1$ *the marginal likelihoods for* Y_i *under the spike/slab priors* π_0 *and* π_1*. For* $\eta \in (0,1)$*, we define a mixing weight*

$$\Delta_\eta(Y_i) = \frac{\eta m_1(Y_i)}{\eta m_1(Y_i) + (1-\eta)m_0(Y_i)}. \tag{2.8}$$

Then the Bayesian predictive density under the prior (2.7) *is a mixture, i.e.*

$$\hat{p}(\widetilde{Y}_i \,|\, Y_i) = \Delta_\eta(Y_i)\hat{p}_1(\widetilde{Y}_i \,|\, Y_i) + [1 - \Delta_\eta(Y_i)]\hat{p}_0(\widetilde{Y}_i \,|\, Y_i) \tag{2.9}$$

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

Proof. This follows readily from the definition $\hat{p}(\widetilde{Y}_i \,|\, Y_i) = \frac{\int \pi(\widetilde{Y}_i \,|\, \mu)\pi(\mu \,|\, Y_i)\mathrm{d}\mu}{\int \pi(\mu \,|\, Y_i)\mathrm{d}\mu}$.

This elegant decomposition provides useful insights into the workings of the mixture prior. The predictive density will be dominated by the slab predictive density when $\Delta_\eta(Y_i)$ is close to one, i.e. the observation Y_i 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 Y_i 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 Y_i. In other words, we can rewrite the mixing weight as a functional of the Bayes factor

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

where $BF(Y_i; 0, 1) = \frac{m_0(Y_i)}{m_1(Y_i)}$ 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 [20]. These authors proposed a class of univariate predictive density estimators that are analogs of hard thresholding in that they glue together two densities depending on the magnitude (signal detectability) of observed data $|Y_i|$. These estimators, however, do not have a fully Bayesian motivation and are not smooth with respect to Y_i. 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* $\Lambda(\theta_i) = \int \Delta_\eta(Y_i)\pi(Y_i \,|\, \theta_i)\mathrm{d}x$ *and using the notation in Lemma 2, the prediction risk under the prior* (2.7) *satisfies*

$$\rho(\theta_i, \hat{p}) < \Lambda(\theta_i)\rho(\theta_i, \hat{p}_1) + (1 - \Lambda(\theta_i))\rho(\theta_i, \hat{p}_0). \tag{2.10}$$

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

$$\begin{aligned}\rho(\theta_i, \hat{p}) &< \int \pi(Y_i \,|\, \theta_i) \left\{ \int \pi(\widetilde{Y}_i \,|\, \theta_i) E \left[\log \frac{\pi(\widetilde{Y}_i \,|\, \theta_i)}{\gamma \hat{p}_1(\widetilde{Y}_i \,|\, Y_i) + (1-\gamma)\hat{p}_0(\widetilde{Y}_i \,|\, Y_i)} \right] \mathrm{d}\widetilde{Y}_i \right\} \mathrm{d}Y_i \\ &= \Lambda(\theta_i)\rho(\theta_i, \hat{p}_1) + (1 - \Lambda(\theta_i))\rho(\theta_i, \hat{p}_0).\end{aligned}$$

This elegant upper bound shows how the risk is dominated either by the spike predictive density (for parameter values θ_i such that $\Lambda(\theta_i)$ is small) or the slab predictive density (for parameter values θ_i such that $\Lambda(\theta_i)$ is large). This bound is perhaps more intuitive than useful, however. Our analysis rests on a more precise characterization of the risk. In Lemma 4 below, we obtain a risk decomposition for $\rho(\theta_i, \hat{p})$ but for *general* mixtures (2.7).

Lemma 4. *The univariate risk for the Spike-and-Slab prior* (2.7) *satisfies for* $\theta \in \mathbb{R}$

$$\rho(\theta, \hat{p}) = \rho(\theta, \hat{p}_1) + E \log N_{\theta,1}^{SS}(Z) - E \log N_{\theta,v}^{SS}(Z), \tag{2.11}$$

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 9.2

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 [20].*

Lemma 4 is a simple generalization of Theorem 2.1 in [20], 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 [25] studied in Section 2.2.2. 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.

2.2.1 Dirac Spike and Laplace Slab

In Section 2.1, we convinced ourselves that a single Laplace prior will not be able to yield rate-optimal predictive distributions. We now show that the Laplace slab $\pi_1(\theta\,|\lambda)$ *in concert with* a Dirac spike $\pi_0(\theta)$ within (2.7) does.

Theorem 3. *Assume the Spike-and-Slab prior* (2.7) *with a Dirac spike* $\pi_0(\theta_i) = \delta_0(\theta_i)$ *and a Laplace slab* $\pi_1(\theta_i\,|\,\lambda)$ *in* (2.3). *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* $r \in (0,\infty)$

$$\sup_{\theta \in \Theta(s_n)} \rho_n(\theta, \hat{p}) \leq \frac{5}{1+r} s_n \log(n/s_n) + \widetilde{C}(r) \tag{2.12}$$

where $\widetilde{C}(r)$ *is either* (5.17) *or* (5.18).

Proof. Section 5.

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

[24]. Of course, our result in Theorem 3 is only rate-minimax, where the multiplication constant may not optimal. Up until now, minimax (rate-optimality) property was established for discrete (grid) slab priors which may be simpler to treat analytically [21]. The only other continuous slab result in the literature so far is in [21] who analyzed *uniform* slabs over a finite parameter domain depending on s_n (which are perhaps not as widely used).

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 [25]. 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* $|Y_i| > \lambda + \sqrt{2v\log(n/s_n)}$, *with i.e. the observed data is above the usual detection threshold rescaled by* v *and shifted by* λ, *and (2) when* $|Y_i| \leq \lambda + \sqrt{2v\log(n/s_n)}$. *In the first case, the slab predictive density takes over and the reverse is true for case (2). [20] used a related regime switching idea to construct an estimator by pasting two predictive densities depending on the size of* $|Y_i|$. *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 $\theta_i \neq 0$. 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 $|\theta_i| \to \infty$, 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 $|\theta_i|$. Notably, there are certain values of θ_i for which the risk is inflated due to uncertainty whether or not the underlying parameter θ_i arrived more likely from the slab/spike.

2.2.2 The Spike-and-Slab LASSO

The Spike-and-Slab LASSO prior [25] has become popular for sparsity recovery due to its self-adaptive shrinkage properties. It generalizes the Laplace prior (2.3) by deplying a two-point Laplace mixture (2.7) where

$$\pi_0(\theta_i) = \lambda_0/2\mathrm{e}^{-\lambda_0|\theta_i|} \quad \text{and} \quad \pi_1(\theta_i) = \lambda_1/2\mathrm{e}^{-\lambda_1|\theta_i|} \quad \text{with } \lambda_1 < \lambda_0. \tag{2.13}$$

This prior has been successfully implemented in high-dimensional regression [27], graphical models [8], factor analysis [26], biclustering [19], 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 [25], 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 [22]. 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 (2.7) with (2.13) 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*

$$\sup_{\theta \in \Theta(s_n)} \rho(\theta, \hat{p}) \sim \frac{s_n}{1+r} \log(n/s_n) \tag{2.14}$$

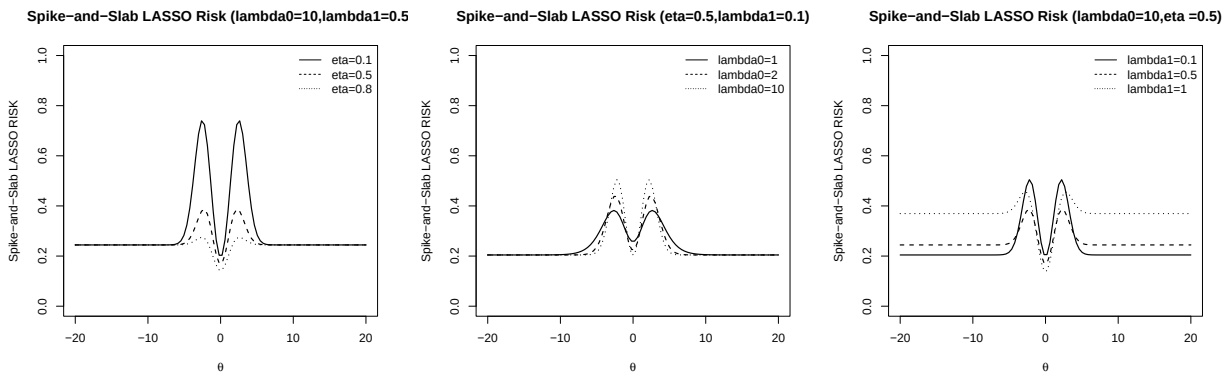

Figure 2: Spike-and-Slab LASSO prediction risk for various calibrations. (Left) Varying η for fixed $\lambda_0 = 10, \lambda_1 = 0.5$; (Middle) Varying λ_0 for fixed $\eta = 0.1, \lambda_1 = 0.1$; (Right) Varying λ_1 for fixed $\eta = 0.1, \lambda_0 = 10$

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

Proof. Section 8.

Remark 4. *(Calibration) Ročková [25] 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 (8.3). 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 8 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 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 [25].*

2.3 Hierarchical Spike-and-Slab Priors

All results in [20, 21] 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 2.2.1. However, now we assume a *hierarchical* version (not an independent product), i.e.

$$\pi(\theta) = \int_\eta \prod_{i=1}^n [(1-\eta)\delta_0 + \eta\pi_1(\theta_i)]\pi(\eta)\mathrm{d}\eta \quad \text{and} \quad \pi(\eta) \sim Beta(a,b) \tag{2.15}$$

for some $a, b > 0$. In Section 2.2, 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 (2.15) 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 [25, 27]. 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

$$\hat{p}(\widetilde{Y} \mid Y) = \int_\eta \prod_{i=1}^n \left[\Delta_\eta(Y_i)\hat{p}_1(\widetilde{Y}_i \mid Y_i) + (1-\Delta_\eta(Y_i))\hat{p}_0(\widetilde{Y}_i \mid Y_i)\right] \mathrm{d}\pi(\eta), \tag{2.16}$$

where $\hat{p}_1$ and $\hat{p}_0$ are the slab and spike univariate posterior predictive densities, $\Delta_\eta(Y_i)$ is the mixing weight (2.8) 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 $\hat{p}(\widetilde{Y} \mid Y, \eta)$ where the average is taken over the posterior $\pi(\eta \mid Y)$, i.e.

$$\hat{p}(\widetilde{Y} \mid Y) = E_{\eta \mid Y}\hat{p}(\widetilde{Y} \mid Y, \eta). \tag{2.17}$$

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* (2.15) *satisfies* $L(\theta, \hat{p}(\cdot \,|\, Y)) \leq E_{\eta \,|\, Y} L(\theta, \hat{p}(\cdot \,|\, Y, \eta))$.

Proof. This follows from rewriting the posterior predictive distribution (2.16) as (2.17) and by applying Jensen's inequality $E \log X < \log EX$ to obtain

$$\begin{aligned} L(\theta, \hat{p}(\cdot \,|\, Y)) &= \int \pi(\widetilde{Y} \,|\, \theta) \log \pi(\widetilde{Y} \,|\, \theta) - \int \pi(\widetilde{Y} \,|\, \theta) \log E_{\eta \,|\, Y} \hat{p}(\widetilde{Y} \,|\, Y, \eta) \mathrm{d}Y \\ &\leq E_{\eta \,|\, Y} L(\theta, \hat{p}(\cdot \,|\, Y, \eta)). \end{aligned} \tag{2.18}$$

Lemma 5 shows that it is the conditional distribution $\pi(\eta \,|\, Y)$ 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 $\pi(\eta \,|\, Y)$ 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* (2.15) *satisfies for* $\lambda > 2$

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

for a suitable constant $C(\lambda, v) > 0$ *defined in* (9.6) *and* $D = 1 + 2/(a-1)$, *where* $Y_{\setminus i}$ *denotes the vector* Y *without the* i^{th} *coordinate.*

Proof. Section 9.3

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* $\pi(\theta)$ *in* (2.15) *with* $a, b > 0$. *Under the Gaussian model* $Y \sim N_n(\theta, I)$, *the posterior distribution* $\pi(\eta \,|\, Y)$ *satisfies*

$$E \left(\frac{\eta}{1-\eta} \,|\, Y \right) \leq \frac{a + E[s_n(\theta) \,|\, Y] + 1}{b - 1} \quad \textit{and} \quad E \left(\frac{1-\eta}{\eta} \,|\, Y \right) \leq E \left(\frac{b+n}{s_n(\theta) + a - 1} \,|\, Y \right)$$

where $s_n(\theta) = \|\theta\|_0$.

Proof. Section 9.4.

Remark 5. *This Lemma shows that the usual calibration [7] with $a = 1$ and $b = n + 1$ implies $E\left(\frac{\eta}{1-\eta} \,|\, Y\right) \lesssim E[s_n(\theta)/n \,|\, Y]$ and $E\left(\frac{1-\eta}{\eta} \,|\, Y\right) \lesssim E[n/s_n(\theta) \,|\, Y]$. 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 $s_n(\theta)/n$ and $n/s_n(\theta)$, respectively. In particular, for $\lambda > 2$ we obtain from (9.9) that $E\left(\frac{\eta}{1-\eta} \,|\, Y\right) > \frac{a+E[s_n(\theta \,|\, Y]}{b-1}$.*

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 $Y \sim N_n(\theta, I)$. 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 $Y \sim N_n(\theta, I)$ and the hierarchical Spike-and-Slab prior (2.15) with $a = 1$ and $b = n + 1$. Then for some suitable $M > 0$ we have*

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

Proof. Section 9.5

Lemma 8 takes care of the "noise" part of the risk bound in Lemma 6 where $(n - s_n)E_{Y_{\setminus i} \,|\, \theta} E[\eta/(1-\eta) \,|\, Y_{\setminus i}] \lesssim s_n$. In order to bound the "signal" part of the risk bound, we need to show that $s_n E_{Y \,|\, \theta} E[(1-\eta)/\eta \,|\, Y] \lesssim s_n \log(n/s_n)$. From Lemma 7, we need to make sure that the posterior $\pi(\eta \,|\, Y)$ 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

$$\Theta_n(s_n, \widetilde{M}) = \Theta_n(s_n) \cap \left\{\theta \in \mathbb{R}^n : \min_{i:\theta_i \neq 0} |\theta_i| > \widetilde{M}\sqrt{\log n}\right\}. \tag{2.19}$$

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 $Y \sim N_n(\theta, I)$ and the hierarchical Spike-and-Slab prior (2.15) 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*

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

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*

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

Proof. Section 9.6

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 (2.19)). The near-minimax performance is guaranteed over the entire parameter space $\Theta_n(s_n)$.

Theorem 5. *Assume the hierarchical prior* (2.15) *with a Laplace slab* (2.3) *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* (2.19) *then we have for* $c > 2$

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

Proof. The proof follows directly from Lemma 6, 7, 8 and 9. The proposed calibrations ascertain that the term $C(\lambda, v)$ defined in (9.6) 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 [25]. The next section extends some of the findings to the more practical case of high-dimensional regression.

3 High-dimensional Sparse Regression

Predictive inference in a low-dimensional (sparse) regression was studied by [11] who established sufficient conditions for minimaxity and dominance of a Bayesian estimator under the uniform prior over a plug-in estimator. Here, we consider a high-dimensional scenario $p > n$ with subset selection uncertainty and an *unknown* and possibly diverging sparsity level s_n. Rather than targeting the exact minimax risk, we will establish risk rates for the Spike-and-Slab prior with either a uniform slab (mirroring the setup in [11]) as well as a Laplace slab. We observe $Y = (Y_1, \dots, Y_n)'$ from

$$Y_i = X_i'\beta_0 + \varepsilon_i \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$$

where X_i is a vector of $(p \times 1)$ covariate values and where $\beta_0 \in \Theta_p(s_n)$, i.e. $\|\beta_0\|_0 \leq s_n$, and $p > n$. The goal is to predict $\widetilde{Y} = (\widetilde{Y}_1, \dots, \widetilde{Y}_m)'$ where

$$\widetilde{Y}_i = \widetilde{X}_i'\beta_0 + \widetilde{\varepsilon}_i \quad \widetilde{\varepsilon}_i \overset{\text{iid}}{\sim} N(0,1),$$

where $\widetilde{X}_i$ is a vector of the same p covariates but with possibly different values and where Y and $\widetilde{Y}$ are conditionally independent given $\beta_0 \in \mathbb{R}^p$. Throughout this section, we assume that the columns of X and $\widetilde{X}$ have been suitably normalized so that $\|X^j\|_2 = \sqrt{n}$ and $\|\widetilde{X}^j\|_2 = \sqrt{m}$. We will also denote $\|X\| = \sup_{1 \leq j \leq p} \|X^j\|_2$. The Spike-and-Slab prior is underpinned by a subset indicator $S \subset \{1, \dots, p\}$ and can be written (for some $0 \leq R \leq p$) as

$$\pi(\beta) = \sum_{s=0}^{R} \pi(s) \sum_{S:|S|=s} \frac{1}{\binom{p}{s}} \times \pi_{S,\lambda}(\beta_S) \otimes_{i \notin S} \delta_0(\beta_i) \tag{3.1}$$

where $\pi_{S,\lambda}(\beta_S)$ is the slab portion for the subset β_S of active coefficients β inside S and where (as in Assumption 1 in [6]) for some $A_1, A_2, A_3, A_4 > 0$

$$A_1 p^{-A_3} \pi(s-1) \leq \pi(s) \leq A_2 p^{-A_4} \pi(s-1), \quad s = 1, \dots, p. \tag{3.2}$$

The predictive density is a mixture of conditional predictive densities, given S, mixed by the posterior $\pi(S \mid Y)$ and can be written as (see e.g. (37) in [11])

$$\hat{p}(\widetilde{Y} \mid Y) = \sum_{S} \pi(S \mid Y) \hat{p}(\widetilde{Y} \mid Y, S). \tag{3.3}$$

We denote with $S_\beta = \{i : \beta_i \neq 0\}$ the support of a vector $\beta \in \mathbb{R}^p$. When the posterior concentrates on $S_0 = S_{\beta_0}$, the posterior predictive density will be dominated by $\hat{p}(\widetilde{Y} \,|\, Y, S_0)$ for which the uniform prior dominates the plug-in density estimator $\pi(\widetilde{Y} \mid \hat{\beta}_{MLE})$ and for which the scaled harmonic prior will be minimax when $|S_0| \geq 3$ [11]. We leverage model selection consistency to obtain predictive risk bounds in high-dimensional scenarios $p > n$ where both S_0 and s_n are unknown. First, we find a convenient characterization of the KL risk. Denote

$$\Delta_{n,\beta,\beta_0}(Y, X) \equiv \frac{\pi(Y \mid \beta)}{\pi(Y \mid \beta_0)} = \mathrm{e}^{-1/2\|X(\beta-\beta_0)\|_2^2 + (Y - X\beta_0)' X(\beta-\beta_0)} \tag{3.4}$$

the normalized likelihood and its marginal version by

$$\Lambda_{n,\beta_0,\pi}(Y, X) = \int_\beta \Delta_{n,\beta,\beta_0}(Y, X) \pi(\beta) \mathrm{d}\beta.$$

This notation helps us compactly characterize the KL prediction risk.

Lemma 10. *Writing* $\bar{X} = (X', \widetilde{X}')'$ *and* $Z = (Y', \widetilde{Y})'$, *the KL predictive risk* $\rho_{n,m}(\beta_0, \hat{p}) \equiv \mathbb{E}_{Y \mid \beta_0} KL(\pi(\widetilde{Y} \mid \beta_0), \hat{p}(\widetilde{Y} \mid Y))$ *for the regression prediction problem satisfies*

$$\rho_{n,m}(\beta_0, \hat{p}) = \mathbb{E}_{Y \mid \beta_0} \log \Lambda_{n,\beta_0,\pi}(Y, X) - \mathbb{E}_{Y \mid \beta_0} \mathbb{E}_{\widetilde{Y} \mid \beta_0} \log \Lambda_{n+m,\beta_0,\pi}(Z, \bar{X}).$$

Proof. We have

$$\begin{aligned}
\hat{p}(\widetilde{Y} \mid Y) &= \pi(\widetilde{Y} \mid \beta_0) \frac{\int_\beta \Delta_{m,\beta,\beta_0}(\widetilde{Y}, \widetilde{X}) \times \Delta_{n,\beta,\beta_0}(Y, X) \pi(\beta) \mathrm{d}\beta}{\int_\beta \Delta_{n,\beta,\beta_0}(Y, X) \pi(\beta) \mathrm{d}\beta} \\
&= \pi(\widetilde{Y} \mid \beta_0) \frac{\int_\beta \Delta_{m+n,\beta,\beta_0}(Z, \bar{X}) \pi(\beta) \mathrm{d}\beta}{\int_\beta \Delta_{n,\beta,\beta_0}(Y, X) \pi(\beta) \mathrm{d}\beta}.
\end{aligned}$$

The rest follows from $\rho_{n,m}(\beta_0, \hat{p}) = \mathbb{E}_{Y \mid \beta_0} KL(\pi(\widetilde{Y} \mid \beta_0), \hat{p}(\widetilde{Y} \mid Y))$.

It is useful to elaborate on the KL divergence $KL(\pi(\widetilde{Y} \mid \beta_0), \hat{p}(\widetilde{Y} \mid Y))$ for the mixture estimator $\hat{p}(\widetilde{Y} \mid Y)$ from (3.3). The following Lemma helps us understand the connection between the concentration of the posterior predictive distributions and the concentration of the parameter posteriors.

Lemma 11. *For a given set of observations Y we have for $\hat{p}(\widetilde{Y} \mid Y)$ from (3.3)*

$$KL(\pi(\widetilde{Y} \mid \beta_0), \hat{p}(\widetilde{Y} \mid Y)) \leq 1/2\, E_{\beta \mid Y} \|\widetilde{X}(\beta - \beta_0)\|_2^2 + \sqrt{m} E_{\beta \mid Y} \|\widetilde{X}(\beta - \beta_0)\|_2. \tag{3.5}$$

Proof. For $\Delta_{n,\beta,\beta_0}(Y, X)$ defined in (3.4) we have

$$KL(\pi(\widetilde{Y} \mid \beta_0), \hat{p}(\widetilde{Y} \mid Y)) = -\int \pi(\widetilde{Y} \mid \beta_0) \log E_{S \mid Y} E_{\beta \mid S,Y} \Delta_{m,\beta,\beta_0}(\widetilde{Y}, \widetilde{X})$$

which implies an upper bound from Jensen's inequality $E \log X \leq \log EX$

$$\mathbb{E}_{\widetilde{Y} \mid \beta_0} E_{S \mid Y} E_{\beta \mid S,Y} \left[1/2 \|\widetilde{X}(\beta - \beta_0)\|_2^2 + (\widetilde{Y} - \widetilde{X}\beta_0)' \widetilde{X}(\beta - \beta_0) \right].$$

Applying the Cauchy-Schwartz inequality we have $\mathbb{E}_{\widetilde{Y} \mid \beta_0} E_{\beta \mid Y} (\widetilde{Y} - \widetilde{X}\beta_0)' \widetilde{X}(\beta - \beta_0) \leq E_{\beta \mid Y} \|\widetilde{X}(\beta - \beta_0)\|_2 \mathbb{E}_{\widetilde{Y} \mid \beta_0} \|\widetilde{\varepsilon}\|_2$. Because $\|\widetilde{\varepsilon}\|_2^2$ has a χ_m^2 distribution with m degrees of freedom, the square root of this variable has an expectation $\sqrt{2}\Gamma((m+1)/2)/\Gamma(m/2) \leq \sqrt{m}$.

Remark 6. *Lemma 11 invites the possibility of quickly bounding the predictive risk using concentration rate results for $\|\widetilde{X}(\beta - \beta_0)\|_2$. In the normal means model from Section 2, one can use Theorem 2.5 in [7] which shows that the typical posterior probability that $\|\beta - \beta_0\|_2$ exceeds the multiple of the minimax rate $r_n = \sqrt{s_n \log(n/s_n)}$ has a sub-Gaussian tail to conclude that $\mathbb{E}_{Y \mid \beta_0} E_{\beta \mid Y} \|\beta - \beta_0\|_2^l \lesssim r_n^l$ for $l \in \mathbb{N}$. Together with the fact that $\|\widetilde{\varepsilon}\|_2$ has an expectation bounded by $\sqrt{r \times n}$ in the setup from Section 2, we obtain a loose upper bound $s_n \log(n/s_n) + \sqrt{rns_n \log(n/s_n)}$ which does not match the minimax rate $1/(1+r)s_n \log(n/s_n)$. This justifies the somewhat more delicate treatment of the predictive risk in Section 2.*

Lemma 11 applies to *any* Spike-and-Slab prior and indicates, for example, that if one can bound the *average* posterior mean of $\|\beta - \beta_0\|_1$ one can directly obtain an upper bound on the prediction risk (despite perhaps quite loose). [6] showed that average posterior probability that the distance $\|\beta - \beta_0\|_1$ exceeds the near-minimax rate $s_n\sqrt{1/n \log p}$ (with multiplication constants depending on suitable compatibility numbers) goes to zero. Using the inequality $\|\widetilde{X}(\beta - \beta_0)\|_2 \leq \|\widetilde{X}\| \|\beta - \beta_0\|_1$ one can quickly deduce a rate $m/n \times s_n^2 \log p$ from (3.5) when $\sqrt{n} \leq s_n\sqrt{\log p}$. This rate, however, has an unjustifiable extra factor s_n and may be improved by transporting the ℓ_2 concentration rate under some suitable of

the largest eigenvalue of $\widetilde{X}'\widetilde{X}$. One way or another, however, Theorem 2 in [6] *does not* immediately imply a bound on the posterior mean of $\|\beta-\beta_0\|_1$ or $\|\beta-\beta_0\|_2$, even for most typical Y. This prevents us from quickly leveraging existing posterior concentration results in regression to bound the predictive risk. We will pursue a different direction to obtain a more concrete upper bound on the KL risk, total-variation (TV) risk and a typical KL distance. We now focus on two particular choices of $\pi_{S,\lambda}(\cdot)$ in (3.1).

3.1 The Uniform Slab

Rather than establishing exact minimaxity, we want to understand the rate at which the KL risk $\rho_{n,m}(\beta_0,\hat{p})$ increases allowing for both p and s_n to grow with n. Since the uniform prior was shown to dominate the plug-in estimator [11], it is natural to consider a version of the Spike-and-Slab prior (3.1) with a uniform slab.

Lemma 12. *Consider a Spike-and-Slab prior* (3.1) *with* $\pi_{S,\lambda}(\beta_S)\propto(\lambda/2)^s$ *where* $s=|S|$ *and with* $R=\min\{n,p\}$. *Assume that* $\pi(s)\propto c^{-s}p^{-as}$ *for some* $a,c>0$. *Assume that* $\|X\|=\sqrt{n}$ *and* $\|\widetilde{X}\|=\sqrt{m}$. *Then*

$$\sup_{\beta_0\in\Theta_p(s_n)}\rho_{n,m}(\beta_0,\hat{p})\le s_n\log(\mathrm{e}\,c\,p^{a+1})+s_n/2\log[2/\pi(n+m)/\lambda^2].$$

Proof. We use Lemma 10. Jensen's inequality $E\log X\le\log EX$ and the fact that $E\mathrm{e}^{\mu+\sigma X}=\mathrm{e}^{\mu^2+\sigma^2/2}$ for a standard normal r.v. X implies

$$\mathbb{E}_{Y\,|\,\beta_0}\log\Lambda_{n,\beta_0,\pi}(Y,X)\le\log\int_\beta\mathrm{e}^{-1/2\|X(\beta-\beta_0)\|_2^2}\mathbb{E}_{Y\,|\,\beta_0}\mathrm{e}^{(Y-X\beta_0)'X(\beta-\beta_0)}\pi(\beta)\mathrm{d}\beta=0.$$

Next, changing variables $b(S)\equiv\beta_S-\beta_{0,S}$ where $\beta_S=\{\beta_i:i\in S\}$ we find that (using similar arguments as in the proof of Theorem 6 in [6]) for X_S a submatrix of columns of X inside S and $s_0=|S_0|$

$$\begin{aligned}\Lambda_{n+m,\beta_0,\pi}(Z,\bar{X})&\ge\frac{\pi(s_0)}{\binom{p}{s_0}}(\lambda/2)^{s_0}\int\mathrm{e}^{-1/2\|\bar{X}b(S_0)\|_2^2}\mathrm{d}b(S_0)\ge\frac{\pi(s_0)(\lambda/2)^{s_0}}{\binom{p}{s_0}}\frac{(2\pi)^{s_0/2}}{|\bar{X}'_{S_0}\bar{X}_{S_0}|^{1/2}}\\&\gtrsim\frac{c^{-s_0}p^{-as_0}s_0^{s_0}}{(\mathrm{e}p)^{s_0}}\frac{(\pi\lambda^2/2)^{s_0/2}}{\|\bar{X}\|^{s_0}}\ge\left(\frac{s_0}{\mathrm{e}\,c\,p^{a+1}}\right)^{s_0}\frac{(\pi\lambda^2/2)^{s_0/2}}{(n+m)^{s_0/2}}\end{aligned}$$

which yields the desired conclusion.

Remark 7. *While the scaling of the uniform prior* $\pi_{S,\lambda}(\beta) \propto (\lambda/2)^s$ *is immaterial for coefficient estimation, it does influence posterior model probabilities. [6] showed that such a prior yields a posterior that is asymptotically indistinguishable from the posterior obtained under the Laplace prior when* λ *is small (e.g.* $\lambda = \sqrt{n}/p$*). This choice would yield an upper bound of the order* $s_n[\log(p) \vee \log(1 + m/n)]$*. For a fixed* λ*, we obtain an upper bound on the risk of the rate* $s_n[\log(p) \vee \log(n + m)]$*.*

3.2 The Laplace Slab

In their Theorem 6, [6] show that posteriors under the uniform slab and the Laplace slab are asymptotically indistinguishable in terms of TV distance under suitable design compatibility conditions when λ is small. Since the posterior predictive distribution is a functional of the posterior, we can expect the posterior predictive distributions to be indistinguishable as well. We obtain risk bounds for the total-variation predictive risk

$$\rho^{TV}_{n,m}(\beta_0, \hat{p}) = \mathbb{E}_{Y \mid \beta_0} \|\pi(\widetilde{Y} \mid \beta_0) - \hat{p}(\widetilde{Y} \mid Y)\|_{TV}$$

as well as upper bounds on the typical KL distance between $\hat{p}(\widetilde{Y} \mid Y)$ and $\pi(\widetilde{Y} \mid \beta_0)$. The rate that we are targeting will be somewhat compatible with the rate of the typical point prediction error $\|\widetilde{X}\beta - \widetilde{X}\beta_0\|_2^2$. Theorem 2 in [6] established the concentration rate for $\|X\beta - X\beta_0\|_2^2$ as proportional to $c(S_0)s_n \log p$ where $c(S_0)$ depends on the properties of the design sub-matrix X_{S_0} of the true model S_0 (such as the compatibility number $\phi(S_0)$ defined in Definition 8 and the smallest scaled singular value $\widetilde{\phi}(s)$ defined in Definition 9) for a penalty parameter λ satisfying $\frac{\|X\|}{p} \leq \lambda \leq 2\bar{\lambda}$ where $\bar{\lambda} = 2\|X\|\sqrt{\log p}$. Similarly as in Lemma 12, our target rate will be $s_n[\log(p) \vee \log(1 + m/n)]$. Our predictive inference risk analysis will focus on sparse β_0 vectors which allow for consistent estimation. To this end, we will impose certain identifiability conditions.

Definition 6. *(Beta-min Condition) We define* s_n*-sparse vectors* β_0 *with support* $S_0 = S_{\beta_0}$ *and strong-enough signals as (for some* $M > 0$*)*

$$\widetilde{\Theta}(s_n, M) = \left\{ \beta_0 \in \mathbb{R}^p : \|\beta_0\|_0 \leq s_n \text{ and } \inf_{i \in S_0} |\beta_{0i}| \geq \beta_{min} \equiv \frac{M}{\widetilde{\psi}(S_0)^2} \frac{\sqrt{|S_0| \log p}}{\|X\| \phi(S_0)} \right\}. \quad (3.6)$$

where $\widetilde{\psi}(S_0)$ *is the smallest scaled sparse eigenvalue of "small models" defined as*

$$\widetilde{\psi}(S) = \widetilde{\phi}\left((2 + 3/A_4 + 33/\phi(S)^2\lambda/\bar{\lambda})|S|\right). \tag{3.7}$$

Corollary 1 in [6] (rephrased as Theorem 10 in the Supplement) concluded consistent estimation of S_0 uniformly for all $\beta_0 \in \widetilde{\Theta}(s_n, M)$ for which $\widetilde{\psi}(S_0)$ and $\phi(S_0)$ are bounded away from zero. In addition, for our risk analysis we impose a mild design assumption.

Assumption 8. *(Design Assumption) For* $S_0 = \{i : \beta_{0i} \neq 0\}$ *with* $s_0 = |S_0|$*, denote with* $X_0 = X_{S_0}$ *and assume* $\|X_0\| = \sqrt{n}$ *and that for some* $0 < b < 1$ *and* $d > 0$

$$\frac{4 \times \chi^2_{s_0,1/2}}{(1-b)^2\beta^2_{min}} < \lambda_{min}(X_0'X_0) < \lambda_{max}(X_0'X_0) \lesssim n(\log p)^d.$$

where $\chi^2_{s_0,1/2}$ *is the median of the Chi-squared distribution with* s_0 *degrees of freedom and* β_{min} *is the signal threshold from* (3.6)*.*

Remark 9. *Since* $s_0 - 1 < \chi^2_{s_0,1/2} < s_0$ *for* $s_0 \geq 2$*, the smallest eigenvalue part of Assumption 8 essentially requires that* $\lambda_{min}(X_0'X_0)$ *is at least of the order* $n/\log p$*. The upper bound allows another logarithmic deviation from the orthogonal design. From an upper bound* $\lambda_{max}(X_0'X_0) \leq \max_{j=1,\ldots,s_0} \|(X_0'X_0)^j\|_1 \leq ns_0$ *(which holds due to the Cauchy-Schwartz inequality and because* $\|X_0\| = \sqrt{n}$*) we see that the upper bound requirement is satisfied when* $s_0 \lesssim (\log p)^d$*.*

The next Theorem characterizes the TV predictive risk as well as the concentration rate for the KL distance for high-dimensional sparse regression with Spike-and-Slab priors with Laplace slab.

Theorem 7. *Consider a Spike-and-Slab prior* (3.1) *with* $\pi_{S,\lambda}(\beta) = (\lambda/2)^s \mathrm{e}^{-\lambda\|\beta_S\|_1}$ *where* $\lambda = \sqrt{n}/p$ *and with* $R = p$ *where* $\pi(s)$ *satisfies* (3.2) *with* $A_4 > 2$*. Under Assumption 8 we have for* $s_n \leq n = o(p/\sqrt{\log p})$*, some suitable* $M > 0$*,* $d > 0$ *and for any* $c_0 > 0$

$$\sup_{\beta \in \widetilde{\Theta}(s_n): \phi(S_0) \wedge \widetilde{\psi}(S_0) \geq c_0} \left[\rho^{TV}_{n,m}(\beta_0, \hat{p})\right]^2 \lesssim \mu_n$$

and

$$\sup_{\beta \in \widetilde{\Theta}(s_n): \phi(S_0) \wedge \widetilde{\psi}(S_0) \geq c_0} \mathbb{P}_{Y \mid \beta_0}\left(KL(\pi(\widetilde{Y} \mid \beta_0), \hat{p}(\widetilde{Y} \mid Y)) \gtrsim \mu_n\right) = o(1).$$

for $\mu_n = s_n[(\log p)^{d-1 \vee 1} \vee \log(1 + m/n)]$*.*

Proof. Section 12

The proof of Theorem 7 relies on subset selection consistency and utilizes a novel upper bound on the marginal likelihood under the Laplace prior. Similarly as in Section 2.3, the rate in Theorem 7 is again *adaptive*, i.e. the prior does not depend on s_n, and (for $m/n \lesssim \log p$) is proportional to the minimax estimation rate $s_n \log(p/s_n)$ up to a logarithmic factor. Such adaptation is a collateral benefit of the fully Bayesian treatment of sparse regression. The results from Section 2 cannot be quickly concluded as a special case with $X = \widetilde{X} = I$. Our analysis of the sparse normal-means model in Section 2 was sharper, where under the beta-min condition we were able to obtain the actual minimax rate $1/(1+r)s_n \log(n/s_n)$ without any additional logarithmic factors.

4 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. By considering a hierarchical extension, we show that adaptation to s_n is possible with the usual beta-binomial prior on the sparsity pattern. Finally, we obtain an adaptive rate for predictive density estimation in high-dimensional regression which mirrors the estimation rate of sparse regression coefficients.

5 Proof of Theorem 3

We recall the risk decomposition in Remark 1 (or equivalently in Lemma 3 in [20])

$$\rho(\theta, \hat{p}) = Eg(Z, \theta, v), \quad \text{where} \quad g(z, \theta, v) = \theta^2/(2r) + \log N_{\theta,1}(z) - \log N_{\theta,v}(z), \tag{5.1}$$

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

$$I_1^v = \sqrt{v} \frac{\Phi(\mu_1/\sqrt{v})}{\phi(\mu_1/\sqrt{v})} \quad \text{and} \quad I_2^v = \sqrt{v} \frac{\Phi(-\mu_2/\sqrt{v})}{\phi(-\mu_2/\sqrt{v})} \tag{5.2}$$

and

$$\mu_1 = z\sqrt{v} + \theta - v\lambda \quad \text{and} \quad \mu_2 = z\sqrt{v} + \theta + v\lambda. \tag{5.3}$$

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

5.1 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 (5.1) depending on the magnitude of $|z+\theta|$ (which has the same distribution as $|Y|$). In Section 2.2 we emphasized that depending on $|Y|$, 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 [20].

5.1.1 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 Y, we would then expect the risk to be dominated by the slab risk. We can write

$$\log \frac{N_{\theta,1}(z)}{N_{\theta,v}(z)} \leq \log\left(\frac{R_1}{R_v}\right) + \log\left(\frac{\frac{1}{R_1} + \frac{\eta}{1-\eta}\lambda}{\frac{1}{R_v} + \frac{\eta}{1-\eta}\frac{\sqrt{v}}{2}\lambda}\right). \tag{5.4}$$

This expression, in fact, brings us back to the risk decomposition in Lemma 2.11 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 (12.2)). Indeed, $\theta^2/(2r) + \log(R_1/R_v)$ bounds the LASSO risk in Lemma 1 and the second summand in (12.2) corresponds to $E\log N^{SS}_{\theta,1}(z) - E\log N^{SS}_{\theta,v}(z)$ in Lemma 2.11. The goal is to show that the second summand in (12.2) is small. Indeed, we have

$$\log\left(\frac{1}{R_1} + \frac{\eta}{1-\eta}\lambda\right) - \log\left(\frac{1}{R_v} + \frac{\eta}{1-\eta}\frac{\sqrt{v}}{2}\lambda\right) \leq \log\left(\frac{2}{\sqrt{v}}\right) + \log\left[1 + \frac{1}{\lambda}\left(\frac{1-\eta}{\eta}\right)^{1-v}\right]. \tag{5.5}$$

Next, we have

$$\frac{R_1}{R_v} < \frac{1}{\Phi(-t_v)} \exp\left\{\frac{\lambda^2(1-v)}{2} + \lambda|z|(1-\sqrt{v}) + \frac{\theta^2}{2}\left(1 - \frac{1}{v}\right) + z\theta\left(1 - \frac{1}{\sqrt{v}}\right)\right\}. \tag{5.6}$$

Now we focus on the term $1/\Phi(-t_v)$ in (5.6). We have $-t_v > -\lambda\sqrt{v}$ and (because $\lambda\sqrt{v} > 0$ we can apply Lemma 17)

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

This yields (using Lemma 16)

$$\log[1/\Phi(-t_v)] \leq \log\sqrt{2\pi} + \frac{\lambda^2 v}{2} + \frac{4}{\lambda^2} + \log(\lambda). \tag{5.7}$$

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

$$R(x_\eta) = \left[\frac{1-\eta}{\eta}\right]^v \quad \text{for some} \;\; h^L_\eta < x_\eta < h^U_\eta < 0 \tag{5.8}$$

where

$$h_\eta^U = -\sqrt{2v \log\left[(\pi 2)^{-v/2}(1-\eta)/\eta\right]} \quad \text{and} \quad h_\eta^L = -\sqrt{2v \log\left[(\pi/2)^{-v/2}(1-\eta)/\eta\right]}. \tag{5.9}$$

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

$$g_1(z,\theta,v) = -\log \Phi(-t_v) + \frac{\lambda^2(1-v)}{2} + \lambda|z|(1-\sqrt{v}) + z\theta\left(1 - \frac{1}{\sqrt{v}}\right) \tag{5.10}$$

$$+ \log(4/\sqrt{v}) + (1-v)\log\left(\frac{1-\eta}{\eta}\right) + \log(1+1/\lambda). \tag{5.11}$$

5.1.2 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

$$g_2(z,\theta,v) = \frac{2\lambda^2 + |z|^2}{r} + 4(1-v)\log\left(\frac{1-\eta}{\eta}\right) + \log(1+\lambda). \tag{5.12}$$

5.1.3 Combining the Cases

We combine the two bounds in (5.11) and (5.12) to obtain

$$g(z,\theta,v) \le g_1(z,\theta,v)\mathbb{I}(z \in \mathcal{A}_1) + g_2(z,\theta,v)\mathbb{I}(z \in \mathcal{A}_2). \tag{5.13}$$

Here, we are pasting two risk bounds depending on the two scenarios for $|z+\theta|$ as opposed to pasting two estimators [20]. 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 (5.11), (5.12) and (5.7) and using the fact that $|z|$ has a folded normal distribution with $E|z| = \sqrt{2/\pi}$ and $E|z|^2 = 1$

$$Eg(z,\theta,v) \le \log\left(4\sqrt{2\pi/v}\right) + \frac{4}{\lambda^2} + \frac{\lambda^2}{2} + \frac{2\lambda^2+1}{r} + \lambda(1-\sqrt{v})\sqrt{2/\pi} \tag{5.14}$$

$$+ 5(1-v)\log\left(\frac{1-\eta}{\eta}\right) + \log(2\lambda+\lambda^2+1). \tag{5.15}$$

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 may 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

$$\rho(\theta,\hat{p}) \le 5(1-v)\log(n/s_n) + \widetilde{C}_r^*, \tag{5.16}$$

where

$$\widetilde{C}_r^* = \log(8\sqrt{\pi}) + 10 + 2C_r^*(1-v) + C_r^*\sqrt{2/\pi} + \log(2\sqrt{C_r^*(1-v)} + C_r^*(1-v)). \quad (5.17)$$

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

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

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

$$\widetilde{C}_r = \log(8\sqrt{2\pi}) + 9 + \frac{C_r}{r+1}(1/2+4) + \sqrt{C_r}\sqrt{1/(8\pi)} + \log(2\sqrt{C_r} + C_r). \quad (5.18)$$

5.1.4 The case when $\theta = 0$

This step is analogous to Section 2 in [20]. Because $N_{\theta,v}(z) > 1$ and from Jensen's inequality and the fact that $E\mathrm{e}^{cZ} = \mathrm{e}^{-c^2/2}$

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

we have for $\theta = 0$

$$\rho(0, \hat{p}) < \log\left(1 + \frac{\eta}{1-\eta}\right) < \frac{\eta}{1-\eta}. \quad (5.19)$$

The statement (2.14) of Theorem 3 follows from the fact that, for separable (independent product) priors,

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

Plugging the inequalities (5.16) and (5.19) 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. *Journal of Computational and Graphical Statistics 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] Karp, D. and S. Sitnik (2009). Inequalities and monotonicity of ratios for generalized hypergeometric function. *Journal of Approximation Theory 161*, 337–352.

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

[16] 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.

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

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

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

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

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

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

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

[24] 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.

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

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

[27] Ročková, V. and E. George (2018). The Spike-and-Slab LASSO. *Journal of the American Statistical Association 113*, 431–444.

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

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

[30] 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

The Appendix contains proofs of all remaining theorems as well as auxiliary lemmata.

6 Proof of Theorem 1

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

6.1 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 (2.3), 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 (5.2). Recall also the definition of μ_1 and μ_2 in (5.3). 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 $\{z \leq -\theta/\sqrt{v}\}$.

6.1.1 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

$$E \log N_{\theta,v}^{LASSO}(Z) \leq \log(\lambda\sqrt{2\pi v}) + T_1^v(\theta) + T_2^v(\theta) \tag{6.1}$$

where

$$T_1^v(\theta) = E\,\mathbb{I}(z > -\theta/\sqrt{v})\mu_1^2/(2v) \quad \text{and} \quad T_2^v(\theta) = E\,\mathbb{I}(z \leq -\theta/\sqrt{v})\mu_2^2/(2v). \tag{6.2}$$

We find that

$$T_1^v(\theta) + T_2^v(\theta) = E\frac{(|z + \theta/\sqrt{v}| - \lambda\sqrt{v})^2}{2} = \frac{\theta^2}{2v} + \frac{1 + \lambda^2 v}{2} - \lambda E|z\sqrt{v} + \theta|. \tag{6.3}$$

6.1.2 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

$$E \log N_{\theta,v}^{LASSO}(z) > \log(\lambda\sqrt{\pi v/2}) + T_1^v(\theta) + T_2^v(\theta) + T_v^3(\theta) \tag{6.4}$$

where $T_1^v(\theta)$ and $T_1^v(\theta)$ were defined earlier in (6.2) and where (because in the case (a) $\mu_1 > -\lambda v$ and in the case (b) $-\mu_2 > -\lambda v$)

$$T_v^3(\theta) = [P(\mu_1 > 0) + P(\mu_2 < 0)] \log 1/2 + P\left(|z + \frac{\theta}{\sqrt{v}}| < \lambda\sqrt{v}\right) \log \Phi(-\lambda\sqrt{v}). \tag{6.5}$$

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 17 we have

$$T_v^3(\theta) > \log 1/2 + P\left(|z + \frac{\theta}{\sqrt{v}}| < \lambda\sqrt{v}\right) \left(-\log\sqrt{2\pi} - \frac{\lambda^2 v}{2} - \log \frac{\sqrt{\lambda^2 v + 4} + \lambda\sqrt{v}}{2}\right). \tag{6.6}$$

6.1.3 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 16 we obtain the desired upper bound on $\rho(\theta, \hat{p})$.

6.2 The case when $\theta = 0$.

By 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 $\rho(0, \hat{p}) \leq -E \log N_{0,v}^{LASSO}(z)$. To find a lower bound for $E \log N_{0,v}^{LASSO}(z)$, we use the notation introduced I_1^v and I_2^v in (5.2) but now for the special case when $\theta = 0$. Similarly as in Section 6.1 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 17 which yields

$$\frac{2}{\sqrt{\mu_2^2/v + 4} + \mu_2/\sqrt{v}} < I_2^v = \frac{1 - \Phi(\mu_2/\sqrt{v})}{\phi(\mu_2/\sqrt{v})} < \frac{2}{\sqrt{\mu_2^2/v + 2} + \mu_2/\sqrt{v}}. \tag{6.7}$$

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

$$\frac{2}{\sqrt{\mu_1^2/v + 4} - \mu_1/\sqrt{v}} < I_1^v = \frac{1 - \Phi(-\mu_1/\sqrt{v})}{\phi(\mu_1/\sqrt{v})} < \frac{2}{\sqrt{\mu_1^2/v + 2} - \mu_1/\sqrt{v}}. \tag{6.8}$$

While here we only need the lower bounds, in the next Section 7 we utilize also the upper bounds. This yields

$$E \log \left[\frac{\lambda}{2}(I_1^v + I_2^v)\right] > \log(\lambda\sqrt{v}) - E \log(\lambda\sqrt{v} + |z|) - E \log \frac{\sqrt{\frac{4}{(\lambda\sqrt{v}+|z|)^2} + 1} + 1}{2} \tag{6.9}$$

We use Jensen's inequality $E \log(\lambda\sqrt{v} + |z|) \leq \log(\lambda\sqrt{v} + \sqrt{2/\pi})$ and Lemma 16 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

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

7 Lower Bound for the LASSO

We know that for separable priors (such as the Laplace product prior (2.3)) we have

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

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

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

Lemma 13. *Consider the Laplace prior (2.3) 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. First, we recall the lower bound for $E\log N_{1,0}^{LASSO}(z)$ in (6.9) obtained in Section 6.2.

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

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 6.2. Recall the notation I_1^v and I_2^v in (5.2) and μ_1 and μ_2 in (5.3). 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 (6.8). 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 (6.7). 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 16 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 $a>0$

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

8 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 (5.2) and (5.3) 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

$$\rho(\theta, \hat{p}) = \rho(\theta, \hat{p}_1) + E\log N_{\theta,1}^{SS}(z) - E\log N_{\theta,v}^{SS}(z) \tag{8.1}$$

where

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

Alternatively, we could write

$$\rho(\theta, \hat{p}) = \rho(\theta, \hat{p}_0) + E\log \widetilde{N}_{\theta,1}^{SS}(z) - E\log \widetilde{N}_{\theta,v}^{SS}(z) \tag{8.2}$$

where

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

We will utilize both of these characterizations. The decomposition (8.1) is useful when the observed data Y 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 Y is small. While the upper bound on the risk in Lemma 3 uses an average mixing weight $\Delta_\eta(Y)$, we pause with averaging over the distribution $\pi(Y \mid \theta)$ and bound the Kullback-Leibler loss in two

regimes depending on the magnitude of Y. 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. Throughout this section, we denote the observed data Y simply with x and thereby assume $x \sim N(\theta, 1)$.

8.1 The case when $\theta = 0$.

We utilize the expression (8.2) and (using the risk expression at $\theta = 0$ for Bayesian LASSO from Section 6.2) we find that

$$\rho(0,\hat{p}) \leq \frac{\sqrt{2}}{\lambda_0\sqrt{\pi v}} + \frac{4}{\lambda_0^2 v} + \log\left(1 + \frac{\eta}{1-\eta} E_{x\,|\,\theta=0} \frac{m_1(x)}{m_0(x)}\right) \tag{8.3}$$

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

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

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 17 (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 $-\lambda_0 < x = z \le 0$ we have $I_1^1(\lambda) \le I_2^1(\lambda)$ and (using Lemma 17)

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

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

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

(4) When $x = z \le -\lambda_0 < -\lambda_1$ 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 $\rho(0,\hat{p}) \lesssim s_n/n$.

8.2 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

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

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

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

Recall (from (9.3)) that under the Laplace prior with a parameter λ, the KL loss equals (using the expression $x + y/r = z/\sqrt{v} + \theta/v$)

$$K_\lambda(\theta, z) = \theta^2/(2r) + \log[\lambda/2(I_1^1(\lambda) + I_2^1(\lambda))] - E_{y\,|\,\theta} \log[\lambda/2(I_1^v(\lambda) + I_2^v(\lambda))]. \tag{8.4}$$

This means

$$\begin{aligned}\rho(\theta, \hat{p}) \leq & P[A_\eta(\theta, A, d)^c] \log\left[1 + \left(\frac{\eta}{1-\eta}\right)\frac{\lambda_1}{\lambda_0}(n/s_n)^d/A\right] + \int_{z \in A_\eta^c(\theta,A,d)} K_{\lambda_0}(\theta, z)\phi(z)\mathrm{d}z \\ & + P[A_\eta(\theta, A, d)] \log\left[1 + \left(\frac{1-\eta}{\eta}\right)\frac{\lambda_0}{\lambda_1}A(n/s_n)^{-d}\right] + \int_{z \in A_\eta(\theta,A,d)} K_{\lambda_1}(\theta, z)\phi(z)\mathrm{d}z.\end{aligned} \tag{8.5}$$

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

Lemma 14. *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 17)

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

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 8.1) 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

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

where $\widetilde{x}_c = \lambda_1 + \sqrt{2d \log[n/s_n]}$ 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 $\widetilde{x}_c < \lambda_0$ when $\lambda_0 = n/s_n$, we conclude that $A_\eta(\theta, A, d)^c \subset \{x : |x| < \widetilde{x}_c\}$. □

Now we continue with the proof of Theorem 4.

8.2.1 When $r \in (0, 1)$

Going back to (8.5), 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 6.1.3, 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 $\rho(\theta, \hat{p}) \lesssim (1 - v)\log(n/s_n)$.

8.2.2 When $r \in [1, \infty)$

Using (8.5), 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 14 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 6.1.3 we find

$$\begin{aligned} K_\lambda(z) &\leq \frac{\theta^2}{2r} + \log(2/\sqrt{v}) + \frac{(|z+\theta| - \lambda)^2}{2} - \frac{(|z+\theta/\sqrt{v}| - \lambda\sqrt{v})^2}{2} - \log\Phi(-\lambda\sqrt{v}) \\ &\leq \frac{\theta^2}{2r} + \log(\lambda 4\sqrt{2\pi/v}) + \frac{(|z+\theta| - \lambda)^2}{2} - \frac{(|z+\theta/\sqrt{v}| - \lambda\sqrt{v})^2}{2} + \lambda^2 v/2 + \frac{1}{\lambda^2}. \end{aligned} \tag{8.6}$$

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 (8.5) 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 $(1-v)\log(n/s_n)$.

9 Proofs of Lemmata

9.1 Proof of Lemma 1

We have the following risk definition

$$\rho(\theta, \hat{p}) = \int \pi(Y \mid \theta) \int \pi(\widetilde{Y} \mid \theta) \log[\pi(\widetilde{Y} \mid \theta)/\hat{p}(\widetilde{Y} \mid Y)]\mathrm{d}\widetilde{Y}\mathrm{d}Y. \tag{9.1}$$

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

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

and

$$\frac{\pi(\widetilde{Y} \mid \theta)}{\hat{p}(\widetilde{Y} \mid Y)} = \frac{\exp(\widetilde{Y}\theta/r - \theta^2/2r) \int \exp\{\mu Y - \mu^2/2\}\pi_1(\mu \mid \lambda)\mathrm{d}\mu}{\int \exp\{\mu(\widetilde{Y}/r + Y) - \mu^2/2(1 + 1/r)\}\pi_1(\mu \mid \lambda)\mathrm{d}\mu}. \tag{9.2}$$

Then

$$\begin{aligned}\log \frac{\pi(\widetilde{Y} \mid \theta)}{\hat{p}(\widetilde{Y} \mid Y)} =& \widetilde{Y}\theta/r - \theta^2/2r + \log \int \exp\{\mu Y - \mu^2/2\}\pi_1(\mu \mid \lambda)\mathrm{d}\mu \\ &- \log \int \exp\{\mu(\widetilde{Y}/r + Y) - \mu^2/2(1 + 1/r)\}\pi_1(\mu \mid \lambda)\mathrm{d}\mu. \end{aligned} \tag{9.3}$$

The expectation of the first term with respect to $\widetilde{Y} \sim N(\theta, r)$ is $\theta^2/(2r)$. Since $Y \sim N(\theta, 1)$ and $\widetilde{Y} \sim N(\theta, r)$, the expectation of the other two terms can be taken with respect to $Y + \widetilde{Y}/r \sim N(\theta/v, 1/v)$ 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.

9.2 Proof of Lemma 4

The Kullback-Leibler loss can be written as

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

Next, note that

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

and

$$\frac{1 - \Delta_\eta(Y)}{\Delta_\eta(Y)} \frac{\hat{p}_0(\widetilde{Y} \mid Y)}{\hat{p}_1(\widetilde{Y} \mid Y)} = \frac{(1 - \eta)}{\eta} \frac{m_0(Y)}{m_1(Y)} \frac{\hat{p}_0(\widetilde{Y} \mid Y)}{\hat{p}_1(\widetilde{Y} \mid Y)}.$$

Next,

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

To obtain $\rho(\theta, \hat{p})$ we take an expectation over $Y \sim N(\theta, 1)$ since $Y + \widetilde{Y}/r \sim N(\theta(1 + 1/r), 1 + 1/r)$ 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 $-E_{Y\,|\,\theta} \log \Delta_\eta(Y) = E_z \log N^{SS}_{\theta,1}(z)$ we obtain the desired expression.

9.3 Proof of Lemma 6

Because, given η, the Kullback-Leibler loss is separable

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

taking the expectation of both sides of (2.18) over the distribution $\pi(Y \,|\, \theta)$ yields

$$\rho(\theta, \hat{p}) \le (n - s_n) E_{Y\,|\,\theta} E_{\eta\,|\,Y} L(0, \hat{p}(\cdot \,|\, Y_i, \eta)) + E_{Y\,|\,\theta} E_{\eta\,|\,Y} \sum_{i:\theta_i \neq 0} L(\theta_i, \hat{p}(\cdot \,|\, Y_i, \eta)). \tag{9.4}$$

We use a similar expression as in the proof of Lemma 4 in Section 9.2. We first focus on the case when $\theta_i \neq 0$. From now on, we will be using simpler notation $x = Y$ and $y = \widetilde{Y}$. We will also denote with θ the unknown true parameter value and with μ the random vector used to estimate θ.

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

Then

$$\widetilde{g}(x, y, \theta, \eta) = \log \frac{\phi((y-\theta)/\sqrt{r})}{\phi(y/\sqrt{r})} - \log[1 - \Delta_\eta(x)] - \log\left[1 + \frac{\Delta_\eta(x)}{1 - \Delta_\eta(x)} \frac{\hat{p}_1(y \,|\, x)}{\hat{p}_0(y \,|\, x)}\right]. \tag{9.5}$$

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 5.1.1 and 5.1.2 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 5.1.1, 5.1.2 and 5.1.3 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

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

where the term $\widetilde{C}(\lambda,v,z)$ contains aspects of the bounds (5.11) and (5.12) that do not involve η. After taking the expectation over $Z \sim N(0,1)$ we arrive at a version of the (5.15), only with $\log[(1-\eta)/\eta]$ replaced by its conditional expectation. It also follows from (5.15) that

$$C(\lambda,v) \equiv E_z\widetilde{C}(\lambda,v,z) = \log\left(4\sqrt{\frac{2\pi}{v}}\right) + \frac{1}{\lambda^2} + \frac{\lambda^2}{2} + \frac{2\lambda^2+1}{r} + \lambda(1-\sqrt{v})\sqrt{\frac{2}{\pi}} + \log(2\lambda+\lambda^2+1)..$$
(9.6)

When $\theta_i = 0$, we have from (9.5)

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

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

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

The conditional expectation $E\left(\frac{\eta}{1-\eta}\,|\,Y\right)$ will be (for large n) very similar to the conditional expectation $E\left(\frac{\eta}{1-\eta}\,|\,Y_{\setminus i}\right)$, where $Y_{\setminus i}$ denotes the sub-vector of Y after excluding the i^{th} coordinate. For $\lambda > 2$, we can use the sandwich relation (9.9) to find that

$$\frac{E\left(\frac{\eta}{1-\eta}\,|\,Y\right)}{E\left(\frac{\eta}{1-\eta}\,|\,Y_{\setminus i}\right)} \le \frac{a+E[s_n(\mu)\,|\,Y]+1}{a+E[s_n(\mu_{\setminus i})\,|\,Y_{\setminus i}]} \le \frac{a+E[s_n(\mu)\,|\,Y]+1}{a+E[s_n(\mu)\,|\,Y]-1} < 1+\frac{2}{a-1}.$$

Then we can write

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

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

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

9.4 Proof of Lemma 7

Since the parameter η is hierarchically separated from the data by $\mu\sim\pi(\mu)$, we can write

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

Then

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

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

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

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

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

Then

$$E\left(\frac{\eta}{1-\eta}\,|\,Y\right)=E_{\theta\,|\,Y}E\left(\frac{\eta}{1-\eta}\,|\,\theta\right)=\int_\mu\pi(\mu\,|\,Y)\frac{\int_\eta\frac{\eta^{n+1}}{1-\eta}\pi(\eta)\left[1+\frac{1-\eta}{\eta}\frac{2}{\lambda}\right]^{n-s_n(\mu)}}{\int_\eta\eta^n\pi(\eta)\left[1+\frac{1-\eta}{\eta}\frac{2}{\lambda}\right]^{n-s_n(\mu)}}. \tag{9.7}$$

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

$$F_2(a',b',c';z)=\frac{\Gamma(c')}{\Gamma(b')\Gamma(c'-b')}\int_0^1\eta^{b'-1}(1-\eta)^{c'-b'-1}(1-\eta z)^{-a'}\mathrm{d}\eta. \tag{9.8}$$

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

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

In Lemma 15 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 [14] who considered a different integer shift on the second and third argument. We use the notation $f_\delta(a', b', c'; z)$ in (10.1) and use Lemma 15 to find that (when $z = \lambda/2 - 1 < 0$)

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

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

$$\frac{a+s}{b-1} < E\left(\frac{\eta}{1-\eta} \,|\, \mu\right) \le \frac{a+s+1}{b} < \frac{a+s+1}{b-1} \tag{9.9}$$

Similarly, we have

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

Applying Lemma 15 we find that the ratio above is monotone *decreasing* in $z = \lambda/2 - 1$ for $z > -1$. Thereby

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

□

9.5 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 [7] $\pi(s_n(\mu) = s) \leq \pi(s_n(\mu) = s-1)C_\pi$ for some $0 < C_\pi < 1$ which can be shown as in their Example 2.2). We can thereby apply Theorem 2.1 in [7]. Denoting $B_n(M) = \{\mu \in \mathbb{R}^n : \|\mu\|_0 \leq M\, s_n\}$, this theorem yields that for some suitable $M > 0$, the posterior concentrates on sparse vectors in the sense that $P_{Y\,|\,\theta}\Pi(B_n^c \,|\, Y) = o(1)$ for any $\theta \in \Theta_n(s_n)$. Lemma 7 then yields that for $a = 2$ and $b = n+1$

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

9.6 Proof of Lemma 9

We rewrite the model $Y \sim N(\theta, I)$ as

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

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 [6]). Then for $\theta \in \Theta_n(s_n, \widetilde{M})$, 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

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

The one-dimensional marginal likelihood under the Laplace prior satisfies

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

where

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

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

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

and

$$\frac{\pi(Y \mid S)}{\pi(Y \mid S^+)} = \frac{2}{\lambda} \frac{1}{I_1(Y_j) + I_2(Y_j)}.$$

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

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

Because of Lemma 17 and Lemma 16 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 $(a+b)^2 > a^2/2 - b^2$)

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

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

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

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 $C = 8\sqrt{2\pi}$

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

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 $c > 2 + d$

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

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

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

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

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

10 Auxiliary Results for Sparse Normal Means

Lemma 15. *For the Gauss hypergeometric function $F_2(a', b', c'; z)$ defined in (9.8), the ratio*

$$f_\delta(a', b', c'; z) = \frac{F_2(a', b' + \delta, c'; -z)}{F_2(a', b', c'; -z)} \tag{10.1}$$

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

Proof. We will prove this analogously as in Theorem 1 of [14]. We denote with $A_\delta = \frac{\Gamma(c')}{\Gamma(b'+\delta)\Gamma(c'-b'-\delta)}$ and write

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

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

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

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 ([18] Chapter IX, formula (1.1)). □

Lemma 16. *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 17. *(Mills Ratio Bounds) For the Gaussian Mills ratio $R(x) = (1 - \Phi(x))/\phi(x)$ we have for any $x > 0$*

$$\frac{2}{\sqrt{x^2 + 4} + x} < \frac{1 - \Phi(x)}{\phi(x)} < \frac{2}{\sqrt{x^2 + 2} + x}. \tag{10.2}$$

Proof. This result shown is in [4].

11 Auxiliary Results for Sparse Regression

We will be using the following notation $\|X\| = \max_{1 \le j \le p} \|X^j\|_2$, where X^j is the j^{th} column of the matrix X. For a vector $\beta \in \mathbb{R}^p$ and a set of indices $S \subseteq \{1, \dots, p\}$ we denote with $\beta_S = \{\beta_i : i \in S\}$ the active subset. Similarly, with $S_\beta = \{i : \beta_i \neq 0\}$ we denote the support of the vector $\beta \in \mathbb{R}^p$. Lastly, with X_S we denote the sub-matrix $[X^j : j \in S]$ of columns that belong to S.

Definition 8. *(Definition 2.1 in [6]) The compatibility number of a model $S \subset \{1, \dots, p\}$ is given by*

$$\phi(S) = \inf \left\{ \frac{\|X\beta\|_2 |S|^{1/2}}{\|X\| \|\beta_S\|_1} : \|\beta_{S^c}\|_1 \le 7 \|\beta_S\|_1, \beta_S \neq 0 \right\}.$$

Definition 9. *(Definition 2.3 in [6]) The smallest scaled singular value of dimension s is defined as*

$$\widetilde{\phi}(s) = \inf \left\{ \frac{\|X\beta\|_2}{\|X\| \|\beta\|_2} : 0 \neq |S_\beta| \le s \right\}.$$

Theorem 10. *(Consistent Model Selection) Assume the prior (3.2) with $A_4 > 1$ and $a < A_4 - 1$ where $s_0 \le p^a$. For any $c_0 > 0$ and $s_n \lambda \sqrt{\log p}/\|X\| \to 0$ there exists a set $\mathcal{D}_n \subset \mathbb{R}^n$ and $a_1 > 0$ such that for any $Y \in \mathcal{D}_n$*

$$\sup_{\beta_0 \in \widetilde{\Theta}(s_n, M) : \phi(S_0) \ge c_0, \widetilde{\psi}(S_0) \ge c_0} \mathbb{P}(\beta : S_\beta \neq S_0 \,|\, Y) \lesssim \frac{1}{p^{a_1}}$$

where $P(\mathcal{D}_n^c) \lesssim p^{-c_1}$ for some $c_1 > 0$.

Proof. This follows from the proof of Corollary 1 in [6].

Lemma 18. *Assume that X_0 has been normalized such that $\|X_0\| = n$. Given the Laplace prior $\pi \equiv \pi_{\lambda,S_0}(\beta) = (\lambda/2)^s \mathrm{e}^{-\lambda\|\beta\|_1}$ supported on S_0 with $s = |S_0|$, we have*

$$\log \Lambda_{n,\beta_0,\pi}(Y,X) \geq -\lambda\|\beta_0\|_1 - 1/2 - \lambda/n + s\log(\lambda/n) - \log s!$$

almost surely.

Proof. Analogous to Lemma 2 in [6].

Lemma 19. *Assume the Laplace prior $\pi \equiv \pi_{\lambda,S_0}(\beta) = (\lambda/2)^s \mathrm{e}^{-\lambda\|\beta\|_1}$ supported on S_0 of size $s = |S_0|$. Under the Assumption 8 we have for $\beta_0 \in \widetilde{\Theta}(s_n, M)$*

$$E_{Y\sim N(X\beta_0,I)} \log \Lambda_{n,\beta_0,\pi}(Y,X) \leq \frac{s}{2}\left[1 + 2\lambda^2 n(\log p)^d + C_0\sqrt{s/n\log p} + \log\left(\frac{8\pi}{n}\right)\right] - \lambda\|\beta_0\|_1.$$

where C_0 depends on $\widetilde{\psi}(S_0), \phi(S_0)$ and b.

Proof. For a vector $\beta \in \mathbb{R}^s$, we denote with $sign(\beta)$ an $(s \times 1)$ vector of $sign(\beta_i)$. Next, for $\pi \equiv \pi_{\lambda,S_0}(\beta)$ and $\Lambda_{n,\beta_0,\pi}(Y,X) \equiv \int_\beta \Delta_{n,\beta,\beta_0}(Y,X)\pi_{S_0,\lambda}(\beta)\mathrm{d}\beta$ with $\Delta_{n,\beta,\beta_0}(Y,X)$ defined in (3.4) we can write

$$\Lambda_{n,\beta_0,\pi}(Y,X) = \mathrm{e}^{-1/2\|X\beta_0\|_2^2 - \beta_0' X'(Y - X\beta_0)}(\lambda/2)^s \times \mathcal{I}, \tag{11.1}$$

where, using a shorthand notation $X_0 = X_{S_0}$ and $\beta = \beta_{S_0}$,

$$\mathcal{I} = \int \mathrm{e}^{-1/2\|X_0\beta\|_2^2 + \beta'[X_0'Y - \lambda\, sign(\beta)]}\mathrm{d}\beta. \tag{11.2}$$

For $\beta \in \mathbb{R}^s$, there are 2^s patterns $\{\pm 1\}^s$ of $sign(\beta)$ and we denote this set with Ξ. We write $\mathbb{R}^s = \cup_{\kappa\in\Xi}\mathcal{O}_\kappa$ where $\mathcal{O}_\kappa$ corresponds to an orthant that corresponds to the sign pattern indexed by κ. For each $\kappa \in \Xi$ we define a shrinkage estimator

$$\mu^\kappa = (X_0'X_0)^{-1}(X_0'Y - \lambda \mathbf{1}_\kappa)$$

where $\mathbf{1}_\kappa$ corresponds to the ± 1 pattern based on the orthant κ. We denote with $\phi(x;\mu,\Sigma)$ the density of an s-variate normal distribution with mean μ and variance matrix Σ. Then we decompose the integral (11.2) into

$$\mathcal{I} = \sum_{\kappa\in\Xi} J_\kappa, \quad \text{where} \quad J_\kappa = \frac{\int_{\mathcal{O}_\kappa}\phi(\beta;\mu^\kappa,(X_0'X_0)^{-1})\mathrm{d}\beta}{\phi(\mu^\kappa;0,(X_0'X_0)^{-1})}. \tag{11.3}$$

Next, we denote with $A_{\kappa^*} = \{Y : J_{\kappa^*} = \max_\kappa J_\kappa\}$ and write

$$E_{Y\sim N(X\beta_0,I)} \log \mathcal{I} \leq E_{Y\sim N(X\beta_0,I)} \sum_{\kappa^*\in\Xi} \mathbb{I}(Y \in A_{\kappa^*}) \log(2^s J_{\kappa^*}).$$

Denote with $\mu = (X_0'X_0)^{-1}X_0'Y$ the MLE estimator for the true model S_0, then we have

$$J_\kappa \leq \frac{1}{\phi(\mu^\kappa; 0, (X_0'X_0)^{-1})} = \frac{\mathrm{e}^{\mu'(X_0'X_0)\mu/2 - \lambda \mathbf{1}_\kappa'\mu + \lambda^2 \mathbf{1}_\kappa'(X_0'X_0)\mathbf{1}_\kappa/2}}{(2\pi)^{-s/2}|X_0'X_0|^{1/2}}.$$

Then since $E\mu'(X_0'X_0)\mu = tr(X_0(X_0'X_0)^{-1}X_0'EYY') = \mathrm{rank}(X_0)/2 + \|X\beta_0\|_2^2/2$ we have

$$\begin{aligned} E_{Y\sim N(X\beta_0,I)} \log I \leq & s/2\log(8\pi) - s/2\log n + s/2 + \|X\beta_0\|_2^2/2 + \lambda^2 \times ns(\log p)^d \\ & - E_{Y\sim N(X\beta_0,I)} \sum_{\kappa^*\in\Xi} \mathbb{I}(Y \in A_{\kappa^*}) \lambda \mathbf{1}_{\kappa^*}'\mu, \end{aligned} \tag{11.4}$$

where we used the fact that $\mathbf{1}_\kappa'(X_0'X_0)\mathbf{1}_\kappa \leq \lambda_{max}(X_0'X_0)s \leq ns(\log p)^d$ under Assumption 8. Next, recall the definition of $\kappa^* = \arg\max_{\kappa\in\Xi} J_\kappa$ which is a random variable in Y. Now, we inspect the occurrence of an event $\{\mathbf{1}_{\kappa^*} = \mathrm{sign}(\mu)\}$. We can rewrite J_κ as

$$J_\kappa = \mathrm{e}^{\mu'(X_0'X_0)\mu/2} \int_{\mathcal{O}_\kappa} \mathrm{e}^{-(\beta-\mu)'(X_0'X_0)(\beta-\mu)/2 - \lambda\|\beta\|_1} \mathrm{d}\beta.$$

Since the function $\mathrm{e}^{-\lambda|\beta|}$ is symmetrical around the origin, its integral for each orthant $\mathcal{O}_\kappa$ is the same and equals $1/\lambda^s$. The orthant intergals change after reweighting by the Gaussian kernel $\mathrm{e}^{-(\beta-\mu)'(X_0'X_0)(\beta-\mu)/2}$. Filtering the Gaussian "likelihood" $N(\mu, (X_0'X_0)^{-1})$ through the Laplace prior has the effect of squashing the Gaussian distribution towards (not across) coordinate axes and, depending on the magnitude of λ, creating ridgelines at the coordinate axes. In other words, multiplying the Gaussian likelihood by a Laplace density does not change the ordering of Gaussian orthant probabilities. The orthant $\mathcal{O}_{\kappa^*}$ which has the highest Gaussian orthant probability for the distribution $N(\mu, (X_0'X_0)^{-1})$ will also be the one for which J_κ is the largest. In addition, the orthant $\mathcal{O}_{\kappa^*}$ will very often be the one containing the mode μ. This will *always* be the case when $X_0'X_0 = I_s$. It is reasonable to expect that when the correlation among the columns in X_0 is not too strong and/or when the signal β_0 is strong enough, this event will happen with overwhelmingly large probability. Below, we conclude that Assumption 6 and 8 are sufficient conditions for this to happen.

Denote the ellipse $\mathcal{E}_\mu(\chi^2_{s,1/2}) = \{\beta \in \mathbb{R}^s : (\beta-\mu)'X_0'X_0(\beta-\mu) \le \chi^2_{s,1/2}\}$ where $\chi^2_{s,1/2}$ is the median of the Chi-squared distribution with s degrees of freedom. This ellipse contains $1/2$ of the Gaussian mass $N(\mu, (X_0'X_0)^{-1})$. If $\mathcal{E}_\mu(\chi^2_{s,1/2}) \cap \mathcal{O}_\kappa = \mathcal{E}_\mu(\chi^2_{s,1/2})$, then orthant $\mathcal{O}_\kappa$ has the largest orthant probability and thereby also the integral J_κ, i.e. $\kappa = \kappa^*$. This happens, for instance, when

$$|\mu_j| > c \equiv 2\sqrt{\chi^2_{s,1/2}/\lambda_{min}(X_0'X_0)} \tag{11.5}$$

because the axes of the ellipse are of the length $c_j = 2\sqrt{\widetilde{\lambda}_j \times \chi^2_{s,1/2}}$, where $\widetilde{\lambda}_j$ is the j^{th} eigenvalue of $(X_0'X_0)^{-1}$. This means that

$$\{|\mu_j| > c\} \subset \{sign(\mu_j) = 1_{\kappa^* j}\} \quad \forall j \in \{1, \ldots, s\}. \tag{11.6}$$

On the other hand, when $\mathcal{E}_\mu(\chi^2_{s,1/2})$ crosses the j^{th} coordinate axis, then $sign(\mu_j)$ may or may not correspond to $1_{\kappa^* j}$.

Next, we slice $\mathbb{R}^n$ into 2^s mutually exclusive sets depending on whether or not $|\mu_j| \ge c$. Define by I_m the set of all subsets of $\{1, \ldots, s\}$ of size m. Then we have

$$\mathbb{R}^n = \bigcup_{m=0}^{s} \bigcup_{I \in I_m} A(I),$$

where $A(I) = \{Y : |\mu_j| > c \text{ for } j \in I \text{ and } |\mu_j| \le c \text{ for } j \notin I\}$. Due to (11.6), for each $I \in I_m$ we have on the event $\{Y \in A_{\kappa^*}\} \cap A(I)$

$$\mathbf{1}'_{\kappa^*}\mu \ge \sum_{j\in I}|\mu_j| - \sum_{j\notin I}|\mu_j| > \sum_{j\in I}|\beta_{0i}| - \sum_{j\in I}|\varepsilon^*_i| - \sum_{j\notin I} c,$$

where we have used the fact that $|\mu_j| = |\beta_{0j} + \varepsilon^*_j| > |\beta_{0j}| - |\varepsilon^*_j|$ for $\varepsilon^* = (X_0'X_0)^{-1}X_0'\varepsilon$. Going back to (11.4), the term

$$E_{Y\sim N(X\beta_0,I)} \sum_{\kappa^*\in\Xi} \mathbb{I}(Y \in A_{\kappa^*})\mathbf{1}'_{\kappa^*}\mu = E_{Y\sim N(X\beta_0,I)} \sum_{m=0}^{s} \sum_{I\in I_m} \sum_{\kappa^*\in\Xi} \mathbb{I}(Y \in A_{\kappa^*} \cap A(I))\mathbf{1}'_{\kappa^*}\mu$$

can be lower-bounded by

$$\sum_{m=0}^{s} \sum_{I\in I_m} \sum_{\kappa^*\in\Xi} P[Y \in A_{\kappa^*}\cap A(I)] \left(\sum_{j\in I}|\beta_{0j}| - \sum_{i\notin I} c\right) - E\|\varepsilon^*\|_1 > \sum_{j=1}^{s}|\beta_{0i}|P(|\mu_j| > c) - s\,c - E\|\varepsilon^*\|_1. \tag{11.7}$$

Each $|\varepsilon_j^*|$ has a folded normal distribution with an expectation $\sqrt{2\sigma_j^2/\pi}$ where σ_j^2 is the j^{th} diagonal element of $(X_0'X_0)^{-1}$ and satisfies $\sigma_j^2 \leq \lambda_{max}[(X_0'X_0)^{-1}] = 1/\lambda_{min}(X_0'X_0)$. This yields

$$-E_{Y\sim N(X\beta_0,I)} \sum_{\kappa^*\in\Xi} \mathbb{I}(Y \in A_{\kappa^*}) \mathbf{1}_{\kappa^*}'\mu \leq s\, c + s\sqrt{2/(\pi\lambda_{min}(X_0'X_0))} - \|\beta_0\|_1 + \sum_{j=1}^{s} |\beta_{0i}| P(|\mu_j| \leq c).$$

Now, we inspect $P(|\mu_j| \leq c)$. Because $|\mu_j| > |\beta_{0j} + \varepsilon_j^*| > |\beta_{0j}| - |\varepsilon_j^*|$ we have from Assumption 6 and 8 that $|\beta_{0j}| > \beta_{min} > c/(1-b)$ and thereby $|\beta_{0j}| - c > b|\beta_{0j}|$. Then using a Gaussian tail bound we obtain

$$P(|\mu_j| \leq c) \leq P(|\varepsilon_j^*| \geq |\beta_{0j}| - c) \leq 2\mathrm{e}^{-b^2\beta_{0j}^2/(2\sigma_j^2)}.$$

Then because $|x|\exp(-x^2/2) \leq 1$ we have

$$|\beta_{0j}| P(|\mu_j| \leq c) \leq |\beta_{0j}| \mathrm{e}^{-b^2\beta_{0j}^2/(2\sigma_j^2)} \leq \sigma_j/b \leq \frac{1}{b\sqrt{\lambda_{min}(X_0'X_0)}}.$$

It is known [**?**] that $s - 1 < \chi_{s,1/2}^2 < s$ which, from our assumptions, yields for $M_0 = M/[\widetilde{\psi}(S_0)^2\phi(S_0)]$ and $s \geq 2$

$$\frac{1}{\sqrt{\lambda_{min}(X_0'X_0)}} < \frac{M_0(1-b)}{2\sqrt{\chi_{s,1/2}^2}} \sqrt{s/n\log p} < \frac{M_0(1-b)}{2}\sqrt{1/n\log p}.$$

A similar bound could also be obtained for $s = 1$ using the approximation $\chi_{s,1/2}^2 \approx s(1 - 2/(9s))^3$. This implies (using the definition of c in (11.5))

$$\begin{aligned} -E_{Y\sim N(X\beta_0,I)} \sum_{\kappa^*\in\Xi} \mathbb{I}(Y \in A_{\kappa^*}) \mathbf{1}_{\kappa^*}'\mu &\leq s\left[2\sqrt{s} + \sqrt{2/\pi} + 1/b\right] \frac{M_0(1-b)}{2}\sqrt{1/n\log p} - \|\beta_0\|_1 \\ &\leq C_0 s\sqrt{s/n\log p} - \|\beta_0\|_1. \end{aligned}$$

Together with (11.4) and (11.1), this concludes the Lemma.

12 Proof of Theorem 7

We first focus on the set $\mathcal{D}_n$ from Theorem 10. Choosing $\lambda = \sqrt{n}/p$ (as in Example 9 in [6]) the assumptions in Theorem 10 are satisfied for $a = 1$ and $A_4 > 2$ because $s\lambda\sqrt{\log p}/\sqrt{n} \leq$

$n/p\sqrt{\log p} = o(1)$ where $s = |S_0|$. The set $\mathcal{D}_n$ is defined as an intersection of the set (6.12) in [6] and the set from the proof of Theorem 5 in [6]. Define $\mathcal{B} = \{\beta \in \mathbb{R}^p : S_\beta = S_0\}$. Due to Pinsker's inequality we note that

$$\rho^{TV}_{n,m}(\beta_0, \hat{p})^2 \leq 4 \times P(\mathcal{D}_n) + 1/2 \times \mathbb{E}_{Y \mid \beta_0} \mathbb{I}(Y \in \mathcal{D}_n) KL(\pi(\widetilde{Y} \mid \beta_0), \hat{p}(\widetilde{Y} \mid Y)). \tag{12.1}$$

The second summand is the typical KL distance for a set of largely probable Y and can be written as

$$1/2 \times \mathbb{E}_{Y \mid \beta_0} \mathbb{E}_{\widetilde{Y} \mid \beta_0} \mathbb{I}(Y \in \mathcal{D}_n) \log \frac{\Lambda_{n,\beta_0,\pi}(Y, X)}{\Lambda_{n+m,\beta_0,\pi}(Z, \bar{X})}.$$

Under the Spike-and-Slab prior with a Laplace slab, we can rewrite this term as follows

$$\mathbb{E}_{Y \mid \beta_0} \mathbb{I}(Y \in \mathcal{D}_n) \mathbb{E}_{\widetilde{Y} \mid \beta_0} \log \frac{\int_{\mathcal{B}} \Delta_{n,\beta,\beta_0}(Y, X)\pi(\beta)\mathrm{d}\beta}{\int \Delta_{n+m,\beta,\beta_0}(Z, \bar{X})\pi(\beta)\mathrm{d}\beta} + \mathbb{E}_{Y \mid \beta_0} \mathbb{I}(Y \in \mathcal{D}_n) \log \frac{1}{P(\mathcal{B} \mid Y)}. \tag{12.2}$$

Next, we write (using Jensen's inequality $E \log X \leq \log EX$) using Theorem 10

$$\begin{aligned} \mathbb{E}_{Y \mid \beta_0} \mathbb{I}(Y \in \mathcal{D}_n) \log \frac{1}{P(\mathcal{B} \mid Y)} &= P(\mathcal{D}_n) + \mathbb{E}_{Y \mid \beta_0} \mathbb{I}(Y \in \mathcal{D}_n) \frac{P(\mathcal{B}^c \mid Y)}{1 - P(\mathcal{B}^c \mid Y)} \\ &< P(Y \in \mathcal{D}_n)(1 + 1/(p^{a_1} - 1)). \end{aligned}$$

The first term in (12.2) can be upper bounded by the KL risk (restricted to $\mathcal{D}_n$) obtained under the oracle prior $\pi_{S_0,\lambda}(\beta)$

$$\rho^{O}_{n,m}(\beta_0, \hat{p}) \equiv \mathbb{E}_{Y \mid \beta_0} \mathbb{I}(Y \in \mathcal{D}_n) \mathbb{E}_{\widetilde{Y} \mid \beta_0} \log \frac{\int \Delta_{n,\beta,\beta_0}(Y, X)\pi_{S_0,\lambda}(\beta)\mathrm{d}\beta}{\int \Delta_{n+m,\beta,\beta_0}(Z, \bar{X})\pi_{S_0,\lambda}(\beta)\mathrm{d}\beta}$$

The term $\rho^{O}_{n,m}(\beta_0, \hat{p})$ integrates over the high-probability Y's inside $\mathcal{D}_n$. For the log-numerator, we apply Lemma 19 which presents an upper bound for the (normalized) marginal likelihood under the Laplace prior. From the proof of Lemma 19 it can be seen that restricting the integration $\mathcal{D}_n$ *does not* increase the upper bound on the entire marginal likelihood. We can thus apply this upper bound when confined to $\mathcal{D}_n$. For the log-denominator, we apply the lower bound Lemma 18 which holds uniformly (and almost surely) for all Y. Altogether, this yields

$$\begin{aligned} \rho^{O}_{n,m}(\beta_0, \hat{p}) \leq & \frac{s}{2}\left[1 + 2\widetilde{C}\lambda^2 n \log^d p + C_0\sqrt{s/n \log p} + \log\left(\frac{8\pi}{n}\right)\right] \\ & + 1/2 + \lambda/(n+m) + s\log[(n+m)/\lambda] + \log s!. \end{aligned}$$

With $\lambda = \sqrt{n}/p$ and because $n/p \lesssim 1/\sqrt{\log(p)}$ we have $\lambda^2 n \log^d p \lesssim \log^{d-1} p$ we have

$$\rho^O_{n,m}(\beta_0, \hat{p}) \lesssim s[\log^{d-1\vee 1} p \vee \log(1 + m/n)]$$

Combined with (12.1) we conclude that $\rho^{TV}_{n,m}(\beta_0, \hat{p})^2 \lesssim \eta_n$. Moreover, throughout the proof we showed that uniformly on $\mathcal{D}_n$, the KL divergence can be bounded by a multiple of η_n. The probability of the KL divergence exceeding a multiple of η_n is thus smaller than the probability of the complement of set $\mathcal{D}_n$.

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

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

[16] 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.

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

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

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

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

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

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

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

[24] 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.

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

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

[27] Ročková, V. and E. George (2018). The Spike-and-Slab LASSO. *Journal of the American Statistical Association 113*, 431–444.

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

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

[30] 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.