

ON MINIMAX OPTIMALITY OF SPARSE BAYES PREDICTIVE DENSITY ESTIMATES

BY GOURAB MUKHERJEE¹ AND IAIN M. JOHNSTONE²

¹Department of Data Sciences and Operations, University of Southern California, gourab@usc.edu

²Departments of Statistics and of Biomedical Data Science, Stanford University, imj@stanford.edu

We study predictive density estimation under Kullback–Leibler loss in ℓ_0 -sparse Gaussian sequence models. We propose proper Bayes predictive density estimates and establish asymptotic minimaxity in sparse models. Fundamental for this is a new risk decomposition for sparse, or spike-and-slab priors.

A surprise is the existence of a phase transition in the future-to-past variance ratio r . For $r < r_0 = (\sqrt{5} - 1)/4$, the natural discrete prior ceases to be asymptotically optimal. Instead, for subcritical r , a ‘bi-grid’ prior with a central region of reduced grid spacing recovers asymptotic minimaxity. This phenomenon seems to have no analog in the otherwise parallel theory of point estimation of a multivariate normal mean under quadratic loss.

For spike-and-uniform slab priors to have any prospect of minimaxity, we show that the sparse parameter space needs also to be magnitude constrained. Within a substantial range of magnitudes, such spike-and-slab priors can attain asymptotic minimaxity.

1. Introduction and main results. Predictive density estimation is a fundamental problem in statistical prediction analysis [1, 9]. Here, it is studied in a high dimensional Gaussian setting under sparsity assumptions on the unknown location parameters. Fuller references and background for the problem are given after a formulation of our main results.

We consider a simple Gaussian model for high dimensional prediction:

$$(1) \quad X \sim N_n(\theta, v_x I), \quad Y \sim N_n(\theta, v_y I), \quad X \perp\!\!\!\perp Y | \theta.$$

Our goal is to predict the distribution of a future observation Y on the basis of the “past” observation vector X . In this model, the past and future observations are independent, but are linked by the common mean parameter θ which is assumed to be unknown. The variances v_x and v_y may differ and are assumed to be known.

The true probability densities of X and Y are denoted by $p(x|\theta, v_x)$ and $p(y|\theta, v_y)$, respectively. We seek estimators $\hat{p}(y|x)$ of the future observation density $p(y|\theta, v_y)$, and study their risk properties under sparsity assumptions on θ as dimension n increases to ∞ .

To evaluate the performance of such a predictive density estimator (prde), we use Kullback–Leibler “distance” as loss function:

$$L(\theta, \hat{p}(\cdot|x)) = \int p(y|\theta, v_y) \log \frac{p(y|\theta, v_y)}{\hat{p}(y|x)} dy.$$

The corresponding KL risk function follows by averaging over the distribution of the past observation:

$$\rho(\theta, \hat{p}) = \int L(\theta, \hat{p}(\cdot|x)) p(x|\theta, v_x) dx.$$

Received July 2017; revised April 2021.

MSC2020 subject classifications. Primary 62C12; secondary 62C25, 62F10, 62J07.

Key words and phrases. Predictive density, asymptotic minimaxity, proper Bayes rule, sparsity, high dimensional, least favorable prior, spike and slab.

Now, given a prior measure $\pi(d\boldsymbol{\theta})$, the average or integrated risk is

$$(2) \quad B(\pi, \hat{p}) = \int \rho(\boldsymbol{\theta}, \hat{p})\pi(d\boldsymbol{\theta}).$$

For any prior measure $\pi(d\boldsymbol{\theta})$, proper or improper, such that the posterior $\pi(d\boldsymbol{\theta}|\mathbf{x})$ is well defined, the Bayes predictive density is given by

$$(3) \quad \hat{p}_\pi(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\boldsymbol{\theta}, v_y)\pi(d\boldsymbol{\theta}|\mathbf{x}).$$

The Bayes predictive density in (3) minimizes both the posterior expected loss, denoted by $\int L(\boldsymbol{\theta}, \hat{p}(\cdot|\mathbf{x}))\pi(d\boldsymbol{\theta}|\mathbf{x})$, and the integrated risk $B(\pi, \hat{p})$, when the latter is finite, among all density estimates. The minimum is the Bayes KL risk:

$$(4) \quad B(\pi) := \inf_{\hat{p}} B(\pi, \hat{p}).$$

We study the predictive risk $\rho(\boldsymbol{\theta}, \hat{p})$ in a high dimensional setting under an ℓ_0 -sparsity condition on the parameter space. This “exact sparsity” condition has been widely used in statistical estimation problems, for example, [19], Chapter 8. With $\|\boldsymbol{\theta}\|_0 = \#\{i : \theta_i \neq 0\}$, consider the parameter set:

$$\Theta_n[s] = \{\boldsymbol{\theta} \in \mathbb{R}^n : \|\boldsymbol{\theta}\|_0 \leq s\}.$$

The minimax KL risk for estimation over Θ is given by

$$(5) \quad R_N(\Theta) = \inf_{\hat{p}} \sup_{\boldsymbol{\theta} \in \Theta} \rho(\boldsymbol{\theta}, \hat{p}),$$

the infimum being taken over *all* predictive density estimators $\hat{p}(\mathbf{y}|\mathbf{x})$. We often write prde for predictive density estimate. The notation $a_n \sim b_n$ denotes $a_n/b_n \rightarrow 1$ as $n \rightarrow \infty$ and $a_n = O(b_n)$ denotes $|a_n/b_n|$ is bounded for all large n .

1.1. Main results. Henceforth, we assume $v_x = 1$. As the problem is scale equivariant, results for general v_x will easily follow. A key parameter is the future-to-past variance ratio

$$(6) \quad r = v_y/v_x = v_y, \quad v = (1 + r^{-1})^{-1}.$$

Here, v is the “oracle variance,” which would be the variance of the UMVUE for $\boldsymbol{\theta}$, if both \mathbf{X} and \mathbf{Y} were observed. The variance ratio r determines not only the magnitude of the minimax risk but also the construction of minimax optimal prdes. In our asymptotic model, the dimension $n \rightarrow \infty$ and the sparsity $s = s_n$ may depend on n , but the variance ratio r remains fixed.

In the sparse limit $\eta_n = s_n/n \rightarrow 0$, for any fixed $r \in (0, \infty)$, Mukherjee and Johnstone [34] evaluated the minimax risk to be

$$(7) \quad R_N(\Theta_n[s_n]) \sim \frac{1}{1+r} s_n \log(n/s_n) = \frac{1}{1+r} n \eta_n \log \eta_n^{-1},$$

and a thresholding based prde was shown to attain the minimax risk.

By their nature, thresholding rules are not smooth functions of the data. This paper develops proper Bayes prdes—necessarily smooth functions—that are asymptotically minimax in sparse regimes. Our constructions begin with sparse *univariate* symmetric priors

$$(8) \quad \pi[\eta] = (1 - \eta)\delta_0 + \frac{1}{2}\eta(v^+ + v^-),$$

where δ_0 is unit mass at 0, and $\eta \in [0, 1]$ is the sparsity parameter, while v^+ is a probability measure on $(0, \infty)$ and v^- is its reflection on $(-\infty, 0)$.

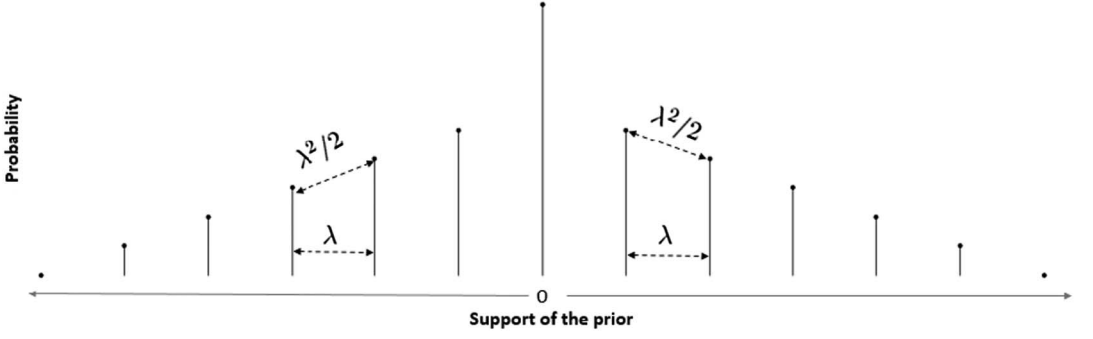


FIG. 1. Schematic for the grid prior. The uniform spacing λ between the support points is shown on the x-axis. The probabilities of the support points are shown on the y-axis using a logarithmic scale, hence the decay appears linear.

For such sparse priors, we introduce a new risk decomposition, Theorem 2.1, that takes the degenerate prior δ_0 as starting point, instead of the more commonly used uniform prior. This risk decomposition is fundamental for all proofs in the paper.

Priors on vector θ are built from i.i.d. draws

$$(9) \quad \pi_n(d\theta) = \prod_{i=1}^n \pi[\eta_n](d\theta_i),$$

where $\eta_n = s_n/n$ relates the multivariate sparsity s_n to the univariate parameter η_n . The Bayes prde based on prior π_n is the product density estimate:

$$(10) \quad \hat{p}_\pi(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^n \hat{p}_\pi(y_i|x_i).$$

The notation often drops the data suffixes and uses \hat{p}_π for both the univariate and the multivariate Bayes predictive density when the context is clear.

We begin with a discrete “grid prior” ν_G^+ in which the support points have equal spacing

$$(11) \quad \lambda = \lambda(\eta) = \sqrt{2 \log \eta^{-v}},$$

and geometric mass decay at rate $\eta^v = e^{-\lambda^2/2}$. More precisely,

$$\nu_G^+ = c_G \sum_{j=1}^{\infty} \eta^{(j-1)v} \delta_{\lambda j}, \quad c_G = 1 - \eta^v.$$

The corresponding sparse grid prior $\pi_G[\eta]$ built via (8) has a schematic illustration in Figure 1. Such “Mallows” discrete priors are a natural starting point for our predictive setting given their optimality properties in point estimation, recalled in the next subsection.

The choice π_G can also be motivated directly with three observations. The first, stated precisely in Section 3.5, is that among symmetric univariate three point priors with $\nu^+ = \delta_{a\lambda}$, $a > 0$, only the choice $a = 1$ is asymptotically least favorable. Second, the convex hull of $\text{supp}(\nu^+)$ must be unbounded, lest the risk function of \hat{p}_π grow without bound for large θ . Third, the probability decay rate $\eta^v = \exp(-\lambda^2/2)$ as a function of spacing λ is similar to the geometric decay used in [18] for minimax sparse point-estimation using discrete priors. Among discrete univariate priors, then, the grid prior π_G is perhaps the simplest choice compatible with these remarks.

Our first result gives a precise description of the first-order asymptotic maximum risk of the Bayes prde \hat{p}_G based on the multivariate product prior $\pi_{G,n}(d\theta) = \prod_{i=1}^n \pi_G[\eta_n](\theta_i)$, where

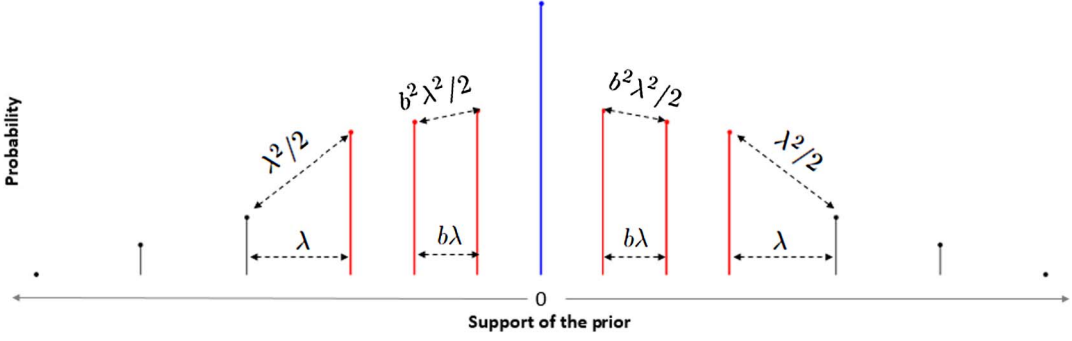


FIG. 2. Schematic for the bi-grid prior. The x -axis now shows the two spacings, and the y -axis (again on a logarithmic scale) the two different rates of log-linear decay of the prior probabilities.

$\eta_n = s_n/n$. Define

$$(12) \quad \begin{aligned} h_r &= (1+2r)(1+r)^{-2}(1-2r-4r^2)/4 \leq 1/4, \\ h_r^+ &= \max(h_r, 0). \end{aligned}$$

Let $r_0 = (\sqrt{5} - 1)/4$ be the positive root of the equation $4r^2 + 2r - 1 = 0$, and note that $h_r^+ > 0$ iff $r < r_0$.

THEOREM 1.1. As $\eta_n = s_n/n \rightarrow 0$, for any fixed $r \in (0, \infty)$ we have

$$\sup_{\Theta_n[s_n]} \rho(\theta, \hat{p}_G) = R_N(\Theta_n[s_n])(1 + h_r^+ + o(1)) \quad \text{as } n \rightarrow \infty.$$

Thus for all $r \geq r_0$, \hat{p}_G is exactly minimax optimal, while for all $r < r_0$, it is minimax suboptimal but still attains the minimax rate, and has maximum risk at most 1.25 times the minimax value, whatever be the value of r .

As the future-to-past variance ratio r decreases, the difficulty of the predictive density estimation problem increases, as we have to estimate the future observation density based on increasingly noisy past observations. Theorem 1.1 shows that rules which are minimax optimal for higher values of r can be suboptimal for lower values of r . This phenomenon was seen with threshold density estimates in [34], Section S.2, Lemma S.2.1, as well as in the recent work of [29] on nonsparse prediction.

To obtain asymptotic minimaxity for all r , we need to modify the prior. The Bi-grid π_B prior is obtained from π_G by selecting an “inner zone” on which the spacing of the prior atoms is reduced from λ to $b\lambda$, where

$$(13) \quad b = \min\{4r(1+r)(1+2r)^{-1}, 1\}.$$

Note that $b < 1$ iff $r < r_0$. The decay ratio in the inner zone is increased from $\eta^v = e^{-\lambda^2/2}$ to $\eta^{vb^2} = e^{-b^2\lambda^2/2}$. See Figure 2 for a schematic depiction. Section 3.3 explains why the reduced spacing in the inner zone is needed. In brief, the narrower grid “pulls down” the maximum risk of the Bayes rule for π_B to the asymptotically minimax level.

More precisely, $\pi_B[\eta]$ is a univariate sparse symmetric prior of form (8) with

$$v_B^+ = c_B \left[\sum_{k=1}^K \eta^{(k-1)vb^2} \delta_{v_k} + \eta^{(K-1)vb^2} \sum_{j=1}^{\infty} \eta^{jv} \delta_{\mu_j} \right].$$

The normalization $c_B = c_B(\eta)$ is at (36). The support points fall in two zones:

- (i) Inner zone: $v_k = \lambda + (k - 1)b\lambda$ for $k = 1, \dots, K$,
- (ii) Outer zone: $\mu_j = v_K + j\lambda$ for $j = 1, 2, \dots$

The cardinality of the inner zone is

$$(14) \quad K = 1 + \lceil 2b^{-3/2} \rceil.$$

In fact, any integer $K \in [1 + \lceil 2b^{-3/2} \rceil, \infty]$ works; see Section 3.6. For definiteness and minimal departure from π_G , we use (14).

How do the mass distributions of π_B and π_G compare? A crude continuous approximation (see the Supplementary Material) says that the “density ratio” $d\pi_B/d\pi_G(x)$ increases exponentially in x in the inner zone. In the outer zone, it is a constant greater than one, that is, π_B has more mass in the tails.

A main result of the paper is that the Bayes predictive density estimate \hat{p}_B based on the product prior $\pi_{B,n}(d\theta) = \prod_{i=1}^n \pi_B[\eta_n](d\theta_i)$ is asymptotically minimax optimal.

THEOREM 1.2. *For each fixed $r \in (0, \infty)$, as $\eta_n = s_n/n \rightarrow 0$, we have*

$$\sup_{\Theta_n[s_n]} \rho(\theta, \hat{p}_B) = R_N(\Theta_n[s_n])(1 + o(1)) \quad \text{as } n \rightarrow \infty.$$

The following theorem shows that the bi-grid prior $\pi_{B,n}$ is also asymptotically least favorable.

THEOREM 1.3. *If $s_n \rightarrow \infty$ and $s_n/n \rightarrow 0$, then*

$$B(\pi_{B,n}) = R_N(\Theta_n[s_n]) \cdot (1 + o(1)).$$

Unlike Theorem 1.2, we need the assumption that $s_n \rightarrow \infty$. It ensures that $\pi_{B,n}$ actually concentrates on $\Theta_n[s_n]$, namely that $\pi_{B,n}(\Theta_n[s_n]) \rightarrow 1$ as $n \rightarrow \infty$. For the case where s_n does not diverge to ∞ an asymptotically least favorable prior can be constructed from a sparse prior built from “independent blocks.” The construction is discussed in Section 3.4.

1.2. Discussion. A fully Bayesian approach is a natural route to prdes with good properties [2, 15], with advantages over “plug-in” or thresholding based density estimates. Indeed, a coordinatewise threshold rule $\hat{p}_T(y|x) = \prod_{i=1}^n \hat{p}_T(y_i|x_i)$ is typically built from univariate prdes which combine two Bayes prdes, for example, based on uniform \hat{p}_U and cluster priors \hat{p}_{CL} , as in [34], equation (14):

$$\hat{p}_T(y_i|x_i) = \begin{cases} \hat{p}_U(y_i|x_i) & \text{if } |x_i| > v^{-1/2}\lambda, \\ \hat{p}_{CL}(y_i|x_i) & \text{if } |x_i| \leq v^{-1/2}\lambda. \end{cases}$$

This is manifestly discontinuous as a function of the data \mathbf{x} .

The bi-grid Bayes rule achieves the same purposes as the hybrid \hat{p}_T . Indeed, the close spacing $b\lambda$ in the inner section of π_B yields the same risk control as the (unevenly spaced) cluster prior for small and moderate θ , while the uniform λ spacing in the outer section of π_B controls risk for large θ in the same way as the uniform prior.

Decision theoretic parallels between predictive density estimation and the point estimation of a Gaussian mean under quadratic loss have been established by [4, 10, 11, 13, 14, 21, 23, 43] for unconstrained θ , and by [42], [7], [27] and [34] for various constraint sets Θ .

The phase transition seen in Theorems 1.1 and 1.2 seems however to have no parallel in point estimation. Indeed, it follows from [18] that a first-order minimax rule for quadratic loss in the sparse setting is derived from the Mallows prior [28], with $v_Q^+ =$

$(1 - \eta) \sum_{j=1}^{\infty} \eta^{j-1} \delta_{\lambda_e j}$. Here, $\lambda_e = \sqrt{2 \log \eta^{-1}} = v^{-1/2} \lambda$ so that the predictive setting involves a reduced spacing in the prior. More significantly, there is no analog in point estimation of the inner section with its further reduced spacing for $r < r_0$.

Our main technical contribution lies in sharp methods for bounding the global KL risk for general bi-grid priors; see Lemmas 3.1 and 3.2, and for spike-and-slab priors, Section 4. The sharp predictive risk bounds established here provide new asymptotic perspectives in the information geometric framework of [22, 24, 44] and augment new sparse prediction techniques for general multivariate predictive density estimation theory developed in [4, 10, 23, 25, 26, 30, 32].

1.3. Minimax risk of spike and slab priors. Some of the most popular Bayesian variable selection techniques are built on the “spike and slab” priors [12, 16, 31]. Such priors and their computationally tractable extensions have found success in variable selection in high-dimensional sparse regression models, for example, [3, 17, 37–40] and the references therein. While this is a well-established methodological research area [36], optimality of their respective predictive density estimates has so far not been studied.

Here, we consider simple “spike and slab” prior distributions in the flavor of the foundational paper [31]. Begin with a sparse univariate prior, a special case of (8),

$$(15) \quad \pi_S[\eta, \ell] = (1 - \eta) \delta_0 + \eta / (2\ell) I\{\mu \in [-\ell, \ell]\} d\mu.$$

In parallel with (9), build a multivariate product prior $\pi_{S,n}$ from n i.i.d. copies of $\pi_S[\eta_n, \ell]$, where as before $\eta_n = s_n/n$. We might consider multivariate Bayes predictive densities $\hat{p}_S[\ell]$ based on $\pi_{S,n}$.

It is intuitively clear that such Bayes prdes are necessarily asymptotically suboptimal: for any fixed $\ell \in [0, \infty)$, for all $s_n > 0$, we have

$$(16) \quad \left\{ \sup_{\Theta_n[s_n]} \rho(\boldsymbol{\theta}, \hat{p}_S[\ell]) \right\} / R_N(\Theta_n[s_n]) = \infty$$

for each fixed n . Indeed, the support of π_S is restricted to $[-\ell, \ell]$, and the corresponding prde has large risk away from the support. A formal proof follows Lemma 4.1.

Consider therefore bounded subsets of the sparse parameter sets $\Theta_n[s_n]$:

$$\Theta_n[s, t] = \{\boldsymbol{\theta} \in \mathbb{R}^n : \|\boldsymbol{\theta}\|_0 \leq s \text{ and } |\theta_i| \leq t \text{ for all } i = 1, \dots, n\}.$$

We allow $t = t_n$ to increase with n , and note next that the increase must be at least as fast as $\lambda_n = \lambda(\eta_n)$ (cf. (11)) to have minimax risk equivalent to $\Theta_n[s_n]$.

LEMMA 1.4. *For all t_n , there is a simple bound*

$$R_N(\Theta_n[s_n, t_n]) \leq s_n t_n^2 / (2r).$$

If $t_n > \lambda_n = \sqrt{2 \log \eta_n^{-1}}$, then

$$(17) \quad R_N(\Theta_n[s_n, t_n]) \sim s_n \lambda_n^2 / (2r) \sim R_N(\Theta_n[s_n]).$$

The following result exhibits a substantial range of magnitude constraints t_n for which $\hat{p}_S[t_n]$ is asymptotically minimax over $\Theta_n[s_n, t_n]$. All proofs for this subsection, along with a figure and high-level overview of the strategy, appear in Section 4.

THEOREM 1.5. *As $\eta_n = s_n/n \rightarrow 0$, suppose that $t_n / (\log \eta_n^{-1})^{1/2} \rightarrow \infty$ but $\log t_n / (\log \eta_n^{-1}) \rightarrow 0$. Then as $n \rightarrow \infty$,*

$$\sup_{\Theta_n[s_n, t_n]} \rho(\boldsymbol{\theta}, \hat{p}_S[t_n]) \sim R_N(\Theta_n[s_n, t_n]).$$

Note that if $t_n \rightarrow \infty$ at a rate slower than $(\log \eta_n^{-1})^{1/2}$ then, by Lemma 1.4, $R_N(\Theta_n[s_n, t_n])$ is no longer equivalent to $R_N(\Theta_n[s_n])$ as $n \rightarrow \infty$. At the other extreme, we show next that if t_n grows at rate $\eta_n^{-\beta}$ or higher for any $\beta > 0$, then no spike and uniform slab procedure can be minimax optimal.

THEOREM 1.6. *If $\eta_n = s_n/n \rightarrow 0$ and $\log t_n = \beta \log \eta_n^{-1}$ for some $\beta > 0$, then*

$$\min_{\ell > 1} \sup_{\Theta[s_n, t_n]} \rho(\theta, \hat{p}_S[\ell]) \geq (1 + \beta) R_N(\Theta_n[s_n, t_n]) (1 + o(1)) \quad \text{as } n \rightarrow \infty.$$

We emphasize that Theorem 1.6 shows that, even for true parameters within the support of the uniform slab, risk can exceed the minimax bound. Informally, the proof shows that if the slab is small, $\log \ell < \beta \lambda_n^2$, then the risk at $\theta = t_n$ is unacceptably large, while if it is large, $\log \ell \geq \beta \lambda_n^2$, there is poor risk at $\theta = \sqrt{1 + \beta} \lambda_n$.

1.4. Organization of the paper. Section 2 presents the fundamental risk decomposition, its proof and some discussion. Section 3 presents the risk properties of the grid and bi-grid prior based prdes and proofs of the main results. Section 4 proves the spike-and-slab results. Section 5 compares the performance of the prdes through simulation experiments. The Appendix and Sections 1, 2 of the Supplementary Material contain the proofs of the lemmas.

Notations. The standard normal density and cumulative distribution are denoted by ϕ and Φ ; $\tilde{\Phi} = 1 - \Phi$. For sequences $a_n \sim b_n$ means $\lim_{n \rightarrow \infty} a_n/b_n = 1$.

2. A risk decomposition for spike and slab priors. *Univariate problem.* We focus on priors with i.i.d. components (9), so that the predictive density then has product form (10). The predictive risk is then additive

$$(18) \quad \rho(\theta, \hat{p}_\pi) = \sum_{i=1}^n \rho(\theta_i, \hat{p}_\pi).$$

[We use \hat{p}_π for both univariate and multivariate Bayes predictive densities: the context will make clear which is used.]

For our sparse parameter sets $\Theta_n[s]$ and $\Theta_n[s, t]$, there is an easy reduction of the maximum multivariate risk of a product rule to a univariate risk maximum. Indeed, (18) yields

$$(19) \quad s_n \sup_{|\theta| \leq t_n} \rho(\theta, \hat{p}) \leq \sup_{\Theta_n[s_n, t_n]} \rho(\theta, \hat{p}) \leq n(1 - \eta_n) \rho(0, \hat{p}) + s_n \sup_{\theta \in \mathbb{R}} \rho(\theta, \hat{p}).$$

Sparse priors. Now suppose that $X|\theta \sim N(\theta, 1)$ and $Y|\theta \sim N(\theta, r)$ and that the past and future observations X, Y are independent given θ . Consider a sparse proper prior of the form

$$(20) \quad \pi(d\mu) = (1 - \eta)\delta_0 + \eta\nu(d\mu),$$

for probability measure ν on \mathbb{R} and $\eta \in [0, 1]$. The associated (univariate) Bayes predictive density estimate \hat{p}_π is given by (3).

The following risk decomposition is fundamental. It will be applied to study discrete priors in Section 3 and uniform slab priors in Section 4.

THEOREM 2.1. *With the preceding definitions, let $Z \sim \mathcal{N}(0, 1)$ and $v = (1 + r^{-1})^{-1}$. For a sparse prior (20),*

$$(21) \quad \begin{aligned} \rho(\theta, \hat{p}_\pi) &= \rho(\theta, \hat{p}_{\delta_0}) - \mathbb{E} \log N_{\theta, v}(Z) + \mathbb{E} \log D_\theta(Z) \\ &= \theta^2/(2r) - \mathbb{E} \log N_{\theta, v}(Z) + \mathbb{E} \log D_\theta(Z), \end{aligned}$$

where $D_\theta(Z) = N_{\theta,1}(Z)$ and

$$(22) \quad N_{\theta,v}(Z) = 1 + \frac{\eta}{1-\eta} \int \exp\left\{\frac{\mu Z}{\sqrt{v}} + \frac{\mu\theta}{v} - \frac{\mu^2}{2v}\right\} v(d\mu).$$

Decomposition (21) takes the degenerate prior δ_0 as starting point for comparison of the risk $\rho(\theta, \hat{p}_\pi)$ of a Bayes prde. This is natural for sparse priors (20) and might be contrasted with the representation George, Liang and Xu [10], Lemma 2, which takes the uniform prior prde as point of departure.

PROOF. The decomposition (21) compares $\rho(\theta, \hat{p}_\pi)$ to $\rho(\theta, \hat{p}_{\delta_0}) = \theta^2/(2r)$, the KL risk of $\hat{p}_{\delta_0}(y|x) = \phi(y|0, r)$ corresponding to $\pi = \delta_0$ and $\eta = 0$. Accordingly, using (3), write the Bayes predictive density as

$$(23) \quad \hat{p}_\pi(y|x) = \frac{\int \phi(y|\mu, r) \phi(x - \mu) \pi(d\mu)}{\int \phi(x - \mu) \pi(d\mu)} = \phi(y|0, r) \frac{N(x, y)}{D(x)},$$

after rewriting numerator and denominator in the first ratio respectively as

$$\pi_0 \phi(y|0, r) \phi(x) N(x, y), \quad \text{and} \quad \pi_0 \phi(x) D(x),$$

where $\pi_0 = \pi(\{0\}) = 1 - \eta$. After simple algebra, we find

$$(24) \quad N(x, y) = \int \exp\left\{\mu\left(x + \frac{y}{r}\right) - \frac{\mu^2}{2}\left(1 + \frac{1}{r}\right)\right\} \frac{\pi(d\mu)}{\pi_0}$$

and $D(x)$ is analogous, but without terms in y and r . Note also that

$$\rho(\theta, \hat{p}_{\delta_0}) = \mathbb{E}_\theta \log\left(\frac{\phi(Y|\theta, r)}{\phi(Y|0, r)}\right) = \mathbb{E}_\theta \left[\frac{\theta Y}{r} - \frac{\theta^2}{2r}\right] = \frac{\theta^2}{2r}.$$

Hence, from (23) and the definition of predictive loss

$$L(\theta, \hat{p}_\pi(\cdot|x)) = \mathbb{E}_\theta \log\left(\frac{\phi(Y|\theta, r)}{\hat{p}_\pi(Y|x)}\right) = \frac{\theta^2}{2r} - \mathbb{E}_\theta \log N(x, Y) + \log D(x).$$

To obtain $\rho(\theta, \hat{p}_\pi)$, take expectation also over $X \sim N(\theta, 1)$. Since $Y \sim N(\theta, r)$ independently of X , the random variable $X + Y/r \sim \mathcal{N}(\theta/v, 1/v)$ may be expressed in the form $\theta/v + Z/\sqrt{v}$. Recalling the sparse prior form $\pi(d\mu) = (1 - \eta)\delta_0 + \eta v$, we get

$$N(X, Y) \stackrel{\mathcal{D}}{=} 1 + \frac{\eta}{1-\eta} \int \exp\left\{\frac{\mu Z}{\sqrt{v}} + \frac{\mu\theta}{v} - \frac{\mu^2}{2v}\right\} v(d\mu) = N_{\theta,v}(Z).$$

Similarly, $D(X) \stackrel{\mathcal{D}}{=} D_\theta(Z)$ and the lemma follows from the previous two displays. \square

Clearly, $N_{\theta,v}(Z), D_\theta(Z) \geq 1$, and so we have the simple but useful “basic lower” and “basic upper” risk bounds

$$(25) \quad \frac{\theta^2}{2r} - \mathbb{E} \log N_{\theta,v}(Z) \leq \rho(\theta, \hat{p}_\pi) \leq \frac{\theta^2}{2r} + \mathbb{E} \log D_\theta(Z).$$

From Jensen’s inequality,

$$(26) \quad \mathbb{E} \log N_{\theta,v}(Z) \leq \log(\mathbb{E} N_{\theta,v}(Z)),$$

and since $\mathbb{E} \exp(\zeta Z) = \exp(\zeta^2/2)$, by Fubini’s theorem

$$(27) \quad \mathbb{E} N_{\theta,v}(Z) = 1 + \frac{\eta}{1-\eta} \int \exp\left(\frac{\mu\theta}{v}\right) v(d\mu),$$

and, in particular,

$$\mathbb{E}D_0(Z) = \mathbb{E}N_{0,1}(Z) = (1 - \eta)^{-1}.$$

Consequently, from the right-hand side of (25), then (26) (for $v = 1$) and the previous display,

$$(28) \quad \rho(0, \hat{p}_\pi) \leq \log(1 - \eta)^{-1} = \eta(1 + o(1)) \quad \text{as } \eta \rightarrow 0.$$

3. Risk properties for discrete priors. The bulk of this section is devoted to the proof of Theorems 1.1 and 1.2. We first outline the approach. First, return to the univariate reduction (19). From (28), it is clear that $n\rho(0, \hat{p}_1) \leq n\eta_n(1 + o(1)) = s_n(1 + o(1))$. So for the minimaxity results of Theorems 1.1 (for $r > r_0$), 1.2 and 1.5, it suffices to show the univariate bound

$$(29) \quad \sup_{\theta \in \mathbb{R}} \rho(\theta, \hat{p}_1) \leq \lambda_n^2/(2r) + o(\lambda_n^2),$$

for then, with $\Theta_n = \Theta_n[s_n]$ or $\Theta_n[s_n, t_n]$,

$$\sup_{\Theta_n} \rho(\theta, \hat{p}_{\pi_n}) \leq s_n[\lambda_n^2/(2r) + o(\lambda_n^2)].$$

To establish (29), we use the key risk decomposition of Proposition 2.1. For this, we introduce a class of discrete sparse priors that includes both grid and bi-grid priors. We develop lower and upper bounds, respectively, for $\mathbb{E} \log N_{\theta,v}(Z)$ and $\mathbb{E} \log D_\theta(Z)$ in (21). These bounds are combined to yield an upper estimate

$$\rho(\theta, \hat{p}_D) \leq (2r)^{-1} \lambda^2 \sigma(\theta) + O(\lambda),$$

for some function σ . In Section 3.3, we first provide heuristics—Figure 3—and then a formal proof of conditions under which $\sigma(\theta) \leq 1$ for all θ , establishing Theorems 1.1 and 1.2.

3.1. *A class of discrete sparse priors.* For $0 < b \leq 1$ and $r > 0$, let

$$(30) \quad \pi_D[\eta; b, r] = \sum_{j \in \mathbb{Z}} \pi_j \delta_{\mu_j}$$

where $\mu_{-j} = -\mu_j$, $\pi_{-j} = \pi_j$. The support points satisfy $\mu_0 = 0$ and $\mu_j = \lambda \alpha_j$ for $j > 0$, where the piecewise linear spacing function

$$(31) \quad \alpha_j = \begin{cases} 1 + b(j-1) & 1 \leq j \leq K, \\ \alpha_K + j - K & j > K \end{cases}$$

has increments $\dot{\alpha}_j = \alpha_{j+1} - \alpha_j = b$ or 1 according as $j \leq K$ or $j > K$. Set $\zeta = \eta^v$. The prior masses are given by

$$(32) \quad \pi_0 = 1 - \eta, \quad \pi_j = c(\eta) \eta \zeta^{\beta_j - 1},$$

for $j \geq 1$. The decay function in the prior probabilities

$$(33) \quad \beta_j = \begin{cases} 1 + b^2(j-1) & 1 \leq j \leq K, \\ \beta_K + j - K & j > K \end{cases}$$

has the same form as α_j with b replaced by b^2 . This choice is crucial for Lemma 3.1 below and its consequent risk bounds. In particular, note that $\beta_j \leq \alpha_j$ and that the increments $\dot{\beta}_j = \beta_{j+1} - \beta_j$ satisfy

$$(34) \quad \dot{\beta}_j = \dot{\alpha}_j^2 \quad \text{all } j \geq 1.$$

In addition, $l \rightarrow g_l = \alpha_l^2 - \beta_l$ is increasing for $l \geq 1$, as

$$(35) \quad g_{l+1} - g_l = \dot{\alpha}_l(\alpha_{l+1} + \alpha_l) - \dot{\beta}_l = 2\dot{\alpha}_l\alpha_l > 0.$$

The normalizing constant $c(\eta) = c_B(\eta)/2$, where

$$(36) \quad \frac{1}{c_B(\eta)} = \sum_{j=1}^{\infty} \zeta^{\beta_j-1} = \frac{1 - \eta^{b^2 v K}}{1 - \eta^{b^2 v}} + \frac{\eta^{b^2 v(K-1)+v}}{1 - \eta^v}.$$

3.2. Risk component bounds for discrete priors. Since π_D is a sparse prior, we may apply the decomposition of predictive risk given in Proposition 2.1. Inserting the discrete measure (30), we obtain

$$(37) \quad N_{\theta,v}(Z) = 1 + \sum_{j \neq 0} N_j,$$

$$(38) \quad N_j = \pi_0^{-1} \pi_j \exp \left\{ v^{-1/2} \mu_j Z + v^{-1} \left(\mu_j \theta - \frac{1}{2} \mu_j^2 \right) \right\}.$$

In the special case $v = 1$, it will be helpful to write $D_\theta(Z) = N_{\theta,1}(Z)$ as

$$(39) \quad D_\theta(Z) = 1 + \sum_{j \neq 0} D_j,$$

$$(40) \quad D_j = \pi_0^{-1} \pi_j \exp \left\{ \mu_j Z + \mu_j \theta - \frac{1}{2} \mu_j^2 \right\}.$$

The probability ratio π_j/π_0 can also be written in exponential form. To this end, introduce $c_1(\eta) = c(\eta)(1 - \eta)^{-1}$. Recall that $v^{-1} = 1 + r^{-1}$ and $\zeta = \eta^v = \exp(-\lambda^2/2)$ and then rewrite $\eta = \zeta^{v^{-1}} = \exp\{-\frac{1}{2}\lambda^2(1 + r^{-1})\}$. Using (32), we arrive at

$$(41) \quad \pi_0^{-1} \pi_j = c_1(\eta) \exp \left\{ -\frac{1}{2} \lambda^2 (\beta_j + r^{-1}) \right\}.$$

We can therefore, for example, rewrite

$$(42) \quad \begin{aligned} D_j &= c_1(\eta) \exp \{ \mu_j Z - G(\mu_j; \theta) \}, \\ G(\mu_j; \theta) &= \frac{1}{2} \mu_j^2 - \mu_j \theta + \frac{1}{2} \lambda^2 (\beta_j + r^{-1}). \end{aligned}$$

To obtain an upper bound for $\rho(\theta, \hat{p}_D)$ we use (22). It turns out to be enough to focus on (the logs of) two consecutive terms N_j, N_{j+1} in (37); ignoring all other terms trivially yields a lower bound for $N_{\theta,v}$. For the upper bound for D_θ , a single (suitably chosen) term D_j in (39) suffices, but more care is needed to show that the neglected terms are negligible.

Bring in a coordinate system (l, ω) for θ : each $\theta \geq 0$ can be uniquely written in the form

$$\theta = \lambda(\alpha_l + \omega), \quad l \in \mathbb{N}, \omega \in [0, \dot{\alpha}_l).$$

We can therefore write $l = l(\theta)$ and $\omega = \omega(\theta)$.

We argue heuristically that $l(\theta)$ is an appropriate choice of index for our bounds. Indeed, from (38) and (41),

$$(43) \quad \mathbb{E} \log N_j = c - \frac{1}{2} \{ (\mu_j - \theta)^2 / v - \lambda^2 \beta_j \}$$

after collecting terms not involving j into c . Hence, for $\theta \in [\mu_l, \mu_{l+1})$, the choice $j = l$ or $l + 1$ will minimize or nearly minimize the quadratic, and these suffice for the lower bound. For D_θ , we have from (42) that $\mathbb{E} \log D_j = \log c_1(\eta) - G(\mu_j; \theta)$. We show in the

Appendix (in the proof of Lemma 3.1) that $j \rightarrow G(\mu_j; \theta)$ is indeed minimized at $j = l$ for each $\theta \in [\mu_l, \mu_{l+1})$.

Focus therefore on the terms $N_{l(\theta)}$ and $D_{l(\theta)}$. When $\theta = \lambda(\alpha_l + \omega)$,

$$\mu_j \theta - \frac{1}{2} \mu_j^2 = \frac{1}{2} \lambda^2 (2\alpha_j(\alpha_l + \omega) - \alpha_j^2).$$

Combining this with (41), for $j = l, l + 1$, we can write

$$\begin{aligned} N_l &= c_1(\eta) \exp \left\{ \frac{1}{2} \lambda^2 n(l, \omega) + \alpha_l \lambda Z / \sqrt{v} \right\}, \\ N_{l+1} &= c_1(\eta) \exp \left\{ \frac{1}{2} \lambda^2 \check{n}(l, \omega) + \alpha_{l+1} \lambda Z / \sqrt{v} \right\}, \\ D_l &= c_1(\eta) \exp \left\{ \frac{1}{2} \lambda^2 d(l, \omega) + \alpha_l \lambda Z \right\} \end{aligned} \quad (44)$$

in terms of three linear functions of ω :

$$\begin{aligned} n(l, \omega) &= v^{-1}(\alpha_l^2 + 2\alpha_l \omega) - \beta_l - r^{-1}, \\ d(l, \omega) &= \alpha_l^2 + 2\alpha_l \omega - \beta_l - r^{-1} \end{aligned} \quad (45)$$

and, corresponding to N_{l+1} ,

$$\check{n}(l, \omega) = n(l, \omega) + 2v^{-1}\dot{\alpha}_l \omega - (1 + v^{-1})\dot{\alpha}_l^2. \quad (46)$$

We now state our key uniform bounds on the risk components of (21).

LEMMA 3.1. *For any fixed $r \in (0, \infty)$ and $b \in (0, 1]$, with λ defined in (11), uniformly in $\theta = \lambda(\alpha_l + \omega) \geq \lambda$, we have the following bounds:*

$$\begin{aligned} \mathbb{E} \log N_{\theta, v}(Z) &\geq \frac{1}{2} \lambda^2 (n \vee \check{n})(l, \omega) + O(1), \\ \mathbb{E} \log D_{\theta}(Z) &\leq \frac{1}{2} \lambda^2 d^+(l, \omega) + O(\lambda). \end{aligned}$$

For $0 \leq \theta < \lambda$, we just have $\mathbb{E} \log N_{\theta, v}(Z) \geq 0$, and $\mathbb{E} \log D_{\theta}(Z) \leq O(\lambda)$.

The proof is given in the **Appendix**. The appearance of the positive part of $d(l, \omega)$ in the upper bound may be understood this way: if $d(l, \omega) < 0$, we cannot expect the term D_l to dominate $D_0 = 1$ in (39).

In the reverse direction, we need only a bound for θ lying in a subset of $[\mu_1, \mu_2)$ in our proofs of Theorems 1.1 and 1.2.

LEMMA 3.2. *For any fixed $r \in (0, \infty)$, $b \in (0, 1]$, with λ defined in (11) and setting $\omega_1 = b(1 + v)/2$, uniformly in $\theta \in \lambda[\alpha_1, \alpha_1 + \omega_1]$, we have*

$$\mathbb{E} \log N_{\theta, v}(Z) \leq \frac{1}{2} \lambda^2 n(1, \omega) + O(\lambda).$$

3.3. Proof of Theorems 1.1 and 1.2. Inserting the bounds of Lemma 3.1 in risk decomposition (21), we get

$$\begin{aligned} \rho(\theta, \hat{p}_D) &\leq (2r)^{-1} \lambda^2 \sigma(l, \omega) + O(\lambda), \\ \sigma(l, \omega) &= \begin{cases} \omega^2 & \text{if } l = 0, \\ (\alpha_l + \omega)^2 - r(n \vee \check{n})(l, \omega) + r d^+(l, \omega) & \text{if } l \geq 1. \end{cases} \end{aligned} \quad (47)$$

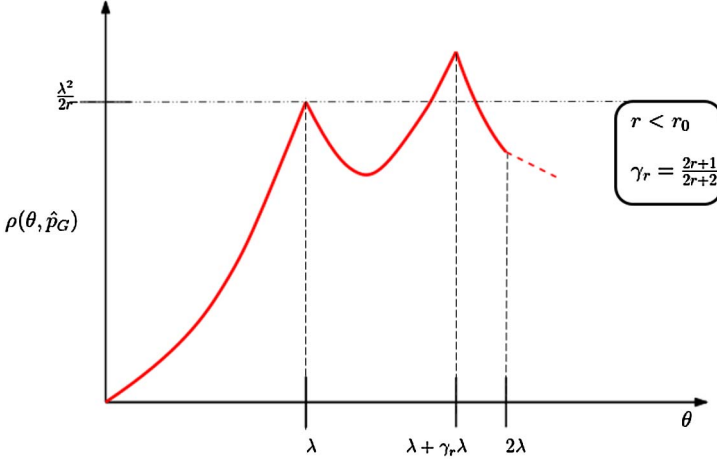


FIG. 3. Schematic for the risk bound (47) for $\theta \rightarrow \rho(\theta, \hat{p}_G)$ for the grid prior; being asymptotically minimax when the second peak is no higher than the first.

Our task is to bound $\sigma(l, \omega)$; more specifically for Theorem 1.1 to show that $\sigma(l, \omega) \leq 1 + h_r^+$ for the grid prior and for Theorem 1.2 to ensure that $\sigma(l, \omega) \leq 1$ for the bi-grid prior with b in (13). Figure 3 shows the idea of the main part of the proof. We argue below that the maximum of $\sigma(\theta) = \sigma(l, \omega)$ falls in the interval $[\lambda, \lambda\alpha_2]$, which in the case of the uniform grid prior is just $[\lambda, 2\lambda]$. The function $\sigma(\theta)$ is argued to be piecewise quadratic with

$$\max_{\lambda \leq \theta \leq 2\lambda} \sigma(\theta) = \max\{1, 1 + \gamma_r(\gamma_r - 2r)\}.$$

The second maximum is attained at $\theta_* = \lambda + \gamma_r \lambda$, with $\gamma_r = (2r + 1)/(2r + 2)$. It will then follow that the grid prior estimate \hat{p}_G is asymptotically minimax if and only if $\gamma_r \leq 2r$, which translates to $r \geq r_0 = (\sqrt{5} - 1)/4$.

For $r < r_0$, the maximum risk can be “pulled down” by reducing the spacing between λ and the next support point $\lambda + b\lambda$ (we set $\dot{\alpha}_l = b$). For the bi-grid prior, the second maximum then satisfies

$$\theta_* = \lambda + \gamma_r b\lambda, \quad \sigma(\theta_*) = 1 + \gamma_r b(\gamma_r b - 2r) \leq 1$$

exactly when b is no larger than the value (13).

To begin the proof, observe first that by symmetry we can reduce to $\theta \geq 0$. For $l = 0$, control on the risk is immediate from (25), and so, from now on consider $l \geq 1$. We make some observations on $\sigma(l, \omega)$. When $d(l, \omega) \geq 0$, from (45), $r(n - d) = \alpha_l^2 + 2\alpha_l \omega$, and so

$$(48) \quad \sigma(l, \omega) \leq (\alpha_l + \omega)^2 - r(n - d) = \omega^2 \leq 1.$$

Back in the general case, from (46), both $n(l, \omega)$ and $\check{n}(l, \omega)$ are linear for $\omega \in [0, \dot{\alpha}_l]$, intersecting at $\omega_l = \dot{\alpha}_l(1 + v)/2 < \dot{\alpha}_l$. Now $n(l, 0) > \check{n}(l, 0)$ while \check{n} has a larger positive slope. Hence $n \vee \check{n}$ equals n on $[0, \omega_l]$ and \check{n} on $[\omega_l, \dot{\alpha}_l]$. A calculation shows that

$$\begin{aligned} \check{n}(l, \dot{\alpha}_l) &= v^{-1}(\alpha_l^2 + 2\alpha_l \dot{\alpha}_l + \dot{\alpha}_l^2) - \dot{\alpha}_l^2 - \beta_l - r^{-1} \\ &= v^{-1}\alpha_{l+1}^2 - \beta_{l+1} - r^{-1} = n(l + 1, 0). \end{aligned}$$

Similarly, $d(l, \dot{\alpha}_l) = d(l + 1, 0)$ and so $\theta \rightarrow d(\theta)$ is piecewise linear, continuous and strictly increasing from $d(1, 0) = -r^{-1} < 0$ to $+\infty$ as $\theta \rightarrow \infty$. Consequently, there is a unique $\theta_* = (l_*, \omega_*)$ at which $d(\theta_*) = 0$.

From these remarks, it follows that $\sigma(l, \omega)$ is piecewise quadratic and convex for $\omega \in [\alpha_l, \alpha_{l+1}]$. Hence its maxima can only occur among the join points $\omega = 0, \omega_l, 1$ and ω_* in the

single case $l = l_*$. However, since $d(\theta_*) = 0$, it follows from (48) that $\sigma(l_*, \omega_*) \leq 1$, so we can safely ignore this case. Consequently, from (47) and noting that $\sigma(0, 1-) = 1$, we have

$$(49) \quad \|\sigma\|_\infty := \max_{\theta \geq 0, \theta = (l, \omega)} \sigma(l, \omega) = 1 \vee \max_{l \geq 1} \{\sigma(l, 0), \sigma(l, \omega_l), \sigma(l, \dot{\alpha}_l)\}.$$

Now suppose that $0 \leq \omega \leq \omega_l$ and that $d(l, \omega) \leq 0$. In this case, since $n \leq \check{n}$ and $1 - rv^{-1} = -r$, we have

$$(50) \quad \begin{aligned} \sigma(l, \omega) &= (\alpha_l + \omega)^2 - rn(l, \omega) \\ &= \omega^2 + (\alpha_l^2 + 2\alpha_l\omega)(1 - rv^{-1}) + r\beta_l + 1 \\ &= 1 + \omega^2 + r(\beta_l - \alpha_l^2) - 2r\alpha_l\omega \\ &\leq 1 + \omega^2 - 2r\omega \end{aligned}$$

say, where we used $\alpha_l \geq 1$ and $\alpha_l^2 - \beta_l \geq \alpha_1^2 - \beta_1 = 0$, from (35).

In particular $\sigma(l, 0) \leq 1$, and combining with (48), this holds for *all* l . Also, for $l \in \mathcal{L} = \{l : d(l, \omega_l) < 0\}$, we have $\sigma(l, \omega_l) \leq 1 + \omega_l(\omega_l - 2r)$, while for $l \notin \mathcal{L}$, again from (48), $\sigma(l, \omega_l) \leq 1$. Now, (49) simplifies to

$$(51) \quad \|\sigma\|_\infty \leq 1 + \max_{l \in \mathcal{L}} \omega_l(\omega_l - 2r)_+.$$

For the grid prior, $b = 1$. We have $\omega_l = (1 + v)/2 = (2r + 1)/(2r + 2)$, and

$$\omega_l(\omega_l - 2r) = (1 + 2r)(1 + r)^{-2}(1 - 2r - 4r^2)/4 = h_r$$

and we have established the upper bound in Theorem 1.1.

For the lower bound, it suffices to look at the risk at a single point. In view of Figure 3 and the discussion preceding (49), we try $\theta_1 = \lambda(1 + \omega_1)$. Look at the risk at θ_1 . Apply Lemma 3.2 using $n(1, \omega_1) = 2v^{-1}\omega_1$, to get from (25),

$$\rho(\theta_1, \hat{p}_G) \geq (2r)^{-1}\lambda^2\{(1 + \omega_1)^2 - 2rv^{-1}\omega_1\} + O(\lambda) = (2r)^{-1}\lambda^2(1 + h_r^+) + O(\lambda),$$

since the quantity in braces equals $1 + \omega_1^2 - 2r\omega_1 = \sigma(1, \omega_1) = 1 + h_r^+$. This completes the proof of Theorem 1.1.

We now turn our attention to proving Theorem 1.2. We first verify that if $b \leq \min\{1, 4r\}$, then $l \geq 1 + \lceil 2b^{-3/2} \rceil$ necessarily implies $d(l, \omega_l) \geq 0$.

From the monotonicity (35), along with $\omega \geq 0$, we have

$$d(l, \omega_l) + r^{-1} \geq \alpha_l^2 - \beta_l \geq \alpha_K^2 - \beta_K \geq b^2(K - 1)^2 \geq 4b^{-1} \geq r^{-1},$$

using $b \leq 2$ for the third, $K \geq 1 + \lceil 2b^{-3/2} \rceil$ for the fourth and $b \leq 4r$ for the fifth inequalities.

Now, return to (51): if $l \in \mathcal{L}$ then $d(l, \omega_l) < 0$ and so, from the previous paragraph necessarily $l < 1 + \lceil 2b^{-3/2} \rceil$, which by definition entails $\dot{\alpha}_l = b$ so long as $K \geq 1 + \lceil 2b^{-3/2} \rceil$. Now $\omega_l = b(1 + v)/2 \leq 2r$ is equivalent to $b \leq 4r/(1 + v)$. So, in this case, $\|\sigma\| = 1$, and so for *all* θ we have $\rho(\theta, \hat{p}_D) \leq (2r)^{-1}\lambda^2 + O(\lambda)$, which establishes (29), and hence Theorem 1.2.

3.4. Proof of Theorem 1.3. By Theorem 1.2, it suffice to prove a lower bound on the Bayes risk. As $\pi_{B,n}$ is i.i.d. and due to the product structure of the problem, its Bayes risk simplifies

$$B(\pi_{B,n}, \hat{p}_B) = nB(\pi_B, \hat{p}_B).$$

For the univariate problem the Bayes risk of the prior π_B is

$$\begin{aligned} B(\pi_B, \hat{p}_B) &\geq \eta_n c(\eta_n) \{\rho(\lambda_n, \hat{p}_B) + \rho(-\lambda_n, \hat{p}_B)\} \\ &= 2\eta_n c(\eta_n) \rho(\lambda_n, \hat{p}_B) \geq 2\eta_n c(\eta_n) [\lambda_n^2/(2r) - \mathbb{E} \log N_{\lambda_n, v}(Z)], \end{aligned}$$

where the equality above follows by symmetry and the inequality by (25). From (36), we have $2c(\eta_n) = c_B(\eta) \geq 1 - O(\eta_n^{b^2 v})$. Lemma 3.2 shows that $\mathbb{E} \log N_{\lambda, v}(Z) = O(\lambda)$ because $n(1, 0)$ [defined in (45)] equals 0. Hence $B(\pi_B, \hat{p}_B) \geq \eta_n \lambda_n^2 / (2r) \cdot (1 + o(1))$ and the proof is done.

3.5. Three point priors. Let $\pi_a = \pi_a[\eta]$ be a sparse symmetric three point prior given by (8) with $v^+ = \delta_{a\lambda}$ for $a > 0$. In Section 6, we prove the following.

LEMMA 3.3. *Let \hat{p}_a be the prde corresponding to π_a . Then, as $\eta \rightarrow 0$,*

$$(52) \quad \rho(a\lambda, \hat{p}_a) \leq (2r)^{-1} \lambda^2 \tau(a) + O(\lambda)$$

$$(53) \quad \tau(a) = \begin{cases} a^2 & a^2 \leq 1, \\ [1 - r(a^2 - 1)]_+ & a^2 \geq 1. \end{cases}$$

In particular, as $\eta \rightarrow 0$, the prior π_a is least favorable only when $a = 1$:

$$(54) \quad B(\pi_a, \hat{p}_a) \sim \eta \rho(a\lambda, \hat{p}_a) \sim (2r)^{-1} \eta \lambda^2 \tau(a).$$

3.6. Remarks. 1. When $K = \infty$, the bi-grid prior π_B has support points (in \mathbb{R}^+) separated by $(1, b, b, \dots)$. We denote this special case $\pi_{B'}$, and we emphasize that it is still a bi-grid prior (unless $b = 1$), though it may be seen as simpler than π_B . The proof of Theorem 1.2 shows that with $b = 4r(1+r)(1+2r)^{-1}$, prior $\pi_{B'}$ is asymptotically minimax for $r \leq r_0$.

However, there is no choice of b for which $\pi_{B'}$ is asymptotically minimax for all r . Indeed, if b be fixed, simply choose r small enough that $b > 4r(1+r)(1+2r)^{-1} = 4r/(1+v)$, and then from (50), we have

$$\|\sigma\|_\infty \geq \sigma(1, \omega_1) = 1 + \omega_1(\omega_1 - 2r) > 1.$$

2. When s_n does not diverge to ∞ , an “independent blocks” sparse prior using π_B is asymptotically least favorable, along the lines of [19], Chapter 8.6. Let $\pi_S(\tau; m)$ denote a single spike prior of scale τ on \mathbb{R}^m . This chooses an index $I \in \{1, \dots, m\}$ at random and sets $\theta = \tau e_I$, where e_i is a unit length vector in the i th coordinate direction. We randomly draw τ from $(v_B^+ + v_B^-)/2$. However, instead of (11), we choose $\lambda = v^{1/2}(t_m - \log t_m)$ where $t_m = \sqrt{2 \log m}$. The independent blocks prior $\pi_{B,n}$ on $\Theta[s_n]$ is built by dividing $\{1, \dots, n\}$ into s_n contiguous blocks B_j , each of length $m = m_n = \lfloor n/s_n \rfloor$. Independently, for each block B_j , draw components according to $\pi_S(\cdot; m)$ and set $\theta_i = 0$ for the remaining $n - m_n s_n$ coordinates. This prior is supported on $\Theta[s_n]$ as any draw from $\pi_{B,n}$ has exactly s_n nonzero components. The proof that it is least favorable is then analogous to that of Theorem 6 in [34].

4. Risk properties of spike and slab procedures. We again use the risk decomposition provided by Lemma 2.1, now with the univariate spike and slab prior $\pi_S[\eta, \ell]$. We use $N_{\theta, v}^S(Z)$ and $D_\theta^S(Z)$ to denote the associated risk components of Lemma 2.1 for the spike and slab predictive density estimates $\hat{p}_S[\ell]$ based on the prior $\pi_S[\eta, \ell]$ for some $\ell > 0$ (the dependence on ℓ is kept implicit in the notation).

Figure 4 gives a schematic showing the strategy for the proof of Theorems 1.5 and 1.6. Separate risk bounds for $\hat{p}_S[\ell]$ are established below for θ lying in intervals roughly corresponding to $[0, \lambda_n]$, $[\lambda_n, \lambda_e]$ and $[\lambda_e, \ell]$ where $\lambda_e = v^{-1/2} \lambda_n$; a threshold used in sparse point estimation. The critical interval is $[\lambda_n, \lambda_e]$, and the risk bound there suffices for asymptotic minimaxity if $\log \ell = o(\lambda_n^2)$, which leads to Theorem 1.5 if $\log t_n = o(\lambda_n^2)$ and we take $\ell = t_n$.

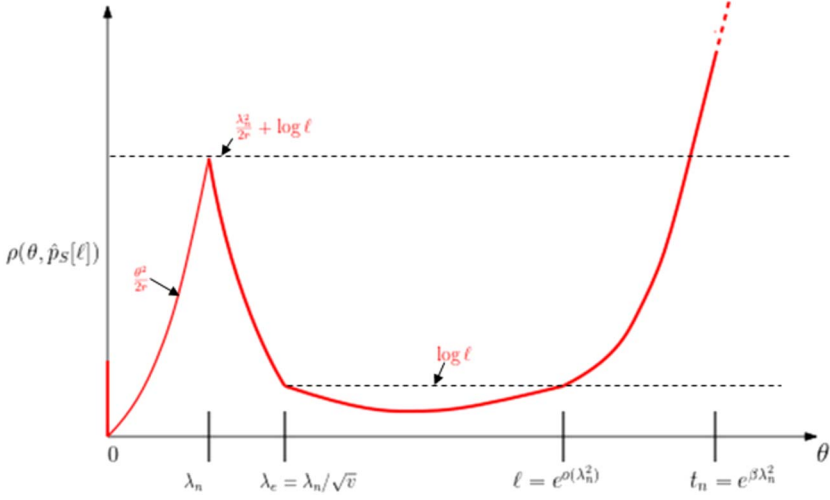


FIG. 4. Schematic for risk bounds (56) for uniform slab prior $\pi_S[\eta, \ell]$ and estimate $\hat{p}_S[\ell]$.

If, however, $\log t_n \sim \beta \lambda_n^2$, then no uniform slab width works: if $\log \ell \geq \beta \lambda_n^2/(2v)$, roughly, then the maximum at approximately $\theta = \sqrt{1 + \beta} \lambda_n$ is too high, while for $\log \ell < \beta \lambda_n^2/(2v)$, the maximum risk is too large near the right endpoint, $\theta = t_n$.

To ease notation, we often drop the suffixes from λ_n and η_n , particularly while discussing univariate *prdes*. Their risk functions are calculated in the regime $\lambda \rightarrow \infty$ as $\eta \rightarrow 0$.

PROOF OF LEMMA 1.4. For the first upper bound, simply take $v = \delta_0$ in Lemma 2.1; the corresponding $\pi = \delta_0$ has $\rho(\theta, \hat{p}_{\delta_0}) = \theta^2/(2r)$. The bound now follows from (19). For the second statement, we claim that whenever $t_n > \lambda_n$, then as $n \rightarrow \infty$,

$$(55) \quad R_N(\Theta_n[s_n, t_n]) \sim R_N(\Theta_n[s_n]) \sim s_n \lambda_n^2/(2r).$$

Indeed, the independent blocks prior π_n^{IB} constructed in [34], Theorem 6, to show that $R_N(\Theta_n[s_n]) \sim s_n \lambda_n^2/(2r)$ is actually, by its very definition, supported on $\Theta_n[s_n, v_n]$, where $v_n < \sqrt{v} \sqrt{2 \log[n/s_n]} \leq \lambda_n < t_n$. Since obviously $\Theta_n[s_n, v_n] \subset \Theta_n[s_n, t_n] \subset \Theta_n[s_n]$, the conclusion (55) follows. \square

For lower bounds on risk of its predictive density estimate, the following convexity inequality is helpful. It is proved in the Supplementary Material.

LEMMA 4.1. *If $\eta \leq \frac{1}{2}$ and $\theta \ell/v \geq 1$, then*

$$\mathbb{E} \log N_{\theta, v}^S(Z) \leq \theta \ell/v.$$

The *proof of* (16) follows easily from the above lemma. From the left-hand side of (25) and Lemma 4.1,

$$\rho(\theta, \hat{p}_S[\ell]) \geq \frac{\theta^2}{2r} - \frac{\theta \ell}{v}, \quad \text{for } \theta \geq \frac{v}{\ell}.$$

Hence, from (18),

$$\sup_{\Theta_n[s_n]} \rho(\theta, \hat{p}_S[\ell]) \geq s_n \sup_{\theta \in \mathbb{R}} \rho(\theta, \hat{p}_S[\ell]) = \infty,$$

while $R_N(\Theta_n[s_n])$ is finite for each n , for example, [34], so (16) follows.

4.1. *Proof of Theorem 1.5.* Recall that $\lambda = \sqrt{2v \log \eta^{-1}}$ and define $\tilde{\lambda} = \lambda/\sqrt{v} + \sqrt{2 \log \lambda}$. We will show a piecewise risk bound

$$(56) \quad \rho(\theta, \hat{p}_S[\ell]) \leq \begin{cases} \theta^2/(2r) + O(\lambda \log \lambda) & 0 < \theta < \lambda, \\ \lambda^2/(2r) + \log \ell + O(\lambda \log \lambda) & \lambda \leq \theta < \tilde{\lambda}, \\ \log \ell + O(\lambda) & \tilde{\lambda} \leq \theta \leq \ell. \end{cases}$$

For $0 < \theta < \lambda$, simply use the basic upper bound (25) along with the following bound for D_θ^S , shown in the Supplementary Material: for each $r > 0$,

$$(57) \quad \mathbb{E} \log D_\theta^S(Z) \leq \begin{cases} O(\lambda \log \lambda) & 0 < \theta < \tilde{\lambda}, \\ \theta^2/2 - \lambda^2/(2v) + O(\lambda) & \theta \geq \tilde{\lambda}. \end{cases}$$

For the remaining two cases, that is, for $\theta > \lambda$, we use the full decomposition (21) of Lemma 2.1. To this end, an alternative representation for $N_{\theta,v}^S$ will be useful. Completing the square in (22), we get

$$(58) \quad N_{\theta,v}^S(Z) = 1 + c(\eta)\sqrt{v} \exp\left(\frac{1}{2}Z_{\theta,v}^2\right) \Phi_{\ell,v},$$

where we have set $Z_{\theta,v} = Z + \theta/\sqrt{v}$ and

$$\Phi_{\ell,v} = \Phi(v^{-1/2}(\ell - \theta) - Z) - \Phi(v^{-1/2}(-\ell - \theta) - Z).$$

In the Supplementary Material, we show that, uniformly in $v \in (0, 1)$, $\ell \geq 1$ and $|\theta| \leq \ell$,

$$(59) \quad \mathbb{E} \log \Phi_{\ell,v} \geq a_0 := \log \phi(0) + 2/3.$$

The constant $c(\eta) = \eta(1 - \eta)^{-1}\{2\ell\phi(0)\}^{-1}$ satisfies

$$(60) \quad -\log \ell - \lambda^2/(2v) \leq \log\{(1 - \eta)c(\eta)\} = \log \phi(0) - \log \ell - \lambda^2/(2v)$$

From the preceding three displays and $\mathbb{E}Z_{\theta,v}^2 = 1 + \theta^2/v$, we obtain

$$(61) \quad \begin{aligned} -\mathbb{E} \log N_{\theta,v}^S(Z) &\leq -\log c(\eta) - \frac{1}{2} \log v - \frac{1}{2} \mathbb{E}Z_{\theta,v}^2 - \mathbb{E} \log \Phi_{\ell,v} \\ &\leq \log \ell + \lambda^2/(2v) - \theta^2/(2v) + O(1). \end{aligned}$$

Now observe from (61) and $v^{-1} = r^{-1} + 1$ that

$$\theta^2/(2r) - \mathbb{E} \log N_{\theta,v}^S(Z) \leq \lambda^2/(2r) - (\theta^2 - \lambda^2)/2 + \log \ell + O(1).$$

Combining this with the bounds in (57) yields the remaining two bounds.

For any $\ell \geq 1$ such that $\log \ell = o(\lambda^2)$, we conclude that as $\lambda \rightarrow \infty$,

$$\sup_{\theta \leq \ell} \rho(\theta, \hat{p}_S[\ell]) \leq \frac{\lambda^2}{2r} (1 + o(1)).$$

This completes the proof of (29) and, as remarked there, the proof of Theorem 1.5.

4.2. *Proof of Theorem 1.6.* We use the basic lower risk bound (25), and show that for suitable θ that $\mathbb{E} \log N_{\theta,v}^S Z$ cannot be large enough to offset the leading term $\theta^2/(2r)$. To obtain a result uniform over all slab widths ℓ , we need two different types of upper bound on $N_{\theta,v}^S$.

Define t_λ and $\tilde{t}_\lambda = o(t_\lambda)$ by setting $\log t_\lambda = \beta\lambda^2/(2v)$ and $\log \tilde{t}_\lambda = \log t_\lambda - \lambda$. We look first at large values of ℓ , using representation (58). Observe first that for $\ell > \tilde{t}_\lambda$, the right-hand side of (60) yields

$$\sqrt{v}c(\eta) \leq C \exp\{-\log \tilde{t}_\lambda - \lambda^2/(2v)\} = C \exp\{-\tilde{\theta}^2/(2v)\}$$

for a constant $C = C(v)$ if we set $\tilde{\theta}^2 = \lambda^2 + 2v \log \tilde{t}_\lambda$. Using now (58) and $\Phi_{\ell,v} < 1$, we have

$$\begin{aligned} \log N_{\tilde{\theta},v}^S(Z) &\leq \log\{1 + C \exp[-\tilde{\theta}^2/(2v) + (Z + \tilde{\theta}/\sqrt{v})^2/2]\} \\ &\leq \log 2 + \log(1 + C) + Z^2/2 + |Z|\tilde{\theta}/\sqrt{v}. \end{aligned}$$

Consequently, $\mathbb{E} \log N_{\tilde{\theta},v}^S(Z) \leq k_1 + k_2 \tilde{\theta}$ where $k_i = k_i(v)$. Hence, from the left-hand side of risk bound (25),

$$\rho(\tilde{\theta}, \hat{p}_S[\ell]) \geq \frac{\tilde{\theta}^2}{2r} - k_1 \tilde{\theta} - k_2.$$

Now observe from the definition of \tilde{t}_λ that $\tilde{\theta}^2 = (1 + \beta)\lambda^2 - 2v\lambda$ and that $\tilde{\theta} < t_\lambda$ for large λ . We conclude that for large λ ,

$$(62) \quad \inf_{\ell > \tilde{t}_\lambda} \sup_{\theta \in [0, t_\lambda]} \frac{\rho(\theta, \hat{p}_S[\ell])}{\lambda^2/(2r)} \geq 1 + \beta + O(\lambda^{-1}).$$

For $\ell \leq \tilde{t}_\lambda$, we set $\theta = t_\lambda$ and use the left-hand side of (25), then Lemma 4.1:

$$\sup_{\theta \leq t_\lambda} \rho(t_\lambda, p_S[\ell]) \geq \frac{t_\lambda^2}{2r} - \frac{t_\lambda \ell}{v} \geq \frac{t_\lambda^2}{2r} - \frac{t_\lambda \tilde{t}_\lambda}{v} \geq \frac{t_\lambda^2}{2r} (1 + o(1)),$$

where in the last inequality we used $\tilde{t}_\lambda = o(t_\lambda)$. Consequently,

$$(63) \quad \inf_{\ell \leq \tilde{t}_\lambda} \sup_{\theta \in [0, t_\lambda]} \frac{\rho(\theta, \hat{p}_S[\ell])}{\lambda^2/(2r)} \geq \frac{t_\lambda^2}{\lambda^2} (1 + o(1)).$$

Combining (62) with (63) and then using (19) to go over to the multivariate problem, we obtain

$$\min_{\ell > 1} \sup_{\Theta_n[s_n, t_n]} \rho(\boldsymbol{\theta}, \hat{p}_S[\ell]) \geq (1 + \beta) s_n \lambda_n^2 / (2r) (1 + o(1)).$$

Theorem 1.6 now follows from (17) of Lemma 1.4.

5. Numerical experiments. We turn to the numerical effectiveness of our asymptotic results under different levels of sparsity η_n , with special focus on moderate values. The product structure and the good bounds (19) relating maximal multivariate and univariate risks allow us to concentrate on the univariate prdes. We use a constrained prior space

$$\mathbf{m}_\ell(\eta) = \{\pi \in \mathcal{P}(\mathbb{R}) : \pi(\theta = 0) \geq 1 - \eta, \pi(|\theta| > \ell) = 0\},$$

and set $\ell = 5\lambda = 5\sqrt{2 \log \eta^{-v}}$. We consider three sparsity levels: (a) Moderate: $\eta = 0.1$, (b) High: $\eta = 0.001$, (c) Very High: $\eta = 10^{-10}$.

We compare the following prdes:

- Hard threshold Plug-in prde (H-Plugin): [34], equation (31),

$$\hat{p}_H(y|x) = p(y|\hat{\theta}_H, v_y) \text{ where } \hat{\theta}_H(x) = x I\{|x| > (v_x/v)^{1/2}\lambda\}.$$

- Cluster prior and Thresholding (C-Thresh) based asymptotically minimax prde \hat{p}_T proposed in [34], equations (12)–(14),
- Bayes prdes based on the grid and bi-grid priors (Grid, Bi-Grid) rescaled on \mathbf{m}_ℓ : \hat{p}_G, \hat{p}_B
- Spike and slab predictive density estimator (SS): $\hat{p}_S[\ell]$.

TABLE 1

Numerical evaluation of the maximum risk for the different univariate predictive densities over $[-\ell, \ell]$ as the degree of sparsity (η) and predictive difficulty r varies. Here, we have chosen $\ell = 5\lambda$, where λ is defined in (11). In the “Asymp” column, we report the asymptotic minimax risk $\lambda^2/(2r)$. In the other columns, we report the maximum risk of the estimators as quotients of the “Asymp” risk

Sparsity	r	Asymp	H-Plugin	C-Threshold	Grid	Bi-Grid	SS
0.1	1	1.1513	120.4%	82.5%	88.3%	88.3%	105.8%
	0.5	1.5351	173.6%	108.8%	104.9%	104.9%	118.0%
	0.25	1.8421	278.5%	128.0%	127.0%	129.0%	132.3%
	0.1	2.0933	588.1%	145.2%	165.4%	155.9%	146.5%
0.001	1	3.4539	109.1%	70.7%	70.8%	70.8%	86.2%
	0.5	4.6052	162.1%	85.9%	84.6%	84.6%	96.9%
	0.25	5.5262	267.6%	89.9%	100.2%	96.8%	106.9%
	0.1	6.2798	582.8%	107.2%	115.6%	113.4%	118.0%
1E-10	1	11.5129	123.9%	150.4%	78.6%	78.6%	86.9%
	0.5	15.3506	185.4%	87.9%	87.1%	87.1%	93.9%
	0.25	18.4207	308.4%	94.6%	98.1%	96.3%	100.1%
	0.1	20.9326	677.0%	101.8%	110.5%	101.7%	106.3%

Table 1 reports the maximum value of the risk plots for these predictive estimators (Supplementary Material, Table 1 shows the locations of the respective maxims). Figure 5 plots $\theta \rightarrow \rho(\theta, \hat{p})$, showing however the rescaled value $\rho(0, \hat{p})(1 - \eta)/\eta$ at $\theta = 0$. [The hard threshold plug-in density estimator \hat{p}_H is omitted, as has poor maximum risk in Table 1 and confuses the plots.]

The tables and plots show that the bi-grid prior Bayes prde \hat{p}_B and the C-Threshold prde \hat{p}_T have similar worst case performance. For each r , the maximal risks of \hat{p}_B and \hat{p}_T lie near or below the asymptotic level of $\log \eta^{-1}/(1 + r)$ under high and very high sparsity, and at worst moderately above the asymptotic level for moderate sparsity. However, \hat{p}_T has substantially higher risk at the origin than the other prdes considered here, particularly for moderate sparsity. Differences in the performances of the grid and bi-grid prior based prdes appear under high sparsity; for further comparisons, see Figures 2 and 3 in the Supplementary Material. The maximal risk of the spike and slab procedure is higher than that of \hat{p}_T or \hat{p}_B but does not exceed the asymptotic minimax level by much. Finally, the basic features of the risk plots are unchanged even under moderate sparsity.

6. Discussion and future work. Product priors based on infinite cluster priors $\pi_\infty[\eta, r]$ of [34], Section 6, will lead to minimax optimal Bayes prdes. Details, which do not follow directly from those for the bi-grid prior, are provided in [8].

Our discussion of spike and slab priors was confined to uniform slabs. Theorem 2.1 can be used to show that Gaussian slabs are suboptimal, while Bayes prdes based on heavier-tailed slabs in the range from Laplace to Cauchy are minimax optimal. The tools to bound the maximal risk of continuous priors differ from those used here and will be detailed separately [33].

Our results are based on known sparsity levels. We make a few remarks on adaptation to unknown sparsity from theoretical and computational perspectives. A manuscript in preparation considers adaptivity for continuous slabs with Laplace and Cauchy tails. Adaptation to minimax risk is possible up to multiplicative constants and an additive logarithmic term. Both exact sparsity (ℓ_0) and approximate sparsity ($\ell_p, 0 < p < 2$) are considered.

Recently, computationally tractable Bayesian methods which adapt to unknown sparsity levels and possibly dense signals have been developed for point estimation [3, 5, 40]. In our

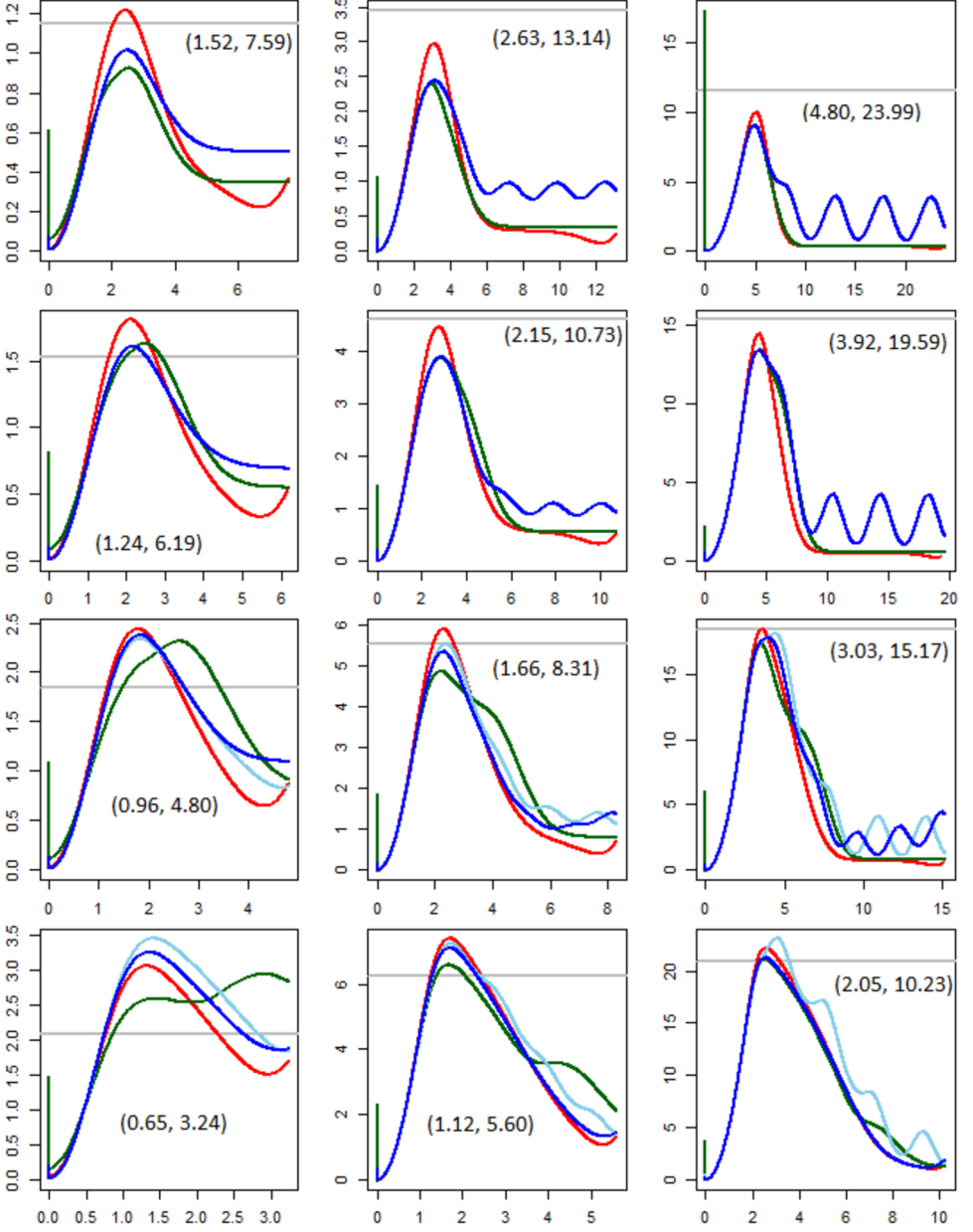


FIG. 5. Risk plots $\rho(\theta, \cdot) \{(1 - \eta)/\eta\}^{I(\theta=0)}$ for univariate predictive density estimators \hat{p}_T (dark green), \hat{p}_G (skyblue), \hat{p}_B (blue) and \hat{p}_S (red) versus $\theta \in [0, \ell]$, for $\ell = 5\lambda$. Columns vary with moderate, high and very high sparsity, $\eta = 0.1, 0.001, 10^{-10}$, left to right. Rows vary $r = 1, 0.5, 0.25$ and 0.1 from top to bottom. The horizontal line shows the asymptotic univariate minimax risk of $\log \eta^{-1}/(1+r) = \lambda^2/(2r)$, with $\lambda = \sqrt{2 \log \eta^{-v}}$ and ℓ shown in the insets. Note that, \hat{p}_G (skyblue) and \hat{p}_B (blue) overlap exactly in plots for the first two rows.

sequence model (1), under unknown sparsity level $\eta_n = s_n/n$, there exist fast procedures for estimating posteriors from spike-and-slab priors that are mixtures of a Dirac measure at 0 and a continuous distribution [6, 20, 41].

APPENDIX

We present proofs of the risk bounds in Lemmas 3.1 and 3.3. The proof of Lemma 3.2 uses tools similar to Lemma 3.1 and appears in the Supplementary Material.

Proof of Lemma 3.1. We do the easy lower bounds involving $N_{\theta,v}(Z)$ first. Indeed, the bound for $\theta < \lambda$ follows just from $N_{\theta,v}(Z) \geq 1$. For $\theta = \lambda(\alpha_l + \omega) \geq \lambda$, from (44) using $\mathbb{E}(Z) = 0$ we get

$$\mathbb{E} \log N_l = \log c_1(\eta) + \frac{1}{2} \lambda^2 n(l, \omega), \quad \text{and,}$$

$$\mathbb{E} \log N_{l+1} = \log c_1(\eta) + \frac{1}{2} \lambda^2 \check{n}(l, \omega).$$

But $\log c_1(\eta) = \log c(\eta) - \log(1 - \eta)^{-1} = O(1)$ as $\lambda \rightarrow \infty$. Hence, the proof of the lower bound is completed by using

$$\mathbb{E} \log N_{\theta,v}(Z) \geq \max\{\mathbb{E} \log N_l, \mathbb{E} \log N_{l+1}\}.$$

The proof of the upper bound on $\mathbb{E} \log D_\theta(Z)$ is more involved, and we first outline the approach. From (39) and $1 + x + y < (1 + x)(1 + y/x)$, we have

$$(64) \quad \log D_\theta(Z) \leq \log(1 + D_l) + \log(1 + \check{D}_l),$$

where we set $\check{D}_l = \sum_{i \notin \{0, l\}} D_i / D_l$. Henceforth in the proof, we make the choice $l = l(\theta)$ except that when $0 \leq \theta < \mu_1$ we set $l = 1$.

For the first term (henceforth we call it the main term) in (64), we will show

$$(65) \quad \mathbb{E} \log(1 + D_l) \leq \begin{cases} \frac{1}{2} \lambda^2 d^+(l, \omega) + O(\lambda) & \text{for } l \geq 1, \\ O(1) & \text{if } 0 \leq \theta < \lambda \end{cases}$$

with $O(\lambda)$ being uniform in l . For the other term in (64), we will show that it is $O(\lambda)$ for all l (and so, henceforth we call it the remainder term). For that purpose, we write $D_{i,l} = D_i / D_l$ and decompose

$$\check{D}_l = \sum_{k=1}^{\infty} D_{l+k,l} + \sum_{k=1}^{l-1} D_{l-k,l} + \sum_{j=1}^{\infty} D_{-j,l}.$$

We use the elementary inequality $\log(1 + \sum \gamma_m) \leq \sum \log(1 + \gamma_m)$ to obtain that $\mathbb{E} \log(1 + \check{D}_l)$ is bounded above by

$$\mathbb{E} \log \left(1 + \sum_{k=1}^{\infty} D_{l+k,l} \right) + \mathbb{E} \log \left(1 + \sum_{k=1}^{l-1} D_{l-k,l} \right) + \mathbb{E} \log \left(1 + \sum_{j=1}^{\infty} D_{-j,l} \right).$$

Now, note that $D_{-j} \stackrel{\mathcal{D}}{=} D_j \exp\{-2\mu_j \theta\} \leq D_j$ since $\mu_j = -\mu_j$, $\pi_{-j} = \pi_j$ and $\mathcal{L}(Z)$ is symmetric. Hence

$$\sum_{j=1}^{\infty} D_{-j,l} \stackrel{\mathcal{D}}{\leq} \sum_{j=1}^{\infty} D_{j,l} = \sum_{k=1}^{l-1} D_{l-k,l} + 1 + \sum_{k=1}^{\infty} D_{l+k,l}.$$

Combining the above two displays and again using the aforementioned inequality on log sums, we obtain

$$(66) \quad \mathbb{E} \log(1 + \check{D}_l) \leq 2\mathbb{E} \log \left(1 + \sum_{k=1}^{\infty} D_{l+k,l} \right) + \log 2 + 2\mathbb{E} \log \left(1 + \sum_{k=1}^{l-1} D_{l-k,l} \right).$$

We will later show that the two main right-hand side terms are each $O(\lambda)$. This concludes the outline; we now turn to detailed analysis.

The main term in (64). We first dispose of the case $0 \leq \theta < \lambda$. From (40) and (41),

$$D_1 = c_1(\eta) \exp \left\{ \lambda Z + \lambda \theta - \frac{1}{2} \lambda^2 (2 + r^{-1}) \right\}.$$

Since $\theta < \lambda$ and $c_1(\eta) < (1 - \eta)^{-1}$, and using $\log(1 + x) \leq \log 2 + (\log x)_+$,

$$\log(1 + D_1) \leq \log 2 + \log(1 - \eta)^{-1} + \lambda(Z - 2^{-1}r^{-1}\lambda)_+$$

and hence $\mathbb{E} \log(1 + D_1) \leq O(1)$. This last bound uses an inequality we also need later: from the two term bound on Mills ratio (e.g., [19], Exercise 8.1),

$$(67) \quad \mathbb{E}(Z - x)_+ = \phi(x) - x\tilde{\Phi}(x) \leq x^{-2}\phi(x).$$

Now suppose that $\theta = \lambda(\alpha_l + \omega) \geq \lambda$ and use representation (44) for D_l . Abbreviating $\frac{1}{2}\lambda^2 d(l, \omega)$ as $d_{l\omega}$, we obtain

$$\begin{aligned} \mathbb{E} \log(1 + D_l) &= \mathbb{E} \log D_l + \mathbb{E} \log(1 + D_l^{-1}) \\ &= \log c(\eta) + \log(1 - \eta)^{-1} + d_{l\omega} + \log 2 + \mathbb{E}(\log D_l^{-1})_+. \end{aligned}$$

Symmetry of $\mathcal{L}(Z)$ about 0 implies that $\log D_l^{-1} \stackrel{\mathcal{D}}{=} -\log c(\eta) + \log(1 - \eta) + \mu_l Z - d_{l\omega}$. As $c(\eta) < 1$, we have

$$\mathbb{E}(\log D_l^{-1})_+ \leq -\log c(\eta) + \mathbb{E}(\mu_l Z - d_{l\omega})_+.$$

From the previous two displays and $\log(1 - \eta)^{-1} = O(\eta)$, we have

$$(68) \quad \mathbb{E} \log(1 + D_l) \leq d_{l\omega} + \mathbb{E}(\mu_l Z - d_{l\omega})_+ + O(1).$$

We now bound the expectation on the right-hand side. Consider first those l for which $\alpha_l \leq 2 + r^{-1}$, and thus $\mu_l \leq (2 + r^{-1})\lambda$. Noting that

$$\mathbb{E}(\mu_l Z - d_{l\omega})_+ \leq -d_{l\omega} I\{d_{l\omega} \leq 0\} + \mu_l \mathbb{E} Z_+,$$

we then conclude that

$$d_{l\omega} + \mathbb{E}(\mu_l Z - d_{l\omega})_+ \leq (d_{l\omega})_+ + (2 + r^{-1})\phi(0)\lambda.$$

Now consider the remaining l , with $\alpha_l \geq 2 + r^{-1}$, for which we claim that

$$(69) \quad \alpha_l^2 - \beta_l - r^{-1} \geq \frac{1}{2}\alpha_l^2.$$

We verify this via the equivalent form $\alpha_l^2 - 2\beta_l \geq 2r^{-1}$. Indeed, since $\beta_l \leq \alpha_l$, we have

$$\alpha_l^2 - 2\beta_l \geq \alpha_l(\alpha_l - 2) \geq (2 + r^{-1})r^{-1} \geq 2r^{-1}.$$

Since $\omega \geq 0$, we have from (45) and (69),

$$d_{l\omega} \geq \frac{1}{2}\lambda^2[\alpha_l^2 - \beta_l - r^{-1}] \geq \frac{1}{4}(\lambda\alpha_l)^2 = \frac{1}{4}\mu_l^2.$$

From the bound (67), we calculate

$$\mathbb{E}(\mu_l Z - d_{l\omega})_+ \leq \mu_l \mathbb{E}(Z - \mu_l/4)_+ \leq 16 \frac{\phi(\mu_l/4)}{\mu_l}$$

uniformly in $\lambda \geq 1$ and l such that $\alpha_l \geq 2 + r^{-1}$. The right-hand side is uniformly bounded in l . Combining the two cases with (68), we have proven the bound (65) on the first term of (64).

We turn now to bounding the remainder (66). This depends on the decay between successive terms D_j , so we start by using (42) to derive a useful representation for D_{j+1}/D_j . Indeed, using $\mu_j = \lambda\alpha_j$ and $\theta = \lambda(\alpha_l + \omega)$, we define

$$\begin{aligned}\Delta_j &= \Delta(j; l, \omega) = (2/\lambda^2)[G(\mu_{j+1}; \theta) - G(\mu_j; \theta)] \\ &= \dot{\alpha}_j[\alpha_{j+1} + \alpha_j - 2\alpha_l - 2\omega] + \dot{\beta}_j\end{aligned}$$

and arrive at, for $j \geq 1$,

$$(70) \quad \frac{D_{j+1}}{D_j} = \exp\left\{\lambda\dot{\alpha}_j Z - \frac{1}{2}\lambda^2\Delta_j\right\}.$$

We now show that Δ_j crosses zero at $j = l$, meaning $\Delta_j \geq 0$ for $j \geq l$ and $\Delta_j \leq 0$ for $j < l$. This will also verify the claim in Section 2 that $j \rightarrow G(\mu_j; \theta)$ is minimized at $j = l(\theta)$ for each $\theta \in [\mu_l, \mu_{l+1})$. The argument splits into two largely parallel cases.

Suppose first that $j \geq l$, so that $j = l + k$ for $k \geq 0$. Using $\alpha_l + \omega < \alpha_{l+1}$, then $\dot{\beta}_{l+k} = \dot{\alpha}_{l+k}^2$ and finally $\dot{\alpha}_{l+k} + \alpha_{l+k} = \alpha_{l+k+1}$, we have for any $k \geq 0$,

$$(71) \quad \Delta_{l+k} > \dot{\alpha}_{l+k}[\alpha_{l+k+1} + \alpha_{l+k} - 2\alpha_{l+1}] + \dot{\alpha}_{l+k}^2 = 2\dot{\alpha}_{l+k}(\alpha_{l+k+1} - \alpha_{l+1}) \geq 0,$$

with the last inequality being strict for $k \geq 1$.

Suppose now that $j < l$, so that $j = l - k - 1$ for $k \geq 0$. Using $\alpha_l + \omega \geq \alpha_l$, then $\dot{\beta}_{l-k-1} = \dot{\alpha}_{l-k-1}^2$, and finally $\dot{\alpha}_{l-k-1} + \alpha_{l-k-1} = \alpha_{l-k}$, we have

$$\begin{aligned}\Delta_{l-k-1} &\leq \dot{\alpha}_{l-k-1}[\alpha_{l-k} + \alpha_{l-k-1} - 2\alpha_l] + \dot{\alpha}_{l-k-1}^2 \\ &= 2\dot{\alpha}_{l-k-1}(\alpha_{l-k} - \alpha_l) \leq 0,\end{aligned}$$

with strict inequality when $k \geq 1$.

As final preparation, we record a useful bound whose proof is provided in the Supplementary Material.

LEMMA A.1. *If a_1, a_2, \dots are positive, then for each $n \geq 1$,*

$$(72) \quad \log\left(1 + \sum_{k=1}^{n+1} a_k\right) < \log(1 + a_1) + \sum_{k=1}^n \frac{a_{k+1}}{a_k}.$$

We next concentrate on bounding the *first term of (66)*. Noting that D_j s are positive, use (72) with $a_k = D_{l+k}/D_l$ and $\log(1 + a_1) \leq \log 2 + (\log a_1)_+$ to write

$$(73) \quad \mathbb{E} \log\left(1 + \sum_{k=1}^{\infty} D_{l+k,l}\right) \leq \log 2 + \mathbb{E}\left(\log \frac{D_{l+1}}{D_l}\right)_+ + \mathbb{E}\left[\sum_{k=1}^{\infty} \frac{D_{l+k+1}}{D_{l+k}}\right].$$

In (70) with $j = l$, we have seen that $\Delta_l \geq 0$ and so

$$\mathbb{E}\left(\log \frac{D_{l+1}}{D_l}\right)_+ \leq \lambda\dot{\alpha}_l \mathbb{E}Z_+ \leq \lambda\phi(0).$$

When $j = l + k$, observe from (71) that $\Delta_{l+k} \geq 2\dot{\alpha}_{l+k}^2 + 2\dot{\alpha}_{l+k}(\alpha_{l+k} - \alpha_{l+1})$. From (70), now with $j = l + k$ for $k \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\frac{D_{l+k+1}}{D_{l+k}}\right] &= \exp\left\{\frac{1}{2}\lambda^2[\dot{\alpha}_{l+k}^2 - \Delta_{l+k}]\right\} \\ &\leq \exp\left\{-\frac{1}{2}\lambda^2[\dot{\alpha}_{l+k}^2 + 2\dot{\alpha}_{l+k}(\alpha_{l+k} - \alpha_{l+1})]\right\} \\ &\leq \exp\left\{-\frac{1}{2}\lambda^2 b^2 - \lambda^2 b^2(k-1)\right\},\end{aligned}$$

so that the right-hand side of (73) is $O(\lambda) + O(e^{-\lambda^2 b^2/2}) = O(\lambda)$. The last inequality in the above display uses $j \rightarrow \dot{\alpha}_j$ is increasing and $\dot{\alpha}_j \geq b$.

Second term of (66). Now use (72) with $a_k = D_{l-k}/D_l$:

$$(74) \quad \mathbb{E} \log \left(1 + \sum_{k=1}^{l-1} D_{l-k,l} \right) \leq \log 2 + \mathbb{E} \left(\log \frac{D_{l-1}}{D_l} \right)_+ + \mathbb{E} \left\{ \sum_{k=1}^{l-2} \frac{D_{l-k-1}}{D_{l-k}} \right\}.$$

In (70) with $j = l - 1$, we have seen that $\Delta_{l-1} \leq 0$ and so

$$\mathbb{E} \left(\log \frac{D_{l-1}}{D_l} \right)_+ \leq \lambda \dot{\alpha}_{l-1} \mathbb{E} Z_+ \leq \lambda \phi(0).$$

From (70), now with $j = l - k - 1$,

$$\begin{aligned} \mathbb{E} \left\{ \frac{D_{l-k-1}}{D_{l-k}} \right\} &= \mathbb{E} \left\{ \exp \left\{ -\lambda \dot{\alpha}_{l-k-1} Z + \frac{1}{2} \lambda^2 \Delta_{l-k-1} \right\} \right\} \\ &\leq \exp \left\{ \frac{1}{2} \lambda^2 \dot{\alpha}_{l-k-1} [\dot{\alpha}_{l-k-1} + 2(\alpha_{l-k} - \alpha_l)] \right\}, \end{aligned}$$

as $\Delta_{l-k-1} \leq 2\dot{\alpha}_{l-k-1}(\alpha_{l-k} - \alpha_l)$. Again, using $j \rightarrow \dot{\alpha}_j$ is increasing and $\dot{\alpha}_j \geq b$, we have

$$\dot{\alpha}_{l-k-1} + 2(\alpha_{l-k} - \alpha_l) \leq \dot{\alpha}_{l-k} - 2(\alpha_l - \alpha_{l-k+1}) - 2\dot{\alpha}_{l-k} \leq -b - 2(k-1)b.$$

Using $\dot{\alpha}_{l-k-1} \geq b$ again, we conclude that

$$\mathbb{E} \left\{ \sum_{k=1}^{l-2} \frac{D_{l-k-1}}{D_{l-k}} \right\} \leq \sum_{k=1}^{\infty} \exp \left\{ -\frac{1}{2} \lambda^2 b^2 - \lambda^2 b^2 (k-1) \right\} = O(e^{-\lambda^2 b^2/2}).$$

Thus, we have proved the desired bound on the second term. This completes the proof of the lemma.

Proof of Lemma 3.3. The argument borrows some steps from the proof of Lemma 3.1, but is simpler, though not a special case. The three point prior corresponds, in (30) to choices $\pi_0 = 1 - \eta$, $\pi_1 = \eta/2$, $\mu_1 = a\lambda$. From (37)–(38), we have $N_{a\lambda, v}(Z) = 1 + N_1 + N_{-1}$, with

$$(75) \quad N_1 = c_1(\eta) \exp \left\{ v^{-1/2} a\lambda Z + (2v)^{-1} a^2 \lambda^2 - \frac{1}{2} \lambda^2 (1 + r^{-1}) \right\},$$

where $c_1(\eta) = 2^{-1}(1 - \eta)^{-1}$ and $N_{-1} = N_1 \exp(-2v^{-1/2} a\lambda Z - 2v^{-1} a^2 \lambda^2)$. Correspondingly $D_\theta(Z) = 1 + D_1 + D_{-1}$, where D_1 and D_{-1} are obtained from N_1 and N_{-1} by replacing v with 1. From Theorem 2.1, and $\log(1 + D_1 + D_{-1}) \leq \log(1 + D_1) + \log(1 + D_{-1}/D_1)$,

$$\begin{aligned} \rho(a\lambda, \hat{p}_a) &= (2r)^{-1} a^2 \lambda^2 - \mathbb{E} \log(1 + N_1 + N_{-1}) + \mathbb{E} \log(1 + D_1 + D_{-1}) \\ &\leq (2r)^{-1} a^2 \lambda^2 - \mathbb{E} \log(1 + N_1) + \mathbb{E} \log(1 + D_1) + \mathbb{E} \log(1 + D_{-1}/D_1) \\ &\leq (2r)^{-1} a^2 \lambda^2 - (\mathbb{E} \log N_1)_+ + 2 \log 2 + \mathbb{E}(\log D_1)_+ + \mathbb{E}[\log(D_{-1}/D_1)]_+. \end{aligned}$$

From (75), and its analog for D_1 , we have, on setting $\epsilon(\eta) = \log c_1(\eta) < 0$, recalling that $rv^{-1} = r + 1$, and using (67),

$$\begin{aligned} (\mathbb{E} \log N_1)_+ &= [\epsilon(\eta) + (2v)^{-1} (a^2 - 1) \lambda^2]_+ \geq \epsilon(\eta) + (2r)^{-1} \lambda^2 (r + 1) (a^2 - 1)_+, \\ \mathbb{E}(\log D_1)_+ &\leq \epsilon(\eta) + a\lambda \mathbb{E} Z_+ + (2r)^{-1} \lambda^2 (ra^2 - r - 1)_+, \\ \mathbb{E}[\log(D_{-1}/D_1)]_+ &= 2a\lambda \mathbb{E}(Z - a\lambda)_+ = 2\phi(a\lambda)/a\lambda = O(\lambda). \end{aligned}$$

Combine the last four displays to get

$$\rho(a\lambda, \hat{p}_a) \leq (2r)^{-1} \lambda^2 \tilde{\tau}(a) + O(\lambda),$$

where

$$\tilde{\tau}(a) = a^2 - (r+1)(a^2 - 1)_+ + (ra^2 - r - 1)_+ = \tau(a).$$

Acknowledgments. The authors thank the Associate Editor and three referees for especially stimulating comments that improved the presentation.

Funding. GM was supported in part by the Zumberge individual award from the University of Southern California's James H. Zumberge faculty research and innovation fund and by NSF Grant DMS-1811866.

IMJ was supported in part by NSF Grants DMS-1407813, 1418362 and 1811614 and thanks the Australian National University for hospitality while working on this paper.

SUPPLEMENTARY MATERIAL

Supplementary Materials to “On minimax optimality of sparse Bayes predictive density estimates” (DOI: [10.1214/21-AOS2086SUPP](https://doi.org/10.1214/21-AOS2086SUPP); .pdf). The supplement [35] proves Lemma 3.2 and all the inequalities and lemmas used in Section 4. It also contains results from additional numerical experiments and further discussions on the risk properties of prdes.

REFERENCES

- [1] AITCHISON, J. and DUNSMORE, I. R. (1975). *Statistical Prediction Analysis*. Cambridge Univ. Press, Cambridge-Melbourne. [MR0408097](#)
- [2] ASLAN, M. (2006). Asymptotically minimax Bayes predictive densities. *Ann. Statist.* **34** 2921–2938. [MR2329473](#) <https://doi.org/10.1214/0090536060000000885>
- [3] BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *J. Amer. Statist. Assoc.* **110** 1479–1490. [MR3449048](#) <https://doi.org/10.1080/01621459.2014.960967>
- [4] BROWN, L. D., GEORGE, E. I. and XU, X. (2008). Admissible predictive density estimation. *Ann. Statist.* **36** 1156–1170. [MR2418653](#) <https://doi.org/10.1214/07-AOS506>
- [5] CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. [MR2650751](#) <https://doi.org/10.1093/biomet/asq017>
- [6] CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. [MR3059077](#) <https://doi.org/10.1214/12-AOS1029>
- [7] FOURDRINIER, D., MARCHAND, É., RIGHI, A. and STRAWDERMAN, W. E. (2011). On improved predictive density estimation with parametric constraints. *Electron. J. Stat.* **5** 172–191. [MR2792550](#) <https://doi.org/10.1214/11-EJS603>
- [8] GANGOPADHYAY, U. and MUKHERJEE, G. (2021). On discrete priors and sparse minimax optimal predictive densities. *Electron. J. Stat.* **15**. [MR4255304](#) <https://doi.org/10.1214/21-ejs1818>
- [9] GEISSER, S. (1993). *Predictive Inference: An Introduction*. Monographs on Statistics and Applied Probability **55**. CRC Press, New York. [MR1252174](#) <https://doi.org/10.1007/978-1-4899-4467-2>
- [10] GEORGE, E. I., LIANG, F. and XU, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *Ann. Statist.* **34** 78–91. [MR2275235](#) <https://doi.org/10.1214/0090536060000000155>
- [11] GEORGE, E. I., LIANG, F. and XU, X. (2012). From minimax shrinkage estimation to minimax shrinkage prediction. *Statist. Sci.* **27** 82–94. [MR2953497](#) <https://doi.org/10.1214/11-STS383>
- [12] GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* 339–373.
- [13] GEORGE, E. I. and XU, X. (2008). Predictive density estimation for multiple regression. *Econometric Theory* **24** 528–544. [MR2391619](#) <https://doi.org/10.1017/S0266466608080213>

- [14] GHOSH, M., MERGEL, V. and DATTA, G. S. (2008). Estimation, prediction and the Stein phenomenon under divergence loss. *J. Multivariate Anal.* **99** 1941–1961. MR2466545 <https://doi.org/10.1016/j.jmva.2008.02.002>
- [15] HARTIGAN, J. A. (1998). The maximum likelihood prior. *Ann. Statist.* **26** 2083–2103. MR1700222 <https://doi.org/10.1214/aos/1024691462>
- [16] ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158 <https://doi.org/10.1214/009053604000001147>
- [17] ISHWARAN, H. and RAO, J. S. (2005). Spike and slab gene selection for multigroup microarray data. *J. Amer. Statist. Assoc.* **100** 764–780. MR2201009 <https://doi.org/10.1198/016214505000000051>
- [18] JOHNSTONE, I. M. (1994). On minimax estimation of a sparse normal mean vector. *Ann. Statist.* **22** 271–289. MR1272083 <https://doi.org/10.1214/aos/1176325368>
- [19] JOHNSTONE, I. M. (2013). Gaussian estimation: Sequence and wavelet models. Available at <https://imjohnstone.su.domains>.
- [20] JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135 <https://doi.org/10.1214/009053604000000030>
- [21] KOBAYASHI, K. and KOMAKI, F. (2008). Bayesian shrinkage prediction for the regression problem. *J. Multivariate Anal.* **99** 1888–1905. MR2466542 <https://doi.org/10.1016/j.jmva.2008.01.014>
- [22] KOMAKI, F. (1996). On asymptotic properties of predictive distributions. *Biometrika* **83** 299–313. MR1439785 <https://doi.org/10.1093/biomet/83.2.299>
- [23] KOMAKI, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika* **88** 859–864. MR1859415 <https://doi.org/10.1093/biomet/88.3.859>
- [24] KOMAKI, F. (2006). Shrinkage priors for Bayesian prediction. *Ann. Statist.* **34** 808–819. MR2283393 <https://doi.org/10.1214/009053606000000010>
- [25] KUBOKAWA, T., MARCHAND, É. and STRAWDERMAN, W. E. (2015). On predictive density estimation for location families under integrated squared error loss. *J. Multivariate Anal.* **142** 57–74. MR3412739 <https://doi.org/10.1016/j.jmva.2015.07.013>
- [26] KUBOKAWA, T., MARCHAND, É. and STRAWDERMAN, W. E. (2017). On predictive density estimation for location families under integrated absolute error loss. *Bernoulli* **23** 3197–3212. MR3654804 <https://doi.org/10.3150/16-BEJ842>
- [27] KUBOKAWA, T., MARCHAND, É., STRAWDERMAN, W. E. and TURCOTTE, J.-P. (2013). Minimaxity in predictive density estimation with parametric constraints. *J. Multivariate Anal.* **116** 382–397. MR3049911 <https://doi.org/10.1016/j.jmva.2013.01.001>
- [28] MALLOWS, C. (1978). Minimizing an integral. *SIAM Rev.* **20** 183–183. <https://doi.org/10.1137/1020016>
- [29] MARUYAMA, Y., MATSUDA, T. and OHNISHI, T. (2019). Harmonic Bayesian prediction under α -divergence. *IEEE Trans. Inf. Theory* **65** 5352–5366. MR4009238 <https://doi.org/10.1109/TIT.2019.2915245>
- [30] MATSUDA, T. and KOMAKI, F. (2015). Singular value shrinkage priors for Bayesian prediction. *Biometrika* **102** 843–854. MR3431557 <https://doi.org/10.1093/biomet/asv036>
- [31] MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. MR0997578
- [32] MUKHERJEE, G. (2013). *Sparsity and Shrinkage in Predictive Density Estimation*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Stanford University. MR4187552
- [33] MUKHERJEE, G. (2021). Minimax adaptive predictive density estimation for nonparametric regression. In preparation.
- [34] MUKHERJEE, G. and JOHNSTONE, I. M. (2015). Exact minimax estimation of the predictive density in sparse Gaussian models. *Ann. Statist.* **43** 937–961. MR3346693 <https://doi.org/10.1214/14-AOS1251>
- [35] MUKHERJEE, G. and JOHNSTONE, I. M. (2022). Supplement to “On Minimax Optimality of Sparse Bayes Predictive Density Estimates.” <https://doi.org/10.1214/21-AOS2086SUPP>
- [36] O’HARA, R. B. and SILLANPÄÄ, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Anal.* **4** 85–117. MR2486240 <https://doi.org/10.1214/09-BA403>
- [37] PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001 <https://doi.org/10.1198/016214508000000337>
- [38] ROČKOVÁ, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Ann. Statist.* **46** 401–437. MR3766957 <https://doi.org/10.1214/17-AOS1554>
- [39] ROČKOVÁ, V. and GEORGE, E. I. (2014). Negotiating multicollinearity with spike-and-slab priors. *Metron* **72** 217–229. MR3233150 <https://doi.org/10.1007/s40300-014-0047-y>
- [40] ROČKOVÁ, V. and GEORGE, E. I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.* **113** 431–444. MR3803476 <https://doi.org/10.1080/01621459.2016.1260469>

- [41] VAN ERVEN, T., SZABO, B. et al. (2021). Fast exact Bayesian inference for sparse signals in the normal sequence model. *Bayesian Anal.* **16** 933–960. [MR4303874](#) <https://doi.org/10.1214/20-BA1227>
- [42] XU, X. and LIANG, F. (2010). Asymptotic minimax risk of predictive density estimation for non-parametric regression. *Bernoulli* **16** 543–560. [MR2668914](#) <https://doi.org/10.3150/09-BEJ222>
- [43] XU, X. and ZHOU, D. (2011). Empirical Bayes predictive densities for high-dimensional normal models. *J. Multivariate Anal.* **102** 1417–1428. [MR2819959](#) <https://doi.org/10.1016/j.jmva.2011.05.008>
- [44] YANO, K. and KOMAKI, F. (2017). Information criteria for prediction when the distributions of current and future observations differ. *Statist. Sinica* **27** 1205–1223. [MR3699701](#)