The Annals of Statistics
2015, Vol. 43, No. 3, 937–961
DOI: 10.1214/14-AOS1251

EXACT MINIMAX ESTIMATION OF THE PREDICTIVE DENSITY IN SPARSE GAUSSIAN MODELS[1]

BY GOURAB MUKHERJEE AND IAIN M. JOHNSTONE

University of Southern California and Stanford University

We consider estimating the predictive density under Kullback–Leibler loss in an ℓ_0 sparse Gaussian sequence model. Explicit expressions of the first order minimax risk along with its exact constant, asymptotically least favorable priors and optimal predictive density estimates are derived. Compared to the sparse recovery results involving point estimation of the normal mean, new decision theoretic phenomena are seen. Suboptimal performance of the class of plug-in density estimates reflects the predictive nature of the problem and optimal strategies need diversification of the future risk. We find that minimax optimal strategies lie outside the Gaussian family but can be constructed with threshold predictive density estimates. Novel minimax techniques involving simultaneous calibration of the sparsity adjustment and the risk diversification mechanisms are used to design optimal predictive density estimates.

1. Introduction.

Statistical prediction analysis aims to use past data to choose a probability distribution that will be good in predicting the behavior of future samples. This well-established subject [Aitchison and Dunsmore (1975), Geisser (1993)] finds application in game theory, econometrics, information theory, machine learning, mathematical finance, etc.

In this paper we study predictive density estimation in a high-dimensional setting and, in particular, explore the consequences of sparsity assumptions on the unknown parameters.

1.1. *Main results.*

We begin by describing some of our main results: fuller references, background and interpretation follow in Section 1.2.

We work in the simplest Gaussian model for high-dimensional prediction:

$$X \sim N_n(\theta, v_x I), \qquad Y \sim N_n(\theta, v_y I), \qquad X \perp\!\!\!\perp Y|\theta. \tag{1}$$

On the basis of the "past" observation vector X, we seek to predict the distribution of a future observation Y. The past and future observations are independent, but are linked by the common mean parameter θ, assumed to be unknown. Note, however,

Received January 2014; revised June 2014.

[1] Supported in part by NSF Grant DMS-09-06812 and NIH Grant R01 EB001988.

MSC2010 subject classifications. Primary 62C20; secondary 62M20, 60G25, 91G70.

Key words and phrases. Predictive density, risk diversification, minimax, sparsity, high-dimensional, mutual information, plug-in risk, thresholding.

that the variances, assumed here to be known, may differ. We write $p(x|\theta, v_x)$ and $p(y|\theta, v_y)$ for the probability densities of X and Y, respectively.

We seek estimators $\hat{p}(y|x)$ of the future observation density $p(y|\theta, v_y)$, and to compare their performance under sparsity assumptions on θ. We recall two natural ways of generating large classes of estimators. Perhaps simplest are the "plug-in" or estimative densities: given a point estimate $\hat{\theta}(X)$, simply set $\hat{p}(y|x) = p(y|\hat{\theta})$. We often use the abbreviation $p[\hat{\theta}]$. Second, given any prior measure $\pi(d\theta)$, proper or improper, such that the posterior $\pi(d\theta|x)$ is well defined, the Bayes predictive density is

$$\hat{p}_\pi(y|x) = \int p(y|\theta, v_y)\pi(d\theta|x). \tag{2}$$

The important case of a uniform prior measure $\pi(d\theta) = d\theta$ leads to predictive density $\hat{p}_U(y|x)$, easily seen to correspond to $N_n(x, (v_x + v_y)I)$.

We will examine similarities and differences between high-dimensional prediction and high-dimensional estimation. In particular, $\hat{p}_U(y|x)$ plays in prediction the role of the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}(x) = x$ in the multinormal mean estimation setting. In contrast to the corresponding plug-in estimate $p[\hat{\theta}_{\text{MLE}}]$, the density $\hat{p}_U$ incorporates the variability of the location estimate which leads to a flattening of the estimator: $v_x + v_y > v_y$.

To evaluate the performance of a predictive density estimator $\hat{p}(y|x)$, we use the familiar 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 K–L risk function follows by averaging over the distribution of the past observation:

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

Given a prior measure $\pi(d\theta)$, the average or integrated risk is

$$B(\pi, \hat{p}) = \int \rho(\theta, \hat{p})\pi(d\theta). \tag{3}$$

The Bayes predictive density (2) can be shown to minimize both the posterior expected loss $\int L(\theta, \hat{p}(\cdot|x))\pi(d\theta|x)$ and the integrated risk $B(\pi, \hat{p})$ in the class of all density estimates. This is a general fact in statistical decision theory [Brown (1974)], the resulting minimum the Bayes K–L risk:

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

Our main focus is on how to optimize the predictive risk $\rho(\theta, \hat{p})$ in a high-dimensional setting under an ℓ_0-sparsity condition on the parameter space. Thus, let $\|\theta\|_0 = \#\{i : \theta_i \neq 0\}$ and

$$\Theta_n[s] = \{\theta \in \mathbb{R}^n : \|\theta\|_0 \leq s\}. \tag{5}$$

This "exact" sparsity condition has been widely used in estimation; in this paper we initiate study of its implications for predictive density estimation.

The minimax K–L risk for estimation over Θ is given by

$$R_N(\Theta) = \inf_{\hat{p}} \sup_{\theta \in \Theta} \rho(\theta, \hat{p}), \tag{6}$$

where the infimum is taken over *all* measurable predictive density estimators $\hat{p}(y|x)$. For comparison, we write $R_{\mathcal{E}}(\Theta) = \inf_{\hat{\theta}} \sup_{\Theta} \rho(\theta, p[\hat{\theta}])$ for the minimax risk restricted to the sub-class $\mathcal{E}$ of plug-in or "estimative" densities.

To state our main results, henceforth we will assume $v_x = 1$ and introduce the key parameters

$$r = v_y/v_x = v_y, \qquad v_w = (1 + r^{-1})^{-1}. \tag{7}$$

Here v_w is the "oracle variance" which would be the variance of the UMVUE for θ, were *both* X and Y observed.

In our asymptotic model, the dimensionality $n \to \infty$ and the sparsity $s = s_n$ may depend on n, but the variance ratio r remains fixed. The notation $a_n \sim b_n$ denotes $a_n/b_n \to 1$ as $n \to \infty$.

THEOREM 1A. *Fix* $r \in (0, \infty)$. *If* $\eta_n = s_n/n \to 0$, *then*

$$R_N(\Theta_n[s_n]) \sim \frac{1}{1+r} s_n \log(n/s_n) = \frac{1}{1+r} n\eta_n \log \eta_n^{-1}. \tag{8}$$

The minimax risk is proportional to the sparsity s_n, with a logarithmic penalty factor. The case where $s_n \equiv s$ remains constant is included. The expression is quite analogous to that obtained for point estimation with quadratic loss, namely, $2s_n \log(n/s_n)$ [Donoho and Johnstone (1994), Donoho et al. (1992) and Johnstone (2013), Chapter 8.8, hereafter cited as Johnstone (2013)]. However, we shall see that quite different phenomena emerge in the predictive density setting.

Indeed, the future-to-past variance ratio r is an important parameter of the predictive estimation problem. The minimax risk increases as r decreases: we need to estimate the future observation density based on increasingly noisy past observations (in relative terms, $r = v_y/v_x$), and so the difficulty of the density estimation problem increases. However, the *rate* of convergence with n in (8) does not depend on r, and so exact determination of the constants is needed to show the role of r in this prediction problem.

The inefficiency of plug-in estimators is an immediate consequence of Theorem 1A. Let $q(\theta, \hat{\theta}) = E\|\hat{\theta}(X) - \theta\|^2$ denote the risk of point estimator $\hat{\theta}$ under *squared-error* loss. It is straightforward to show for a plug-in density estimate $p[\hat{\theta}]$ that $\rho(\theta, p[\hat{\theta}]) = q(\theta, \hat{\theta})/(2r)$. Hence, from the point estimation minimax risk just cited,

$$R_{\mathcal{E}}(\Theta_n[s_n]) \sim \frac{1}{r} s_n \log(n/s_n) \sim \left(1 + \frac{1}{r}\right) R_N(\Theta_n[s_n]).$$

The inefficiency of plug-in estimators thus equals the oracle precision,

$$1/v_w = 1 + 1/r,$$

and becomes arbitrarily large as the variance ratio $r \to 0$.

We turn now to the asymptotically least favorable priors and optimal estimators in Theorem 1A. Let δ_λ denote unit point mass at λ and

$$\pi[\eta, \lambda] = (1 - \eta)\delta_0 + \eta\delta_\lambda \tag{9}$$

be a univariate two-point prior: this is a sparse prior when η is small and λ large. Let

$$\lambda_e = \sqrt{2 \log \eta_n^{-1}(1 - \eta_n)}, \qquad \lambda_f = \sqrt{v_w}\lambda_e. \tag{10}$$

In point estimation based on X, we recall that λ_e is essentially the threshold of detectability corresponding to sparsity $\eta_n = s_n/n$. Although Y is not yet observed, we will see that in the prediction setting the UMVUE scaled threshold $\lambda_f < \lambda_e$ plays a partly analogous role.

Build a sparse high-dimensional prior from i.i.d. draws:

$$\pi_n^{\mathrm{IID}}(d\theta) = \prod_{i=1}^{n} \pi[\eta_n, \lambda_f](d\theta_i). \tag{11}$$

If the sparsity s_n increases without bound with n, then this i.i.d. prior with scale λ_f is asymptotically least favorable:

THEOREM 1B. *If $s_n \to \infty$ and $s_n/n \to 0$, then*

$$B(\pi_n^{\mathrm{IID}}) = R_N(\Theta_n[s_n]) \cdot (1 + o(1)).$$

The assumption that $s_n \to \infty$ ensures that π_n^{IID} concentrates on $\Theta[s_n]$, namely, that $\pi_n^{\mathrm{IID}}(\Theta[s_n]) \to 1$ as $n \to \infty$. This hypothesis is not needed for Theorem 1A; indeed, a sparse prior built from "independent blocks" is asymptotically least favorable assuming only $s_n/n \to 0$. This more elaborate prior is described in Section 5.

Some of the novel aspects of the predictive density estimation problem appear in the description of optimal *estimators*, that is, ones that asymptotically attain the minimax bound in Theorem 1A. In point estimation, the simplest asymptotically minimax rule for sparsity s_n is given by co-ordinatewise hard thresholding $\hat{\theta}_i(x) = x_i I\{|x_i| \geq \lambda_e\}$. For prediction, we consider the following class of univariate density estimators as analogs of hard thresholding:

$$\hat{p}_T(y_1|x_1) = \begin{cases} \hat{p}_\pi(y_1|x_1), & \text{if } |x_1| \leq \lambda_e, \\ \hat{p}_U(y_1|x_1), & \text{if } |x_1| > \lambda_e. \end{cases} \tag{12}$$

The univariate density estimates are combined to form a multivariate predictive density estimate via a product rule

$$\hat{p}_T(y|x) = \prod_{i=1}^{n} \hat{p}_T(y_i|x_i). \tag{13}$$

The threshold λ_e in (12) is that corresponding to estimation based on X at sparsity $\eta_n = s_n/n$. Above the threshold, the uniform prior predictive density $\hat{p}_U$ corresponds to the (unbiased) MLE. Below threshold, we shall need the flexibility of the Bayes predictive density (2). Indeed, as explained in Section 4, it does not suffice to use $\pi = \delta_0$, point mass at 0, which would be the predictive analog of thresholding to zero in point estimation.

Instead, we use a sparse univariate cluster prior $\pi = \pi_{\mathrm{CL}}[\eta, r]$ given by

$$\pi = (1-\eta)\delta_0 + \frac{\eta}{2K} \sum_{k=1}^{K} (\delta_{\mu_k} + \delta_{-\mu_k}). \tag{14}$$

The points $\mu_k = \mu_k(r)$ for $k = 1, \dots, K$ are geometrically spaced to cover an interval $[\nu_\eta, \lambda_e + a]$ containing $[\lambda_f, \lambda_e]$, as described in more detail below. The key point is that it is necessary to "diversify" the predictive risk by introducing prior support points to cover $[-\lambda_e, -\lambda_f] \cup [\lambda_f, \lambda_e]$.

More specifically, for a parameter $a = a_\eta$ given below, let μ_η be the positive root of the overshoot equation

$$\mu^2 + 2a\mu = \lambda_e^2, \tag{15}$$

that occurs in sparse minimax point estimation [e.g., Johnstone (2013), equation (8.48)], and then set $\nu_\eta = \sqrt{v_w}\mu_\eta$: since $\mu_\eta < \lambda_e$, we have $\nu_\eta < \lambda_f$. The support points

$$\mu_1 = \nu_\eta, \qquad \mu_{k+1} = (1+2r)^k \nu_\eta, \qquad k \geq 1, \tag{16}$$

with $K = \max\{k : \mu_k \leq \lambda_e + a\}$. We choose $a_\eta = \sqrt{2 \log \lambda_f}$.

THEOREM 1C. *Assume $\eta_n = s_n/n \to 0$. Let $\hat{p}_{T,\mathrm{CL}}(y|x)$ be the product predictive threshold estimator defined by (12) and (13) using the cluster prior $\pi_{\mathrm{CL}}[\eta_n, r]$. Then $\hat{p}_{T,\mathrm{CL}}$ is asymptotically minimax*:

$$\max_{\Theta_n[s_n]} \rho(\theta, \hat{p}_{T,\mathrm{CL}}) = R_N(\Theta_n[s_n])(1 + o(1)).$$

Note that the number of positive support points in the cluster prior $K = K_\eta$ increases as r decreases. For any fixed η, the cluster prior contains in total $(2K_\eta + 1)$ support points. Also, for any fixed $r \in (0, \infty)$ as $\eta \to 0$, we have

$$K(r) = \lim_{\eta \to 0} K_\eta = \left\lfloor \frac{\log(1 + r^{-1})}{2\log(1+2r)} \right\rfloor.$$

TABLE 1
Number $K(r)$ of positive support points in the cluster prior $\pi_{\mathrm{CL}}[\eta, r]$ as r varies

r	**0.1073**	**0.1235**	**0.1465**	**0.1826**	**0.2485**	**0.4196**	**>0.4196**
$K(r)$	7	6	5	4	3	2	1

Thus, $K(r)$ is a piecewise constant, right continuous function with jumps as shown in Table 1.

The results presented above assume $v_x = 1$. These results can be easily extended to the general case by noting that the minimax risk remains invariant and the scale of past observations and parameter is divided by $\sqrt{v_x}$.

1.2. *Background and previous work.* The relative entropy predictive risk $\rho(\theta, \hat{p})$ measures the exponential rate of divergence of the joint likelihood ratio over a large number of independent trials [Larimore (1983)]. The minimal predictive risk estimate maximizes the expected growth rate in repeated investment scenarios [Cover and Thomas (1991), Chapters 6, 15]. In data compression, $L(\theta, \hat{p}(\cdot|x))$ reflects the excess average code length that we need if we use the conditional density estimate $\hat{p}$ instead of the true density to construct a uniquely decodable code for the data Y given the past x [McMillan (1956)]. Following Bell and Cover (1980), ℓ_0-constrained minimax optimal predictive density estimates in on our model can be used for construction of optimal predictive schemes for gambling, sports betting, portfolio selection and sparse coding [Mukherjee (2013), Chapter 1.3].

Aitchison (1975), Murray (1977) and Ng (1980) showed that in most parametric models there exist Bayes predictive density estimates which are decision theoretically better than the maximum likelihood plug-in estimate. An important issue in predictive inference has always been to compare the performance of the class $\mathcal{E}$ of point estimation (PE) based plug-in density estimates [Barndorff-Nielsen and Cox (1996)] with that of the optimal predictive density estimate. In parameter spaces of fixed dimension, large sample attributes of the predictive risk of efficient plug-in and Bayes density estimates have been studied by Komaki (1996), Hartigan (1998) and Aslan (2006).

The high-dimensional predictive density estimation problem studied in this paper is relevant to a number of contemporary applications, including data compression, sequential investment with side information and sports betting (SM).

Analogy with point estimation. Decision theoretic parallels between predictive density estimation under Kullback–Leibler loss and point estimation under quadratic loss have been explored in our Gaussian model by George, Liang and Xu (2006), Ghosh, Mergel and Datta (2008), Komaki (2004), Xu and Zhou (2011) and George, Liang and Xu (2012). For *unconstrained* parameter spaces $\Theta = \mathbb{R}^n$, fundamental ideas in Gaussian point estimation theory can be extended to yield optimal predictive density estimates [Brown, George and Xu (2008), Fourdrinier et al. (2011), Komaki (2001)]. For *ellipsoids*, Xu and Liang (2010) established an analog of the theorem of Pinsker (1980) by proving that the class of all linear predictive density estimates [see (17)] is minimax optimal.

For *sparse estimation*, instead of parallels, we found contrasts. Minimax risks in the predictive density problem depend on r, but this dependence is not emphasized in the admissibility results in unrestricted spaces. As we have seen, under sparsity construction of optimal minimax estimators requires the notion of *diversification* of the future risk over the interval $[\lambda_f, \lambda_e]$ in a way strongly dependent on r. Thus, efficiency of the prediction schemes depend on careful calibration of the sparsity adjustment and the risk diversification mechanisms.

1.3. *Further results.* *Other classes of estimators.* The class of *linear* estimates $\mathcal{L}$ are Bayes rules based on conjugate product normal priors. The resulting estimators

$$\hat{p}_{L,\alpha} = \prod_{i=1}^{n} N(\alpha_i X_i, \alpha_i + r), \qquad \alpha_i \in [0, 1], \tag{17}$$

are still Gaussian but have larger variance than the future density $p(y|\theta, r) = \phi(y|\theta, r)$. We choose the name "linear" because the conjugate prior implies linearity of the posterior mean in X.

The class $\mathcal{G}$ contains all product *Gaussian density estimates* $p[\hat{\theta}, \hat{d}] = \prod_{i=1}^{n} N(\hat{\theta}_i, \hat{d}_i)$. Clearly, $\mathcal{G}$ contains both $\mathcal{L}$ and $\mathcal{E}$, the latter introduced after (6). The minimax risks $R_{\mathcal{L}}(\Theta)$ and $R_{\mathcal{G}}(\Theta)$ are defined by restricting the infimum in (6) to $\mathcal{L}$ and $\mathcal{G}$, respectively.

We have seen after Theorem 1A that $R_{\mathcal{E}}(\Theta_n[s_n]) \sim (1 + r^{-1}) R_N(\Theta_n[s_n])$. It turns out that extending $\mathcal{E}$ to $\mathcal{G}$ does not help, while, as is typical for sparse estimation, the class of linear estimators $\mathcal{L}$ performs very poorly.

PROPOSITION 1. *Fix $r \in (0, \infty)$. If $s_n/n \to 0$, then*

$$\begin{aligned} R_{\mathcal{L}}(\Theta_n[s_n]) &= (n/2)\log(1 + r^{-1}), \\ R_{\mathcal{L}}(\Theta_n[s_n])/R_N(\Theta_n[s_n]) &\to \infty \quad \textit{and} \\ R_{\mathcal{G}}(\Theta_n[s_n]) &\sim R_{\mathcal{E}}(\Theta_n[s_n]). \end{aligned}$$

Univariate prediction problem. The product structure of our high-dimensional model (1), estimators (13) and priors (11), along with concentration of measure, implies that many aspects of our multivariate results can be understood and proved through an associated univariate prediction problem.

In the univariate setting, assume that the past observation $X|\theta \sim N(\theta, 1)$ and the future observation $Y|\theta \sim N(\theta, r)$. Assume that X and Y are independent given θ.

In addition, suppose that θ is random with distribution $\pi(d\theta)$, assumed to belong to

$$\mathfrak{m}(\eta) = \{\pi \in \mathcal{P}(\mathbb{R}) : \pi(\theta \neq 0) \leq \eta\}, \tag{18}$$

where $\mathcal{P}(\mathbb{R})$ is the collection of all probability measures in $\mathbb{R}$.

A predictive density estimator $\hat{p}(y|x)$ is evaluated through its integrated risk $B(\pi, \hat{p})$ defined at (3). The minimax risk for this univariate prediction problem is given by

$$\beta(\eta, r) := \inf_{\hat{p}} \sup_{\pi \in \mathfrak{m}(\eta)} B(\pi, \hat{p}), \tag{19}$$

and we study sparsity through the asymptotic regime $\eta \to 0$. Recall definition (10) of the scaled threshold $\lambda_f = \lambda_{f,\eta}$.

THEOREM 2. A. *Fix $r \in (0, \infty)$. As $\eta \to 0$,*

$$\beta(\eta, r) = \frac{1}{2r} \eta \lambda_f^2 \bigl(1 + o(1)\bigr). \tag{20}$$

B. *An asymptotically least favorable prior is given by the two-point distribution $\pi[\eta, \lambda_f(\eta)]$ of* (9).

C. *An asymptotically minimax estimator is given by the thresholding construction* (12) *combined with sparse univariate cluster prior $\pi = \pi_{\mathrm{CL}}[\eta, r]$ defined at* (14).

1.4. *Organization of the paper.* The main results of the paper are *multivariate*, Theorems 1A, 1B and 1C. However, the main technical issues in the proofs are best handled in the *univariate* setting of Theorem 2, whose parts A, B and C correspond to Theorems 1A, 1B and 1C, respectively. Section 2 has an overview: it first reviews some connections between the multivariate and univariate settings, then gives heuristic derivations for the lower and upper bounds of univariate Theorem 2. Section 3 and Section 4, respectively, contain the technical proofs for the lower and upper bound on the univariate minimax risk, Theorem 2B and 2C, respectively. Together, they complete the proof of Theorem 2. Proofs of the multivariate results in Theorems 1A, 1B and 1C are completed in Section 5. This section also contains a heuristic proof of Proposition 1 whose rigorous proof is presented in the supplementary material [Mukherjee and Johnstone (2015)].

Glossary. [The notation (6)+2 refers to text 2 lines after equation (6)].

Estimators: Bayes $\hat{p}_\pi$ (2), Uniform prior $\hat{p}_U$ (2)+2, Threshold $\hat{p}_T$ (12), Multivariate product $\hat{p}(y|x)$ (13); Univariate $\hat{p}(y_1|x_1)$.

Classes of estimators and multivariate minimax risks: all nonlinear N, R_N (6), estimative $\mathcal{E}$, $R_{\mathcal{E}}$ (6)+2, "linear" $\mathcal{L}$, $R_{\mathcal{L}}$ (17), Gaussian $\mathcal{G}$, $R_{\mathcal{G}}$ (17)+4.

Univariate minimax risk: β (19).

Parameter spaces: multivariate $\Theta_n[s]$ (5); univariate $\mathfrak{m}(\eta)$ (18).

Priors: Univariate: two point $\pi[\eta,\lambda]$ (9), cluster $\pi_{\mathrm{CL}}[\eta,r]$ (14), Multivariate: π_n^{IID} (11).

Parameters: variance ratio $r=v_y/v_x$, oracle variance v_w (7), sparsity η (9), thresholds λ_e,λ_f (10), cluster prior: overshoot a (15), ν_η (15)+2.

2. Proof overview and interpretation.

2.1. *Connections between multivariate and univariate settings.* Many aspects of the multivariate theorem may be understood, and in part proved, through a discussion of the univariate prediction problem of Theorem 2. An obvious connection between the univariate and multivariate approaches runs as follows: suppose that a multivariate predictive estimator is built as a product of univariate components

$$\hat{p}(y|x)=\prod_{i=1}^{n}\hat{p}_1(y_i|x_i). \tag{21}$$

Suppose also that to a vector $\theta=(\theta_i)$ we associate a univariate (discrete) distribution $\pi_n^e=n^{-1}\sum_{i=1}^n\delta_{\theta_i}$. Since the true multivariate future density $p(Y|\theta,r)$ is also a product of univariate components, it is then readily seen that the multivariate and univariate Bayes K–L risks are related by

$$\rho(\theta,\hat{p})=\sum_{i=1}^{n}\rho(\theta_i,\hat{p}_1)=nB(\pi_n^e,\hat{p}_1). \tag{22}$$

The sparsity condition $\Theta_n[s_n]$ in the multivariate problem corresponds to requiring that the prior $\pi=\pi_n^e$ in the univariate problem satisfies

$$\pi\{\theta_1\neq 0\}\le s_n/n=\eta_n,$$

and thus belongs to the class $\mathfrak{m}(\eta)$ defined in (18). Next, we outline the minimax risk calculations for the sparse predictive density estimation problem.

As a first illustration, to which we return later, consider the maximum risk of a product rule over $\Theta_n[s_n]$: using (22) and (3), we have

$$\sup_{\Theta_n[s_n]}\rho(\theta,\hat{p})=n\Big[(1-\eta_n)\rho(0,\hat{p}_1)+\eta_n\sup_{\theta\in\mathbb{R}}\rho(\theta,\hat{p}_1)\Big]. \tag{23}$$

In the univariate problem, using $\hat{p}_1$, we have the somewhat parallel bound

$$\sup_{\mathfrak{m}(\eta)}B(\pi,\hat{p})=(1-\eta)\rho(0,\hat{p}_1)+\eta\sup_{\theta\in\mathbb{R}}\rho(\theta,\hat{p}_1). \tag{24}$$

Consequently, a careful study of the two univariate quantities

$$\begin{array}{ll}\text{risk at zero:} & \rho(0,\hat{p}_1),\\ \text{maximum risk:} & \sup_\theta\rho(\theta,\hat{p}_1)\end{array} \tag{25}$$

is basic for upper bounds for both univariate and multivariate cases.

2.2. *Theorem* 2B: *Univariate lower bound heuristics.* To understand the apperance of λ_f^2 in the minimax risks, we turn to a heuristic discussion of the lower bound, first in the univariate case.

We use the two point priors (9) and the definition (19):

$$\beta(\eta, r) \geq B\big(\pi[\eta, \lambda]\big) = (1-\eta)\rho(0, \hat{p}_\pi) + \eta\rho(\lambda, \hat{p}_\pi) \geq \eta\rho(\lambda, \hat{p}_\pi), \tag{26}$$

and look for a good bound for $\rho(\lambda, \hat{p}_\pi)$ for a suitable choice of λ.

The key is a mixture representation for predictive risk of a Bayes estimator in terms of quadratic risk, where the weighted mixture is over noise levels $v \in [v_w, 1]$, with v_w being the oracle variance, (7). Brown, George and Xu (2008), Theorem 1, show that the predictive risk of the Bayes predictive density estimate $\hat{p}_\pi$ is

$$\rho(\theta, \hat{p}_\pi) = \frac{1}{2}\int_{v_w}^{1} q(\theta, \hat{\theta}_{\pi,v}; v)\frac{dv}{v^2}, \tag{27}$$

where $q(\theta, \hat{\theta}_{\pi,v}; v) = E_\theta[\hat{\theta}_{\pi,v}(W) - \theta]^2$ is the quadratic risk of the Bayes location estimate $\hat{\theta}_{\pi,v}$ for prior π when $W \sim N(\theta, v)$. In point estimation with quadratic loss, it is known [Johnstone (2013), Chapter 8], that as $\eta \to 0$ an approximately least favorable prior in the class $\mathfrak{m}(\eta)$ is given, for noise level $v = 1$, by the sparse two-point prior $\pi[\eta, \lambda_e(\eta)]$ defined in (9) and $\lambda_e(\eta) = \sqrt{2\log\eta^{-1}(1-\eta)}$. This prior has the remarkable property that points $\theta \leq \lambda_e$ are "invisible" in the sense that even when θ is true, the Bayes estimator $\hat{\theta}_\pi = \hat{\theta}_{\pi,1}$ effectively estimates 0 rather than θ and so makes a mean squared error

$$q(\theta, \hat{\theta}_\pi; 1) \sim \theta^2 \qquad \text{for } 0 \leq \theta \leq \lambda_e. \tag{28}$$

Two issues arise as the noise level v varies. First, the region of invisibility will scale, becoming $0 \leq \theta \leq \sqrt{v}\lambda_e$ at scale v. As v varies in $[v_w, 1]$, the intersection of all regions of invisibility will be $0 \leq \theta \leq \sqrt{v_w}\lambda_e = \lambda_f$ as defined at (10). The second issue is that for a given prior π and predictive Bayes rule $\hat{p}_\pi$ in (27), the Bayes rules $\hat{\theta}_{\pi,v}$ vary with v. We return to this second point in the next section; for now we can hope that for all $v \in [v_w, 1]$,

$$q(\lambda_f, \hat{\theta}_{\pi,v}; v) \gtrsim \lambda_f^2, \tag{29}$$

and so, from mixture representation (27),

$$\rho(\lambda_f, \hat{p}_\pi) \gtrsim \frac{\lambda_f^2}{2}\int_{v_w}^{1}\frac{dv}{v^2} = \frac{\lambda_f^2}{2r},$$

since the integral evaluates to $v_w^{-1} - 1 = r^{-1}$. From this we can conjecture that for $\pi = \pi[\eta, \lambda_f] \in \mathfrak{m}(\eta)$,

$$B(\pi) > \eta\rho(\lambda_f, \hat{p}_\pi) \gtrsim \eta\frac{\lambda_f^2}{2r}. \tag{30}$$

A full proof, with slightly modified definitions, is given in Section 3.

2.3. *Theorem* 2C: *Univariate upper bound heuristics*. We now turn to a heuristic discussion of constructing a density estimate to show that the lower bound (30) is asymptotically correct. Pursuing the analogy with point estimation, we know that in that setting optimal estimators can be found within the family of hard thresholding rules $\hat{\theta}(x) = xI\{|x| > \lambda\}$. The natural analog for predictive density estimation would have the form

$$\hat{p}_{T,\pi_0}[\lambda](y|x) = \begin{cases} \hat{p}_U(y|x), & |x| > \lambda, \\ \hat{p}_{\pi_0}(y|x), & |x| \leq \lambda. \end{cases} \tag{31}$$

To see this, note that $\hat{p}_U$ is the predictive Bayes rule corresponding to the uniform prior $\pi(d\theta) = d\theta$, which leads to the MLE $\hat{\theta}(x) = x$ in point estimation, while $\hat{p}_{\pi_0}(y|x)$ denotes the predictive Bayes rule corresponding to a prior concentrated entirely at 0, so that

$$\hat{p}_{\pi_0}(y|x) = \phi(y|0, r) \tag{32}$$

is a normal density with mean zero and variance r.

For the upper bound, according to definition (19), we seek an estimator $\hat{p}_1$ for which $\sup_{\mathfrak{m}(\eta)} B(\pi, \hat{p}_1) \sim \eta\lambda_f^2/(2r)$ as $\eta \to 0$. In bound (24), the first component is the risk at zero, $\rho(0, \hat{p}_1)$, and it turns out that this determines the possible values of the threshold λ in (31). Thus, in order that

$$\rho(0, \hat{p}_{T,\pi_0}[\lambda]) = o(\eta\lambda_f^2),$$

it follows [see (51)] that the threshold λ should be chosen as $\lambda = \lambda_e \sim (2\log\eta^{-1})^{1/2}$ and not smaller.

Turning to the second part of (25), we seek an estimator $\hat{p}_1$ with

$$\sup_\theta \rho(\theta, \hat{p}_1) = \frac{\lambda_f^2}{2r} \cdot (1 + o(1)). \tag{33}$$

We first argue that the hard thresholding analog (31) cannot work. Decompose the predictive risk of a univariate threshold estimator $\hat{p}_T$ with threshold λ_e into contributions due to X above and below the threshold

$$\begin{aligned} \rho(\theta, \hat{p}_T) &= E_\theta L(\theta, \hat{p}(\cdot|X)) \\ &= E_\theta[L(\theta, \hat{p}_U(\cdot|X)), |X| > \lambda_e] + E_\theta[L(\theta, \hat{p}_\pi(\cdot|X)), |X| \leq \lambda_e] \\ &= \rho_A(\theta) + \rho_B(\theta), \end{aligned} \tag{34}$$

say. With the "zero prior," the K–L loss is just quadratic in θ,

$$L(\theta, \hat{p}_{\pi_0}(Y|X)) = E_\theta \log\frac{\phi(Y|\theta, r)}{\phi(Y|0, r)} = \frac{\theta^2}{2r},$$

and so, in particular, for $\theta \leq \lambda_e$ we see that

$$\rho(\theta, \hat{p}_{T,\pi_0}) \geq \rho_B(\theta) \gtrsim \frac{\theta^2}{2r} P_\theta[|X| \leq \lambda_e] \tag{35}$$

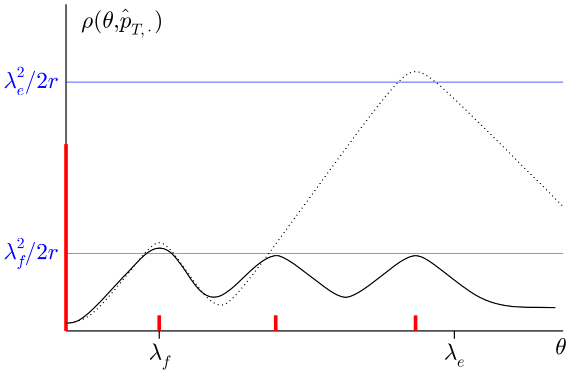

FIG. 1. *Schematic diagram of the risk of univariate threshold density estimates for $\theta \geq 0$. The dotted line is the risk of density estimator $\hat p_{T,\mathrm{LF}}$ based on the 3-point prior $\pi_3[\eta, \lambda_f]$. The addition of appropriately spaced prior mass points (shown in red) up to λ_e pulls down the risk function of the cluster prior-based density estimate $\hat p_{T,\mathrm{CL}}$ below $\lambda_f^2/(2r)$ until the effect of thresholding at λ_e takes over.*

could be as large as $\lambda_e^2/(2r)$, and hence larger than our target risk $\lambda_f^2/(2r)$.

Bearing in mind the role that two-point priors play in the lower bound, it is perhaps natural to ask next if the threshold rule $\hat p_{T,\mathrm{LF}}$ with π_0 in (31) replaced by the (symmetrized) two-point prior $\pi[\eta, \lambda_f]$ could cut off the growth of the quadratic $\theta^2/(2r)$ for $|\theta| \geq \lambda_f$. The 3-point prior $\pi_3[\eta, \lambda_f] \in \mathfrak{m}(\eta)$ places probability $\eta/2$ at the two nonzero atoms at $\pm\lambda_f$. Remarks in Section 3 show that $\pi_3[\eta, \lambda_f]$ is also asymptotically least favorable for the univariate prediction problem as $\eta \to 0$. Indeed, it can be shown (see Section 4) that for this prior and for $\lambda_f \leq |\theta| \leq \lambda_e$,

$$\begin{aligned}\rho(\theta, \hat p_{T,\mathrm{LF}}) &\sim \rho_B(\theta) \\ &\leq \frac{1}{2r}\{\lambda_f^2 - (|\theta| - \lambda_f)[(1+2r)\lambda_f - |\theta|]\} + o(\lambda_f^2).\end{aligned} \tag{36}$$

Consequently, the risk bound dips below $\lambda_f^2/(2r)$ for $\lambda_f \leq |\theta| \leq (1+2r)\lambda_f$ but increases thereafter. So, $\hat p_{T,\mathrm{LF}}$ is minimax optimal if $\lambda_e < (1+2r)\lambda_f$, which occurs if r is sufficiently large, $r > 0.4196$ in Table 1. However, the upper bound exceeds our target risk $\lambda_f^2/(2r)$ if $r \leq 0.4196$. Section S.2 of the supplementary material [Mukherjee and Johnstone (2015)] shows rigorously that $\hat p_{T,\mathrm{LF}}$ is indeed minimax suboptimal for low values of r.

As $\pi_3[\eta, \lambda_f]$ fails to produce minimax optimal density estimates, the strategy then is to introduce extra support points $|\mu_k| \leq \lambda_e$ into the prior chosen to "pull down" the risk $\rho_B(\theta) = E_\theta[L(\theta, \hat p_\pi(\cdot|X)), |X| \leq \lambda_e]$ below $\lambda_f^2/(2r)$ whenever it would otherwise exceed this level. The schematic diagram in Figure 1 illustrates this bounding of the maximum risk. The extra support points added in $[\lambda_f, \lambda_e]$ and $[-\lambda_e, -\lambda_f]$ distribute the predictive risk across that range—"risk diversification"—and keep the maximum risk below $\lambda_f^2/(2r)(1 + o(1))$.

To prove that this works, we obtain upper bounds on $\rho_B(\theta)$ for $\hat{p}_{T,\mathrm{CL}}$ by focusing, when $\theta \in [\mu_k, \mu_{k+1}]$, only on the prior support point μ_k. The main inequality is obtained in (50), namely,

$$\rho_B(\theta) \le \frac{1}{2r}\Big[\lambda_f^2 + \min_k q_k(\theta)\Big] + o(\lambda_f^2),$$

where $q_k(\theta)$ is a quadratic polynomial that is $O(\lambda_f)$ on $[\mu_k, \mu_{k+1}]$. Putting together this and other bounds, we can then finally establish the uniform bound (33). The details are in Section 4.

3. Theorem 2B: Univariate lower bound proof. This section is devoted to a proof of the lower bound part of Theorem 2. The heuristic discussion of the last section indicated the importance of two-point sparse priors and the invisibility property (28). To formulate a precise statement about the upper limit of invisibility, we start with noise level 1 and bring in the positive solution μ_η of the overshoot equation (15), namely, $\mu^2 + 2a\mu = \lambda_e^2$. Here the "overshoot" parameter $a = a_\eta$ should satisfy both $a_\eta \to \infty$ and $a_\eta = o(\mu_\eta)$; we make the specific choice $a_\eta = \sqrt{2\log\lambda_{f,\eta}}$.

In preparation for the range of variance scales in mixture representation (27), we consider the collection of two-point priors $\pi[\eta, \mu]$ for $0 \le \mu \le \mu_\eta$. Using a temporary notation for this section, let $\hat{\theta}_\mu(x) = E[\theta|x]$ be the Bayes rule for squared error loss for the prior $\pi[\eta, \mu]$. The next result shows that when the true parameter is actually μ, and this nonzero support point $\mu \le \mu_\eta$, then the Bayes rule for $\pi[\eta, \mu]$ "gets it wrong" by effectively estimating 0 and making an error of size μ^2, uniformly in $\mu \le \mu_\eta$.

LEMMA 3. *There exists $\varepsilon_\eta \searrow 0$ as $\eta \to 0$ such that for all μ in $[0, \mu_\eta]$,*

$$q(\mu, \hat{\theta}_\mu; 1) \ge \mu^2[1 - \varepsilon_\eta].$$

PROOF. Using standard calculations for the two-point prior, the Bayes rule $\hat{\theta}_\mu = \mu p(\mu|x) = \mu/[1 + m(x)]$, with

$$m(x) = \frac{p(0|x)}{p(\mu|x)} = \frac{1-\eta}{\eta}\frac{\phi(x)}{\phi(x-\mu)} = \exp\left\{\frac{1}{2}\lambda_e^2 - x\mu + \frac{1}{2}\mu^2\right\}. \tag{37}$$

Consequently,

$$\begin{aligned} q(\mu, \hat{\theta}_\mu; 1) &= E_\mu[\hat{\theta}_\mu - \mu]^2 = \mu^2 E_\mu[(1 + m(X))^{-1} - 1]^2 \\ &= \mu^2 E_0[1 + m^{-1}(\mu + Z)]^{-2}, \end{aligned}$$

where $Z \sim N(0, 1)$, and from (37), $m^{-1}(\mu + z) = \exp\{\frac{1}{2}(\mu^2 + 2\mu z - \lambda_e^2)\}$.

Now, using definition (15) of μ_η, for $0 \le \mu \le \mu_\eta$, we have

$$\mu^2 + 2\mu z - \lambda_e^2 \le \mu_\eta^2 + 2\mu_\eta z_+ - \lambda_e^2 = -2\mu_\eta(a - z_+),$$

so that for $0 \le \mu \le \mu_\eta$,

$$\mu^{-2} q(\mu, \hat{\theta}_\mu; 1) \ge E_0\{[1 + \exp(-\mu_\eta(a - Z_+))]^{-2}, Z < a\} = 1 - \varepsilon_\eta,$$

say. For each fixed z, we have $\mu_\eta(a - z_+) \to \infty$ since $a \to \infty$, and so from the dominated convergence theorem we conclude that $\varepsilon(\eta) \to 0$. □

With these preparations, we return to the lower bound in the prediction problem. As $\eta \to 0$, an asymptotically least favorable distribution is given by a sparse two-point prior with the nonzero support point scaled using the oracle standard deviation $v_w^{1/2}$. We shall prove the following:

LEMMA 4. *Let μ_η be the positive solution to overshoot equation* (15) *with $a_\eta = \sqrt{2 \log \lambda_{f,\eta}}$. Set $\nu_\eta = v_w^{1/2} \mu_\eta$ and consider the two-point prior $\pi[\eta, \nu_\eta]$. Then as $\eta \to 0$,*

$$\beta(\eta, r) \ge B(\pi[\eta, \nu_\eta]) \ge \frac{\eta \lambda_f^2}{2r}(1 + o(1)).$$

We note here that since $a_\eta = o(\mu_\eta)$, the overshoot equation implies that

$$\mu_\eta \sim \lambda_{e,\eta} \quad \text{and} \quad \nu_\eta \sim \lambda_{f,\eta}. \tag{38}$$

A stronger conclusion, used in the next section, also follows from the overshoot equation, namely,

$$\lambda_{f,\eta}^2 - \nu_\eta^2 = v_w(\lambda_{e,\eta}^2 - \mu_\eta^2) = v_w \cdot 2a\mu_\eta \le 2a v_w \lambda_{e,\eta} = 2a\sqrt{v}_w \lambda_{f,\eta}. \tag{39}$$

PROOF OF LEMMA 4. Recall (26) and (27) in the heuristic discussion. We now clarify the dependence on scale v of the Bayes rule $\hat{\theta}_{\pi,v}$ in the mixture representation (27). Passing from noise level v to noise level 1 by dividing parameters and estimates by $v^{1/2}$, we obtain the invariance relation

$$q(\theta, \hat{\theta}_{\pi[\eta,\lambda],v}; v) = v q(v^{-1/2}\theta, \hat{\theta}_{\pi[\eta, v^{-1/2}\lambda]}; 1).$$

Now set $\theta = \nu_\eta$ and substitute into (27) to obtain, for $\pi = \pi[\eta, \nu_\eta]$,

$$\rho(\nu_\eta, \hat{p}_\pi) = \frac{1}{2} \int_{v_w}^1 q(v^{-1/2}\nu_\eta, \hat{\theta}_{\pi[\eta, v^{-1/2}\nu_\eta]}; 1) \frac{dv}{v}. \tag{40}$$

Now apply Lemma 3 with $\mu = v^{-1/2}\nu_\eta$ being bounded above by $v_w^{-1/2}\nu_\eta = \mu_\eta$. For all $v \in [v_w, 1]$ we obtain

$$q(v^{-1/2}, \hat{\theta}_{v^{-1/2}\nu_\eta}; 1) \ge v^{-1}\nu_\eta^2[1 - \varepsilon_\eta].$$

Putting this into the mixture representation, we get

$$\rho(\nu_\eta, \hat{p}_\pi) \ge \frac{1}{2}\nu_\eta^2[1 - \varepsilon_\eta] \int_{v_w}^1 \frac{dv}{v^2} = \frac{\nu_\eta^2}{2r}[1 - \varepsilon_\eta].$$

Taking into account both (26) and (38), we have established the lemma. □

Based on the discussion in Section 2, the above lemma establishes a lower bound on the asymptotic minimax risk $\beta(\eta, r)$ in Theorem 2. Similarly, the symmetric 3-point prior

$$\pi_3[\eta, \nu_\eta] = (1-\eta)\delta_0 + (\eta/2)\{\delta_{\nu_\eta} + \delta_{-\nu_\eta}\}$$

will also be asymptotically least favorable over $\mathfrak{m}(\eta)$ as $\eta \to 0$.

4. Theorem 2C: Univariate upper bound proof. The upper bound on the predictive minimax risk $\beta(\eta, r)$ is derived from the upper bound on the maximum Bayes risk of $\hat{p}_{T,\mathrm{CL}}$ over $\mathfrak{m}(\eta)$. In this section we will prove the following lemma which along with Lemma 4 completes the proof of Theorem 2.

LEMMA 5. *For any $r \in (0, \infty)$ we have, as $\eta \to 0$,*

$$\sup_{\pi \in \mathfrak{m}(\eta)} B(\pi, \hat{p}_{T,\mathrm{CL}}) \le \frac{\eta \lambda_f^2}{2r}(1 + o(1)).$$

We consider a threshold predictive density estimate $\hat{p}_T$ which uses the Bayes predictive density estimate from prior π below the threshold λ_e and $\hat{p}_U$ above the threshold λ_e. We bound the maximum predictive risk over $\mathfrak{m}(\eta)$:

$$\sup_{\pi \in \mathfrak{m}(\eta)} B(\pi, \hat{p}_T) \le (1-\eta)\rho(0, \hat{p}_T) + \eta \sup_{\theta} \rho(\theta, \hat{p}_T). \tag{41}$$

Next, as in (34), we decompose the predictive risk of $\hat{p}_T$ into contributions due to X above and below the threshold. We calculate explicit expressions for ρ_A and ρ_B. The predictive loss of $\hat{p}_U$ (see Appendix A.2) is given by

$$L\big(\theta, \hat{p}_U(\cdot|x)\big) = a_{1r} + a_{2r}(\theta - x)^2 \tag{42}$$

with $a_{1r} = \frac{1}{2}[\log(1 + r^{-1}) - (1+r)^{-1}]$ and $a_{2r} = \frac{1}{2}(1+r)^{-1}$. Hence, the above threshold term

$$\rho_A(\theta) = a_{1r} P_\theta\big(|X| > \lambda_e\big) + a_{2r} E_\theta\big[(X-\theta)^2, |X| > \lambda_e\big]. \tag{43}$$

As $\rho_B(\theta)$ depends on the prior π used below the threshold, we restrict our attention to the specific choice of the cluster prior. The risk functions of the hard threshold density estimate $\hat{p}_{T,\pi_0}$ and that of $\hat{p}_{T,\mathrm{LF}}$ can be easily derived from the calculations with the cluster prior.

According to (58) in the Appendix, the Bayes predictive density for a discrete prior $\pi = \sum_{k=-K}^{K} \pi_k \delta_{\mu_k}$ is given by

$$\hat{p}_\pi(y|x) = \sum_{-K}^{K} \phi(y|\mu_k, r)\pi_k \phi(x - \mu_k)/m(x), \tag{44}$$

where $m(x) = \sum_k \pi_k \phi(x - \mu_k)$ denotes the marginal density of π. The K–L loss of $\hat{p}_\pi(\cdot|x)$ is given by

$$L(\theta, \hat{p}_\pi(\cdot|x)) = \mathrm{E}_\theta \log \frac{\phi(Y|\theta, r)}{\hat{p}_\pi(Y|x)}.$$

A simple but informative upper bound for the K–L loss is obtained by retaining only the kth term in (44):

$$\begin{aligned} L(\theta, \hat{p}_\pi(\cdot|x)) &\le \mathrm{E}_\theta \log \frac{\phi(Y|\theta, r)}{\phi(Y|\mu_k, r)} - \log \frac{\pi_k \phi(x - \mu_k)}{\pi_0 \phi(x)} + \log \frac{m(x)}{\pi_0 \phi(x)} \\ &= \frac{1}{2r}(\theta - \mu_k)^2 + \frac{1}{2}(\mu_k^2 - 2x\mu_k) - \log \frac{\pi_k}{\pi_0} + d(x), \end{aligned} \tag{45}$$

where we have set $d(x) = \log[m(x)/(\pi_0 \phi(x))]$.

We are now ready to analyze the bound (41). We follow the steps recalled in the quadratic loss case [see Section S.4 of Mukherjee and Johnstone (2015)] and evaluate the predictive risk at the origin and the maximum risk of the threshold density estimate $\hat{p}_T$. This organization helps to make clear the new features of the predictive loss setting.

Risk at zero. It is easy to show that $\rho(0, \hat{p}_T) = O(\eta\lambda_f)$. First, from (43), we have

$$\rho_A(0) = 2a_{1r}\tilde{\Phi}(\lambda_e) + a_{2r}q_A(0) = O(\eta\lambda_f),$$

where $q_A(0)$ is defined in (S.4.2) and the above calculation follows by using $\tilde{\Phi}(\lambda_e) \le \lambda_e^{-1}\phi(\lambda_e) = O(\lambda_e^{-1}\eta)$ and the quadratic risk-at-zero bound (S.4.4).

For the below-threshold term, we set $k = 0$ in (45), note that $\mu_0 = 0$ and apply Jensen's inequality to obtain

$$\rho_B(0) = E_0[L(0, \hat{p}_\pi(\cdot|X)), |X| \le \lambda] \le E_0[d(X)] \le \log E_0[m(X)/(\pi_0\phi(X))].$$

Since $E_0[m(X)/\phi(X)] = \int m(x)\,dx = 1$ and $\pi_0 = 1 - \eta$, we obtain that

$$\rho_B(0) \le -\log(1 - \eta) \le \eta.$$

Consequently, $\rho_B(0) = O(\eta)$ and so $\rho(0, \hat{p}_{T,\mathrm{CL}}) = O(\eta\lambda_f)$. Note that the above calculations hold for any $\hat{p}_{T,\pi}$ with π being a discrete prior in $\mathfrak{m}(\eta)$.

Maximum risk. From decomposition (41), our goal is to show that

$$\sup_\theta \rho(\theta, \hat{p}_{T,\mathrm{CL}}) = (2r)^{-1}\lambda_f^2(1 + o(1)). \tag{46}$$

We first isolate the main term in the contributions from $\rho_A(\theta)$ and $\rho_B(\theta)$. From (43), clearly $\rho_A(\theta) \le a_{1r} + a_{2r} = O(1)$, which does not contribute. We turn to

$$\rho_B(\theta) = E_\theta[L(\theta, \hat{p}_\pi(\cdot|X)), |X| \le \lambda_e]$$

and returning to (45), we begin by claiming that for $|x| \le \lambda_e$ the final term $d(x) \le \log 2$. Indeed,

$$\frac{m(x)}{\pi_0 \phi(x)} = 1 + \sum_{|k|=1}^{K} \frac{\pi_k}{\pi_0} \frac{\phi(x-\mu_k)}{\phi(x)} = 1 + \sum_{|k|=1}^{K} \frac{\pi_k}{\pi_0} \exp\left\{x\mu_k - \frac{\mu_k^2}{2}\right\}. \tag{47}$$

For $|x| \le \lambda_e$, we have

$$x\mu_k - \mu_k^2/2 \le \lambda_e|\mu_k| - \mu_k^2/2 \le \lambda_e^2/2 = \log \eta^{-1}(1-\eta).$$

Since $\pi_0 = 1 - \eta$, we arrive at

$$E_\theta\left[d(X), |X| \le \lambda_e\right] \le \log 2. \tag{48}$$

The dependence of (45) on θ may then be seen by writing $x = \theta + z$. The first two terms in (45) then take the form

$$\frac{1}{2r}\{[\theta - (1+r)\mu_k]^2 - (r^2 + r)\mu_k^2\} - \mu_k z,$$

while, after recalling that $\pi_k = \eta/(2K)$ and that $\lambda_e^2 = 2\log(1-\eta)\eta^{-1}$, the third term becomes

$$\frac{1}{2}\lambda_e^2 + \log(2K) = \frac{1}{2r}(1+r)\lambda_f^2 + \log(2K).$$

We may therefore rewrite (45) as

$$L(\theta, \hat{p}_\pi(\cdot|x)) \le \frac{1}{2r}[\lambda_f^2 + q_k(\theta)] - \mu_k(x-\theta) + \log(2K) + d(x), \tag{49}$$

where the kth quadratic polynomial

$$q_k(\theta) = [\theta - (1+r)\mu_k]^2 - r^2\mu_k^2 + r(\lambda_f^2 - \mu_k^2).$$

Denote the last three terms of (49) by $J_k(x,\theta)$. From (16) and (48) we see that

$$E_\theta[J_k, |X| \le \lambda_e] \le \mu_k + \log(2K) + \log 2 \le \lambda_e + a + \log(4K) = o(\lambda_f^2).$$

Consequently, we obtain the key bound

$$\rho_B(\theta) \le \frac{1}{2r}\left[\lambda_f^2 + \min_k q_k(\theta)\right] + o(\lambda_f^2). \tag{50}$$

Now we use the geometric structure of the support points μ_k, defined at (16). We bound $\min_k q_k(\theta)$ above by considering the quadratic polynomial $q_k(\theta)$ on $I_k = [\mu_k, \mu_{k+1}]$ and observe that these $2K$ intervals cover the range $(-\lambda_e - a, -\lambda_f) \cup (\lambda_f, \lambda_e + a)$ of interest. See Figure 2. Note that $q_k(\theta)$ achieves its maximum on I_k at both endpoints and that

$$q_k(\mu_{k+1}) = q_k((1+2r)\mu_k) = q_k(\mu_k) = r(\lambda_f^2 - \mu_k^2).$$

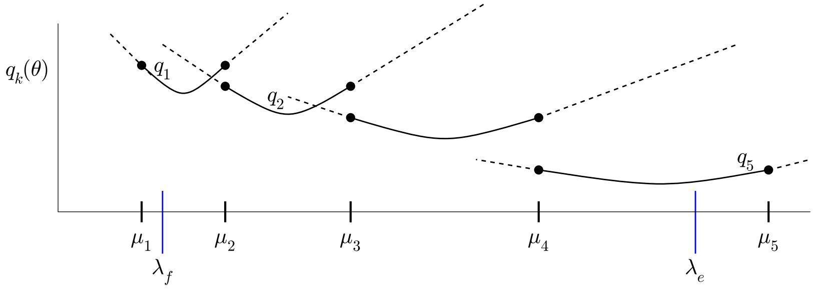

FIG. 2. *Schematic diagram demonstrating the behavior of the quadratic polynomials $q_k(\theta)$ in the interval $[\mu_1, \mu_{K+1}]$. Here $K = 4$. The maximum of $\min_k q_k(\theta)$ for $\theta \in [\mu_1, \mu_{K+1}]$ is bounded by $q_1(\mu_1)$.*

These maxima decrease with k and so are bounded by $q_1(\nu_\eta) = r(\lambda_f^2 - \nu_\eta^2)$. Appealing now to bound (39), we have for $\lambda_f \le |\theta| \le \lambda_e + a$,

$$\min_k q_k(\theta) \le r(\lambda_f^2 - \nu_\eta^2) \le 2r\sqrt{v_w}a\lambda_f.$$

Returning to (50), we now see that the last two terms are each $o(\lambda_f^2)$ and so the final bound (46) is proven. This completes the proof of Lemma 5.

These calculations apply to threshold density estimates based on Bayes estimates of discrete priors. In particular, for $\hat{p}_{T,\mathrm{LF}}$ which is based on the 3-point prior $\pi_3[\eta, \nu_\eta]$, we have $K = 1$ and the bound (36). Thus, the difference $\rho_B(\theta) - \lambda_f^2/2r$ in this case is negligible when $|\theta| \le \mu_2$.

Similarly, the asymptotic risk function of the hard threshold plug-in density estimate $\hat{p}_{T,\pi_0}$ (for which $K = 0$ in our calculations above) exceeds the minimax risk $\beta(\eta, r)$ for $|\theta| \in [\lambda_f, \lambda_e]$ and so is minimax suboptimal for any fixed r. Figure 3 shows the numerical evaluation of the risk functions for the different univariate threshold density estimates.

Also, note that any threshold estimate $\hat{p}_T[\lambda]$ with threshold size λ less than λ_e will be minimax suboptimal, as its risk at the origin will not be negligible as compared to $\beta(\eta, r)$. By (34) and (43) we have

$$\begin{aligned}\rho(0, \hat{p}_T[\lambda]) &\ge 2a_{2r}E[Z^2 I\{Z > \lambda\}] = 2a_{2r}\{\lambda\phi(\lambda) + 2\tilde{\Phi}(\lambda)\} \\ &\ge \lambda\phi(\lambda)/(1 + r),\end{aligned} \tag{51}$$

and so for any fixed $\varepsilon > 0$,

$$\liminf_{1 \le \lambda < \lambda_e(\eta) - \varepsilon} \frac{\rho(0, \hat{p}_T[\lambda])}{\beta(\eta, r)} \to \infty.$$

Thus, $\hat{p}_T[\lambda]$ is suboptimal unless $\lambda \ge \lambda_e$.

5. Theorem 1: Multivariate minimax risk.

Here we will use the univariate minimax results developed in the previous sections to evaluate the asymptotic multivariate minimax risk $R_n = R_N(\Theta_n[s_n])$ over the sparse parameter space $\Theta_n[s_n]$.

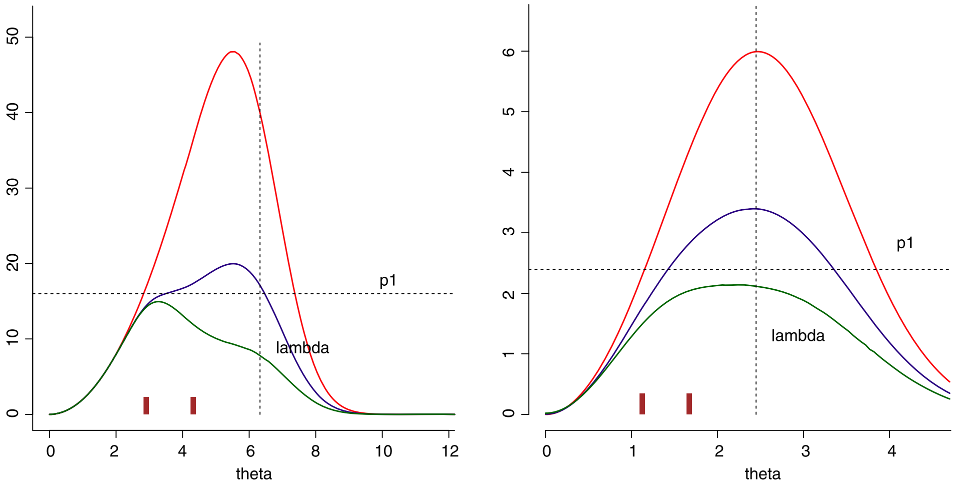

FIG. 3. *Numerical evaluation of the asymptotic risk $\rho_B(\theta)$ for $r = 0.25$ of univariate threshold density estimates: hard threshold plug-in estimate $\hat{p}_{T,\pi_0}$ (red), $\hat{p}_{T,\mathrm{LF}}$ (green) and the cluster prior-based minimax optimal estimate $\hat{p}_{T,\mathrm{CL}}$ (blue). The brown boxes show the nonzero support point of the cluster prior and the univarate asymptotic minimax risk $\beta(\eta, r) = (2r)^{-1}\lambda_f^2$ and the threshold λ_e are respectively denoted by dotted horizontal and vertical lines. The plot on left has $\eta = e^{-20}$ (very high sparsity), $\lambda_f = 2.83$, $\lambda_e = 6.32$ and the right one has $\eta = 0.05$ (moderate sparsity), $\lambda_f = 1.09$, $\lambda_e = 2.45$.*

5.1. *Lower bound proof: Theorem* 1B *and an extension.* We first prove a lower bound for the multivariate minimax risk under only the assumption that $s_n/n \to 0$—without requiring, as in Theorem 1B, that also $s_n \to \infty$. This is done using an "independent blocks" sparse prior, along the lines of Johnstone (2013), Chapter 8.6, that we will show to be asymptotically least favorable. This result establishes the lower bound half of Theorem 1A. At the end of the subsection, we prove Theorem 1B using the simpler i.i.d. prior.

Let $\pi_S(\tau; m)$ denote a single spike prior of scale τ on $\mathbb{R}^m$: choose an index $I \in \{1, \dots, m\}$ at random and set $\theta = \tau e_I$, where e_I is a unit length vector in the ith coordinate direction. We will use a scale $\tau_m = \lambda_m - \log \lambda_m$ which is somewhat smaller than $\lambda_m = \sqrt{2 \log m}$.

The independent blocks prior π^{IB} on $\Theta[s_n]$ is built by dividing $\{1, \dots, n\}$ into s_n contiguous blocks B_j, $j = 1, \dots, m$ each of length $m = m_n = [n/s_n]$. Draw components θ_i in each block B_j according to an independent copy of $\pi_S(\nu_m; m)$ where the scale $\nu_m = \sqrt{v_w}\tau_m$ is matched to the prediction setting. Finally, set $\theta_i = 0$ for the remaining $n - m_n s_n$ components. Thus, π^{IB} is supported on $\Theta[s_n]$ since any draw θ from π^{IB} has exactly s_n nonzero components.

The lower bound half of Theorem 1A follows from the following result, the analog of Theorem 1B for the independent blocks prior.

THEOREM 6. *Fix $r \in (0, \infty)$. If $s_n/n \to 0$, then*

$$R_N\bigl(\Theta_n[s_n]\bigr) \geq B\bigl(\pi_n^{\mathrm{IB}}\bigr) \geq (1+r)^{-1} s_n \log(n/s_n).$$

PROOF. Bounding maximum risk by Bayes risk and using the product structure shows that

$$R_n = R_N\bigl(\Theta_n[s_n]\bigr) \geq B\bigl(\pi_n^{\mathrm{IB}}\bigr) = s_n B\bigl(\pi_S(\nu_m; m)\bigr). \tag{52}$$

Next, using B_Q^v to denote the Bayes risk for noise level v, the multivariate form of the connecting equation and scale invariance enable us to write

$$B\bigl(\pi_S(\nu_m; m)\bigr) = \frac{1}{2}\int_{v_w}^{1} B_Q^v\bigl(\pi_S(\nu_m; m)\bigr)\frac{dv}{v^2} = \frac{1}{2}\int_{v_w}^{1} B_Q\biggl(\pi_S\biggl(\frac{\nu_m}{\sqrt{v}}; m\biggr)\biggr)\frac{dv}{v}.$$

The next lemma, proved in Section S.5 of Mukherjee and Johnstone (2015), provides a uniform lower bound for the quadratic loss Bayes risk of a single spike prior. It is a multivariate analog of Lemma 3.

PROPOSITION 7. *Suppose that $y \sim N_n(0, I)$. Set $\lambda_n = \sqrt{2\log n}$ and $\tau_n = \lambda_n - \log\lambda_n$. Then there exists $\varepsilon_n \to 0$ such that uniformly in $\tau \in [0, \tau_n]$,*

$$B_q\bigl(\pi_S(\tau; n)\bigr) \geq \tau^2(1 - \varepsilon_n).$$

Noting that $v \in [v_w, 1]$ implies that $\nu_m/\sqrt{v} \leq \nu_m/\sqrt{v_w} = \tau_m$, and then applying the proposition,

$$B\bigl(\pi_S(\nu_m; m)\bigr) \geq \frac{(1-\varepsilon_m)}{2}\int_{v_w}^{1} \frac{\nu_m^2}{v^2}\, dv = (1-\varepsilon_m)\frac{\nu_m^2}{2r}.$$

Combining this with (52) and the definition of ν_m, we obtain

$$R_n \geq (1-\varepsilon_m) s_n v_w \tau_m^2/(2r) \sim (1+r)^{-1} s_n \log(n/s_n). \tag{53}$$ □

PROOF OF THEOREM 1B. Note that because of the product structure of the problem and the prior π_n^{IID} we have

$$B\bigl(\pi_n^{\mathrm{IID}}\bigr) = \sum_{i=1}^{n} \beta(\eta_n, r) = n\beta(\eta_n, r),$$

which is asymptotically equal to $R_N(\Theta[s_n])$, using the univariate Theorem 2 [cf. (20)] and

$$(2r)^{-1}\lambda_f^2 = (2r)^{-1} v_w \lambda_e^2 \sim (1+r)^{-1}\log\eta_n^{-1} \qquad \text{as } n \to \infty. \tag{54}$$

Also, as $s_n \to \infty$, $\pi_n^{\mathrm{IID}}(\Theta[s_n]) \to 1$ by application of Chebyshev's inequality and, hence, π_n^{IID} is an asymptotically least favorable prior under the conditions of Theorem 1B. □

5.2. *Upper bound proof*: *Theorem* 1C. First, an upper bound on $R_N(\Theta_n[s_n])$ is derived based on the maximum risk of the multivariate product threshold density estimate $\hat{p}_{T,\mathrm{CL}}$ defined in Theorem 1C. Using the product structure of the threshold estimate as well as that of the unknown future density

$$\hat{p}_{T,\mathrm{CL}}(y|x) = \prod_{i=1}^{n} \hat{p}_{T,\mathrm{CL}}(y_i|x_i) \quad \text{and} \quad p(y|\theta, r) = \prod_{i=1}^{n} p(y_i|\theta_i, r),$$

the risk of our multivariate threshold estimate simplifies as an agglomerative coordinate wise risk of the respective univariate density estimates

$$\rho(\theta, \hat{p}_{T,\mathrm{CL}}) = E_\theta \log \frac{p(y|\theta, r)}{\hat{p}_{T,\mathrm{CL}}(y|x)} = \sum_{i=1}^{n} \rho(\theta_i, \hat{p}_T).$$

Now, maximizing over $\theta \in \Theta_n[s_n]$, we have

$$R_n \le \sup_{\Theta_n[s_n]} \rho(\theta, \hat{p}_{T,\mathrm{CL}}) \le (n - s_n)\rho(0, \hat{p}_{T,\mathrm{CL}}) + s_n \sup_\theta \rho(\theta, \hat{p}_{T,\mathrm{CL}}).$$

From the univariate study, we know that $\rho(0, \hat{p}_{T,\mathrm{CL}}) = O(\eta_n \lambda_f)$, which makes $(n - s_n)\rho(0, \hat{p}_{T,\mathrm{CL}}) = O(s_n \lambda_f)$ negligible relative to

$$s_n \sup_\theta \rho(\theta, \hat{p}_{T,\mathrm{CL}}) = (2r)^{-1} s_n \lambda_f^2 (1 + o(1)),$$

where we used (46). Thus, taking account also of (54), we have the desired upper bound on the minimax risk

$$R_n \le (2r)^{-1} s_n \lambda_f^2 (1 + o(1)) \sim (1 + r)^{-1} s_n \log(n/s_n). \tag{55}$$

Completion of Proof of Theorems 1A, 1B *and* 1C: As the lower bound (53) and upper bound (55) on R_n match asymptotically, the first order asymptotic minimax risk of Theorem 1A is achieved, and the proof of all parts is done.

5.3. *Proof of Proposition* 1. Estimates in $\mathcal{L}$ and $\mathcal{G}$ are products of the form (21) and so $R_{\mathcal{L},n} = R_{\mathcal{L}}(\Theta_n[s_n])$ can be studied using the associated univariate problem and decomposition (23). It is shown in Appendix A.2 that

$$\rho(\theta, \hat{p}_{L,\alpha}) = \frac{1}{2}\log\left(1 + \frac{\alpha}{r}\right) + \frac{(1-\alpha)^2}{2(r+\alpha)}\left[\theta^2 - \frac{\alpha}{1-\alpha}\right]. \tag{56}$$

Thus, $\sup_\theta \rho(\theta, \hat{p}_{L,\alpha})$ is infinite unless $\alpha = 1$, that is, the uniform prior estimate $\hat{p}_U$, in which case $\rho(\theta, \hat{p}_U) \equiv \frac{1}{2}\log(1 + r^{-1})$. Thus,

$$R_{\mathcal{L},n} = \frac{n}{2}\log(1 + r^{-1}) \gg \frac{s_n}{1+r}\log\left(\frac{n}{s_n}\right) \sim R_n.$$

In particular, $R_{\mathcal{L},n}/R_n \to \infty$ when $s_n/n \to 0$.

We turn to the Gaussian class $\mathcal{G}$. Since $\mathcal{E} \subset \mathcal{G}$, clearly $R_{\mathcal{G},n} \le R_{\mathcal{E},n} = (2r)^{-1} n \eta_n \lambda_e^2$. We give here a heuristic argument for the reverse inequality, which gives the idea for the rigorous proof given in Section S.3 of the supplementary material [Mukherjee and Johnstone (2015)]. From the decomposition (23), any near-optimal estimator in $\mathcal{G}$ must have univariate risk at 0 bounded as follows:

$$\rho(0, \hat{p}_1) \le r^{-1}\eta_n\lambda_e^2. \tag{57}$$

Now from (60) we know that the risk at the origin for the univariate Gaussian density estimate $p[\hat{\theta}, \hat{d}]$ is

$$\rho(0, p[\hat{\theta}, \hat{d}]) = 2^{-1}E_0\{\log(r^{-1}\hat{d}) + \hat{d}^{-1}(r + \hat{\theta}^2) - 1\},$$

which for any fixed choice of $\hat{\theta}$ achieves its minimum at $d_{\mathrm{opt}}[\hat{\theta}] = r + \hat{\theta}^2$. Thus, for such an optimal choice of $\hat{d}$,

$$\rho(0, p[\hat{\theta}, d_{\mathrm{opt}}(\hat{\theta})]) = E_0 \log(1 + r^{-1}\hat{\theta}^2),$$

and for this to satisfy (57), we must have $\hat{\theta}(x) \approx 0$ for $|x| \le \lambda_e(1 + o(1))$. Thus, $\hat{p}$ would approximately need to have the threshold structure (31), (32) for $|x| \le \lambda_e$ and so the bound (35) shows that

$$\rho(\theta, \hat{p}_1) \ge \frac{\theta^2}{2r} P_\theta(|X| \le \lambda_e) \sim \frac{\lambda_e^2}{2r}.$$

Returning to decomposition (23), we can now see that $R_{\mathcal{G},n} \gtrsim (2r)^{-1}s_n\lambda_e^2 \sim R_{\mathcal{E},n}$, which completes the heuristic argument.

6. Discussion.

Avoiding thresholding. The asymptotic minimax rules $\hat{p}_T$ described in Theorems 1C and 2C are based on thresholding. It would be desirable to construct a prior π for which the Bayes predictive density $\hat{p}_\pi$ in (2) is itself asymptotically minimax, without any use of the discontinuous thresholding operation.

Consider, then, a symmetric univariate prior $\pi_\infty[\eta, r]$ whose support consists of the origin and infinite number of equidistant clusters each containing $2K$ points in the same spatial alignment as for $\pi_{\mathrm{CL}}[\eta, r]$:

$$\pi_\infty[\eta, r] = (1 - \eta)\delta_0 + \frac{1 - \eta}{2}\sum_{j=0}^{\infty} \eta^{j+1} \sum_{k=1}^{K} q_k(\delta_{\mu_{jk}} + \delta_{-\mu_{jk}}),$$

where $\mu_{jk} = j\lambda_e + \mu_k$ and for $k = 2, \dots, K$ and $\gamma = \log \eta^{-1}$, we have $q_k = \gamma^{-k}$ and $q_1 = 1 - \sum_2^K q_k$.

Based on $\pi_\infty[\eta_n, r]$, one can construct a multivariate prior $\pi_{n,\infty}^{\mathrm{IID}}$ using (11), which heuristic arguments indicate will not only be least favorable but also yield a minimax optimal density estimate. A detailed proof is forthcoming.

Approximate sparsity and other extensions. Starting from Johnstone (2013), Chapters 8 and 13, the ℓ_0 sparsity results presented here can be extended to obtain minimax optimal predictive density estimates over weak and strong ℓ_p sparse parameter spaces. An interesting topic for future work will be whether, as in point estimation [Donoho and Johnstone (1994)], the phenomena seen here can be generalized to a family of loss functions. Simple analogues of the connecting equations [Brown, George and Xu (2008), Theorem 1] between the predictive and quadratic PE regimes do not exist in those cases, though some of the decision theoretic parallels can still be proved particularly for the ℓ_2 loss [Gatsonis (1984)].

APPENDIX

A.1. Bayes density estimate for discrete priors. The posterior distribution for the discrete prior $\pi = \sum_{k=-K}^{K} \pi_k \delta_{\mu_k}$ is given by

$$\pi(\mu_k|x) = \{m(x)\}^{-1}\phi(x|\mu_k, 1)\pi_k \qquad \text{where } m(x) = \sum_k \pi_k \phi(x|\mu_k, 1).$$

So, for the Bayes predictive density based on the prior π,

$$\hat{p}_\pi(y|x) = \sum_{k=-K}^{K} \phi(y|\mu_k, r)\pi(\mu_k|x) = \sum_{k=-K}^{K} \phi(y|\mu_k, r)\frac{\phi(x|\mu_k, 1)\pi_k}{m(x)}. \tag{58}$$

A.2. K–L risk for gaussian and linear density estimates. The predictive risk of the univariate Gaussian density estimate $p[\hat{\theta}, \hat{d}] = N(\hat{\theta}, \hat{d})$ is given by

$$\rho(\theta, p[\hat{\theta}, \hat{d}]) = E_\theta\{\log\phi(Y|\theta, r)\} - E_\theta\{\log\phi(Y|\hat{\theta}(X), \hat{d}(X))\},$$

where the expectation is over $X \sim N(\theta, 1)$ and $Y \sim N(\theta, r)$. Noting that

$$E_\theta\{\log\phi(Y|\hat{\theta}, \hat{d})|X = x\} = -\tfrac{1}{2}\log(2\pi\hat{d}(x)) - (2\hat{d}(x))^{-1}\{r + (\hat{\theta}(x) - \theta)^2\}$$

and $E_\theta \log\phi(Y|\theta, r) = -\frac{1}{2}\log(2\pi r) - \frac{1}{2}$, we obtain

$$L(\theta, \hat{p}(\cdot|x)) = \frac{1}{2}\log(r^{-1}\hat{d}) + \frac{r + (\hat{\theta}(x) - \theta)^2}{2\hat{d}} - \frac{1}{2}, \tag{59}$$

and the following expression for the K–L risk of members in $\mathcal{G}$:

$$\rho(\theta, p[\hat{\theta}, \hat{d}]) = \frac{1}{2}\left[E_\theta \log(r^{-1}\hat{d}) + E_\theta\left\{\frac{r + (\hat{\theta} - \theta)^2}{\hat{d}}\right\} - 1\right]. \tag{60}$$

Consider now "linear" estimators. Starting with the conjugate prior $\theta \sim N(0, \alpha/(1 - \alpha))$ for $0 \le \alpha \le 1$, standard calculations show that the posterior density $\pi(\theta|x)$ is $N(\alpha x, \alpha)$ and the predictive density $\hat{p}_{L,\alpha}$, being the convolution of Gaussians, compare (2), is seen to be $N(\alpha x, r + \alpha)$. Now, using $\hat{d} = r + \alpha$ and $\hat{\theta} = \alpha X$ in (60), we get

$$\rho(\theta, \hat{p}_{L,\alpha}) = \tfrac{1}{2}[\log(1 + r^{-1}\alpha) + (r + \alpha)^{-1}\{r + E_\theta(\alpha X - \theta)^2\} - 1].$$

The linear risk formula (56) now follows from the quadratic risk of αX. Next, we present some details about the risk of the particular linear estimate $\hat{p}_U$.

Proof of (42). The estimator $\hat{p}_U = \hat{p}_{L,1}$ is given by the $N(x, 1+r)$ distribution, and so from (59)

$$L(\theta, \hat{p}_U(\cdot|x)) = \frac{1}{2}\log(1+r^{-1}) + \frac{r + (\theta - x)^2}{2(1+r)} - \frac{1}{2},$$

from which (42) is immediate.

Acknowledgments. We thank the Associate Editor and three referees for constructive suggestions to shorten and improve the paper.

SUPPLEMENTARY MATERIAL

Supplementary material to "Exact minimax estimation of the predictive density in sparse Gaussian models" (DOI: 10.1214/14-AOS1251SUPP; .pdf). The supplement Mukherjee and Johnstone (2015) contains a brief description of the relevance of the predictive density estimation problem in related application areas along with the proof for the suboptimality of the univariate threshold density estimate $\hat{p}_{T,\mathrm{LF}}$ (in Section S.2) and the details of the proof of Proposition 1 (in Section S.3). The arguments for the maximum quadratic risk of hard threshold point estimates are reviewed in Section S.4 and the proof of Proposition 7 is presented in Section S.5. Links to R-codes used in producing Table 1 and Figure 3 are also provided.

REFERENCES

AITCHISON, J. (1975). Goodness of prediction fit. *Biometrika* **62** 547–554. MR0391353

AITCHISON, J. and DUNSMORE, I. R. (1975). *Statistical Prediction Analysis*. Cambridge Univ. Press, Cambridge. MR0408097

ASLAN, M. (2006). Asymptotically minimax Bayes predictive densities. *Ann. Statist.* **34** 2921–2938. MR2329473

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1996). Prediction and asymptotics. *Bernoulli* **2** 319–340. MR1440272

BELL, R. M. and COVER, T. M. (1980). Competitive optimality of logarithmic investment. *Math. Oper. Res.* **5** 161–166. MR0571810

BROWN, L. (1974). Lecture notes on statistical decision theory. Available at http://www-stat.wharton.upenn.edu/~lbrown.

BROWN, L. D., GEORGE, E. I. and XU, X. (2008). Admissible predictive density estimation. *Ann. Statist.* **36** 1156–1170. MR2418653

COVER, T. M. and THOMAS, J. A. (1991). *Elements of Information Theory*. Wiley, New York. MR1122806

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99** 277–303.

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

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

GATSONIS, C. A. (1984). Deriving posterior distributions for a location parameter: A decision theoretic approach. *Ann. Statist.* **12** 958–970. MR0751285

GEISSER, S. (1993). *Predictive Inference: An Introduction. Monographs on Statistics and Applied Probability* **55**. Chapman & Hall, New York. MR1252174

GEORGE, E. I., LIANG, F. and XU, X. (2006). Improved minimax predictive densities under Kullback–Leibler loss. *Ann. Statist.* **34** 78–91. MR2275235

GEORGE, E. I., LIANG, F. and XU, X. (2012). From minimax shrinkage estimation to minimax shrinkage prediction. *Statist. Sci.* **27** 82–94. MR2953497

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

HARTIGAN, J. A. (1998). The maximum likelihood prior. *Ann. Statist.* **26** 2083–2103. MR1700222

JOHNSTONE, I. M. (2013). Gaussian estimation: Sequence and wavelet models. Available at http://www-stat.stanford.edu/~imj.

KOMAKI, F. (1996). On asymptotic properties of predictive distributions. *Biometrika* **83** 299–313. MR1439785

KOMAKI, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika* **88** 859–864. MR1859415

KOMAKI, F. (2004). Simultaneous prediction of independent Poisson observables. *Ann. Statist.* **32** 1744–1769. MR2089141

LARIMORE, W. E. (1983). Predictive inference, sufficiency, entropy and an asymptotic likelihood principle. *Biometrika* **70** 175–181. MR0742987

MCMILLAN, B. (1956). Two inequalities implied by unique decipherability. *IRE Transactions on Information Theory* **2** 115–116.

MUKHERJEE, G. (2013). Sparsity and shrinkage in predictive density estimation. Ph.D. thesis, Stanford Univ. Available at http://purl.stanford.edu/gm306wz2890.

MUKHERJEE, G. and JOHNSTONE, I. M. (2015). Supplement to "Exact minimax estimation of the predictive density in sparse Gaussian models." DOI:10.1214/14-AOS1251SUPP.

MURRAY, G. D. (1977). A note on the estimation of probability density functions. *Biometrika* **64** 150–152. MR0448690

NG, V. M. (1980). On the estimation of parametric density functions. *Biometrika* **67** 505–506. MR0581751

PINSKER, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Inf. Transm.* **16** 120–133. Originally in Russian in *Problemy Peredachi Informatsii* **16** 52–67. MR0624591

XU, X. and LIANG, F. (2010). Asymptotic minimax risk of predictive density estimation for nonparametric regression. *Bernoulli* **16** 543–560. MR2668914

XU, X. and ZHOU, D. (2011). Empirical Bayes predictive densities for high-dimensional normal models. *J. Multivariate Anal.* **102** 1417–1428. MR2819959

DEPARTMENT OF DATA SCIENCES AND OPERATIONS
MARSHALL SCHOOL OF BUSINESS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089-0809
USA
E-MAIL: gourab@usc.edu

DEPARTMENT OF STATISTICS
SEQUOIA HALL, 390 SERRA MALL
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065
USA
E-MAIL: imj@stanford.edu