ARTICLE IN PRESS

Journal of Multivariate Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Empirical Bayes predictive densities for high-dimensional normal models

Xinyi Xu *, Dunke Zhou

Department of Statistics, The Ohio State University, Columbus, OH 43210-2147, United States

ARTICLE INFO

Article history:
Received 11 March 2010
Available online xxxx

AMS 2000 subject classifications:
primary 62C12
secondary 62C20
62J07

Keywords:
Predictive density
Kullback–Leibler loss
Empirical Bayes
Minimaxity
Oracle inequality
Shrinkage estimation

ABSTRACT

This paper addresses the problem of estimating the density of a future outcome from a multivariate normal model. We propose a class of empirical Bayes predictive densities and evaluate their performances under the Kullback–Leibler (KL) divergence. We show that these empirical Bayes predictive densities dominate the Bayesian predictive density under the uniform prior and thus are minimax under some general conditions. We also establish the asymptotic optimality of these empirical Bayes predictive densities in infinite-dimensional parameter spaces through an oracle inequality.

1. Introduction

Constructing accurate predictive strategies is a fundamental problem in both statistics and sciences. Traditional approaches to prediction provide a point forecast of an unknown future quantity and sometimes attach an error bound to convey uncertainty. A more comprehensive approach is to provide a predictive density that assigns probabilities to all possible outcomes. Such complete descriptions of uncertainty lead to sharper risk assessment and better decision making. In the past decades, predictive densities have been widely used in diverse fields, ranging from climatology [42] to financial management [43]. Besides predicting future trends and behavior patterns, predictive densities have also been used in model checking and model diagnostics [33,14,36], missing data analysis [34,15,35,16,28], and data compression and information theory [3,8,27].

In this paper, we consider the problem of estimating the density of a future outcome from a multivariate normal model, the centerpiece of parametric models. Suppose that we observe a p-dimensional normal vector $X \mid \theta \sim N_p(\theta, v_x I_p)$ and would like to predict a (conditionally) independent future outcome $Y \mid \theta \sim N_p(\theta, v_y I_p)$, which is centered at the same unknown mean θ but has a possibly different variance. Assume that $v_x > 0$ and $v_y > 0$ are known or can be independently estimated from the data. By a sufficiency and transformation reduction, this problem seems to be equivalent to observing $X_1, \ldots, X_n \mid \theta$ i.i.d. $\sim N_p(\theta, \Sigma)$ and predicting future outcomes $X_{n+1}, \ldots, X_{n+m}$ from the same data generating process. The (conditional) density of Y can be estimated by a predictive estimate $\hat{p}(y \mid x)$. We measure the closeness of $\hat{p}(y \mid x)$ to the true density $p(y \mid \theta)$ by the Kullback–Leibler (KL) divergence

$$L(p, \hat{p}) = \int p(y \mid \theta) \log \frac{p(y \mid \theta)}{\hat{p}(y \mid x)} \, dy, \tag{1}$$

* Corresponding author.
E-mail addresses: xinyi@stat.osu.edu, xu.214@osu.edu (X. Xu), dunkchou@stat.osu.edu (D. Zhou).

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2011.05.008

Please cite this article in press as: X. Xu, D. Zhou, Empirical Bayes predictive densities for high-dimensional normal models, Journal of Multivariate Analysis (2011), doi:10.1016/j.jmva.2011.05.008

and evaluate the performance of $\hat{p}(y \mid x)$ by its risk function

$$R_{KL}(p, \hat{p}) = \int p(x \mid \theta) L(p, \hat{p}) \mathrm{d}x. \tag{2}$$

For the comparison of two procedures, we say that $\hat{p}_1$ dominates $\hat{p}_2$ if $R_{KL}(p, \hat{p}_1) \leq R_{KL}(p, \hat{p}_2)$ for all θ and with strict inequality for some θ.

There are two widely used approaches to predictive density estimation—the "plug-in" approach and the Bayesian approach. The plug-in approach simply replaces θ by an estimate $\hat{\theta}(x)$ and then use

$$\hat{p}_{\hat{\theta}}(y \mid x) = p(y \mid \theta = \hat{\theta}(x)).$$

Although appealing in its simplicity, this approach ignores the uncertainty in parameter estimation and thus often lead to inferior predictive density estimators (see, for example, [2,26,13,24,41,40]). In contrast, the Bayesian approach integrates θ out with respect to a pre-specified prior distribution π to get

$$\hat{p}_{\pi}(y \mid x) = \frac{\int p(x \mid \theta) p(y \mid \theta) \pi(\theta) \mathrm{d}\theta}{\int p(x \mid \theta) \pi(\theta) \mathrm{d}\theta} = \int p(y \mid \theta) \pi(\theta \mid x) \mathrm{d}\theta.$$

Thus it directly incorporates the parameter uncertainty into the density estimate. However, as the dimension of the parameter grows, prior specification becomes challenging. A standard choice is to use a uniform prior $\pi_U \equiv 1$, under which the Bayesian predictive density is

$$\hat{p}_U(y \mid x) = \frac{1}{\{2\pi(v_x + v_y)\}^{\frac{p}{2}}} \exp\left\{-\frac{\|y - x\|^2}{2(v_x + v_y)}\right\}. \tag{3}$$

It dominates the plug-in predictive density $\hat{p}(y \mid \hat{\theta}_{MLE})$, which substitutes the maximum likelihood estimate (MLE) $\hat{\theta}_{MLE} = x$ for θ [1]. Moreover, it is best invariant and minimax with constant risk [31,32,27], and is admissible when the parameter dimension $p = 1$ or 2 [27,7]. However, when $p \geq 3$, $\hat{p}_U(y \mid x)$ is inadmissible and can be further dominated by the Bayesian predictive densities under the harmonic prior and under the Strawderman prior [25,27,20]. These Bayesian predictive densities are shown to be admissible in high-dimensional spaces [7], however, they are not of the form of normal distributions. In fact, their density representations are very complicated or even do not have closed-forms, which makes it difficult to implement them in practice. A principal purpose of our paper is to construct a class of predictive densities that have simple forms and at the same time retain nice risk properties.

It is interesting to note that the above results closely parallel some key developments concerning multivariate normal mean estimation under quadratic loss. Based on observing $X \mid \theta \sim N(\theta, v_x I)$, that problem is to estimate θ under

$$R_Q(\theta, \hat{\theta}) = E\|\theta - \hat{\theta}\|^2.$$

The maximum likelihood estimator $\hat{\theta}_{MLE} = X$ is best invariant, minimax and admissible when $p = 1$ or 2, but can be further dominated by the Bayesian estimators under the harmonic prior [37] and under the Strawderman prior [39]. Thus $\hat{\theta}_{MLE}$ plays the same role as $\hat{p}_U$ in the predictive density estimation problem. A further connection between $\hat{\theta}_{MLE}$ and $\hat{p}_U$ is revealed by the fact that $\hat{\theta}_{MLE}$ can also be motivated as the Bayesian estimator under the uniform prior $\pi_U \equiv 1$. George et al. [20] and Brown et al. [7] drew out these parallels by showing that there is a fascinating connection between the predictive density estimation problem and the classic point estimation problem, so that we can borrow strength from some important, beautiful and fundamental results in the latter area.

One of the most famous estimators in the normal mean estimation problem is the James–Stein estimator

$$\hat{\theta}_{JS}(x) = \left(1 - \frac{(p-2)v_x}{\|x\|^2}\right) x. \tag{4}$$

It was introduced by James and Stein [23] in 1961 and shocked the statistics community. Its representation is very simple—as shown in (4), $\hat{\theta}_{JS}$ just shrinks the MLE $\hat{\theta}_{MLE} = x$ toward 0 by a multiplicative factor $1 - (p-2)v_x/\|x\|^2$, but it has many excellent risk properties. When $p \geq 3$, it outperforms the best invariant estimator $\hat{\theta}_{MLE}$ for all values of the parameter θ and hence is minimax. Although it is not smooth enough to be admissible, it or its positive part $\hat{\theta}_{JS+} = \max(0, \hat{\theta}_{JS})$ is "close" to a Bayes estimator [10] and is difficult to improve upon [6,30]. The simple form and superior risk properties of the James–Stein estimator have led to its wide applications in many scientific problems, such as estimating baseball batting averages [11,19], assessing seasonal factors [29], and inferring gene networks [22]. It also plays an important role in modern wavelet analysis and nonparametric regression estimation. A key result in these fields is the so-called "oracle inequality", which shows that the quadratic risk of the James–Stein estimator may exceed the "oracle risk" only by a constant. Therefore, the James–Stein estimator is asymptotically minimax as the parameter dimension $p \to \infty$ and can be used to construct adaptive minimax estimators in Sobolev spaces [12]. These results had enormous influence and many papers tried to provide heuristic arguments for the superiority of the James–Stein estimator. A leading argument is the empirical Bayes argument provided

Please cite this article in press as: X. Xu, D. Zhou, Empirical Bayes predictive densities for high-dimensional normal models, Journal of Multivariate Analysis (2011), doi:10.1016/j.jmva.2011.05.008

by Efron and Morris [9], which shows that the James–Stein estimator can be motivated as an empirical Bayes estimator under a normal prior.

Inspired by these works in the point estimation problem, we propose a class of predictive densities in the predictive density estimation problem using an analogous empirical Bayes approach. In Section 2, we detail the derivation of our empirical Bayes predictive densities and show that they have very simple normal forms, centered at shrinkage estimators of θ and having data-dependent variances to incorporate the estimation uncertainty. In Section 3, we prove that analogous to the James–Stein estimator, our empirical Bayes predictive densities dominate the best invariant predictive density $\hat{p}_U$ under some general conditions and thus are minimax. We conduct a numerical analysis to compare the KL risks of two empirical Bayes predictive densities with the risks of $\hat{p}_U$ and the Bayesian predictive density under the harmonic prior. Then in Section 4, we establish an analogous oracle inequality for an empirical Bayes predictive density and show that it is asymptotically minimax in infinite-dimensional parameter spaces. Finally, in Section 5, we summarize the major contributions of this paper and discuss possible extensions.

2. The empirical Bayes predictive densities

In the classic problem of estimating the normal mean under quadratic loss, suppose that the normal mean θ follows a normal prior $N(0, vI_p)$, the resulting Bayesian estimator can be expressed as $\hat{\theta} = v \cdot x/(v_x + v)$. It is a linear estimator of x with the linear coefficient $v/(v_x + v)$. Efron and Morris [9] showed that the James–Stein estimator can be viewed as an empirical Bayes estimator that uses the data to select the prior and thus to determine the linear coefficient. In the predictive density estimation problem, we extend their approach to construct a class of empirical Bayes predictive densities that share many similar properties with the James–Stein estimator.

We analogously first consider the Bayesian predictive densities under normal priors. Under a prior $\pi_v(\theta) \sim N(0, vI_p)$, where $v > 0$ is an unknown constant, the Bayesian predictive density $\hat{p}_{\pi_v}$ can be represented by

$$\hat{p}_v(y \mid x) \sim \left(\frac{v}{v_x + v} x; \left(1 - \frac{v}{v_x + v}\right) v_y + \frac{v}{v_x + v}(v_x + v_y) \right). \tag{5}$$

We call this Bayesian predictive density a "linear predictive density" due to the above connection, although linearity does not have any literal meaning here, and we call $\lambda = v/(v_x + v)$ the "linear coefficient". To determine the optimal value of λ, we note that the KL risk of $\hat{p}_v$ can be expressed as

$$R_{KL}(\theta, \hat{p}_v) = E \log \frac{p(y \mid \theta)}{\hat{p}_v(y \mid x)} = \frac{p}{2} \log \frac{v_x + v_y}{v_y} + \frac{1}{2}\left[p \log \frac{v_w + v}{v_x + v} + \frac{pv_w + \|\theta\|^2}{v_w + v} - \frac{pv_x + \|\theta\|^2}{v_x + v} \right], \tag{6}$$

where $v_w = v_x v_y/(v_x + v_y)$. Ideally this risk could be minimized by taking

$$\lambda^* = \frac{v^*}{v^* + v} = \frac{\|\theta\|^2}{pv_x + \|\theta\|^2} = 1 - \frac{pv_x}{pv_x + \|\theta\|^2}. \tag{7}$$

The predictive density $\hat{p}_{\lambda^*}$ is called the *ideal* predictive density and its KL risk is called the *oracle* risk. However, $\hat{p}_{\lambda^*}$ is not computable in practice because it requires the knowledge of the unknown parameter θ. Therefore, we follow the approach of [9] to estimate λ^* by its unbiased estimator $\hat{\lambda}^*_U = \lambda_{p-2} = 1 - (p-2)v_x/\|x\|^2$. Moreover, to ensure that the variance of the predictive density is positive, we truncate λ_{p-2} at 0. Substituting $\lambda_{p-2,+} = \max(0, \hat{\lambda}^*_{p-2})$ into the predictive density yields a James–Stein type empirical Bayes predictive density

$$\hat{p}_{p-2}(y \mid x) \sim N\left(\left(1 - \frac{(p-2)v_x}{\|x\|^2}\right)_+ x; v_y + \left(1 - \frac{(p-2)v_x}{\|x\|^2}\right)_+ v_x \right).$$

Note that its mean is exactly the positive part James–Stein estimator for θ.

In the above derivation, if we had estimated λ^* by its marginal MLE $\hat{\lambda}^*_{MLE} = \lambda_{p,+} = (1 - pv_x/\|x\|^2)_+$ instead, the resulting empirical Bayes predictive density would be

$$\hat{p}_p(y \mid x) \sim N\left(\left(1 - \frac{pv_x}{\|x\|^2}\right)_+ x; v_y + \left(1 - \frac{pv_x}{\|x\|^2}\right)_+ v_x \right).$$

Estimators such as $\lambda_{p-2,+}$ and $\lambda_{p,+}$ are usually called shrinkage estimators, because they shrink the estimator X toward 0. There exist, in fact, many other shrinkage estimators of X. A general class studied in the point estimation problem is

$$\lambda_k(x) = 1 - kv_x/\|x\|^2, \tag{8}$$

where $k \geq 0$ is a non-negative constant. Similarly truncating λ_k at 0 and substituting $\lambda_{k,+} = \max(0, \lambda_k)$ into the density function leads to the following class of empirical Bayes predictive densities

$$\hat{p}_k(y \mid x) \sim N\left(\left(1 - \frac{kv_x}{\|x\|^2}\right)_+ x; v_y + \left(1 - \frac{kv_x}{\|x\|^2}\right)_+ v_x \right), \quad \text{where } k \geq 0. \tag{9}$$

Please cite this article in press as: X. Xu, D. Zhou, Empirical Bayes predictive densities for high-dimensional normal models, Journal of Multivariate Analysis (2011), doi:10.1016/j.jmva.2011.05.008

It is worth noting that the variance of $\hat{p}_k$ is always greater than or equal to the plug-in variance v_y, hence $\hat{p}_k$ automatically incorporates the estimation uncertainty by inflating the variance. Moreover, for a fixed $k \geq 0$, as $\|x\|^2 \to 0$, the predictive density $\hat{p}_k$ converges to the naive plug-in procedure $N(0;\, v_y)$, which simply replaces the parameter θ by 0; and as $\|x\|^2 \to \infty$, $\hat{p}_k$ converges to the best invariant predictive density $\hat{p}_U \sim N(x;\, v_x+v_y)$. Therefore, $\hat{p}_k$ can be viewed as a shrinkage predictive estimator that "pulls" $\hat{p}_U$ toward 0. As we shall show in the next section, it could provide substantial risk improvement over $\hat{p}_U$ when θ is close to the shrinkage target 0.

Remark 1. Another well-known method for constructing the James–Stein estimator uses the (pseudo-) marginal distribution of X. As shown in [5], any Bayesian estimator $\hat{\theta}_\pi = E_\pi(\theta \mid x)$ in the point estimation problem can be represented by

$$\hat{\theta}_\pi = x + \nabla \log m_\pi(x), \tag{10}$$

where $\nabla = (\partial/\partial x_1, \ldots, \partial/\partial x_p)$ and $m_\pi(x)$ is the marginal likelihood of X under π. Although the positive part James–Stein estimator is not a real Bayesian estimator under any prior, it could be obtained by substituting m_π in (10) with a pseudo-marginal distribution

$$\begin{aligned} m_{JS}(x) &= k_p \|x\|^{-(p-2)}, \quad \text{if } \|x\|^2 \geq (p-2); \\ &= \exp\{-\|x\|^2/2\}, \quad \text{if } \|x\|^2 < (p-2), \end{aligned}$$

where $k_p = (e/(p-2))^{-(p-2)/2}$ [4]. In the predictive density estimation problem, George et al. [20] showed that a Bayesian predictive density estimator $\hat{p}_\pi$ can be represented in an analogous marginal form, namely,

$$\hat{p}_\pi(y \mid x) = \frac{m_\pi(w;\, v_w)}{m_\pi(x;\, v_x)} \hat{p}_U(y \mid x),$$

where $w = (v_y x + v_x y)/(v_x + v_y)$ and $v_w = v_x v_y/(v_x + v_y)$. Therefore, it might seem natural to construct a James–Stein type predictive density $\hat{p}_{JS}$ using a modified pseudo-marginal distribution

$$\begin{aligned} \tilde{m}_{JS}(x;\, v) &= k_p \|x\|^{-(p-2)}, \quad \text{if } \|x\|^2/v \geq (p-2); \\ &= v^{-(p-2)/2} \exp\{-\|x\|^2/2v\}, \quad \text{if } \|x\|^2/v < (p-2). \end{aligned}$$

Unfortunately, we found that such a pseudo-Bayes procedure is not a bona fide predictive distribution, because its integral $\int \hat{p}_{JS}(y \mid x)dy$ varies with x so that it cannot be normalized. Hence this pseudo-marginal approach is not directly applicable to the predictive density estimation problem.

3. Minimaxity of the empirical Bayes predictive densities

Our construction of the empirical Bayes predictive densities suggests that they may serve similar roles in the predictive estimation problem as the James–Stein estimator in the point estimation problem. In this section, we show that under some general conditions, an empirical Bayes predictive density $\hat{p}_k$ dominates the best invariant predictive density $\hat{p}_U$ and thus is minimax. We demonstrate the risk performances of two empirical Bayes predictive densities $\hat{p}_{p-2}$ and $\hat{p}_{p-3}$ through a numerical analysis.

3.1. Sufficient conditions for minimaxity

We first show that the KL risk of a general "linear" predictive density

$$\hat{p}_\lambda(y \mid x) \sim N\left(\lambda x;\, (1-\lambda)v_y + \lambda(v_x + v_y)\right)$$

depends on the variances v_x and v_y only through their ratio $v_r = v_y/v_x$. Note that we put double quotes around "linear" to indicate that λ could be either a constant or a function of x.

Lemma 1. *The KL risk of a "linear" predictive density $\hat{p}_\lambda$ can be represented by*

$$R(\theta, \hat{p}_\lambda) = E\left[\frac{p}{2}\log\frac{\lambda + v_r}{v_r} + \frac{\|\lambda Z - \mu\|^2 - p\lambda}{2(\lambda + v_r)}\right], \tag{11}$$

where $v_r = v_y/v_x$, $z = X/\sqrt{v_x}$ and $\mu = \theta/\sqrt{v_x}$.

Proof. See Appendix. □

Please cite this article in press as: X. Xu, D. Zhou, Empirical Bayes predictive densities for high-dimensional normal models, Journal of Multivariate Analysis (2011), doi:10.1016/j.jmva.2011.05.008

Lemma 1 facilitates the risk comparison between the best invariant predictive density $\hat{p}_U$ and an empirical Bayes predictive density $\hat{p}_k$, because they are both "linear" predictive densities with $\lambda = 1$ and $\lambda = \lambda_{k,+}(x) = \left(1 - kv_x/\|x\|^2\right)_+$, respectively. To simplify the notations, in the rest of the paper we will assume that the "linear coefficients" are always positively truncated and we will use λ_k and $\lambda_{k,+}(x)$ indistinguishably. The next lemma provides a lower bound for the KL risk difference between $\hat{p}_U$ and $\hat{p}_k$.

Lemma 2. *For any $k \geq 0$, the KL risk difference between $\hat{p}_U$ in (3) and $\hat{p}_k$ in (9) has a lower bound*

$$R(\theta, \hat{p}_U) - R(\theta, \hat{p}_k) \geq E\left[\frac{p}{2}\log\frac{1+v_r}{\lambda_k+v_r} + \left(p - 4 - k - \frac{4k+6}{kv_r} - \frac{8}{kv_r^2}\right)\frac{1-\lambda_k}{2(\lambda_k+v_r)}\right], \tag{12}$$

where $\lambda_k = \left(1 - kv_x/\|x\|^2\right)_+$.

Proof. See Appendix. □

This expression (12) plays a key role in the predictive density estimation problem. Based on it, we can now establish the following sufficient conditions for an empirical Bayes predictive density $\hat{p}_k$ to dominate $\hat{p}_U$ and thus be minimax.

Theorem 1. *The empirical Bayes predictive density (9) is minimax under the KL loss if*

$$p \geq p_0(v_r) \quad \textit{and} \quad k \in [k_{\min}(v_r), k_{\max}(v_r, p)], \tag{13}$$

where

$$p_0(v_r) = \frac{4(v_r+1) + 4\sqrt{2+3v_r/2}}{(1+2v_r)v_r/(1+v_r)}, \tag{14}$$

$$k_{\min}(v_r) = \frac{1}{v_r}(2 + \sqrt{2+3v_r/2}), \quad \textit{and} \quad k_{\max}(v_r, p) = \frac{1+2v_r}{1+v_r}p - 4 - \frac{1}{v_r}(4 + 2\sqrt{2+3v_r/2}). \tag{15}$$

Proof. By Taylor expansion, for any $\lambda_k \in [0, 1]$,

$$E\left[\frac{p}{2}\log\frac{1+v_r}{\lambda_k+v_r}\right] = E\left[-\frac{p}{2}\log\left(1 - \frac{1-\lambda_k}{1+v_r}\right)\right] \geq E\left[\frac{p(1-\lambda_k)}{2(1+v_r)}\right] \geq E\left[\frac{v_r}{1+v_r}\cdot\frac{p(1-\lambda_k)}{2(\lambda_k+v_r)}\right].$$

Therefore, the KL risk lower bound (12) can be rewritten as

$$R(\theta, \hat{p}_U) - R(\theta, \hat{p}_k) \geq E\left[\left(\frac{1+2v_r}{1+v_r}p - 4 - k - \frac{4k+6}{kv_r} - \frac{8}{kv_r^2}\right)\frac{1-\lambda_k}{2(\lambda_k+v_r)}\right].$$

If we have

$$\frac{1+2v_r}{1+v_r}p - 4 - k - \frac{4k+6}{kv_r} - \frac{8}{kv_r^2} \geq 0, \tag{16}$$

then obviously $R(\theta, \hat{p}_k) \leq R(\theta, \hat{p}_U)$, and the minimaxity of $\hat{p}_k$ would follow immediately from the minimaxity of $\hat{p}_U$. Direct calculation yields that the inequality (16) holds when

$$k \in \left[\frac{\frac{1+2v_r}{1+v_r}pv_r - 4(v_r+1) - \sqrt{\Delta}}{2v_r}, \frac{\frac{1+2v_r}{1+v_r}pv_r - 4(v_r+1) + \sqrt{\Delta}}{2v_r}\right], \tag{17}$$

where

$$\Delta = \left[\frac{1+2v_r}{1+v_r}pv_r - 4(v_r+1)\right]^2 - 4(6v_r+8)$$

is non-negative iff $p \geq p_0(v_r)$ given in (14). Moreover, it is easy to check that when $p \geq p_0(v_r)$,

$$\Delta \geq \left(\frac{1-2v_r}{1-v_r}\right)^2 v_r^2(p - p_0(v_r))^2.$$

Thus, to obtain the inequality (16), it suffices to replace Δ in (17) by $\left(\frac{1-2v_r}{1-v_r}\right)^2 v_r^2(p - p_0(v_r))^2$, which produces $k \in [k_{\min}(v_r), k_{\max}(v_r, p)]$. □

The above sufficient conditions show that for an empirical Bayes predictive density $\hat{p}_k$ to dominate the best invariant predictive density $\hat{p}_U$, the parameter dimension p needs to be at least p_0. Note that p_0 is a decreasing function of the variance

Please cite this article in press as: X. Xu, D. Zhou, Empirical Bayes predictive densities for high-dimensional normal models, Journal of Multivariate Analysis (2011), doi:10.1016/j.jmva.2011.05.008

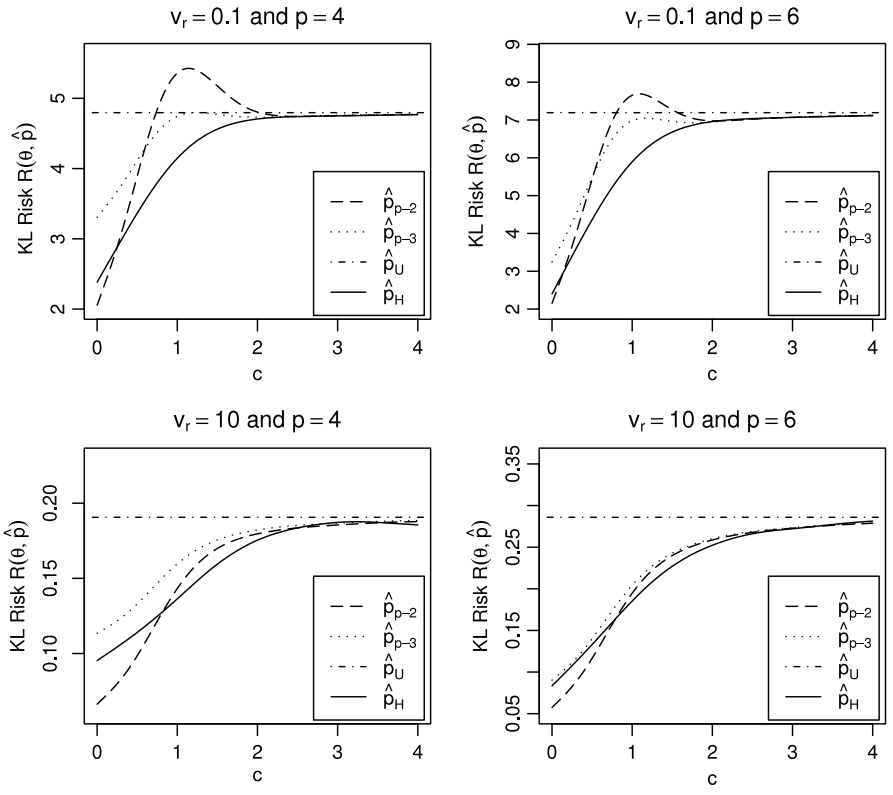

Fig. 1. The KL risks of the empirical Bayes predictive densities $\hat{p}_{p-2}$, $\hat{p}_{p-3}$ and the Bayesian predictive densities $\hat{p}_U$, $\hat{p}_H$ for $\theta = (c, \ldots, c)$.

ratio v_r, and as $v_r \to \infty$, p_0 reaches its lower limit 2. Therefore, for any finite v_r, we can find an empirical Bayes predictive density $\hat{p}_k$ to dominates $\hat{p}_U$ only when the parameter dimension is at least 3. This is consistent with the result that $\hat{p}_U$ is admissible when $p = 1$ or 2 [7].

Examples.

(i) Suppose that $v_r = 1$. Then $p_0 = 10.32$, $k_{\min} = 3.74$ and $k_{\max} = 1.5p - 11.74$. Therefore, the empirical Bayes predictive density $\hat{p}_{p-7}$ is minimax when $p \geq 11$.
(ii) Suppose that $v_r = 3$. Then $p_0 = 4.99$, $k_{\min} = 1.70$ and $k_{\max} = 1.75p - 7.03$. Therefore, the empirical Bayes predictive densities $\hat{p}_{p-3}$ and $\hat{p}_{p-4}$ are minimax when $p \geq 6$.
(iii) Suppose that $v_r = 5$. Then $p_0 = 3.96$, $k_{\min} = 1.23$ and $k_{\max} = 1.83p - 6.03$. Therefore, the empirical Bayes predictive densities $\hat{p}_{p-2}$ and $\hat{p}_{p-3}$ are minimax when $p \geq 5$.
(iv) Suppose that $v_r = 10$. Then $p_0 = 3.17$, $k_{\min} = 0.82$ and $k_{\max} = 1.91p - 5.22$. Therefore, the empirical Bayes predictive densities $\hat{p}_{p-2}$ and $\hat{p}_{p-3}$ are minimax when $p \geq 4$.
(v) Suppose that $v_r = 20$. Then $p_0 = 2.73$, $k_{\min} = 0.57$ and $k_{\max} = 1.95p - 4.77$. Therefore, the empirical Bayes predictive density $\hat{p}_{p-2}$ is minimax when $p \geq 3$.

3.2. Numerical analysis

To demonstrate the risk performances of our empirical Bayes predictive densities, we conduct simulation studies to compare the KL risks of two empirical Bayes predictive densities, $\hat{p}_{p-2}$ and $\hat{p}_{p-3}$, with those of two (generalized) Bayesian predictive densities: (i) $\hat{p}_U$ under the uniform prior $\pi_U \equiv 1$; (ii) $\hat{p}_H$ under the harmonic prior $\pi_H(\theta) \propto \|\theta\|^{-(p-2)}$. Here $\hat{p}_U$ serves as a benchmark for minimax predictive densities because it is best invariant and minimax with constant risk, and $\hat{p}_H$ represents the admissible predictive densities that have excellent risk properties but very complicated functional forms.

In Fig. 1, to illustrate the KL risk functions when θ is high dimensional, we let $\theta = (c, \ldots, c)'$, where $c \in [0, 4]$. Four different combinations of the variance ratio v_r and the parameter dimensions p are considered. In the upper two plots, v_r and p do not satisfy the minimax sufficient conditions in Theorem 1, so the KL risk of $\hat{p}_{p-2}$ is larger than that of $\hat{p}_U$ for some θ values. In the bottom two plots, the minimax sufficient conditions are satisfied, so $\hat{p}_{p-2}$ and $\hat{p}_{p-3}$ uniformly improve $\hat{p}_U$ for every θ value.

In all these scenarios, when θ is close to 0, the empirical Bayes predictive densities $\hat{p}_{p-2}$, $\hat{p}_{p-3}$ and the Bayesian predictive density $\hat{p}_H$ have substantially smaller KL risks than the best invariant estimator $\hat{p}_U$, and the risk reduction offered by $\hat{p}_{p-2}$

Please cite this article in press as: X. Xu, D. Zhou, Empirical Bayes predictive densities for high-dimensional normal models, Journal of Multivariate Analysis (2011), doi:10.1016/j.jmva.2011.05.008

is the greatest. This is because the normal prior π_v and the harmonic prior π_H are unimodal at 0, so the corresponding predictive densities "shrink" $\hat{p}_U$ toward 0, and $\hat{p}_{p-2}$ shrinks the most among these three. When θ is far from 0, the risks of $\hat{p}_{p-2}$, $\hat{p}_{p-3}$ and $\hat{p}_H$ asymptote to the risk of $\hat{p}_U$ and are very close to each other. At the same time, risks reductions are larger for larger p at each fixed v_r.

Overall speaking, the KL risk of $\hat{p}_{p-2}$ is comparable to that of $\hat{p}_H$ when the minimaxity sufficient conditions in Theorem 1 are satisfied, and $\hat{p}_{p-2}$ could offer even larger risk reductions when θ is close to the shrinkage target 0. Moreover, $\hat{p}_{p-2}$ is much easier to implement in practice because of its simple normal form.

4. An oracle inequality and asymptotic minimaxity

We now turn to investigate the KL risk properties of the empirical Bayes predictive densities in infinite-dimensional spaces. Suppose that the outcome of interest is associated with predictors through a nonparametric regression model, where the unknown function f is in a compact functional space such as a Sobolev space. Xu and Liang [44] showed that estimating the density of a future outcome from this nonparametric regression model based on n observations is equivalent to estimating the density of $Y \sim N_n(\theta, v_yI_n)$ based on observing $X \sim N_n(\theta, v_xI_n)$, where $v_x = v_y = n$ and θ is constrained in an ellipsoidal space. Therefore, constructing an optimal predictive density for the nonparametric regression model is equivalent to constructing an asymptotically optimal predictive density for the infinite-dimensional normal sequence model. We shall show that analogous to the James–Stein estimator, an empirical Bayes predictive density satisfies an oracle inequality, so its KL risk may exceed the oracle risk by only a constant and thus the empirical Bayes predictive density is asymptotically minimax.

We begin with a lemma that provides an explicit representation of the oracle risk, which is the risk of the *ideal* predictive density

$$\hat{p}_{\lambda^*} \sim N\left(\left(1 - \frac{pv_x}{pv_x + \|\theta\|^2}\right)x;\, v_y + \left(1 - \frac{pv_x}{pv_x + \|\theta\|^2}\right)v_x\right),$$

where θ is the true parameter value.

Lemma 3. *The oracle risk for the predictive density estimation problem can be represented by*

$$R(\theta, \hat{p}_{\lambda^*}) = \inf_{\hat{p}\in\mathcal{L}} R(\theta, \hat{p}) = \frac{p}{2}\log\frac{\|\mu\|^2 + (p + \|\mu\|^2)v_r}{(p + \|\mu\|^2)v_r}, \tag{18}$$

where $\mu = \theta/\sqrt{v_x}$ and $\mathcal{L}$ is the collection of all "linear" predictive densities.

Proof. See Appendix. □

Although the oracle risk can only be obtained by an "oracle" who knows the true parameter value θ, the next theorem shows that an empirical Bayes predictive density $\hat{p}_k$ can nearly achieve this optimal risk. To make the proof easier, we will show this result for the empirical Bayes predictive density $\hat{p}_{p-2}$. However, similar conclusions can be easily drawn for other empirical Bayes predictive densities $\hat{p}_k$ following the same steps.

Theorem 2. *When $p \geq 3$, the KL risk of the empirical Bayes estimator $\hat{p}_{p-2}$ satisfies the following "oracle" inequality*

$$R(\theta, \hat{p}_{\lambda^*}) \leq R(\theta, \hat{p}_{p-2}) \leq R(\theta, \hat{p}_{\lambda^*}) + \left(\frac{2}{v_r} + \frac{5}{2v_r^2} + \frac{4}{v_r^3}\right), \tag{19}$$

where $v_r = v_y/v_x$ is the variance ratio.

Proof. The first half of the inequality (19) follows immediately from the facts that $\hat{p}_{p-2}$ is a "linear" predictive density and that $\hat{p}_{\lambda^*}$ has the smallest risk among all "linear" predictive densities.

To prove the second half of the inequality, we note that the KL risk difference between the best invariant estimator $\hat{p}_U$ and $\hat{p}_{\lambda^*}$ an be represented by

$$\begin{aligned}
R(\theta, \hat{p}_U) - R(\theta, \hat{p}_{\lambda^*}) &= \frac{p}{2}\log\frac{1 + v_r}{v_r} - \frac{p}{2}\log\frac{\|\mu\|^2 + (p + \|\mu\|^2)v_r}{(p + \|\mu\|^2)v_r} \\
&= -\frac{p}{2}\log\left(1 - \frac{p}{(1 + v_r)(p + \|\mu\|^2)}\right) \\
&= -\frac{p}{2}\log\left(1 - \frac{p - 2}{(1 + v_r)(p + \|\mu\|^2)}\right) + \frac{p}{2}\log\left(1 + \frac{2}{(1 + v_r)\|\mu\|^2 + pv_r}\right) \\
&\leq -\frac{p}{2}\log\left(1 - \frac{(p - 2)}{(1 + v_r)(p + \|\mu\|^2)}\right) + \frac{p}{2}\log\left(1 + \frac{2}{pv_r}\right)
\end{aligned}$$

Please cite this article in press as: X. Xu, D. Zhou, Empirical Bayes predictive densities for high-dimensional normal models, Journal of Multivariate Analysis (2011), doi:10.1016/j.jmva.2011.05.008

$$= -\frac{p}{2}\log\left(1-\frac{(p-2)}{(1+v_r)(p+\|\mu\|^2)}\right)+\frac{1}{v_r}\log\left(1+\frac{2}{pv_r}\right)^{\frac{pv_r}{2}}$$
$$\le -\frac{p}{2}\log\left(1-\frac{(p-2)}{(1+v_r)(p+\|\mu\|^2)}\right)+\frac{1}{v_r}$$
$$\le E\left[\frac{p}{2}\log\frac{1+v_r}{\lambda_k+v_r}\right]+\frac{1}{v_r},$$

where the second inequality follows from the facts that the function $(1+1/x)^x$ is increasing in x and that $\lim_{x\to\infty}(1+1/x)^x = e$, and the third inequality follows from the Jensen's inequality and that $E(1/\|Z\|^2) \ge 1/(p+\|\mu\|^2)$. Combining this expression with the expression (12) in Lemma 2, we obtain the following upper bound of KL risk difference between $\hat{p}_{p-2}$ and $\hat{p}_{\lambda^*}$

$$\begin{aligned}R(\theta,\hat{p}_{p-2})-R(\theta,\hat{p}_{\lambda^*}) &= \left[R(\theta,\hat{p}_U)-R(\theta,\hat{p}_{\lambda^*})\right]-\left[R(\theta,\hat{p}_U)-R(\theta,\hat{p}_{p-2})\right]\\ &\le E\left[\frac{p}{2}\log\frac{1+v_r}{\lambda_k+v_r}\right]+\frac{1}{v_r}-E\left[\frac{p}{2}\log\frac{1+v_r}{\lambda_k+v_r}-\left(2+\frac{4(p-2)+6}{(p-2)v_r}+\frac{8}{(p-2)v_r^2}\right)\right.\\ &\quad\left.\times\frac{1-\lambda_k}{2(\lambda_k+v_r)}\right]\\ &\le \frac{1}{v_r}+E\left[\left(2+\frac{10}{v_r}+\frac{8}{v_r^2}\right)\frac{1}{2v_r}\right]\\ &\le \frac{2}{v_r}+\frac{5}{v_r^2}+\frac{4}{v_r^3}.\end{aligned}$$

This finishes the proof of the second part of inequality (19). □

In the infinite-dimensional normal sequence model, as shown in [44], the parameter dimension p is equal to the sample size n and the variances $v_x = v_y = n$. Therefore, the oracle inequality (19) becomes

$$R(\theta,\hat{p}_{\lambda^*}) \le R(\theta,\hat{p}_{p-2}) \le R(\theta,\hat{p}_{\lambda^*}) + 17/2. \tag{20}$$

The asymptotic optimality of a predictive density for this model is evaluated under the *average* KL risk $R(\theta,\hat{p})/n$. It can be shown that as $n\to\infty$, the *average oracle* risk reaches the asymptotic minimax risk, and by (20), the *average* KL risk of $\hat{p}_{p-2}$ asymptotes to the *average oracle* risk. Therefore, $\hat{p}_{p-2}$ also reaches the asymptotic minimax risk.

Example. Suppose that θ is restricted in a $\mathcal{L}^2$ ball $\Theta(C) = \{\theta : \sum_{i=1}^n \theta_i^2 \le C\}$. Xu and Liang [44] showed that the asymptotic minimax risk for this space is

$$\lim_{n\to\infty} R(\Theta(C)) = \frac{1}{2}\log\frac{1+2C}{1+C}.$$

Since $\mu = \theta/\sqrt{v_x} = \theta/\sqrt{n}$ satisfies $\|\mu\|^2 \le nC$, by Lemma 3, the *average* oracle risk achieves this asymptotic minimax risk. Moreover, the *average* risk difference between $\hat{p}_{p-2}$ and $\hat{p}_{\lambda^*}$ is bounded by $17/2n$, which goes to 0 as $n\to\infty$. Therefore, the empirical Bayes predictive density $\hat{p}_{p-2}$ also achieves the asymptotic minimax risk and thus is asymptotically minimax.

5. Discussions

In this paper, we consider the problem of predictive density estimation for multivariate normal models under the KL divergence. We propose a class of empirical Bayes predictive densities that have simple normal forms and share many similar properties with the famous James–Stein estimator in the classic point estimation problem. In particular, we show that these empirical Bayes predictive densities dominate the best invariant predictive density and thus are minimax under some general conditions. We also establish an "oracle" inequality that shows the KL risk of an empirical Bayes predictive density may exceed the oracle risk only by a constant. Therefore, the empirical Bayes predictive density is asymptotically minimax and can potentially be used to construct adaptive minimax predictive densities.

The normal priors π_v that we used to construct the empirical Bayes predictive densities $\hat{p}_k$ are centered at 0, so that $\hat{p}_k$ "shrinks" $\hat{p}_U$ toward 0 and offers the greatest reduction around this point. However, if θ is suspected to be close to a different point or even an affine subspace, we can easily re-center the prior to let the risk reduction most pronounced around that region. For example, a popular prior belief is that the coordinates of the parameter θ are similar, i.e., θ is contained in the subspace $S = [1_p] = \{\theta : \theta = u1_p\}$, where u is an unknown constant and $1_p = (1,\ldots,1)'$. This leads to the prior $\pi_{u,v} \sim N(u1, vI_p)$ and a new class of empirical Bayes predictive densities

$$\hat{p}_k'(y\mid x) \sim N\left(\bar{x}1_p + \left(1-\frac{kv_x}{\|x-\bar{x}1_p\|^2}\right)_+(x-\bar{x}1_p);\ v_y+\left(1-\frac{kv_x}{\|x-\bar{x}1_p\|^2}\right)_+ v_x\right),$$

Please cite this article in press as: X. Xu, D. Zhou, Empirical Bayes predictive densities for high-dimensional normal models, Journal of Multivariate Analysis (2011), doi:10.1016/j.jmva.2011.05.008

where $k \geq 0$ and $\bar{x} = \sum_{i=1}^{p} x_i/p$. These new predictive densities $\hat{p}'_k$'s will "shrink" $\hat{p}_U$ toward the subspace S and will offer greatest risk reduction around it. We can analogously establish minimax conditions and oracle inequalities for these $\hat{p}'_k$'s.

Furthermore, in situations where the shrinkage target is unclear, we can vastly enlarge the region of improved performance by constructing multiple shrinkage predictive densities following the approaches in [17–19]. Such a predictive density can be viewed as a convex combination of re-centered $\hat{p}'_k$'s at desired targets. Since the Kullback–Leibler loss is a convex function in $\hat{p}$, by Jensen's inequality,

$$R\left(\theta, \sum_{i=1}^{R} c_i \hat{p}'_{k_i}\right) = E \log \frac{p(Y \mid \theta)}{\sum_{i=1}^{r} c_i \hat{p}'_{k_i}(Y \mid X)} \leq E \sum_{i=1}^{r} c_i \log \frac{p(Y \mid \theta)}{\hat{p}'_{k_i}(Y \mid X)} = \sum_{i=1}^{r} c_i R(\theta, \hat{p}'_{k_i})$$

for any $0 \leq c_1, \ldots, c_r \leq 1$ and $\sum_{i=1}^{r} c_i = 1$. Therefore, if $\hat{p}'_{k_i}, i = 1, \ldots, r$ are minimax, their convex combination $\sum_{i=1}^{r} c_i \hat{p}'_{k_i}$ is also minimax. George [17–19] showed that such a multiple shrinkage estimators can adaptively shrink toward the point or subspace most favored by the data. George and Xu [21] used the same idea to construct multiple shrinkage Bayesian predictive densities for linear regression models.

Another possible extension of this work is to consider different "linear coefficients" for the mean and the variance of the empirical Bayes predictive densities. For example, we could set the mean of the predictive density at $\hat{\theta}_k = \left(1 - kv_x/\|x\|^2\right)_+ x$ and then estimate the optimal variance v_k separately. It is easy to check that the KL risk of $\hat{p}_{k,v_k}$

$$R(\theta, \hat{p}_{k,v_k}) = E\left[-\frac{p}{2} + \frac{p}{2}\left(\log \frac{\hat{v}_k}{v_y} + \frac{v_y}{\hat{v}_k}\right) + \frac{\|\hat{\theta} - \theta\|^2}{2 \cdot \hat{v}_k}\right]$$

is minimized at $v_k^* = v_y + E\|\hat{\theta} - \theta\|^2/p$. Analogous to the approach in Section 2, we could estimate $E\|\hat{\theta} - \theta\|^2$ by its unbiased estimate $\left(1 - [2k(p-2) - k^2]v_x/(p\|x\|^2)\right) p v_x$ and truncate it at 0. This will lead to a new class of predictive densities

$$\hat{p}''_k(y \mid x) \sim N\left(\left(1 - \frac{kv_x}{\|x\|^2}\right)_+ x;\ v_y + \left(1 - \frac{k_2 v_x}{\|x\|^2}\right)_+ v_x\right),$$

where $k_2 = [2k(p-2) - k^2]/p$. When $k \geq \max(0, p-4)$, the new "linear coefficient" $k_2 \leq k$, hence $\hat{p}''_k$ does not shrink as much as $\hat{p}_k$. We can similarly establish minimax sufficient conditions and oracle inequalities for $\hat{p}''_k$'s using the approaches in Sections 3 and 4.

Acknowledgments

This work was supported in part by the National Science Foundation under award number DMS-09-07070. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix

In this Appendix, we provide the proofs of Lemmas 1 and 2 from Section 3 and Lemma 3 from Section 4.

Proof of Lemma 1. Through straightforward calculation, we can write the KL risk of $\hat{p}_\lambda$ as

$$\begin{aligned} R(\theta, \hat{p}_\lambda) = E \log \frac{p(Y \mid \theta)}{\hat{p}_k(Y \mid X)} &= E\left[\frac{p}{2} \log \frac{\lambda v_x + v_y}{v_y} + \frac{\|Y - \lambda X\|^2}{2(\lambda v_x + v_y)} - \frac{p}{2}\right] \\ &= E\left[\frac{p}{2} \log \frac{\lambda v_x + v_y}{v_y} + \frac{\|\lambda X - \theta\|^2 - p\lambda v_x}{2(\lambda v_x + v_y)}\right]. \end{aligned}$$

The expression (11) then follows by dividing the numerator and the denominator in both terms by v_x. □

Proof of Lemma 2. Since $\hat{p}_U$ and $\hat{p}_k$ are "linear" predictive densities with $\lambda = 1$ and $\lambda = \lambda_k$ respectively, by Lemma 1, their KL risk difference can be represented by

$$\begin{aligned} R(\theta, \hat{p}_U) - R(\theta, \hat{p}_k) &= E\left[\frac{p}{2} \log \frac{1 + v_r}{v_r} + \frac{\|Z - \mu\|^2 - p}{2(1 + v_r)} - \frac{p}{2} \log \frac{\lambda_k + v_r}{v_r} - \frac{\|\lambda_k Z - \mu\|^2 - p\lambda_k}{2(\lambda_k + v_r)}\right] \\ &= E\left[\frac{p}{2} \log \frac{1 + v_r}{\lambda_k + v_r} - \frac{\|\lambda_k Z - \mu\|^2 - p\lambda_k}{2(\lambda_k + v_r)}\right], \end{aligned} \tag{21}$$

where $v_r = v_y/v_x$, $Z = X/\sqrt{v_x}$ and $\mu = \theta/\sqrt{v_x}$. We expand the second term in (21) as

$$E\left[-\frac{\|\lambda_k Z - \mu\|^2 - p\lambda_k}{2(\lambda_k + v_r)}\right] = E\left[-\frac{(1 - \lambda_k)^2\|Z\|^2 - 2(1 - \lambda_k)Z(Z - \mu) + \|Z - \mu\|^2 - p\lambda_k}{2(\lambda_k + v_r)}\right], \tag{22}$$

and we next derive a lower bound for it.

Please cite this article in press as: X. Xu, D. Zhou, Empirical Bayes predictive densities for high-dimensional normal models, Journal of Multivariate Analysis (2011), doi:10.1016/j.jmva.2011.05.008

Firstly, note that $1-\lambda_k$ satisfies

$$k/\|z\|^2 \geq 1-\lambda_k = \begin{cases} k/\|z\|^2 & \text{if } \|z\|^2 \geq k \\ 1 & \text{if } \|z\|^2 < k. \end{cases}$$

Therefore, the first term in (22) can be expressed as

$$-E\left[\frac{(1-\lambda_k)^2\|Z\|^2}{2(\lambda_k+v_r)}\right] \geq -E\left[\frac{k(1-\lambda_k)}{2(\lambda_k+v_r)}\right]. \tag{23}$$

Secondly, since λ_k is differentiable at all points except $\|z\|^2 = k$ with

$$\frac{\partial}{\partial z_i}\lambda_k = \begin{cases} 2kz_i/\|z\|^4 & \text{if } \|z\|^2 > k \\ 0 & \text{if } \|z\|^2 < k, \end{cases}$$

by Stein's identity [38], the second term in (22) can be represented by

$$\begin{aligned} E\left[\frac{2(1-\lambda_k)Z(Z-\mu)}{2(\lambda_k+v_r)}\right] &= E\left[\sum_{i=1}^n \frac{\partial}{\partial z_i}\frac{2(1-\lambda_k)Z}{2(\lambda_k+v_r)}\right] \\ &= E\left[\frac{2p(1-\lambda_k)}{2(\lambda_k+v_r)}\right] - E^{\|Z\|^2\geq k}\left[\frac{4k}{2\|Z\|^2(\lambda_k+v_r)} + \frac{4k^2}{2\|Z\|^4(\lambda_k+v_r)^2}\right] \\ &= E\left[\frac{2p(1-\lambda_k)}{2(\lambda_k+v_r)}\right] - E^{\|Z\|^2\geq k}\left[\frac{4(1-\lambda_k)}{2(\lambda_k+v_r)} + \frac{4(1-\lambda_k)^2}{2(\lambda_k+v_r)^2}\right] \\ &\geq E\left[\frac{(2p-4)(1-\lambda_k)}{2(\lambda_k+v_r)} - \frac{4(1-\lambda_k)^2}{2(\lambda_k+v_r)^2}\right]. \end{aligned} \tag{24}$$

Finally, by Stein's identity again, the third term in (22) can be expressed as

$$\begin{aligned} -E\left[\frac{\|Z-\mu\|^2}{2(\lambda_k+v_r)}\right] &= -E\left[\sum_{i=1}^n \frac{\partial}{\partial z_i}\frac{Z-\mu}{2(\lambda_k+v_r)}\right] \\ &= -E\left[\frac{p}{2(\lambda_k+v_r)}\right] + kE^{\|Z\|^2\geq k}\left[\frac{2Z(Z-\mu)}{2\|Z\|^4(\lambda_k+v_r)^2}\right]. \end{aligned} \tag{25}$$

We make an orthogonal transformation that maps Z into a random variable T such that $T_i|\xi_i$ independently $\sim N(\xi_i, 1)$, where $\xi_1 = \|\mu\|$ and $\xi_2 = \xi_3 = \cdots = \xi_p = 0$. Then in (25),

$$E^{\|Z\|^2\geq k}\left[\frac{Z(Z-\mu)}{\|Z\|^4(\lambda_k+v_r)^2}\right] = E^{\|T\|^2\geq k}\left[\frac{T_1(T_1-\xi_1)+T_2^2+\cdots+T_p^2}{\|T\|^4(\lambda_k+v_r)^2}\right] \geq E^{\|T\|^2\geq k}\left[\frac{T_1(T_1-\xi_1)}{\|T\|^4(\lambda_k+v_r)^2}\right].$$

Conditioning on $T_2^2+\cdots+T_p^2 = c \in [0,k]$, the restriction $\|T\|^2 \geq k$ is equivalent to $T_1^2 \geq k-c$, and the conditional expectation

$$\begin{aligned} &E^{T_1^2\geq k-c}\left[\frac{T_1(T_1-\xi_1)}{\|T\|^4(\lambda_k+v_r)^2} \mid T_2^2+\cdots+T_p^2 = c\right] \\ &= -\frac{t_1}{\|t\|^4(\lambda_k+v_r)^2}\,\varphi(t_1-\xi_1)\left(|_{-\infty}^{-\sqrt{k-c}} + |_{\sqrt{k-c}}^{\infty}\right) + E^{T_1^2\geq k-c}\left[\frac{\partial}{\partial T_1}\frac{T_1}{\|T\|^4(\lambda_k+v_r)^2} \mid T_2^2+\cdots+T_p^2 = c\right] \\ &\geq E^{T_1^2\geq k-c}\left[\frac{1-4T_1^2/\|T\|^2}{\|T\|^4(\lambda_k+v_r)^2} - \frac{4kT_1^2/\|T\|^2}{\|T\|^6(\lambda_k+v_r)^3} \mid T_2^2+\cdots+T_p^2 = c\right], \end{aligned} \tag{26}$$

where $\varphi(\cdot)$ is the density function of the standard normal distribution. Combining (25) and (26), we obtain

$$\begin{aligned} -E\left[\frac{\|Z-\mu\|^2}{2(\lambda_k+v_r)}\right] &\geq -E\left[\frac{p}{2(\lambda_k+v_r)}\right] - E^{\|Z\|^2\geq k}\left[\frac{6k}{2\|Z\|^4(\lambda_k+v_r)^2} + \frac{8k^2}{2\|Z\|^6(\lambda_k+v_r)^3}\right] \\ &\geq -E\left[\frac{p}{2(\lambda_k+v_r)} + \frac{6(1-\lambda_k)^2/k}{2(\lambda_k+v_r)^2} + \frac{8(1-\lambda_k)^3/k}{2(\lambda_k+v_r)^3}\right]. \end{aligned} \tag{27}$$

Please cite this article in press as: X. Xu, D. Zhou, Empirical Bayes predictive densities for high-dimensional normal models, Journal of Multivariate Analysis (2011), doi:10.1016/j.jmva.2011.05.008

Now substituting (23), (24) and (27) into (22) yields

$$E\left[-\frac{\|\lambda_k Z-\mu\|^2-p\lambda_k}{2(\lambda_k+v_r)}\right] \geq E\left[\frac{(p-4-k)(1-\lambda_k)}{2(\lambda_k+v_r)}-\frac{(4k+6)(1-\lambda_k)^2/k}{2(\lambda_k+v_r)^2}-\frac{8(1-\lambda_k)^3/k}{2(\lambda_k+v_r)^3}\right]$$
$$\geq E\left[\left(p-4-k-\frac{4k+6}{kv_r}-\frac{8}{kv_r^2}\right)\frac{1-\lambda_k}{2(\lambda_k+v_r)}\right], \tag{28}$$

where the second inequality follows from the fact $(1-\lambda_k)/(\lambda_k+v_r)\leq 1/v_r$ for any $0\leq\lambda_k\leq 1$. Combining from (21) and (28) leads to the KL risk lower bound (12) in Lemma 2. □

Proof of Lemma 3. The *ideal* predictive density $\hat{p}_{\lambda^*}$ is a "linear" density estimator with coefficient

$$\lambda^*=1-\frac{pv_x}{pv_x+\|\theta\|^2}=\frac{\|\mu\|^2}{p+\|\mu\|^2},$$

where $\mu=\theta/\sqrt{v_x}$. Therefore, by Lemma 1, the KL risk of $\hat{p}_{v^*}$ can by represented by

$$R(\theta,\hat{p}_{\lambda^*})=E\left[\frac{p}{2}\log\frac{\lambda^*+v_r}{v_r}+\frac{\|\lambda^*Z-\mu\|^2-p\lambda^*}{2(\lambda^*+v_r)}\right]$$
$$=\frac{p}{2}\log\frac{\|\mu\|^2+(p+\|\mu\|^2)v_r}{(p+\|\mu\|^2)v_r},$$

where the second equation follows from the facts that $E(Z)=\mu$ and $E(Z^2)=p+\|\mu\|^2$. □

References

[1] J. Aitchison, Goodness of prediction fit, Biometrika 62 (1975) 547–554.
[2] J. Aitchison, I.R. Dunsmore, Statistical Prediction Analysis, Cambridge University Press, 1975.
[3] A.R. Barron, J. Rissanen, B. Yu, The minimum description length principle in coding and modeling, IEEE Transaction on Information Theory 44 (1998) 2743–2760.
[4] M.E. Bock, Shrinkage estimators: pseudo-Bayes estimators for normal mean vectors, in: S.S. Gupta, J.O. Berger (Eds.), Statistical Decision Theory IV, Springer-Verlag, New York, 1988, pp. 281–298.
[5] L.D. Brown, Admissible estimators, recurrent diffusions, and insoluble boundary value problems, Annals of Mathematical Statistics 42 (1971) 855–903.
[6] L.D. Brown, Admissibility in discrete and continuous invariant nonparametric estimation problems and in their multinomial analogs, Annals of Statistics 16 (1988) 1567–1593.
[7] L.D. Brown, E.I. George, X. Xu, Admissible predictive density estimation, Annals of Statistics 36 (2008) 1156–1170.
[8] A. Yuan, B. Clarke, An informative criterion for likelihood selections, IEEE Transaction on Information Theory 45 (1999) 562–571.
[9] B. Efron, C.N. Morris, Limiting the risk of Bayes and empirical Bayes estimators—part II: the empirical Bayes case, Journal of the American Statistical Association 67 (1972) 130–139.
[10] B. Efron, C.N. Morris, Stein's estimation rule and its competitors—an empirical Bayes approach, Journal of the American Statistical Association 68 (1973) 117–130.
[11] B. Efron, C.N. Morris, Data analysis using Stein's estimator and its generalizations, Journal of the American Statistical Association 70 (1975) 311–319.
[12] S.Y. Efromovich, M.S. Pinsker, Estimation of square-integrable probability density of a random variable, Problems of Information Transmission 18 (1982) 175–189.
[13] S. Geisser, Predictive Inference: An Introduction, CRC Press, 1993.
[14] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, second ed. Chapman & Hall/CRC Press, Boca Raton, FL, 2004.
[15] A. Gelman, G. King, C. Liu, Multiple imputation for multiple surveys, Journal of the American Statistical Association 93 (1998) 846–874.
[16] A. Gelman, T.E. Raghunathan, Using conditional distributions for missing-data imputation, Statistical Science 15 (2001) 268–269.
[17] E.I. George, Minimax multiple shrinkage estimation, Annals of Statistics 14 (1986) 188–205.
[18] E.I. George, Combining minimax shrinkage estimators, Journal of the American Statistical Association 81 (1986) 437–445.
[19] E.I. George, A formal Bayes multiple shrinkage estimator, Communications in Statistics A – Theory Methods 15 (1986) 2099–2114.
[20] E.I. George, F. Liang, X. Xu, Improved minimax prediction under Kullback–Leibler loss, Annals of Statistics 34 (2006) 78–91.
[21] E.I. George, X. Xu, Predictive density estimation for multiple regression, Econometric Theory 24 (2008) 1–17.
[22] J. Hausser, K. Strimmer, Entropy inference and the James–Stein estimator, with application to nonlinear gene association networks, Journal of Machine Learning Research 10 (2009) 1469–1484.
[23] W. James, C. Stein, Estimation with quadratic loss, in: Proceedings fo the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, Berkeley, 1961, pp. 361–379.
[24] F. Komaki, On asymptotic properties of predictive distributions, Biometrika 83 (1996) 299–313.
[25] F. Komaki, A shrinkage predictive distribution for multivariate normal observations, Biometrika 88 (2001) 859–864.
[26] M.S. Levy, S.K. Perng, An optimal prediction function for the normal linear model, Journal of the American Statistical Association 81 (1986) 196–198.
[27] F. Liang, A. Barron, Exact minimax strategies for predictive density estimation, data compression and model selection, IEEE Information Theory Transactions 50 (2004) 2708–2726.
[28] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, second ed. Wiley, New York, 2002.
[29] D.M. Miller, D. William, Shrinkage estimators of time series seasonal factors and their effect on forecasting accuracy, Crime Forecasting 19 (2003) 669–684.
[30] T. Moore, R.J. Brook, Riske estimation optimality of James–Stein estimators, Annals of Statistics 6 (1978) 917–919.
[31] G.D. Murray, A note on the estimation of probability density functions, Biometrika 64 (1977) 150–152.
[32] V.M. Ng, On the estimation of parametric density functions, Biometrika 67 (1980) 505–506.
[33] I. Pardoe, A Bayesian sampling approach to regression model checking, Journal of Computational and Graphical Statistics 10 (2001) 617–627.
[34] D.B. Rubin, Multiple imputation after 18+ years, Journal of the American Statistical Association 91 (1996) 473–489.
[35] J.L. Schafer, Multiple imputation: a primer, Statistical Methods in Medical Research 8 (1999) 3–15.
[36] S. Sinharay, M.S. Johnson, H.S. Stern, Posterior predictive assessment of item response theory models, Applied Psychological Measurement 30 (2006) 298–321.

Please cite this article in press as: X. Xu, D. Zhou, Empirical Bayes predictive densities for high-dimensional normal models, Journal of Multivariate Analysis (2011), doi:10.1016/j.jmva.2011.05.008

[37] C. Stein, Estimation of the mean of a multivariate normal distribution, in: J. Hajek (Ed.), Proceedings of the Prague Symposium on Asymptotic Statistics, Universita Karlova, Prague, 1974, pp. 345–381.
[38] C. Stein, Estimation of the mean of a multivariate normal distribution, Annals of Statistics 9 (1981) 1135–1151.
[39] W.E. Strawderman, Proper Bayes minimax estimators of the multivariate normal mean, Annals of Mathematical Statistics 42 (1971) 385–388.
[40] F. Tanaka, Generalized Bayesian predictive density operators, in: The 14th Quantum Information Technology Symposium, 2006, pp. 107–110.
[41] F. Tanaka, F. Komaki, Bayesian predictive density operators for exchangeable quantum-statistical models, Physical Review A, American Institute of Physics 71 (2005) 052323.
[42] J.W. Taylor, R. Buizza, Comparing temperature density forecasts from GARCH and atmospheric models, Journal of Forecasting 23 (2004) 337–355.
[43] J. Weinberg, L.D. Brown, J.R. Stroud, Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data, Journal of the American Statistical Association 102 (2007) 1185–1198.
[44] X. Xu, F. Liang, Asymptotic minimax risk of predictive density estimation for nonparametric regression, Bernoulli 16 (2010) 543–560.

Please cite this article in press as: X. Xu, D. Zhou, Empirical Bayes predictive densities for high-dimensional normal models, Journal of Multivariate Analysis (2011), doi:10.1016/j.jmva.2011.05.008