Econometric Theory, **24**, 2008, 528–544. Printed in the United States of America.
DOI: 10.1017/S0266466608080213

PREDICTIVE DENSITY ESTIMATION FOR MULTIPLE REGRESSION

EDWARD I. GEORGE
The Wharton School, University of Pennsylvania

XINYI XU
The Ohio State University

Suppose we observe $X \sim N_m(A\beta,\sigma^2 I)$ and would like to estimate the predictive density $p(y|\beta)$ of a future $Y \sim N_n(B\beta,\sigma^2 I)$. Evaluating predictive estimates $\hat{p}(y|x)$ by Kullback–Leibler loss, we develop and evaluate Bayes procedures for this problem. We obtain general sufficient conditions for minimaxity and dominance of the "noninformative" uniform prior Bayes procedure. We extend these results to situations where only a subset of the predictors in A is thought to be potentially irrelevant. We then consider the more realistic situation where there is model uncertainty and this subset is unknown. For this situation we develop multiple shrinkage predictive estimators and obtain general minimaxity and dominance conditions. Finally, we provide an explicit example of a minimax multiple shrinkage predictive estimator based on scaled harmonic priors.

1. INTRODUCTION

We begin with the canonical normal linear model setup

$$X \sim N_m(A\beta, \sigma^2 I), \tag{1}$$

where X is an $m \times 1$ vector of m observations, A is a full rank, fixed $m \times p$ matrix of p potential predictors where $m \geq p$, and β is a $p \times 1$ vector of unknown regression coefficients. Based on observing $X = x$, we consider the problem of estimating the predictive density $p(y|\beta)$ of a future $n \times 1$ vector Y where

$$Y \sim N_n(B\beta, \sigma^2 I). \tag{2}$$

Here B is a fixed $n \times p$ matrix of the same p potential predictors in A, although with possibly different values. We also assume that X and Y are conditionally independent given β. Finally, we assume that σ^2 is known and without loss of generality set $\sigma^2 = 1$ throughout.

We acknowledge Larry Brown, Feng Liang, Linda Zhao, and three referees for their helpful suggestions. This work was supported by various NSF grants, DMS-0605102 the most recent. Address correspondence to Xinyi Xu, Department of Statistics, The Ohio State University, 1958 Niel Ave., Columbus, OH 43210-1247, USA; e-mail: xinyi@stat.osu.edu.

  0266-4666/08 $15.00

For each value of x, we evaluate a predictive estimate $\hat{p}(y|x)$ of $p(y|\beta)$ by the well-known Kullback–Leibler (KL) loss

$$L(\beta, \hat{p}(y|x)) = \int p(y|\beta) \log \frac{p(y|\beta)}{\hat{p}(y|x)}\, dy. \tag{3}$$

The overall quality of the procedure $\hat{p} = \hat{p}(y|x)$ for each β is then conveniently summarized by the KL risk

$$R_{KL}(\beta, \hat{p}) = \int p(x|\beta) L(\beta, \hat{p}(y|x))\, dx. \tag{4}$$

Letting $\hat{\beta}_x = (A'A)^{-1}A'x$ be the traditional least squares estimate of β based on x, it is tempting to consider the plug-in predictive estimate $\hat{p}_{plug\text{-}in}(y|\hat{\beta}_x)$, which simply substitutes $\hat{\beta}_x$ for β in $p(y|\beta)$. However, as we show in Section 2 by extending the arguments of Aitchison (1975), the formal Bayes predictive estimate

$$\hat{p}_U(y|x) = \frac{\int p(x|\beta) p(y|\beta)\, d\beta}{\int p(x|\beta)\, d\beta} \tag{5}$$

has smaller KL risk than $\hat{p}_{plug\text{-}in}(y|\hat{\beta}_x)$ for every β. Thus, $\hat{p}_{plug\text{-}in}(y|\hat{\beta}_x)$ should be ruled out, and we turn our focus to Bayes procedures.

For a prior π on β, the Bayes predictive estimator $\hat{p}_\pi(y|x)$ is given by

$$\hat{p}_\pi(y|x) = \frac{\int p(x|\beta) p(y|\beta) \pi(\beta)\, d\beta}{\int p(x|\beta) \pi(\beta)\, d\beta}. \tag{6}$$

It also follows from the arguments of Aitchison (1975) that for proper π, $\hat{p}_\pi$ minimizes the average risk $r_\pi(\hat{p}) = \int R_{KL}(\beta, \hat{p}) \pi(\beta)\, d\beta$. Note that $\hat{p}_U$ in (5) is the formal Bayes estimate under the improper uniform "noninformative" density $\pi_U(\beta) \equiv 1$ and would seem to be a good default procedure. Indeed, $\hat{p}_U$ has constant risk and is minimax under KL loss; see Liang (2002) and Liang and Barron (2004). But surprisingly, as we will show, in many cases $\hat{p}_U$ itself can be uniformly dominated in terms of KL risk by other Bayes predictive estimators.

In Section 2, we develop general conditions under which $\hat{p}_\pi$ will be minimax and uniformly dominate $\hat{p}_U$ in terms of the KL risk (4) for the multiple regression prediction problem. Our results can be seen as a substantial generalization of the work of George, Liang, and Xu (2006), who considered the special case of this problem when $X \sim N_m(\mu, \sigma_x^2 I)$ and $Y \sim N_m(\mu, \sigma_y^2 I)$, where μ is the common unknown multivariate normal mean. Moving further

away from this common mean setup, we proceed in Section 3 to extend these results to the setting where only a subset of the p predictors is considered to be potentially irrelevant. In Section 4, we consider the more realistic model uncertainty setting where such a subset is unknown, and we develop minimax multiple shrinkage predictive densities that adaptively shrink toward the model most favored by the data. In Section 5, we conclude by showing how our results can be extended for minimax shrinkage prediction toward any linear subspaces. Although we do not consider the issue of admissibility in this paper, it may be of interest to note that for the preceding multivariate normal prediction problem Brown, George, and Xu (2007) recently established that all admissible predictive densities are Bayes procedures.

2. PRIORS FOR MINIMAX PREDICTIVE ESTIMATION

In this section, we develop general conditions on π for $\hat{p}_\pi$ in (6) to uniformly dominate $\hat{p}_U$ in (5) under KL risk (4). The minimaxity of such $\hat{p}_\pi$ will then follow immediately from the minimaxity of $\hat{p}_U$.

We begin by establishing some convenient notation. As indicated previously, we use $\hat{\beta}_x = (A'A)^{-1}A'x$ to denote the least squares estimate of β based on x. Although y is not observed, it will be useful to use

$$\hat{\beta}_{x,y} = (C'C)^{-1}C'\begin{pmatrix} x \\ y \end{pmatrix} \quad \text{where } C = \begin{pmatrix} A \\ B \end{pmatrix} \tag{7}$$

to denote the least squares estimate of β based on x and y. Note that $\hat{\beta}_x \sim N_p(\beta, \Sigma_A)$ and $\hat{\beta}_{x,y} \sim N_p(\beta, \Sigma_C)$, where for notational convenience throughout we let $\Sigma_A = (A'A)^{-1}$ and $\Sigma_C = (C'C)^{-1}$. It will also be useful to let $RSS_x = \|x - A\hat{\beta}_x\|^2$ and

$$RSS_{x,y} = \left\| \begin{pmatrix} x \\ y \end{pmatrix} - C\hat{\beta}_{x,y} \right\|^2$$

denote the corresponding residual sums of squares (RSS). In terms of this notation, we have the following result.

LEMMA 1. *The uniform prior predictive estimate $\hat{p}_U$ in (5) can be expressed as*

$$\begin{aligned} \hat{p}_U(y|x) &= \frac{1}{(2\pi)^{n/2}} \frac{|C'C|^{-1/2}}{|A'A|^{-1/2}} \exp\left\{ -\frac{RSS_{x,y} - RSS_x}{2} \right\} \\ &= \frac{1}{(2\pi)^{p/2}|\Psi|} \exp\left\{ \frac{(y - B\hat{\beta}_x)'\Psi^{-1}(y - B\hat{\beta}_x)}{2} \right\}, \end{aligned} \tag{8}$$

where $\Psi = I + B\Sigma_A B'$. *Moreover, the KL risk of* $\hat{p}_U$ *is uniformly smaller than that of the plug-in estimator* $\hat{p}_{plug\text{-}in}(y|\hat{\beta}_x)$.

Proof. Because $\hat{\beta}_x|\beta \sim N_p(\beta, \Sigma_A)$, the posterior of β under the uniform prior is $\beta|\hat{\beta}_x \sim N_p(\hat{\beta}_x, \Sigma_A)$. It follows that the posterior of $B\beta$ is $B\beta|\hat{\beta}_x \sim N_p(B\hat{\beta}_x, B\Sigma_A B')$, and thus the predictive estimator is

$$Y|\hat{\beta}_x \sim N_p(B\hat{\beta}_x, I + B\Sigma_A B').$$

To calculate the risk of $\hat{p}_U$, let $\hat{H}_A = A(A'A)^{-1}A'$ denote the hat matrix based on x and $\hat{H}_C = C(C'C)^{-1}C'$ denote the hat matrix based on both x and y. It is easy to see that

$$\begin{aligned}
R_{KL}(\beta, \hat{p}_U) &= \iint p(x|\beta)p(y|\beta)\log\frac{p(y|\beta)}{\hat{p}_U(y|x)}\,dxdy \\
&= \frac{1}{2}\log\frac{|C'C|}{|A'A|} - \frac{n}{2} + \frac{1}{2}\iint p(x|\beta)p(y|\beta)[RSS_{x,y} - RSS_x]\,dxdy \\
&= \frac{1}{2}\log\frac{|C'C|}{|A'A|} - \frac{n}{2} + \frac{1}{2}[\operatorname{trace}(I_{m+n} - \hat{H}_C) - \operatorname{trace}(I_m - \hat{H}_A)] \\
&= \frac{1}{2}\log\frac{|C'C|}{|A'A|} - \frac{n}{2} + \frac{n}{2} \\
&= \frac{1}{2}\sum_{i=1}^{p}\log(e_i + 1),
\end{aligned}$$

where $e_1, \ldots, e_p$ are the eigenvalues of $(A'A)^{-1}B'B$. Moreover,

$$\begin{aligned}
R_{KL}(\beta, \hat{p}_{plug\text{-}in}(y|\hat{\beta}_x)) &= \iint p(x|\beta)p(y|\beta)\log\frac{p(y|\beta)}{\hat{p}_{plug\text{-}in}(y|\hat{\beta}_x)}\,dxdy \\
&= \frac{1}{2}\iint p(x|\beta)p(y|\beta)[\|y - B\hat{\theta}_x\|^2 - \|y - B\theta\|^2]\,dxdy \\
&= \frac{1}{2}\int p(x|\beta)\|B\hat{\theta}_x - B\theta\|^2\,dx \\
&= \frac{1}{2}\operatorname{trace}(B(A'A)^{-1}B') \\
&= \frac{1}{2}\sum_{i=1}^{p} e_i.
\end{aligned}$$

That $\hat{p}_U$ dominates $\hat{p}_{plug\text{-}in}(y|\hat{\beta}_x)$ follows from the fact that $\log(x+1) \leq x$ for any $x > 0$. ■

Risk comparisons of a Bayes predictive density $\hat{p}_\pi$ with $\hat{p}_U$ are greatly facilitated by the following representation of $\hat{p}_\pi$ in terms of $\hat{p}_U$. An analogous representation of the posterior mean in terms of the maximum likelihood estimator (MLE), which simplifies multivariate normal mean estimation under quadratic risk, was proposed by Brown (1971). For our representation, it will be useful to denote the marginal distribution of $Z|\beta \sim N_p(\beta, \Sigma)$ under π by

$$m_\pi(z;\Sigma) = \int p(z|\beta)\pi(\beta)\,d\beta. \tag{9}$$

Thus, the marginal distributions of $\hat{\beta}_x$ and $\hat{\beta}_{x,y}$ under π are denoted by $m_\pi(\hat{\beta}_x, \Sigma_A)$ and $m_\pi(\hat{\beta}_{x,y}, \Sigma_C)$, respectively.

LEMMA 2. *If $m_\pi(z;\Sigma)$ is finite for all z and Σ, then $\hat{p}_\pi(y|x)$ is a proper probability distribution. Furthermore, it can be expressed as*

$$\hat{p}_\pi(y|x) = \frac{m_\pi(\hat{\beta}_{x,y}, \Sigma_C)}{m_\pi(\hat{\beta}_x, \Sigma_A)}\,\hat{p}_U(y|x), \tag{10}$$

where $\hat{p}_U$ is defined by (8).

Proof. When $m_\pi(z;\Sigma)$ is finite for all z and Σ, that $\hat{p}_\pi(y|x)$ is a proper probability distribution follows from integrating with respect to y and switching the order of integration.

Next, straightforward calculation yields

$$\begin{aligned}
&\int p(x|\beta)\pi(\beta)\,d\beta \\
&\quad = \int \frac{1}{(2\pi)^{m/2}} \exp\left\{-\frac{\|x - A\beta\|^2}{2}\right\}\pi(\beta)\,d\beta \\
&\quad = \int \frac{1}{(2\pi)^{m/2}} \exp\left\{-\frac{\|x - A\hat{\beta}_x\|^2 + \|A\hat{\beta}_x - A\beta\|^2}{2}\right\}\pi(\beta)\,d\beta \\
&\quad = \frac{1}{(2\pi)^{(m-p)/2}} \exp\left\{-\frac{\|x - A\hat{\beta}_x\|^2}{2}\right\} \\
&\qquad \times \int \frac{1}{(2\pi)^{p/2}} \exp\left\{-\frac{\|A\hat{\beta}_x - A\beta\|^2}{2}\right\}\pi(\beta)\,d\beta \\
&\quad = \frac{|A'A|^{-1/2}}{(2\pi)^{(m-p)/2}} \exp\left\{-\frac{RSS_x}{2}\right\} m_\pi(\hat{\beta}_x, \Sigma_A).
\end{aligned} \tag{11}$$

Similarly, we obtain

$$\int p(x|\beta)p(y|\beta)\pi(\beta)\,d\beta = \frac{|C'C|^{-1/2}}{(2\pi)^{(m+n-p)/2}} \exp\left\{-\frac{RSS_{x,y}}{2}\right\} m_\pi(\hat{\beta}_{x,y}, \Sigma_C). \tag{12}$$

The representation (10) follows immediately from (6), (11), and (12). ■

The next result provides a representation of the difference between the KL risks of $\hat{p}_U$ and $\hat{p}_\pi$ in terms of the marginal distributions of $\hat{\beta}_x$ and $\hat{\beta}_{x,y}$.

LEMMA 3. *The difference between the KL risks of $\hat{p}_U$ and $\hat{p}_\pi$ is given by*

$$R_{KL}(\beta, \hat{p}_U) - R_{KL}(\beta, \hat{p}_\pi) = E_{\beta,\Sigma_C} \log m_\pi(\hat{\beta}_{x,y}; \Sigma_C) - E_{\beta,\Sigma_A} \log m_\pi(\hat{\beta}_x; \Sigma_A), \tag{13}$$

where $E_{\beta,\Sigma}(\cdot)$ stands for expectation with respect to the $N_p(\beta,\Sigma)$ distribution.

Proof. The KL risk difference between $\hat{p}_U$ and $\hat{p}_\pi$ can be expressed as

$$\begin{aligned} R_{KL}(\beta, \hat{p}_U) - R_{KL}(\beta, \hat{p}_\pi) &= \iint p(x|\beta)p(y|\beta)\log \frac{\hat{p}_\pi(y|x)}{\hat{p}_U(y|x)}\,dxdy \\ &= \iint p(x|\beta)p(y|\beta)\log \frac{m_\pi(\hat{\beta}_{x,y}, \Sigma_C)}{m_\pi(\hat{\beta}_x, \Sigma_A)}\,dxdy, \end{aligned}$$

where the last equality follows from Lemma 2. The result then follows from the change of variable theorem. ■

To exploit the representation (13), we proceed to transform the distributions to canonical form. Because Σ_A and Σ_C are both symmetric and positive definite, there exists an invertible $p \times p$ matrix W such that

$$\Sigma_A = WW' \quad \text{and} \quad \Sigma_C = W\Sigma_D W', \tag{14}$$

where

$$\Sigma_D = \text{diag}(d_1, \ldots, d_p). \tag{15}$$

Because $\Sigma_C = (C'C)^{-1} = (A'A + B'B)^{-1}$ and $B'B$ is nonnegative definite, $d_i \in (0,1]$ for all $1 \le i \le p$ with at least one $d_i < 1$. Finally, let $\mu = W^{-1}\beta$, $\hat{\mu}_x = W^{-1}\hat{\beta}_x$, and $\hat{\mu}_{x,y} = W^{-1}\hat{\beta}_{x,y}$, so that

$$\hat{\mu}_x \sim N_p(\mu, I) \quad \text{and} \quad \hat{\mu}_{x,y} \sim N_p(\mu, \Sigma_D). \tag{16}$$

LEMMA 4. *Let* $\pi_W(\mu) = \pi(W\mu)$. *Then, the difference between the KL risks of* $\hat{p}_U$ *and* $\hat{p}_\pi$ *is given by*

$$R_{KL}(\beta, \hat{p}_U) - R_{KL}(\beta, \hat{p}_\pi) = E_{\mu,\Sigma_D} \log m_{\pi_W}(\hat{\mu}_{x,y}; \Sigma_D) - E_{\mu,I} \log m_{\pi_W}(\hat{\mu}_x; I), \tag{17}$$

where $E_{\mu,\Sigma}(\cdot)$ *stands for expectation with respect to the* $N_p(\mu, \Sigma)$ *distribution.*

Proof. The result follows by transforming the expressions in Lemma 3:

$$\begin{aligned} E_{\beta,\Sigma_A} \log m_\pi(\hat{\beta}_x; \Sigma_A) &= \int p(\hat{\beta}_x|\beta) \log \int p(\hat{\beta}_x|\beta)\pi(\beta)\, d\beta d\hat{\beta}_x \\ &= \int p(\hat{\mu}_x|\mu) \log \int p(\hat{\mu}_x|\mu)\pi_W(\mu)\, d\mu d\hat{\mu}_x \\ &= E_{\mu,I} \log m_{\pi_W}(\hat{\mu}_x; I). \end{aligned}$$

Similarly,

$$E_{\beta,\Sigma_C} \log m_\pi(\hat{\beta}_{x,y}; \Sigma_C) = E_{\mu,\Sigma_D} \log m_{\pi_W}(\hat{\mu}_{x,y}; \Sigma_D).$$

Thus, (17) equals (13). ■

We now proceed to find conditions on m_π for which the risk difference (17) is nonnegative for all μ. Because $\hat{p}_U$ is minimax, this will then imply that $\hat{p}_\pi$ is minimax under the prior π corresponding to π_W. Now for $w \in [0,1]$, let

$$V_w = wI + (1-w)\Sigma_D, \tag{18}$$

where Σ_D is defined as in (15). Next, for $Z \sim N_p(\mu, V_w)$, let

$$h_\mu(V_w) = E_{\mu,V_w} \log m_{\pi_W}(Z; V_w). \tag{19}$$

Thus, we may rewrite (17) as

$$R_{KL}(\beta, \hat{p}_U) - R_{KL}(\beta, \hat{p}_\pi) = h_\mu(V_0) - h_\mu(V_1). \tag{20}$$

Because $h_\mu(w)$ is continuous in w, it suffices to derive conditions on m_π such that $(\partial/\partial w)h_\mu(w) < 0$ for all μ and $w \in [0,1]$. Letting v_i be the ith diagonal element of V_w, we have by the chain rule

$$\frac{\partial}{\partial w} h_\mu = \sum_1^p \frac{\partial h_\mu}{\partial v_i} \frac{\partial v_i}{\partial w} = \sum_1^p (1 - d_i) \frac{\partial h_\mu}{\partial v_i}. \tag{21}$$

The following result provides unbiased estimates of the components of (21) that, when combined with (17), will be seen to play a key role in establishing sufficient conditions on m_π for $\hat{p}_\pi$ to be minimax and to dominate $\hat{p}_U$. As noted by George et al. (2006), these estimates are very similar to the unbiased esti-

mates of risk for the estimation of a multivariate mean under squared error loss; see Stein (1974, 1981).

LEMMA 5. *If $m_{\pi_W}(z;I)$ is finite for all z, then for any $0 \le w \le 1$, $m_{\pi_W}(z;V_w)$ is finite. Moreover,*

$$\frac{\partial}{\partial v_i} h_\mu = E_{\mu,V_w}\left(\frac{\dfrac{\partial^2}{\partial z_i^2} m_{\pi_W}(Z;V_w)}{m_{\pi_W}(Z;V_w)} - \frac{1}{2}\left(\frac{\partial}{\partial z_i} \log m_{\pi_W}(Z;V_w) \right)^2 \right) \tag{22}$$

$$= E_{\mu,V_w}\left(2\, \frac{\dfrac{\partial^2}{\partial z_i^2} \sqrt{m_{\pi_W}(Z;V_w)}}{\sqrt{m_{\pi_W}(Z;V_w)}} \right). \tag{23}$$

Proof. When $m_{\pi_W}(z;I)$ is finite for all z, it is easy to check that for any fixed z and any $0 \le w \le 1$,

$$m_{\pi_W}(z;V_w) \le \left(\prod_{i=1}^{k} d_i^{-1/2} \right) m_{\pi_W}(z;I) < \infty.$$

Next, letting $Z^* = V_w^{-1/2}(Z - \mu) \sim N(0,I)$, we have

$$\frac{\partial}{\partial v_i} h_\mu = \frac{\partial}{\partial v_i} E \log m_{\pi_W}(V_w^{1/2} Z^* + \mu; V_w)$$

$$= E\, \frac{\dfrac{\partial}{\partial v_i} m_{\pi_W}(V_w^{1/2} Z^* + \mu; V_w)}{m_{\pi_W}(V_w^{1/2} Z^* + \mu; V_w)}, \tag{24}$$

where

$$\frac{\partial}{\partial v_i} m_{\pi_W}(V_w^{1/2} z^* + \mu; V_w)$$

$$= \frac{\partial}{\partial v_i} \int \frac{1}{(2\pi)^{p/2} \sqrt{v_1 \ldots v_p}} \exp\left\{ -\sum_{i=1}^{p} \frac{(\sqrt{v_i} z_i^* + \mu_i - \mu_i')^2}{2v_i} \right\} \pi_W(\mu')\, d\mu'$$

$$= \int \left(-\frac{1}{2v_i} + \frac{(z_i - \mu_i')^2}{2v_i^2} - \frac{z_i^{*2}}{2v_i} - \frac{z_i^*(\mu_i - \mu_i')}{2v_i^{3/2}} \right) p(z|\mu')\pi_W(\mu')\, d\mu'$$

$$= \frac{\partial}{\partial v_i} m_{\pi_W}(z;V_w) - \int \frac{(z_i - \mu_i)(z_i - \mu_i')}{2v_i^2}\, p(z|\mu')\pi_W(\mu')\, d\mu'.$$

Making use of the well-known univariate heat equation

$$\frac{\partial}{\partial v_i} m_{\pi_W}(z;V_w) = \frac{1}{2}\frac{\partial^2}{\partial z_i^2} m_{\pi_W}(z;V_w) \tag{25}$$

(see, e.g., Steele, 2001, and the Brown, 1971, representation $E_\pi(\mu_i'|z_i) = z_i + v_i(\partial/\partial z_i)\log m_{\pi_W}(z)$), (22) and (23) can be verified via the same steps as in the proof of Lemma 3 in George et al. (2006). ■

Now we can obtain sufficient conditions for a Bayes procedure $\hat{p}_\pi$ to be minimax by combining (20), (21), and Lemma 5. The following result provides a substantial generalization of Theorem 1 of George et al. (2006).

THEOREM 1. *Suppose that $m_\pi(z;WW')$ is finite for all z with the invertible matrix W defined as in (14). Let $H(f(z_1,\ldots,z_p))$ be the Hessian matrix of f.*

(i) If $\operatorname{trace}\{H(m_\pi(z;WV_wW'))[\Sigma_A - \Sigma_C]\} \le 0$ *for all* $w \in [0,1]$, *then* $\hat{p}_\pi$ *is minimax under* R_{KL}. *Furthermore,* $\hat{p}_\pi$ *dominates* $\hat{p}_U$ *unless* $\pi = \pi_U$.

(ii) If $\operatorname{trace}\{H(\sqrt{m_\pi(z;WV_wW')})[\Sigma_A - \Sigma_C]\} \le 0$ *for all* $w \in [0,1]$, *then* $\hat{p}_\pi$ *is minimax under* R_{KL}. *Furthermore,* $\hat{p}_\pi$ *dominates* $\hat{p}_U$ *if for all* $w \in [0,1]$, *this inequality is strict on a set of positive Lebesgue measure.*

Proof. To prove the minimaxity of $\hat{p}_\pi$ under R_{KL}, it suffices to show that (22) or (23) is nonpositive because by (21) that would imply the nonnegativity of (20). Dominance would further follow by showing that (22) or (23) is also strictly negative on a set of positive Lebesgue measure.

Noting that $m_{\pi_W}(z;V_w) = m_\pi(Wz;WV_wW')$, and letting $Wz = \tilde{z}$, we obtain

$$\begin{aligned}
\sum_{i=1}^{k}(1-d_i)\frac{\partial^2}{\partial z_i^2} m_{\pi_W}(z;V_w) &= \sum_{i=1}^{k}(1-d_i)\frac{\partial^2}{\partial z_i^2} m_\pi(\tilde{z};WV_wW') \\
&= \sum_{i=1}^{k}(1-d_i)\sum_{j=1}^{p}\sum_{k=1}^{p} W_{ji}\frac{\partial^2 m_\pi(\tilde{z};WV_wW')}{\partial\tilde{z}_j\partial\tilde{z}_k} W_{ki} \\
&= \operatorname{trace}\{(I-\Sigma_D)W'H(m_\pi(\tilde{z};WV_wW'))W\} \\
&= \operatorname{trace}\{H(m_\pi(\tilde{z};WV_wW'))W(I-\Sigma_D)W'\} \\
&= \operatorname{trace}\{H(m_\pi(\tilde{z};WV_wW'))[\Sigma_A-\Sigma_C]\}. \qquad (26)
\end{aligned}$$

Similarly,

$$\sum_{i=1}^{k}(1-d_i)\frac{\partial^2}{\partial z_i^2}\sqrt{m_{\pi_W}(z;V_w)} = \operatorname{trace}\{H(\sqrt{m_\pi(\tilde{z};WV_wW')})[\Sigma_A-\Sigma_C]\}. \tag{27}$$

Now (i) and (ii) follow immediately from (22), (23), (26), and (27). ■

The next result follows using the fact that $(\partial^2/\partial z_i^2)m_{\pi_W}(z;V_w) \leq 0$ when $(\partial^2/\partial \mu_i^2)\pi_W(\mu) \leq 0$.

COROLLARY 1. *Suppose that $m_\pi(z;WW')$ is finite for all z. Then $\hat{p}_\pi$ will be minimax if*

$$\text{trace}\{H(\pi(\beta))[\Sigma_A - \Sigma_C]\} \leq 0 \quad a.e.$$

Furthermore, $\hat{p}_\pi$ will dominate $\hat{p}_U$ unless $\pi = \pi_U$.

Example (Scaled harmonic prior)

Suppose that $A = B$. In this case,

$$\begin{aligned}\text{trace}\{H(\pi(\beta))[\Sigma_A - \Sigma_C]\} &= \frac{1}{2}\text{trace}\{H(\pi(\beta))\Sigma_A\} \\ &= \frac{1}{2}\text{trace}\{H(\pi(\beta))WW'\} \\ &= \frac{1}{2}\nabla^2\pi_W(\mu). \end{aligned} \tag{28}$$

Let $\pi_W(\mu) \propto \|\mu\|^{-(p-2)}$ when $p \geq 3$ and $\pi_W(\mu) \propto 1$ when $p < 3$. Note that π_W is harmonic, i.e., $\nabla^2\pi_W(\mu) \equiv 0$, and not equal to π_U when $p \geq 3$. For $p \geq 3$, the corresponding prior on β is a "scaled harmonic prior"

$$\pi(\beta) \propto \|W^{-1}\beta\|^{-(p-2)} = \|\text{diag}(\eta_1^{-1/2},\ldots,\eta_p^{-1/2})\beta\|^{-(p-2)}, \tag{29}$$

where $\eta_1,\ldots,\eta_p > 0$ are the eigenvalues of Σ_A and for $p < 3$, $\pi(\beta) \propto 1$. (The expression (29) is obtained using the fact that there exists an orthonormal matrix O such that $W = O\ \text{diag}(\eta_1^{1/2},\ldots,\eta_p^{1/2})O'$.) By Corollary 1 and (28), the predictive estimator $\hat{p}_\pi$ under this prior is minimax and dominates $\hat{p}_U$ when $p \geq 3$. It is easy to check that these results hold when $A = rB$ for any known constant r.

3. PREDICTIVE DENSITY ESTIMATION NEAR SUBSET MODELS

When a prior centered at 0 such as the scaled harmonic prior (29) is applied to β, the risk reduction of $\hat{p}_\pi$ over $\hat{p}_U$ is greatest when all the components of β are close to 0. Thus, it would be sensible to use this prior if it was felt that all p predictors in A and B were potentially irrelevant. However, such a prior would be ineffectual if only a subset of the p predictors were irrelevant, in other words, if only a subset of the β components were close to 0. In this section, we extend our results for the setting where such a subset is known. This will set the stage

for Section 4, where we develop new results for the more realistic model uncertainty setting where such a subset is unknown.

Let S be the subset of $\{1,\ldots,p\}$ corresponding to the indices of the potentially irrelevant predictors and let $q_S = |S|$ be the number of elements in S. Let β_S be the subvector of β corresponding to the columns of A indexed by S. Similarly, let $\hat{\beta}_{S,x}$ and $\hat{\beta}_{S,x,y}$ be the subvectors of $\hat{\beta}_x$ and $\hat{\beta}_{x,y}$, respectively, corresponding to β_S. Finally, for notational convenience, let $\Sigma_{A,S}$ and $\Sigma_{C,S}$ be the submatrices of Σ_A and Σ_C, respectively, which are the covariance matrices of $\hat{\beta}_{S,x}$ and $\hat{\beta}_{S,x,y}$.

When only the elements of β_S are thought to be close to zero, it would be sensible to consider a prior that is uniform on $\beta_{\bar{S}}$, where $\bar{S}$ is the complement of S. We denote such a prior by π_S and let π_S^* be the restriction of π_S to β_S so that

$$\pi_S(\beta) = \pi_S^*(\beta_S) \tag{30}$$

is a function of β_S only. To exploit the possibility that β_S is close to zero, π_S^* would then be centered around 0.

LEMMA 6. *If $m_{\pi_S}(z;\Sigma)$ is finite for all z and Σ, then $\hat{p}_{\pi_S}(y|x)$ is a proper probability distribution. Furthermore, it can be expressed as*

$$\hat{p}_{\pi_S}(y|x) = \frac{m_{\pi_S^*}(\hat{\beta}_{S,x,y},\Sigma_{C,S})}{m_{\pi_S^*}(\hat{\beta}_{S,x},\Sigma_{A,S})}\,\hat{p}_U(y|x), \tag{31}$$

where $\hat{p}_U$ is defined by (8).

Proof. The first assertion was proved in Lemma 2. Next, proceeding as in the derivation of (11), we obtain

$$\begin{aligned}
&\int p(x|\beta)\pi_S(\beta)\,d\beta \\
&\quad = \frac{1}{(2\pi)^{(m-p)/2}}\exp\left\{-\frac{\|x-A\hat{\beta}_x\|^2}{2}\right\} \\
&\qquad \times \int \frac{1}{(2\pi)^{p/2}}\exp\left\{-\frac{\|A\hat{\beta}_x - A\beta\|^2}{2}\right\}\pi_S^*(\beta_S)\,d\beta \\
&\quad = \frac{|A'A|^{-1/2}}{(2\pi)^{(m-p)/2}}\exp\left\{-\frac{RSS_x}{2}\right\} m_{\pi_S^*}(\hat{\beta}_{S,x},\Sigma_{A,S}).
\end{aligned} \tag{32}$$

Similarly, we obtain

$$\int p(x|\beta)p(y|\beta)\pi_S(\beta)\,d\beta$$

$$= \frac{|C'C|^{-1/2}}{(2\pi)^{(m+n-p)/2}} \exp\left\{-\frac{RSS_{x,y}}{2}\right\} m_{\pi_S^*}(\hat{\beta}_{S,x,y}, \Sigma_{C,S}). \tag{33}$$

The representation (31) follows immediately from (6), (32), and (33). ■

The following results provide sufficient conditions for the minimaxity of $\hat{p}_{\pi_S}$ and for its dominance over $\hat{p}_U$. We omit the proofs, which are obtained using the same arguments leading to Theorem 1 and Corollary 1. Analogously to our previous development there, we let W_S be an invertible $q_S \times q_S$ matrix such that $\Sigma_{A,S} = W_S W_S'$ and $\Sigma_{C,S} = W\Sigma_{D,S}W'$, where $\Sigma_{D,S} = \text{diag}(d_1, \ldots, d_{q_S})$ as in (15). Finally, let $V_{S,w} = wI + (1-w)\Sigma_D$ as in (18).

THEOREM 2. *Suppose that $m_{\pi_S^*}(z; W_S W_S')$ is finite for all z. Let $H(f(z_1, \ldots, z_{q_S}))$ be the Hessian matrix of f.*

(i) If $\text{trace}\{H(m_{\pi_S^*}(z; W_S V_{S,w} W_S'))[\Sigma_{A,S} - \Sigma_{C,S}]\} \leq 0$ *for all* $w \in [0,1]$, *then $\hat{p}_{\pi_S}$ is minimax under R_{KL}. Furthermore, $\hat{p}_{\pi_S}$ dominates $\hat{p}_U$ unless $\pi_S = \pi_U$.*

(ii) If $\text{trace}\{H(\sqrt{m_{\pi_S^*}(z; W_S V_{S,w} W_S')})[\Sigma_{A,S} - \Sigma_{C,S}]\} \leq 0$ *for all* $w \in [0,1]$, *then $\hat{p}_{\pi_S}$ is minimax under R_{KL}. Furthermore, $\hat{p}_{\pi_S}$ dominates $\hat{p}_U$ if for all $w \in [0,1]$, this inequality is strict on a set of positive Lebesgue measure.*

COROLLARY 2. *Suppose that $m_{\pi_S^*}(z; W_S W_S')$ is finite for all z. Then $\hat{p}_{\pi_S}$ will be minimax if*

$$\text{trace}\{H(\pi_S^*(\beta_S))[\Sigma_{A,S} - \Sigma_{C,S}]\} \leq 0 \quad a.e.$$

Furthermore, $\hat{p}_{\pi_S}$ will dominate $\hat{p}_U$ unless $\pi_S = \pi_U$.

Example (continued) (Scaled harmonic prior)

Suppose that $A = B$ so that as in (28),

$$\text{trace}\{H(\pi_S^*(\beta_S))[\Sigma_{A,S} - \Sigma_{C,S}]\} = \frac{1}{2}\nabla^2 \pi_{W_S}(\mu_S), \tag{34}$$

where $\mu_S = W_S^{-1}\beta_S$. Here let $\pi_{W_S}(\mu) \propto \|\mu_S\|^{-(q_S-2)}$ when $q_S \geq 3$ and $\pi_W(\mu_S) \propto 1$ when $q_S < 3$. As before, π_{W_S} is harmonic, i.e., $\nabla^2 \pi_{W_S}(\mu) \equiv 0$, and not equal to π_U when $q_S \geq 3$. For $q_S \geq 3$, the corresponding scaled harmonic prior on β is

$$\pi_S(\beta) = \pi_S^*(\beta_S) \propto \|W_S^{-1}\beta_S\|^{-(q_S-2)} = \|\text{diag}(\eta_1^{-1/2},\ldots,\eta_{q_S}^{-1/2})\beta_S\|^{-(q_S-2)}, \quad (35)$$

where $\eta_1,\ldots,\eta_{q_S} > 0$ are the eigenvalues of $\Sigma_{A,S}$ and for $q_S < 3$, $\pi_S(\beta) \propto 1$. By Corollary 2 and (34), $\hat{p}_{\pi_S}$ here is minimax and dominates $\hat{p}_U$ when $q_S \geq 3$.

4. MINIMAX MULTIPLE SHRINKAGE PREDICTIVE ESTIMATION

We consider the more realistic model uncertainty setting where there is uncertainty about which subset of the p predictors in A and B should be included in the model. For each choice of S, we have obtained general sufficient conditions for $\hat{p}_{\pi_S}$ to be minimax and to dominate π_U. However, such $\hat{p}_{\pi_S}$ will only offer meaningful risk reduction when β is near the region where π_S is largest. For example, under the scaled harmonic prior in (35), such risk reduction occurs when β_S is close to 0. The difficulty then is how to proceed when the subset of irrelevant predictors indexed by S is unknown. Rather than arbitrarily selecting S, an attractive alternative is to use a multiple shrinkage predictive estimator that uses the data to adaptively emulate the most effective $\hat{p}_{\pi_S}$.

The multiple shrinkage procedure here is obtained by using a finite mixture of the contemplated priors. A similar multiple shrinkage construction for parameter estimation under squared error loss was proposed and developed by George (1986a, 1986b, 1986c). Let Ω be the set of all the subsets S under consideration, possibly even the set of all possible subsets. For each $S \in \Omega$, let π_S be the designated prior of the form (30) on β and assign it probability $w_S \in [0,1]$ such that $\sum_{S\in\Omega} w_S = 1$. Thus we construct the mixture prior

$$\pi^*(\beta) = \sum_{S\in\Omega} w_S \pi_S(\beta). \quad (36)$$

This prior yields the multiple shrinkage predictive estimator

$$\hat{p}^*(y|x) = \sum_{S\in\Omega} \hat{p}(S|x)\hat{p}_{\pi_S}(y|x). \quad (37)$$

Here each $\hat{p}_{\pi_S}$ is given by (31) in Lemma 6, and each posterior probability is of the form

$$\hat{p}(S|x) = \frac{w_S m_{\pi_S^*}(\hat{\beta}_{S,x}, \Sigma_{A,S})}{\sum_{S\in\Omega} w_S m_{\pi_S^*}(\hat{\beta}_{S,x}, \Sigma_{A,S})}, \quad (38)$$

which follows from (32).

The form (37) reveals $\hat{p}^*(y|x)$ to be an adaptive convex combination of the individual shrinkage predictive estimates $\hat{p}_{\pi_S}$. Note that through $\hat{p}(S|x)$, $\hat{p}^*$ doubly shrinks $\hat{p}_U(y|x)$ by putting more weight on the $\hat{p}_{\pi_S}$ for which $m_{\pi_S^*}$ is

largest and hence $\hat{p}_{\pi_S}$ is shrinking most. Thus, we expect $\hat{p}^*$ to offer meaningful risk reduction whenever β_S is near the region where π_S is largest for any $S \in \Omega$. For example, if every π_S in π^* is one of the scaled harmonic priors in (35), such risk reduction occurs when β_S is close to 0 for any $S \in \Omega$ for which $q_S \geq 3$. Thus, the potential for risk reduction using $\hat{p}^*$ is far greater than the risk reduction using an arbitrarily chosen $\hat{p}_{\pi_S}$.

We should also note that the allocation of risk reduction by $\hat{p}^*$ is in part determined by the w_S weights in (38). Because each $\hat{p}(S|x)$ is so adaptive through $m_{\pi_S^*}$, choosing the weights to be uniform should be adequate. However, one may also want to consider some of the more refined suggestions for choosing such weights for the multiple shrinkage estimators in George (1986b).

The potential for a multiple shrinkage $\hat{p}^*$ to offer meaningful risk reduction in many different regions of the parameter space is greatly enhanced when it is minimax and therefore can only improve on the "noninformative" minimax $\hat{p}_U$. The following two results show that such minimaxity and dominance of $\hat{p}_U$ can be obtained. We then conclude with an explicit example of such domination.

THEOREM 3. *Suppose for all $S \in \Omega$, $m_{\pi_S^*}(z; W_S W_S')$ is finite for all z. Let $H(f(z_1, \ldots, z_{q_S}))$ be the Hessian matrix of f. If for all $S \in \Omega$,*

$$\text{trace}\{H(m_{\pi_S^*}(z; W_S V_{S,w} W_S'))[\Sigma_{A,S} - \Sigma_{C,S}]\} \leq 0 \quad \text{for all } w \in [0,1], \tag{39}$$

then $\hat{p}^*$ in (37) is minimax under R_{KL}. Furthermore, $\hat{p}^*$ dominates $\hat{p}_U$ unless $\pi^* = \pi_U$.

Proof. From (31), (37), and (38), it is straightforward to show that $\hat{p}^*$ can be reexpressed as

$$\hat{p}^*(y|x) = \frac{\sum_{S \in \Omega} w_S m_{\pi_S^*}(\hat{\beta}_{S,x,y}, \Sigma_{C,S})}{\sum_{S \in \Omega} w_S m_{\pi_S^*}(\hat{\beta}_{S,x}, \Sigma_{A,S})} \hat{p}_U(y|x). \tag{40}$$

Because p^* is of the same form as $\hat{p}_{\pi_S}$ in (31), namely, a ratio of marginals times $\hat{p}_U$, we can apply the same arguments leading to the proofs of Theorems 1 and 2. These steps show that a sufficient condition for the minimaxity and dominance claims is

$$\left\{ \sum_{S \in \Omega} w_S H(m_{\pi_S^*}(z; W_S V_{S,w} W_S'))[\Sigma_{A,S} - \Sigma_{C,S}] \right\} \leq 0 \quad \text{for all } w \in [0,1].$$

This condition is implied if (39) holds for all $S \in \Omega$. ■

The next result follows using the same argument leading to Corollaries 1 and 2.

COROLLARY 3. *Suppose for all* $S \in \Omega$, $m_{\pi_S^*}(z; W_S W_S')$ *is finite for all* z. *Then* $\hat{p}^*$ *in (37) will be minimax if for all* $S \in \Omega$

$$\text{trace}\{H(\pi_S^*(\beta_S))[\Sigma_{A,S} - \Sigma_{C,S}]\} \leq 0 \quad a.e.$$

Furthermore, $\hat{p}^*$ *will dominate* $\hat{p}_U$ *unless* $\pi = \pi_U$.

Example (continued) (Scaled harmonic prior)

For each $S \in \Omega$, let $\pi_S(\beta)$ be the scaled harmonic prior given by (35) when $q_S \geq 3$ and by $\pi_S(\beta) \propto 1$ when $q_S < 3$. When $A = B$, by Corollary 3, $\hat{p}^*$ under these priors will be minimax and will dominate $\hat{p}_U$ if $q_S \geq 3$ for at least one $S \in \Omega$.

5. PREDICTIVE DENSITY ESTIMATION NEAR LINEAR SUBSPACES

The harmonic prior predictive estimator $\hat{p}_{\pi_S}(y|x)$ described in Section 3 and incorporated into the multiple shrinkage predictive estimators $\hat{p}^*(y|x)$ in Section 4 offers risk reduction in the region of the parameter space where β_S is close to 0. This can be seen as a special case of the following general construction of a predictive estimator that obtains risk reduction when β is close to a linear subspace of R^p.

Suppose one would like to obtain a predictive density estimator with greatest risk reduction in the region where β is close to a linear subspace $G \subset R^p$. In the case of $\hat{p}_{\pi_S}(y|x)$, G would be the subspace of all $\beta \in R^p$ for which $\beta_S \equiv 0$. Alternately, if risk reduction was desired, say, when the components of β were close to equal, then one would consider $G = [1]$, the subspace spanned by $(1, \ldots, 1)'$. Let $P_G\beta \equiv \text{argmin}_{g \in G}\|\beta - g\|$ be the projection of β onto G and define $\beta_G \equiv (I - P_G)\beta$ to be the projection of β onto the orthogonal complement of G. For the construction of $\hat{p}_{\pi_S}(y|x)$ in Section 3, $\beta_G = \beta_S$. For $G = [1]$, $\beta_G = (\beta - \bar{\beta})$ where $\bar{\beta}$ is the vector of components all equal to $(1/p)\sum_{i=1}^p \beta_i$.

The main idea behind the general construction is to use a prior that leads to shrinkage of β_G toward 0 while leaving the remainder of β untouched. This can be obtained by using a prior of the form

$$\pi_G(\beta) = \pi_G^*(\beta_G), \tag{41}$$

which is effectively uniform on $(\beta - \beta_G)$. This is a special case of the prior over β_S in (30). Note that because β_G is $q_G \equiv (p - \dim(G))$ dimensional, π_G^* is a function from R^{q_G} to R.

Analogous to the construction in Lemma 6, predictive density estimators $\hat{p}_{\pi_G}$ corresponding to priors of the form π_G in (41) can be expressed as

$$\hat{p}_{\pi_G}(y|x) = \frac{m_{\pi_G^*}(\hat{\beta}_{G,x,y}, \Sigma_{C,G})}{m_{\pi_G^*}(\hat{\beta}_{G,x}, \Sigma_{A,G})}\, \hat{p}_U(y|x), \tag{42}$$

where $\hat{p}_U$ is defined by (8), $\hat{\beta}_{G,x} = (I - P_G)\hat{\beta}_x$ and $\hat{\beta}_{G,x,y} = (I - P_G)\hat{\beta}_{x,y}$ are the projections of $\hat{\beta}_x$ and $\hat{\beta}_{x,y}$ onto the orthogonal complement of G, respectively, and $\Sigma_{A,G}$ and $\Sigma_{C,G}$ are the covariance matrices of $\hat{\beta}_{G,x}$ and $\hat{\beta}_{G,x,y}$, respectively. It is straightforward to see that Theorem 2 and Corollary 2 and their proofs can be extended to obtain conditions on $\pi_G^*(\beta_G)$ for such $\hat{p}_{\pi_G}$ to be minimax and to dominate $\hat{p}_U$. (Simply substitute the symbol G for the symbol S throughout.)

Example (continued)

Extending (35), consider the following scaled harmonic prior on β. For $q_G \geq 3$, let

$$\pi_G(\beta) = \pi_G^*(\beta_G) \propto \|\mathrm{diag}(\eta_1^{-1/2}, \ldots, \eta_{q_G}^{-1/2})\beta_G\|^{-(q_G-2)}, \tag{43}$$

where $\eta_1, \ldots, \eta_{q_G} > 0$ are the eigenvalues of $\Sigma_{A,G}$, and for $q_G < 3$, let $\pi_G(\beta) \propto 1$. Note that when $q_G \geq 3$ the resulting $\hat{p}_{\pi_G}$ shrinks $\hat{p}_U$ toward G, offering reduced risk when β is close to G. By the extension of Corollary 2, such $\hat{p}_{\pi_G}$ will be minimax and dominate $\hat{p}_U$ when $A = B$ and $q_G \geq 3$.

Finally, following the development in Section 4 one can easily incorporate such $\hat{p}_{\pi_G}$ into multiple shrinkage predictor estimators $\hat{p}^*$. Letting Ω be a set of subspaces G under consideration, construct the mixture prior

$$\pi^*(\beta) = \sum_{G \in \Omega} w_G \pi_G(\beta), \tag{44}$$

where for each $G \in \Omega$, π_G is the designated prior of the form (41) and $w_G \in [0,1]$ is such that $\sum_{G \in \Omega} w_G = 1$. This prior yields the multiple shrinkage predictive estimator

$$\hat{p}^*(y|x) = \sum_{G \in \Omega} \hat{p}(G|x)\hat{p}_{\pi_G}(y|x), \tag{45}$$

where each $\hat{p}_{\pi_G}$ is given by (42) and each posterior probability is of the form

$$\hat{p}(G|x) = \frac{w_G m_{\pi_G^*}(\hat{\beta}_{G,x}, \Sigma_{A,G})}{\sum_{G \in \Omega} w_G m_{\pi_G^*}(\hat{\beta}_{G,x}, \Sigma_{A,G})}. \tag{46}$$

Here, $\hat{p}^*(y|x)$ is an adaptive convex combination of the individual shrinkage predictive estimates $\hat{p}_{\pi_G}$ and offers risk reduction whenever β_G is near the region where π_G is largest for any $G \in \Omega$. Thus, the potential for risk reduction using $\hat{p}^*$ is far greater than the risk reduction using an arbitrarily chosen $\hat{p}_{\pi_G}$. It is straightforward to see that Theorem 3 and Corollary 3 and their proofs can be extended to get conditions for such $\hat{p}^*(y|x)$ to be minimax and dominate $\hat{p}_U$. (Simply substitute the symbol G for the symbol S throughout.)

Example (continued) (Scaled harmonic prior)

For each $G \in \Omega$, let $\pi_G(\beta)$ be the scaled harmonic prior given by (43) when $q_G \geq 3$ and by $\pi_G(\beta) \propto 1$ when $q_G < 3$. When $A = B$, by the extension of Corollary 3, $\hat{p}^*$ for these priors will be minimax and will dominate $\hat{p}_U$ if $q_G \geq 3$ for at least one $G \in \Omega$.

REFERENCES

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

Brown, L.D. (1971) Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Annals of Mathematical Statistics* 42, 855–903.

Brown, L.D., E.I. George, & X. Xu (2007) Admissible predictive density estimation. *Annals of Statistics*, forthcoming.

George, E.I. (1986a) Minimax multiple shrinkage estimation. *Annals of Statistics* 14, 188–205.

George, E.I. (1986b) Combining minimax shrinkage estimators. *Journal of the American Statistical Association* 81, 437–445.

George, E.I. (1986c) A formal Bayes multiple shrinkage estimator. *Communications in Statistics: Part A—Theory and Methods*, special issue, *Stein-Type Multivariate Estimation* 15, 2099–2114.

George, E.I., F. Liang, & X. Xu (2006) Improved minimax prediction under Kullback-Leibler Loss. *Annals of Statistics* 34, 78–91.

Liang, F. (2002) Exact minimax procedures for predictive density estimation and data compression. Ph.D. Dissertation, Department of Statistics, Yale University.

Liang, F. & A. Barron (2004) Exact minimax strategies for predictive density estimation, data compression and model selection. *IEEE Information Theory Transactions* 50, 2708–2726.

Steele, J.M. (2001) *Stochastic Calculus and Financial Applications*. Springer-Verlag.

Stein, C. (1974) Estimation of the mean of a multivariate normal distribution. In J. Hajek (ed.), *Proceedings of the Prague Symposium on Asymptotic Statistics*, pp. 345–381. Universita Karlova.

Stein, C. (1981) Estimation of a multivariate normal mean. *Annals of Statistics* 9, 1135–1151.