From Minimax Shrinkage Estimation to Minimax Shrinkage Prediction

Edward I. George, Feng Liang and Xinyi Xu *

December 5, 2011

Abstract

In a remarkable series of papers beginning in 1956, Charles Stein set the stage for the future development of minimax shrinkage estimators of a multivariate normal mean under quadratic loss. More recently, parallel developments have seen the emergence of minimax shrinkage estimators of multivariate normal predictive densities under Kullback-Leibler risk. We here describe these parallels emphasizing the focus on Bayes procedures and the derivation of the superharmonic conditions for minimaxity as well as further developments of new minimax shrinkage predictive density estimators including multiple shrinkage estimators, empirical Bayes estimators, normal linear model regression estimators and non-parametric regression estimators.

Keywords and Phrases: asymptotic minimaxity; Bayesian prediction; empirical Bayes; inadmissibility; multiple shrinkage; prior distributions; superharmonic marginals; unbiased estimates of risk.

*Edward I. George is Professor in the Department of Statistics, The Wharton School, 3730 Walnut Street 400 JMHH, Philadelphia, PA 19104-6340, edgeorge@wharton.upenn.edu. Feng Liang is Associate Professor in the Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright Street, Champaign, IL 61820, liangf@illinois.edu. Xinyi Xu is Assistant Professor in the Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210-1247, xinyi@stat.osu.edu. *This work was supported by NSF grants and DMS-0732276 and DMS-0907070. The authors are grateful for the helpful comments and clarifications of an anonymous referee.*

1 The Beginning of the Hunt for Minimax Shrinkage Estimators

Perhaps the most basic estimation problem in Statistics is the canonical problem of estimating a multivariate normal mean. Based on the observation of a p-dimensional multivariate normal random variable

$$X \mid \mu \sim N_p(\mu, I), \tag{1}$$

the problem is to find a suitable estimator $\hat{\mu}(x)$ of μ. The celebrated result of Stein (1956) dethroned $\hat{\mu}_{MLE}(x) = x$, the maximum likelihood and best location invariant estimator for this problem, by showing that, when $p \geq 3$, $\hat{\mu}_{MLE}$ is inadmissible under quadratic loss

$$R_Q(\mu, \hat{\mu}) = E_\mu \|\hat{\mu}(X) - \mu\|^2. \tag{2}$$

From a decision theory point of view, an important part of the appeal of $\hat{\mu}_{MLE}$ was the protection offered by its minimax property. The worst possible risk R_Q incurred by $\hat{\mu}_{MLE}$ was no worse than the worst possible risk of any other estimator. Stein's result implied the existence of even better estimators that offered the same minimax protection. He had begun the hunt for these better minimax estimators.

In a remarkable series of follow up papers Stein proceeded to set the stage for this hunt. James and Stein (1961) proposed a new closed-form minimax shrinkage estimator

$$\hat{\mu}_{JS}(x) = \left(1 - \frac{p-2}{\|x\|^2}\right) x, \tag{3}$$

the now well-known James-Stein estimator, and showed explicitly that its risk was less than $R_Q(\mu, \hat{\mu}_{MLE}) \equiv p$ for every value of μ when $p \geq 3$, that is, it uniformly dominated $\hat{\mu}_{MLE}$. The appeal of $\hat{\mu}_{JS}$ under R_Q was compelling. It offered the same guaranteed minimax protection as $\hat{\mu}_{MLE}$ while also offering the possibility of doing much better.

Stein (1962), though primarily concerned with improved confidence regions, described a parametric empirical Bayes motivation for (3), describing how $\hat{\mu}_{JS}(x)$ could be seen as a data based approximation to the posterior mean

$$E_\pi(\mu \mid x) = \left(1 - \frac{1}{1+\nu}\right) x, \tag{4}$$

the Bayes rule which minimizes the average risk $E_\pi R_Q(\mu, \hat{\mu})$ when $\mu \sim N_p(0, \nu I)$. He here also proposed the positive-part James-Stein estimator $\hat{\mu}_{JS+} = \max\{0, \hat{\mu}_{JS}\}$, a dominating improvement over $\hat{\mu}_{JS}(x)$, and commented that "it would be even better to use the Bayes estimate with respect to a reasonable prior distribution". These observations served as a clear indication that the Bayesian paradigm was to play a major role in the hunt for these new shrinkage estimators, opening up a new direction that was to be ultimately successful for establishing large new classes of shrinkage estimators.

Dominating fully Bayes shrinkage estimators soon emerged. Strawderman (1971) proposed $\hat{\mu}_a(x) = E_{\pi_a}(\mu \mid x)$, a class of Bayes shrinkage estimators obtained as posterior means under priors $\pi_a(\mu)$ for which

$$\mu \mid s \sim N_p\,(0, s\,I)\,, \quad s \sim (1+s)^{a-2}. \tag{5}$$

Strawderman explicitly showed that $\hat{\mu}_a$ uniformly dominated $\hat{\mu}_{MLE}$ and was proper Bayes, when $p = 5$ and $a \in [.5, 1)$ or when $p \geq 6$ and $a \in [0, 1)$. This was especially interesting because any proper Bayes was necessarily admissible and so could not be improved upon.

Then, Stein (1974, 1981) showed that $\hat{\mu}_H(x)$, the Bayes estimator under the harmonic prior

$$\pi_H(\mu) = E_{\pi_H}(\mu \mid x) = \|\mu\|^{-(p-2)}, \tag{6}$$

dominated $\hat{\mu}_{MLE}$ when $p \geq 3$. A special case of $\hat{\mu}_a$ when $a = 2$, $\hat{\mu}_H$ was only formal Bayes because $\pi_H(\mu)$ is improper. Undeterred, Stein pointed out that the admissibility of $\hat{\mu}_H$ followed immediately from the general conditions for the admissibility of generalized Bayes estimators laid out by Brown (1971). A further key element of the story, was Brown's (1971) powerful result that all such generalized Bayes rules (including the proper ones of course) constituted a complete class for the problem of estimating multivariate normal mean under quadratic loss. It was now clear that the hunt for new minimax shrinkage estimators was to focus on procedures with at least some Bayesian motivation.

Perhaps even more impressive than the fact that $\hat{\mu}_H$ dominated $\hat{\mu}_{MLE}$ was the way Stein proved it. Making further use of the rich results in Brown (1971), the key to his proof was the fact that any posterior mean Bayes estimator under a prior $\pi(\mu)$ can be expressed as

$$\hat{\mu}_\pi(x) = E_\pi(\mu \mid x) = x + \nabla \log m_\pi(x) \tag{7}$$

where

$$m_\pi(x) \propto \int e^{-(x-\mu)^2/2}\,\pi(\mu)\,d\mu \tag{8}$$

is the marginal distribution of X under $\pi(\mu)$. (Here $\nabla = (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_p})'$ is the familiar gradient).

At first glance it would appear that (7) has little to do with the risk. However, Stein noted that insertion of (7) into R_Q, followed by expansion and an integration-by-parts identity, now known as one of Stein's Lemmas, yields the following general expression for the difference between the risks of $\hat{\mu}_\pi$ and $\hat{\mu}_{MLE}$,

$$\begin{aligned} R_Q(\mu, \hat{\mu}_{MLE}) - R_Q(\mu, \hat{\mu}_\pi) &= E_\mu\left[\|\nabla \log m_\pi(X)\|^2 - 2\frac{\nabla^2 m_\pi(X)}{m_\pi(X)}\right] &(9)\\ &= E_\mu\left[-4\nabla^2\sqrt{m_\pi(X)}/\sqrt{m_\pi(X)}\right] &(10)\end{aligned}$$

(Here $\nabla^2 = \sum_i \frac{\partial^2}{\partial x_i^2}$ is the familiar Laplacian).

Because the bracketed terms in (9) and (10) do not depend on μ, (they are unbiased estimators of the risk difference), the domination of $\hat{\mu}_{MLE}$ by $\hat{\mu}_\pi$ would follow whenever m_π was such that these bracketed terms were nonnegative. As Stein noted, this would be the case in (9) whenever m_π was superharmonic, $\nabla^2 m_\pi(x) \leq 0$, and in (10) whenever $\sqrt{m_\pi}$ was superharmonic, $\nabla^2 \sqrt{m_\pi(x)} \leq 0$, a weaker condition.

The domination of $\hat{\mu}_{MLE}$ by $\hat{\mu}_H$ was seen now to be attributable directly to the fact that the marginal (8) under π_H, a mixture of harmonic functions, is superharmonic when $p \geq 3$. However, such an explanation wouldn't work for the domination of $\hat{\mu}_{MLE}$ by $\hat{\mu}_a$, because the marginal (8) under π_a in (5) is not superharmonic for any $a < 1$. Indeed, as was shown later by Fourdrinier, Strawderman and Wells (1998), a super harmonic marginal cannot be obtained with any proper prior. More importantly however, they were able to establish that the domination by $\hat{\mu}_a$ was attributable to the superharmonicity of $\sqrt{m_{\pi_a}}$ under π_a when $p \geq 5$ (and Strawderman's conditions on a). In fact, it also followed from their results that $\sqrt{m_{\pi_a}}$ is superharmonic when $a \in [1, 2)$ and $p \geq 3$, further broadening the class of minimax improper Bayes estimators.

Prior to the appearance of (9) and (10), minimaxity proofs, though ingenious, had all been tailored to suit the specific estimators at hand. The sheer generality of this new approach was daunting in its scope. By restricting attention to priors that gave rise to marginal distributions with particular properties, the minimax properties of the implied Bayes rules would be guaranteed.

2 The Parallels in the Predictive Estimation Problem Emerge

The seminal work of Stein concerned the canonical problem of how to estimate μ based on an observation of $X \,|\, \mu \sim N_p(\mu, I)$. A more ambitious problem is how to use such an X to estimate the entire probability distribution of a future Y from a normal distribution with this same unknown mean μ, the so-called predictive density of Y. Such a predictive density offers a complete description of predictive uncertainty.

To conveniently treat the possibility of different variances for X and Y, we formulate the predictive problem as follows. Suppose $X \,|\, \mu \sim N_p(\mu, v_x I)$ and $Y \,|\, \mu \sim N_p(\mu, v_y I)$ are independent p-dimensional multivariate normal vectors with common unknown mean μ but known variances v_x and v_y. Letting $p(y \,|\, \mu)$ denote the density of Y, the problem is to find an estimator $\hat{p}(y \,|\, x)$ of $p(y \,|\, \mu)$ based on the observation of $X = x$ only. Such a problem arises naturally, for example, for predicting $Y \,|\, \mu \sim N_p(\mu, \sigma^2 I)$ based on the observation of $X_1, \ldots, X_n \,|\, \mu$ iid $\sim N_p(\mu, \sigma^2 I)$ which is equivalent to observing $\bar{X} \,|\, \mu \sim N_p(\mu, (\sigma^2/n) I)$. This is exactly our formulation with $v_x = \sigma^2/n$ and $v_y = \sigma^2$.

For the evaluation of $\hat{p}(y \,|\, x)$ as an estimator of $p(y \,|\, \mu)$, the analogue of quadratic risk R_Q for

the mean estimation problem is the Kullback-Leibler (KL) risk

$$R_{KL}(\mu, \hat{p}) = \int p(x \mid \mu) L(\mu, \hat{p}(\cdot \mid x)) dx, \tag{11}$$

where $p(x \mid \mu)$ denotes the density of X, and

$$L(\mu, \hat{p}(\cdot \mid x)) = \int p(y \mid \mu) \log \frac{p(y \mid \mu)}{\hat{p}(y \mid x)} dy. \tag{12}$$

is the familiar KL loss.

For a (possibly improper) prior distribution π on μ, the average risk $r(\pi, \hat{p}) = \int R_{KL}(\mu, \hat{p}) \pi(\mu) d\mu$ is minimized by the Bayes rule

$$\hat{p}_\pi(y \mid x) = E_\pi[p(y \mid \mu) \mid x] = \int p(y \mid \mu) \pi(\mu \mid x) d\mu, \tag{13}$$

the posterior mean of $p(y|\mu)$ under π, (Aitchison 1975). It follows from (13) that $\hat{p}_\pi(y|x)$ is a proper probability distribution over y whenever the marginal density of x is finite for all z, (integrate wrt y and switch the order of integration). Furthermore, the mean of $\hat{p}_\pi(y \mid x)$ (when it exists) is equal to $E_\pi(\mu \mid x)$, the Bayes rule for estimating μ under quadratic loss, namely the posterior mean of μ. Thus, $\hat{p}_\pi$ also carries the necessary information for that estimation problem. Note also that unless π is a trivial point prior, such $\hat{p}_\pi(y \mid x)$ will not be of the form of $p(y \mid \mu)$ for any μ. The range of the Bayes rules here falls outside the target space of the densities which are being estimated.

A tempting initial approach to this predictive density estimation problem is to use the simple plug-in estimator $\hat{p}_{MLE} \equiv p(y \mid \mu = \hat{\mu}_{MLE})$ to estimate $p(y \mid \mu)$, the so-called estimative approach. This was the conventional wisdom until the appearance of Aitchison (1975). He showed that the plug-in estimator $\hat{p}_{MLE}$ is uniformly dominated under R_{KL} by

$$\hat{p}_U(y \mid x) \equiv E_{\pi_U}[p(y \mid \mu) \mid x] = \frac{1}{\{2\pi(v_x + v_y)\}^{\frac{p}{2}}} \exp\left\{ -\frac{\|y - x\|^2}{2(v_x + v_y)} \right\}, \tag{14}$$

the posterior mean of $p(y \mid \mu)$ with respect to the uniform prior $\pi_U(\mu) = 1$, the so-called predictive approach. In a related vein, Akaike (1978) pointed out that, by Jensen's inequality, the Bayes rule $\hat{p}_\pi(y \mid x)$ would dominate the random plug-in estimator $\hat{p}(y \mid \mu = \hat{\mu})$ when $\hat{\mu}$ is a random draw from π. Strategies for averaging over μ were looking better than plug-in strategies. The hunt for predictive shrinkage estimators had turned to Bayes procedures.

Distinct from $\hat{p}_{MLE}$, $\hat{p}_U$ was soon shown to be the best location invariant predictive density estimator, see Murray (1977) and Ng (1980). That $\hat{p}_U$ is best invariant and minimax also follows from the more recent general results of Liang and Barron (2003), who also showed that $\hat{p}_U$ is admissible when $p = 1$. The minimaxity of $\hat{p}_U$ was also shown directly by George, Liang and Xu (2006). Thus, $\hat{p}_U$, rather than $\hat{p}_{MLE}$, here plays the role played by $\hat{\mu}_{MLE}$ in the mean estimation context. Not surprisingly, $\hat{\mu}_U = x$, the posterior mean under the uniform prior π_U is identical to $\hat{\mu}_{MLE}$ in that context.

The parallels between the mean estimation problem and the predictive estimation problem came into sharp focus with the stunning breakthrough result of Komaki (2001). He proved that when $p \geq 3$, $\hat{p}_U(y \mid x)$ itself is dominated by the Bayes rule

$$\hat{p}_H(y \mid x) = E_{\pi_H}[p(y \mid \mu) \mid x], \tag{15}$$

under the harmonic prior $\pi_H(\mu)$ in (6) used by Stein (1974). Shortly thereafter Liang (2002) showed that $\hat{p}_U(y \mid x)$ is dominated by the proper Bayes rule $p_a(y \mid x)$ under $\pi_a(\mu)$ for which

$$\mu \mid s \sim N_p\left(0, s\, v_0 I\right), \quad s \sim (1+s)^{a-2}, \tag{16}$$

when $v_x \leq v_0$, and when $p = 5$ and $a \in [.5, 1)$ or $p \geq 6$ and $a \in [0, 1)$, the same conditions that Strawderman had obtained for his estimator. Note that $\pi_a(\mu)$ in (16) is an extension of (5) which depends on the constant v_0. As before, $\pi_H(\mu)$ is the special case of $\pi_a(\mu)$ when $a = 2$. Note that $\hat{p}_U$ is now playing the "straw-man" role that was played by $\hat{\mu}_{MLE}$ in the mean estimation problem.

3 A Unified Theory for Minimax Predictive Density Estimation

The proofs of the domination of $\hat{p}_U$ by $\hat{p}_H$ in Komaki (2001) and by $\hat{p}_a$ in Liang (2002) were both tailored to the specific forms of the dominating estimators. They did not make direct use of the properties of the induced marginal distributions of X and Y. From the theory developed by Brown (1971) and Stein (1974) for the mean estimation problem, it was natural to ask if there was a theory analogous to (7)–(10) which would similarly unify the domination results in the predictive density estimation problem.

As it turned out, just such a theory was established in George, Liang and Xu (2006), the main results of which we now proceed to describe. The story begins with a representation, analogous to Brown's representation $\hat{\mu}_\pi(X) = E_\pi(\mu \mid X) = X + \nabla \log m_\pi(X)$ in (7), that is available for posterior mean Bayes rules in the predictive density estimation problem. A key element of the representation is the form of the marginal distributions for our context which we denote by

$$m_\pi(z; v) = \int p(z \mid \mu)\pi(\mu) d\mu. \tag{17}$$

for $Z \mid \mu \sim N_p(\mu, vI)$ and a prior $\pi(\mu)$. In terms of our previous notation (8), $m_\pi(z) = m_\pi(z; 1)$.

Lemma 1. *The Bayes rule $\hat{p}_\pi(y \mid x)$ in (13) can be expressed as*

$$\hat{p}_\pi(y \mid x) = \frac{m_\pi(w; v_w)}{m_\pi(x; v_x)}\, \hat{p}_U(y \mid x) \tag{18}$$

where $\hat{p}_U(y \mid x)$ is the Bayes rule under $\pi_U(\mu) = 1$ given by (14), $m_\pi(x; v_x)$ is the marginal distribution of X, and $m_\pi(w; v_w)$, where $v_w = \frac{v_x v_y}{v_x + v_y}$, is the marginal distribution of $W = \frac{v_y X + v_x Y}{v_x + v_y}$ for independent $X \mid \mu \sim N_p(\mu, v_x I)$ and $Y \mid \mu \sim N_p(\mu, v_y I)$.

Lemma 1 shows how the form of $\hat{p}_\pi(y \mid x)$ is determined entirely by $\hat{p}_U(y \mid x)$ and the form of $m_\pi(x; v_x)$ and $m_\pi(w; v_w)$. The essential step in its derivation is to factor the joint distribution of x and y into terms including a function of the sufficient statistic w. Inserting the representation (18) into the risk R_{KL} leads immediately to the following unbiased estimate for the KL risk difference between $\hat{p}_U(y \mid x)$ and $\hat{p}_\pi(y \mid x)$.

$$\begin{aligned} R_{KL}(\mu, \hat{p}_U) - R_{KL}(\mu, \hat{p}_\pi) &= \int\int p(x \mid \mu)\, p(y \mid \mu) \log \frac{\hat{p}_\pi(y \mid x)}{\hat{p}_U(y \mid x)} dx dy \\ &= E_{\mu, v_w} \log m_\pi(W; v_w) - E_{\mu, v_x} \log m_\pi(X; v_x). \end{aligned} \tag{19}$$

As one can see from (19) and the fact that $v_w = \frac{v_x v_y}{v_x + v_y} < v_x$, $\hat{p}_U(y \mid x)$ would be uniformly dominated by $\hat{p}_\pi(y \mid x)$ whenever $E_{\mu,v} \log m_\pi(Z; v)$ is decreasing in v. As if by magic, the sign of $\frac{\partial}{\partial v} E_{\mu,v} \log m_\pi(Z; v)$ turned out to be directly linked to the same unbiased risk difference estimates (9) and (10) of Stein (1974).

Lemma 2.

$$\frac{\partial}{\partial v} E_{\mu,v} \log m_\pi(Z; v) = E_{\mu,v} \left[\frac{\nabla^2 m_\pi(Z; v)}{m_\pi(Z; v)} - \frac{1}{2} \|\nabla \log m_\pi(Z; v)\|^2 \right] \tag{20}$$

$$= E_{\mu,v} \left[2\nabla^2 \sqrt{m_\pi(Z; v)} / \sqrt{m_\pi(Z; v)} \right] \tag{21}$$

The proof of Lemma 2 relies on Brown's representation, Stein's Lemma, and the fact that any normal marginal distribution $m_\pi(z; v)$ satisfies

$$\frac{\partial}{\partial v} m_\pi(z; v) = \frac{1}{2} \nabla^2 m_\pi(z; v), \tag{22}$$

the well-known heat equation which has a long history in science and engineering, for example, see Steele (2001). Combining (19) and Lemma 2 with the fact that $\hat{p}_U(y \mid x)$ is minimax yields the following general conditions for the minimaxity of a predictive density estimator, conditions analogous to those obtained by Stein for the minimaxity of a normal mean estimator.

Theorem 1. *If $m_\pi(z; v)$ is finite for all z, then $\hat{p}_\pi(y \mid x)$ will be minimax if either of the following hold for all $v_w \le v \le v_x$:*

(i) $m_\pi(z; v)$ is superharmonic.

(ii) $\sqrt{m_\pi(z; v)}$ is superharmonic.

Although condition (i) implies the weaker condition (ii) above, it is included because of its convenience when it is available. Since a superharmonic prior always yields a superharmonic $m_\pi(z; v)$ for all v, the following corollary is immediate.

Corollary 1. *If $m_\pi(z;v)$ is finite for all z, then $\hat{p}_\pi(y\,|\,x)$ will be minimax if $\pi(\mu)$ is superharmonic.*

Because π_H is superharmonic, it is immediate from Corollary 1 that $\hat{p}_H$ is minimax. Because $\sqrt{m_a(z;v)}$ is superharmonic for all v (under suitable conditions on a), it is immediate from Theorem 1 that $\hat{p}_a$ is minimax. It similarly follows that any of the improper superharmonic t-priors of Faith (1978) or any of the proper generalized t-priors of Fourdrinier, Strawderman and Wells (1998) yield minimax Bayes rules.

The connections between the unbiased risk difference estimates for the KL risk and quadratic risk problems ultimately yields the following identity

$$R_{KL}(\mu,\hat{p}_U) - R_{KL}(\mu,\hat{p}_\pi) = \frac{1}{2}\int_{v_w}^{v_x} \frac{1}{v^2}\left[R_Q(\mu,\hat{\mu}_U) - R_Q(\mu,\hat{\mu}_\pi)\right]_v dv, \tag{23}$$

explaining the parallel minimax conditions in both problems. Brown, George and Xu (2008) used this identity to further draw out connections to establish sufficient conditions for the admissibility of Bayes rules under KL loss, conditions analogous to those of Brown (1971) and Brown and Hwang (1982), and to show that all admissible procedures for the KL risk problems are Bayes rules, a direct parallel of the complete class theorem of Brown (1971) for quadratic risk.

4 The Nature of Shrinkage in Predictive Density Estimation

The James-Stein estimator $\hat{\mu}_{JS}(x)$ in (3) provided an explicit example of how risk improvements for estimating μ are obtained by shrinking X toward 0 by the adaptive multiplicative factor $\left(1-\frac{p-2}{\|x\|^2}\right)$. Similarly, under unimodal priors, posterior mean Bayes rules $\hat{\mu}_\pi(x) = E_\pi(\mu\mid x)$ shrink x toward the center of $\pi(\mu)$, the mean of $\pi(\mu)$ when it exists. (Section 6 will describe how multimodal priors yield multiple shrinkage estimators). As we saw earlier, x here plays both the role of $\hat{\mu}_{MLE}(x) = x$ and of the formal Bayes estimator $\hat{\mu}_U(x) = x$.

The representation (18) reveals how $\hat{p}_\pi(y\,|\,x)$ analogously "shrinks" the formal Bayes estimator $\hat{p}_U(y\mid x)$, but not $\hat{p}_{MLE} \neq \hat{p}_U$, by an adaptive multiplicative factor

$$b_\pi(x,y) = \frac{m_\pi(w;v_w)}{m_\pi(x;v_x)}. \tag{24}$$

However, because $\hat{p}_\pi(y\,|\,x)$ must be a proper probability distribution (whenever m_π is always finite), it cannot be the case that $b_\pi(x,y) < 1$ for all y at any x. Thus, "shrinkage" here really refers to a reconcentration of the probability distribution of $\hat{p}_U(y\,|\,x)$. Furthermore, since the mean of $\hat{p}_\pi(y\,|\,x)$ is $E_\pi(\mu\mid x)$, this reconcentration, under unimodal priors, is toward the center of $\pi(\mu)$, as in the mean estimation case.

Consider, for example, what happens under π_H which is symmetric and unimodal about 0. Figure 1 illustrates how this shrinkage occurs for p_H for various values of x when $p = 5$. Figure 1 plots $\hat{p}_U(y\,|\,x)$ and $\hat{p}_H(y\,|\,x)$ as functions of $y = (y_1, y_2, 0, 0, 0)'$ when $v_x = 1$ and $v_y = 0.2$. Note first

that $\hat{p}_U(y\,|\,x)$ is always the same symmetric shape centered at x. When $x = (2,0,0,0,0)'$, shrinkage occurs by pushing the concentration of $\hat{p}_H(y\,|\,x) = b_H(x,y)\,\hat{p}_U(y\,|\,x)$ towards 0. As x moves further from $(0,0,0,0,0)'$ to $(3,0,0,0,0)'$ and $(4,0,0,0,0)'$ this shrinkage diminishes as $\hat{p}_H(y\,|\,x)$ becomes more and more similar to $\hat{p}_U(y\,|\,x)$.

As in the problem of mean estimation, the shrinkage by $\hat{p}_H$ manifests itself in risk reduction over $\hat{p}_U$. To illustrate this, Figure 2 displays the risk difference $[R_{KL}(\mu,\hat{p}_U) - R_{KL}(\mu,\hat{p}_H]$ at $\mu = (c,\ldots,c)'$, $0 \le c \le 4$ when $v_x = 1$ and $v_y = 0.2$ for dimensions $p = 3,5,7,9$. Paralleling the risk reduction offered by $\hat{\mu}_H$ in the mean estimation problem, the largest risk reduction offered by $\hat{p}_H$ occurs close to $\mu = 0$ and decreases rapidly to 0 as $\|\mu\|$ increases. ($R_{KL}(\mu,\hat{p}_U)$ is constant as a function of μ). At the same time, the risk reduction by $\hat{p}_H$ is larger for larger p at each fixed $\|\mu\|$.

5 Many Possible Shrinkage Targets

By a simple shift of coordinates, the modified James-Stein estimator,

$$\hat{\mu}^b_{JS}(x) = b + \left(1 - \frac{p-2}{\|x-b\|^2}\right)(x-b), \tag{25}$$

remains minimax, but now shrinks x towards $b \in R^p$ where its risk function is smallest. Similarly, minimax Bayes shrinkage estimators of a mean or of a predictive density, can be shifted to shrink towards b, by recentering the prior $\pi(\mu)$ to $\pi^b(\mu) = \pi(\mu - b)$. These shifted estimators are easily obtained by inserting the corresponding translated marginal

$$m^b_\pi(z;v) = m_\pi(z-b;v) \tag{26}$$

into (7) to obtain

$$\hat{\mu}^b_\pi(x) = E^b_\pi(\mu\,|\,x) = x + \nabla \log m^b_\pi(x;1), \tag{27}$$

and into (18) to obtain

$$\hat{p}^b_\pi(y\,|\,x) = \frac{m^b_\pi(w;v_w)}{m^b_\pi(x;v_x)}\,\hat{p}_U(y\,|\,x). \tag{28}$$

Recentered unimodal priors such as π^b_H and π^b_a yield estimators that now shrink x and $\hat{p}_U(y\,|\,x)$ towards b rather than towards 0. Since the superharmonic properties of m_π are inherited by m^b_π, the minimaxity of such estimators will be preserved.

In his discussion of Stein (1962), Lindley (1962) noted that the James-Stein estimator could be modified to shrink towards $(\bar{x},...,\bar{x})' \in R^p$, ($\bar{x}$ is the mean of the components of x), by replacing b and $(p-2)$ in (25) by $(\bar{x},...,\bar{x})'$ and $(p-3)$, respectively. The resulting estimator remains minimax as a long as $p \ge 4$ and offers smallest risk when μ is close to the subspace of μ with identical coordinates, the subspace spanned by the vector $1_p = (1,...,1)'$. Note that $(\bar{x},...,\bar{x})'$ is the projection of x into this subspace.

More generally, minimax Bayes shrinkage estimators of a mean or of a predictive density can be similarly modified to obtain shrinkage towards any (possibly affine) subspace $B \subset R^p$, whenever they correspond to spherically symmetric priors. Such priors, which include π_H and π_a, are functions of μ only through $\|\mu\|$. Such a modification is obtained by recentering the prior $\pi(\mu)$ around B via

$$\pi^B(\mu) = \pi(\mu - P_B\mu), \tag{29}$$

where $P_B\mu = \text{argmin}_{b\in B}\|\mu - b\|$ is the projection of μ onto B. Effectively, $\pi^B(\mu)$ puts a uniform prior on $P_B\mu$ and applies a suitably modified version of π to $(\mu - P_B\mu)$. Note that the dimension of $(\mu - P_B\mu)$, namely $(p - \dim(B))$, must be taken into account when determining the appropriate modification for π. For example, recentering the harmonic prior $\pi_H(\mu) = \|\mu\|^{-(p-2)}$ around the subspace spanned by 1_p yields

$$\pi_H^B(\mu) = \|\mu - \bar{\mu}1_p\|^{-(p-3)}, \tag{30}$$

where $\bar{\mu} = \mu' 1_p/p$. Here, the uniform prior is put on $P_B\mu = \bar{\mu}1_p$, and the harmonic prior in dimension $(p - \dim(B)) = (p-1)$ (which is different from the harmonic prior in R^p) is put on $(\mu - \bar{\mu}1_p)$, the orthogonal complement of B.

The marginal m_π^B corresponding to the recentered π^B in (29) can be directly obtained by recentering the spherically symmetric marginal m_π corresponding to π, that is

$$m_\pi^B(z; v) = m_\pi(z - P_B z; v), \tag{31}$$

where $P_B z$ is the projection of z onto B. Analogously to $\pi^B(\mu)$, $m_\pi^B(z; v)$ is uniform on $P_B z$ and applies a suitably modified version of m_π to $(z - P_B z)$. Here too, the dimension of $(z - P_B z)$, namely $(p - \dim(B))$, must be taken into account when determining the appropriate modification for m_π. For example, recentering the marginal m_π around the subspace spanned by 1_p would entail replacing $\|z\|$ by $\|z - \bar{z}1_p\|$, where $\bar{z} = z'1_p/p$, and appropriately modifying m_π to apply to R^{p-1}.

Applying the recentering (29) to priors such as π_H and π_a, which are unimodal around 0, yields priors π_H^B and π_a^B and hence marginals m_H^B and m_a^B, which are unimodal around B. Such recentered marginals yield mean estimators

$$\hat{\mu}_\pi^B(x) = E_\pi^B(\mu \mid x) = x + \nabla \log m_\pi^B(x; 1), \tag{32}$$

and predictive density estimators

$$\hat{p}_\pi^B(y \mid x) = \frac{m_\pi^B(w; v_w)}{m_\pi^B(x; v_x)}\, \hat{p}_U(y \mid x), \tag{33}$$

that now shrink x and $\hat{p}_U(y \mid x)$ towards B rather than towards 0. Shrinkage will be largest when $x \in B$, and will diminish as x moves away from B. These estimators offer smallest risk when $\mu \in B$, but do not improve in any important way over x and $\hat{p}_U(y \mid x)$ when μ is far from B.

A superharmonic m_π will lead to a superharmonic m_π^B as long as $(p - \dim(B))$ is large enough. For example, the recentered marginal m_H^B will be superharmonic only when $(p - \dim(B)) \geq 3$. In such cases, the minimaxity of both $\hat{\mu}_\pi^B$ and $\hat{p}_\pi^B$ will be preserved.

6 Where To Shrink?

Stein's discovery of the existence of minimax shrinkage estimators such as $\hat{\mu}_{JS}^b(x)$ in (25) demonstrated that costless improvements over the minimax $\hat{\mu}_{MLE}$ were available near any target preselected by the statistician. As Stein (1962) put it when referring to the use of such an estimator to center a confidence region, the target "should be chosen ... as one's best guess" of μ. That frequentist considerations had demonstrated the folly of ignoring subjective input was quite a shock to the perceived "objectivity" of the frequentist perspective.

Although the advent of minimax shrinkage estimators of the form $\hat{\mu}_\pi^B$ in (32) and $\hat{p}_\pi^B$ in (33) opened up the possibility of small risk near any preselected (affine) subspace $B \subset R^p$, (this includes the possibility that B is a single point), it also opened up a challenging new problem, how to best choose such a B. From the vast number of possible choices, the goal was to choose B close to the unknown μ, otherwise risk reduction would be negligible. To add to the difficulties, low dimensional B which offered the greatest risk reduction, were also the most difficult to get close to μ.

When faced with a number of potentially good target choices, say $B_1, \ldots, B_N$, rather than choose one of them and proceed with $\hat{\mu}_\pi^B$ or $\hat{p}_\pi^B$, an attractive alternative is to use a minimax multiple shrinkage estimator, George (1986abc). Such estimators incorporate all the potential targets by combining them into an adaptive convex combination of $\hat{\mu}_\pi^{B_1}, \ldots, \hat{\mu}_\pi^{B_N}$ for mean estimation, and of $\hat{p}_\pi^{B_1}, \ldots, \hat{p}_\pi^{B_N}$ for predictive density estimation. By adaptively shrinking towards the more promising targets, the region of potential risk reduction is vastly enlarged while at the same time retaining the safety of minimaxity.

The construction of these minimax multiple shrinkage estimators proceeds as follows, again making fundamental use of the Bayesian formulation. For a spherically symmetric prior $\pi(\mu)$, a set of subspaces $B_1, \ldots, B_N$ of R^p, and a set of nonnegative weights $w_1, ..., w_N$ such that $\sum_1^N w_i = 1$, consider the mixture prior

$$\pi_*(\mu) = \sum_{i=1}^N w_i \, \pi^{B_i}(\mu), \tag{34}$$

where each π^{B_i} is a recentered priors as in (29). To simplify notation, we consider the case where each π^{B_i} is a recentering of the same π, although in principle such a construction could be applied with different priors. The marginal m_* corresponding to the mixture prior π_* in (34) is then simply

$$m_*(z; v) = \sum_1^N w_i \, m_\pi^{B_i}(z; v) \tag{35}$$

where $m_\pi^{B_i}$ are the recentered marginals corresponding to the π^{B_i} as given by (31).

Applying Brown's representation $\hat{\mu}_\pi = x + \nabla \log m_\pi(x;1)$ from (7) with m_* in (35) immediately yields the multiple shrinkage estimator of μ,

$$\hat{\mu}_*(x) = \sum_{i=1}^{N} p(B_i \mid x)\, \hat{\mu}_\pi^{B_i}(x) \tag{36}$$

where

$$p(B_i \mid x) = \frac{w_i m_\pi^{B_i}(x;1)}{\sum_{i=1}^{N} w_i m_\pi^{B_i}(x;1)}. \tag{37}$$

Similarly, applying the representation $\hat{p}_\pi(y \mid x) = \frac{m_\pi(w;v_w)}{m_\pi(x;v_x)}\, \hat{p}_U(y \mid x)$ from (18) with m_* immediately yields the multiple shrinkage estimator of $p(y \mid \mu)$,

$$\hat{p}_*(y \mid x) = \sum_{i=1}^{N} p(B_i \mid x)\, \hat{p}_\pi^{B_i}(y \mid x) \tag{38}$$

where

$$p(B_i \mid x) = \frac{w_i m_\pi^{B_i}(x;v_x)}{\sum_{i=1}^{N} w_i m_\pi^{B_i}(x;v_x)}. \tag{39}$$

The forms (36) and (38) reveal $\hat{\mu}_*$ and $\hat{p}_*$ to be adaptive convex combination of the individual posterior mean estimators $\hat{\mu}_\pi^{B_i}$ and $\hat{p}_\pi^{B_i}$, respectively. The adaptive weights $p(B_i \mid x)$ in (37) and (39) are the posterior probabilities that μ is contained each of the B_i, effectively putting increased weight on those individual estimators which are shrinking most. Note that the uniform prior estimates $\hat{\mu}_U$ and $\hat{p}_U$ are here doubly shrunk by $\hat{\mu}_*$ and $\hat{p}_*(y \mid x)$; in addition to the individual estimator shrinkage they are further shrunk be the posterior probability $\hat{p}(B_i \mid x)$.

The key to obtaining $\hat{\mu}_*$ and $\hat{p}_*(y \mid x)$ which are minimax is simply to use priors which yield super harmonic $m_\pi^{B_1}, \ldots, m_\pi^{B_N}$. If such is the case, then trivially from (35)

$$\nabla^2 m_* = \sum_{1}^{N} w_i \nabla^2 m_\pi^{B_i} \leq 0, \tag{40}$$

so that m_* will be superharmonic, and the minimaxity of $\hat{\mu}_*$ and $\hat{p}_*(y \mid x)$ will follow immediately. Note that marginals whose squareroot is superharmonic will not be adequate, as this argument will fail.

The adaptive shrinkage behavior of $\hat{\mu}_*$ and $\hat{p}_*$ manifests itself as substantial risk reduction whenever μ is near any of $B_1, \ldots, B_N$. Let us illustrate how the happens for the predictive density estimator $\hat{p}_{H^*}$, the multiple shrinkage version of $\hat{p}_H$. Figure 3 illustrates the risk reduction $[R_{KL}(\mu, \hat{p}_U) - R_{KL}(\mu, \hat{p}_{H^*})]$ at various $\mu = (c, \ldots, c)'$ obtained by $\hat{p}_{H^*}$ which adaptively shrinks $\hat{p}_U(y \mid x)$ towards the closer of the two points $b_1 = (2, \ldots, 2)'$ and $b_2 = (-2, \ldots, -2)'$ using equal weights $w_1 = w_2 = 0.5$. As in Figure 2, we considered the case $v_x = 1, v_y = 0.2$ for $p = 3, 5, 7, 9$. As the plot shows, maximum risk reduction occur when μ is close to b_1 or b_2, and goes to 0 as μ moves away from either of these points. At the same time, for each fixed $\|\mu\|$, risk reduction by $\hat{p}_{H^*}$ is larger for larger p. It is impressive that the size of the risk reduction offered by $\hat{p}_{H^*}$ is nearly the

same as each of its single target counterparts. The cost of multiple shrinkage enhancement seems negligible, especially compared to the benefits.

7 Empirical Bayes Constructions

Beyond their attractive risk properties, the the James-Stein estimator $\hat{\mu}_{JS}$ and its positive-part counterpart $\hat{\mu}_{JS+}$ are especially appealing because of their simple closed forms which are easy to compute. As shown by Xu and Zhou (2011), similarly appealing simple closed form predictive density shrinkage estimators can be obtained by the same empirical Bayes considerations that motivate $\hat{\mu}_{JS}$ and $\hat{\mu}_{JS+}$.

The empirical Bayes motivation of $\hat{\mu}_{JS}$, alluded to in Section 1, simply entails replacing $1/(1+\nu)$ in (4) by $(p-2)/\|x\|^2$, its unbiased estimate under the marginal distribution of $X \mid \mu \sim N_p(\mu, I)$ when $\mu \sim N_p(0, \nu I)$. The positive-part $\hat{\mu}_{JS+}$ is obtained by using the truncated estimate $(p-2)/\max\{1, \|x\|^2\}$ which avoids an implicitly negative estimate of the prior variance ν.

Proceeding analogously, Xu and Zhou considered the Bayesian predictive density estimate,

$$\hat{p}_\nu(y \mid x) \sim N_p\left(\left(1 - \frac{v_x}{v_x+\nu}\right)x, \frac{v_x}{v_x+\nu}v_y + \left(1 - \frac{v_x}{v_x+\nu}\right)(v_x+v_y)\right), \tag{41}$$

when $X \mid \mu \sim N_p(\mu, v_x I)$ and $Y \mid \mu \sim N_p(\mu, v_y I)$ are independent, and $\mu \sim N_p(0, \nu I)$. Replacing $v_x/(v_x+\nu)$ by its truncated unbiased estimate $(p-2)v_x/\max\{v_x, \|x\|^2\}$ under the marginal distribution of X, they obtained the empirical Bayes predictive density estimate

$$\hat{p}_{p-2}(y \mid x) \sim N_p\left(\left(1 - \frac{(p-2)v_x}{\|x\|^2}\right)_+ x;\ \ v_y + \left(1 - \frac{(p-2)v_x}{\|x\|^2}\right)_+ v_x\right) \tag{42}$$

where $(\cdot)_+ = \max\{0, \cdot\}$, an appealing simple closed form. Centered at $\hat{\mu}_{JS+}$, $\hat{p}_{p-2}$ converges to the best invariant procedure $\hat{p}_U \sim N(x, v_x + v_y)$ as $\|x\|^2 \to \infty$, and converges to $N(0, v_y)$ as $\|x\|^2 \to 0$. Thus, $\hat{p}_{p-2}$ can be viewed as a shrinkage predictive density estimator that "pulls" $\hat{p}_U$ towards 0, its shrinkage adaptively determined by the data.

To asses the KL risk properties of such empirical Bayes estimators, Xu and Zhou considered the class of estimators $\hat{p}_k$ of the form (42) with $(p-2)$ replaced by a constant k, a class of simple normal forms centered at shrinkage estimators of μ with data-dependent variances to incorporate estimation uncertainty. For this class, they provided general sufficient conditions on k and the dimension p for $\hat{p}_k$ to dominate the best invariant predictive density $\hat{p}_U$ and thus be minimax. Going further, they also established an "oracle" inequality which suggests that the empirical Bayes predictive density estimator is asymptotically minimax in infinite-dimensional parameter spaces and can potentially be used to construct adaptive minimax estimators. It appears that that these minimax empirical Bayes predictive densities may play the same role as the James-Stein estimator in such problems.

It may be of interest to note that a particular pseudo-marginal empirical Bayes construction that works fine for the mean estimation problem, appears not to work for the predictive density estimation problem. For instance, the positive-part James-Stein estimator $\hat{\mu}_{JS+}$ can be expressed as $\hat{\mu}_{JS+} = x + \nabla \log m_{JS+}(x;1)$, where $m_{JS+}(x;v)$ is the function

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

with $k_p = (e/(p-2))^{-(p-2)/2}$ (see Stein 1974). We refer to $m(z;v)$ as a pseudo-marginal because it is not a bona fide marginal obtained by a real prior. Nonetheless, it plays the formal role of a marginal in the mean estimation problem, and can be used to generate further innovations such as minimax multiple shrinkage James-Stein estimators, (see George 1986abc).

Proceeding by analogy, it would seem that $m(z;v)$ could be inserted into the representation (18) from Lemma 1 to obtain similar results under KL loss. Unfortunately, this does not yield a suitable minimax predictive estimator because $\hat{p}_{JS+}(y \mid x)$ is not a proper probability distribution. Indeed, $\int \hat{p}_{JS+}(y|x)dy \neq 1$ and varies with x. What's gone wrong? Because they do not correspond to real priors, such pseudo-marginals are ultimately are at odds with the probabilistic coherence of a valid Bayesian approach. In contrast to the mean estimation framework, the predictive density estimation framework apparently requires stronger fidelity to the Bayesian paradigm.

8 Predictive Density Estimation for Classical Regression

Moving into the multiple regression setting, Stein (1960) considered the estimation of a p-dimensional coefficient vector under suitably rescaled quadratic loss. He there established the minimaxity of the maximum likelihood estimators, and then proved its inadmissibility when $p \geq 3$, by demonstrating the existence of a dominating shrinkage estimator.

In a similar vein, as one might expect, the theory of predictive density estimation presented in Sections 2 and 3 can also be extended to the multiple regression framework. We here describe the main ideas of the development of this extension which appeared in George and Xu (2008). Similar results, developed independently from a slightly different perspective, appeared at the same time in Kobayashi and Komaki (2008).

Consider the canonical normal linear regression setup:

$$X \mid \beta \sim N_m(A\beta, \sigma^2 I), \quad Y \mid \beta \sim N_n(B\beta, \sigma^2 I), \tag{43}$$

where A is a full rank, fixed $m \times p$, B is a fixed $n \times p$ matrix, and β is a common $p \times 1$ unknown regression coefficient. The error variance σ^2 is assumed to be known, and set to be 1 without loss of generality. The problem is to find an estimator of $\hat{p}(y \mid x)$ of the predictive density $p(y \mid \beta)$,

evaluating its performance by KL risk

$$R_{KL}(\beta, \hat{p}) = \int p(x \mid \beta) L(\beta, \hat{p}(\cdot \mid x)) dx \tag{44}$$

where $L(\beta, \hat{p}(\cdot \mid x))$ is the KL loss between the density $p(y \mid \beta)$ and its estimator $\hat{p}(y \mid x)$.

The story begins with the result, analogous to Aitchison's (1975) for the normal mean problem, that the plug-in estimator $p(y \mid \hat{\beta}_x)$, where $\hat{\beta}_x$ is the least squares estimate of β based on x, is dominated under KL risk by the posterior mean of $p(y \mid \beta)$, the Bayes rule under the uniform prior

$$\hat{p}_U(y \mid x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{|A'A + B'B|^{-\frac{1}{2}}}{|A'A|^{-\frac{1}{2}}} \exp\left\{ -\frac{RSS_{x,y} - RSS_x}{2} \right\}. \tag{45}$$

Here too, $\hat{p}_U$ is minimax (Liang, 2002; Liang and Barron, 2004) and plays the straw-man role of the estimator to beat. The challenge was to determine which priors π would lead to Bayes rules which dominated $\hat{p}_U$, and hence would be minimax too. Analogous to the representation (18) in Lemma 1 for the normal mean problem, the following representation for a Bayes rule $\hat{p}_\pi(y \mid x)$ here, was the key to meeting this challenge.

Lemma 3. *The Bayes rule $\hat{p}_\pi(y \mid x) = \int p(y \mid \beta)\pi(\beta)d\beta$ can be expressed as*

$$\hat{p}_\pi(y \mid x) = \frac{m_\pi(\hat{\beta}_{x,y}; \Sigma_C)}{m_\pi(\hat{\beta}_x; \Sigma_A)} \hat{p}_U(y \mid x). \tag{46}$$

where $\Sigma_A = (A'A)^{-1}$, $C = A'A + B'B$, $\Sigma_C = (C'C)^{-1}$, $\hat{\beta}_x$ is the least squares estimates of β based on x, and $\hat{\beta}_{x,y}$ based on x and y, and $m_\pi(z; \Sigma)$ is the marginal distribution of $Z|\beta \sim N_p(\beta, \Sigma)$ under $\pi(\beta)$.

The representation (46) leads immediately to following analogue of (19) for the KL risk difference between $\hat{p}_U(y \mid x)$ and $\hat{p}_\pi(y \mid x)$,

$$R_{KL}(\beta, \hat{p}_U) - R_{KL}(\beta, \hat{p}_\pi) = E_{\beta,\Sigma_C} \log m_\pi(\hat{\beta}_{x,y}; \Sigma_C) - E_{\beta,\Sigma_A} \log m_\pi(\hat{\beta}_x; \Sigma_A). \tag{47}$$

The challenge thus became that of finding conditions on m_π to make this difference positive, a challenge made more difficult than the previous one for (19) because of the complexity of Σ_A and Σ_C. Fortunately this could be resolved by rotating the problem as follows to obtain diagonal forms. Since Σ_A and Σ_C are both symmetric and positive definite, there exists a full rank $p \times p$ matrix W, such that

$$\Sigma_A = WW', \quad \Sigma_C = WDW', \quad D = diag(d_1, \ldots, d_p). \tag{48}$$

Because $\Sigma_C = (\Sigma_A^{-1} + B'B)^{-1}$ where $B'B$ is nonnegative definite, it follows that $d_i \in (0, 1]$ for all $1 \le i \le p$ with at least one $d_i < 1$. Thus, the parameters for the rotated problem become

$$\mu = W^{-1}\beta, \quad \hat{\mu}_x = W^{-1}\hat{\beta}_x \sim N_p(\mu, I), \quad \hat{\mu}_{x,y} = W^{-1}\hat{\beta}_{x,y} \sim N_p(\mu, D). \tag{49}$$

Letting $V_w = wI + (1-w)D$ for $w \in [0,1]$, the risk difference (47) could be reexpressed as

$$\begin{aligned} R_{KL}(\beta, \hat{p}_U) - R_{KL}(\beta, \hat{p}_\pi) &= E_{\mu,D} \log m_{\pi_W}(\hat{\mu}_{x,y}; D) - E_{\mu,I} \log m_{\pi_W}(\hat{\mu}_x; I) \\ &= h_\mu(V_0) - h_\mu(V_1), \end{aligned} \tag{50}$$

where $h_\mu(V_w) = E_{\mu,V_w} \log m_{\pi_W}(Z; V_w)$ and $\pi_W(\mu) = \pi(W\mu)$. The minimaxity of $\hat{p}_\pi$ would now follow from conditions on m_π such that $(\partial/\partial w) h_\mu(w) < 0$ for all μ and $w \in [0,1]$. The following substantial generalizations of Theorem 1 and Corollary 1 provides exactly those conditions.

Theorem 2. *Suppose $m_\pi(z; WW')$ is finite for all z with the invertible matrix W defined as in (48). Let $H(f(z_1, \cdots, z_p))$ be the Hessian matrix of f.*

(i) If $trace\{H(m_\pi(z; WV_wW'))[\Sigma_A - \Sigma_C]\} \leq 0$ for all $w \in [0,1]$, then $\hat{p}_\pi(y \mid x)$ is minimax.

(ii) If $trace\{H(\sqrt{m_\pi(z; WV_wW')})[\Sigma_A - \Sigma_C]\} \leq 0$ for all $w \in [0,1]$, then $\hat{p}_\pi(y \mid x)$ is minimax.

Corollary 2. *Suppose $m_\pi(z; WW')$ is finite for all z. Then $\hat{p}_\pi(y \mid x)$ is minimax if*

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

As a consequence of Corollary 2, the scaled harmonic prior $\pi_H(\beta|W) \propto \|W^{-1}\beta\|^{p-2}$ can be shown to yield minimax predictive density estimators for the regression setting.

Going further, George and Xu (2008) went on to show that the minimax Bayes estimators here can be modified to shrink towards different points and subspaces as in Section 5, and that the minimax multiple shrinkage constructions of Section 6 apply as well. In particular, they obtained minimax multiple shrinkage estimators that naturally accommodate variable selection uncertainty.

9 Predictive Density Estimation for Non-parametric Regression

Moving in another direction, Xu and Liang (2010) considered predictive density estimation in the context of modern non-parametric regression, a context in which the James-Stein estimator has turned out to play an important asymptotic minimaxity role, see Wasserman (2006). Their results pertain to the canonical setup for non-parametric regression

$$Y(t_i) = f(t_i) + \varepsilon_i, \quad i = 1, \cdots, n, \tag{51}$$

where f is an unknown smooth function in $\mathcal{L}^2[0,1]$, $t_i = i/n$, and ε_i's are i.i.d. $N(0,1)$. A central problem here is to estimate f or various functionals of f based on observing $Y = (Y(t_1), \ldots, Y(t_n))$. Transforming the problem with an orthonormal basis, (51) is equivalent to estimating the θ_i's in

$$y_i = \theta_i + e_i, \quad e_i \sim N\Big(0, \frac{1}{n}\Big), \quad i = 1, \cdots, n, \tag{52}$$

known as the Gaussian sequence model. The model above is different from the ordinary multivariate normal model in two aspects: 1) the model dimension n is increasing with the sample size, and 2) under function space assumptions on f, the θ_i's lie in a constrained space, e.g., an ellipsoid $\{\sum_i a_i^2\theta_i^2 \le C, a_i \to \infty\}$.

A large body of literature has been devoted to minimax estimation of f under $\mathcal{L}^2$ risk over certain function spaces, see, for example, Johnstone (2001), Efromovich (1999), and the references therein. As opposed to the ordinary multivariate normal mean problem, exact minimax analysis is difficult for the Gaussian sequence model (52) when a constraint on the parameters is considered. This difficulty has been overcome by first obtaining the minimax risk of a subclass of estimators of a simple form, and then showing that the overall minimax risk is asymptotically equivalent to the minimax risk of the subclass. For example, an important result from Pinsker (1980) is that when the parameter space is constrained to an ellipsoid, the non-linear minimax risk is asymptotically equivalent to the linear minimax risk, namely the minimax risk of the subclass of linear estimators of the form $\hat{\theta}_i = c_i x_i$.

For non-parametric regression, the following analogue between estimation under $\mathcal{L}^2$ risk and predictive density estimation under KL risk was established in Xu and Liang (2010). The prediction problem for non-parametric regression is formulated as follows. Let $\tilde{Y} = (\tilde{Y}(u_1), \ldots, \tilde{Y}(u_m))$ be future observations arising at a set of dense ($m \ge n$) and equally spaced locations $\{u_j\}_{i=1}^m$. Given f, the predictive density $p(\tilde{y}\,|\,f)$ is just a product of Gaussians. The problem is to find an estimator $\hat{p}(\tilde{y}\,|\,y)$ of $p(\tilde{y}\,|\,f)$, where performance is measured by the averaged KL risk

$$R(f, \hat{p}) = \frac{1}{m} E_{Y,\tilde{Y}|f} \log \frac{p(\tilde{Y}\,|\,f)}{\hat{p}(\tilde{Y}\,|\,Y)}. \tag{53}$$

In this formulation, densities are estimated at the m locations simultaneously by $\hat{p}(\tilde{y}\,|\,y)$. As it turned out, the KL risk based on the simultaneous formulation (53) is the analog of the $\mathcal{L}^2$ risk for estimation. Indeed, under the KL risk (53), the prediction problem for a non-parametric regression model can be converted to the one for a Gaussian sequence model.

Based on this formulation of the problem, minimax analysis proceeds as in the general framework for the minimax study of function estimation used by, for example, Pinsker (1980) and Belitser and Levit (1995, 1996). The linear estimators there, which play a central role in their minimax analysis, take the same form as posterior means under normal priors. Analogously, predictive density estimates under the same normal priors turned out to play the corresponding role in the minimax analysis for prediction. (The same family of Bayes rules arises from the empirical Bayes approach in Section 7). Thus, Xu and Liang (2010) were ultimately able to show that the overall minimax KL risk is asymptotically equivalent to the minimax KL risk of this subclass of Bayes rules, a direct analogue of Pinker's Theorem for predictive density estimation in non-parametric regression.

10 Discussion

Stein's (1956) discovery of the existence of shrinkage estimators that uniformly dominate the minimax maximum likelihood estimator of the mean of a multivariate normal distribution under quadratic risk when $p \geq 3$ was the beginning of a major research effort to develop improved minimax shrinkage estimation. In subsequent papers Stein guided this effort towards the Bayesian paradigm by providing explicit examples of minimax empirical Bayes and fully Bayes rules. Making use of the fundamental results of Brown (1971), he developed a general theory for establishing minimaxity based on the superharmonic properties of the marginal distributions induced by the priors.

The problem of predictive density estimation of a multivariate normal distribution under KL risk has more recently seen a series of remarkably parallel developments. With a focus on Bayes rules catalyzed by Aichison (1975), Komaki (2001) provided a fundamental breakthrough by demonstrating that the harmonic prior Bayes rule dominated the best invariant uniform prior Bayes rule. These results suggested the existence of a theory for minimax estimation based on the superharmonic properties of marginals, a theory that was then established in George, Liang and Xu (2006). Further developments of new minimax shrinkage predictive density estimators now abound, including, as described in this article, multiple shrinkage estimators, empirical Bayes estimators, normal linear model regression estimators, and non-parametric regression estimators. Examples of promising further new directions for predictive density estimation can be found in Komaki (2004, 2006, 2009) which includes results for Poisson distributions, for general location-scale models and for Wishart distributions, in Ghosh, Mergel and Datta (2008) which develops estimation under alternative divergence losses, and Kato (2009) which established improved minimax predictive domination for the multivariate normal distribution under KL risk when both the mean and the variance are unknown. Minimax predictive density estimation is now beginning to flourish.

References

Aitchison, J. (1975). Goodness of Prediction Fit. *Biometrika*, 62, 547-554.

Belitser, E.N., and Levit, B.Y. (1995). On minimax filtering over ellipsoids. *Mathematical Methods of Statistics*,3, 259–273.

Brown, L.D. (1971). Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems. *Annals of Mathematical Statistics*, 42, 855-903.

Brown, L.D., George, E.I., and Xu, X. (2008). Admissible Predictive Density Estimation. *Annals of Statistics*, 36, 3, 1156–1170. DOI: 10.1214/07-AOS506.

Brown, L. D. and Hwang, J. (1982). A unified admissibility proof. In *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. O. Berger, eds.) 1 205230. Academic Press, New York. MR0705290

Efromovich, S.Y. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications.* Springer, New York.

Faith, R.E. (1978). Minimax Bayes Point Estimators of a Multivariate Normal Mean. *J. Mult. Anal.*, 8, 372-379.

Fourdrinier, D., Strawderman, W.E., and Wells, M.T. (1998). On the Construction of Bayes Minimax Estimators. *Annals of Statistics*, 26, 660-671.

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, 7, 2099-2114.

George, E.I., Liang, F. and Xu, X. (2006). Improved Minimax Predictive Densities Under Kullback-Leibler Loss. *Annals of Statistics*, 34 1 78-91.

George, E.I. and Xu, X. (2008). Predictive Density Estimation for Multiple Regression. *Econometric Theory*, 24, 528–544. DOI: 10+10170S0266466608080213.

Ghosh, M., Mergel, V. and Datta, G.S. (2008). Estimation, prediction and the Stein phe- nomenon under divergence loss. *Journal of Multivariate Analysis*, 99, 1941–1961.

Johnstone, I.M. (2003). *Function Estimation and Gaussian Sequence Models.* Draft of a Monograph, Department of Statistics, Stanford University.

Kato, K. (2009). Improved prediction for a multivariate normal distribution with unknown mean and variance. *Ann. Inst. Statist. Math.*, 61, 2009, 531-542.

Kobayashi, K. and Komaki, F. (2008). Bayesian shrinkage prediction for the regression problem. *Journal of Multivariate Analysis* 99 , 1888–1905.

Komaki, F. (2001). A Shrinkage Predictive Distribution for Multivariate Normal Observations, *Biometrika*, 88, 859-864.

Komaki, F. (2004). Simutaneous prediction of independent Poisson observables. *Annals of Statistics*, 32, 1744-1769.

Komaki, F. (2006). Shrinkage priors for Bayesian prediction. *Annals of Statistics*, 34, 2, 808-819.

Komaki, F. (2009). Bayesian predictive densities based on superharmonic priors for the 2-dimensional Wishart model. *Journal of Multivariate Analysis*, 100, 2137–2154.

Lehmann, E.L., and Casella, G. (1998). *Theory of Point Estimation, Second Edition*, Springer, New York.

Liang, F. and Barron, A. (2003). Exact Minimax Strategies for Predictive Density Estimation, Data Compression and Model Selection. *IEEE Information Theory Transactions*, to appear.

Liang, F. (2002). *Exact Minimax Procedures for Predictive Density Estimation and Data Compression*. Ph.D. dissertation, Department of Statistics, Yale University.

Lindley (1962). Discussion of "Confidence Sets for the Mean of a Multivariate Normal Distribution by C. Stein". *Journal of the Royal Statistical Society. Series B*, Vol. 24, No. 2, pp. 285–287.

Murray, G.D. (1977), A Note on the Estimation of Probability Density Functions. *Biometrika*, 64, 150-152.

Ng, V.M. (1980). On the Estimation of Parametric Density Functions. *Biometrika*, 67, 505-506.

Pinsker, M.S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transmission*, 2, 120–133.

Steele, J.M. (2001). *Stochastic Calculus and Financial Applications*. Springer, New York.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.* 1, 197–206.

Stein, C. (1960). Multiple regression. In *Contributions to probability and statistics*, I. Olkin (Ed.), Stanford University Press, 264–305

Stein, C. (1962). Confidence Sets for the Mean of a Multivariate Normal Distribution (with discussion). *Journal of the Royal Statistical Society. Series B*, Vol. 24, No. 2, pp. 265–296

Stein, C. (1974). Estimation of the Mean of a Multivariate Normal Distribution. In *Proceedings of the Prague Symposium on Asymptotic Statistics*, Ed. J. Hajek, pp. 345–81. Prague: Universita Karlova.

Stein, C. (1981). Estimation of a Multivariate Normal Mean. *Ann. Statist.* **9**, 1135–51.

Strawderman, W.E. (1971). Proper Bayes Minimax Estimators of the Multivariate Normal Mean. *Annals of Mathematical Statistics*, 42, 385–388.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer, New York.

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

Xu, X. and Zhou, D. (2011). Empirical Bayes predictive densities for high-dimensional normal models. *Journal of Multivariate Analysis*, 102, 1417–1428.

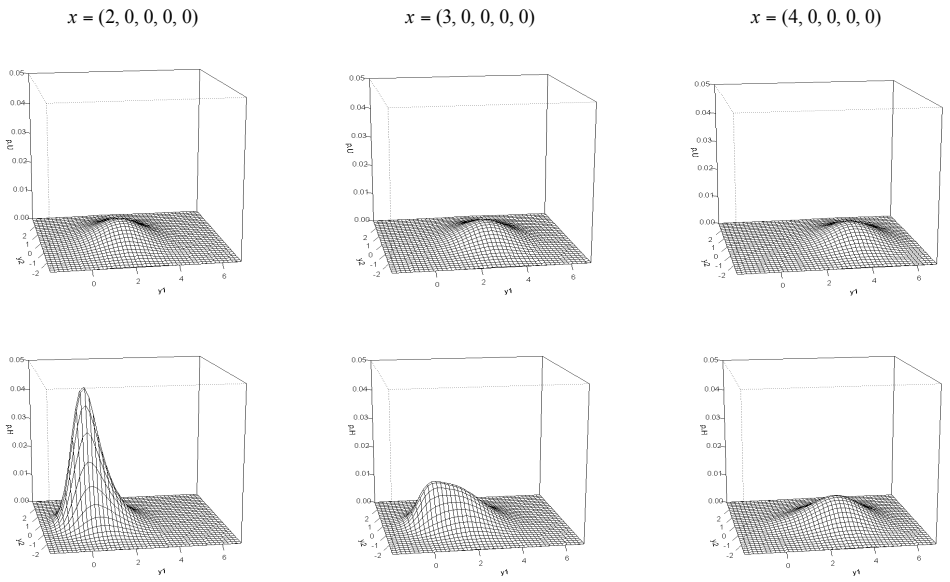

Figure 1: Shrinkage of $\hat{p}_U(y \mid x)$ to obtain $\hat{p}_H(y \mid x)$ when $v_x = 1, v_y = 0.2$ and $p = 5$. Here $y = (y_1, y_2, 0, 0, 0)'$.

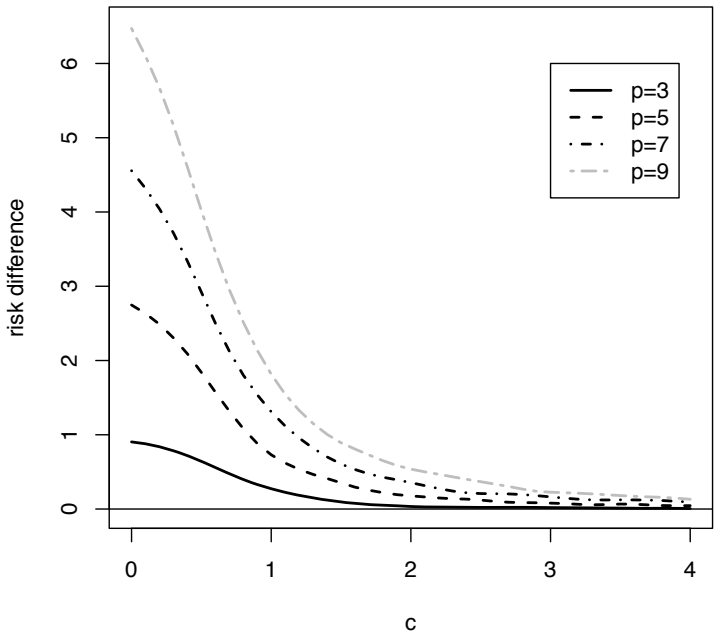

Figure 2: The risk difference between $\hat{p}_U$ and $\hat{p}_H$ when $\mu = (c, \cdots, c)', v_x = 1, v_y = 0.2$.

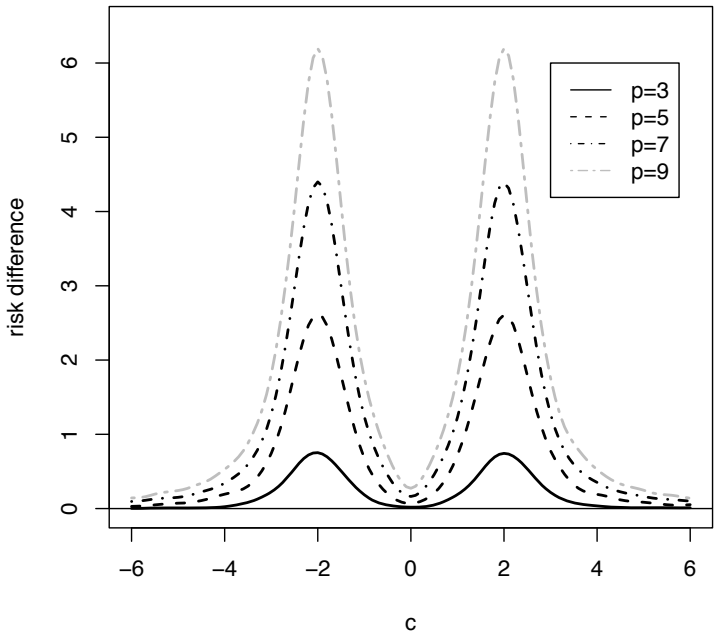

Figure 3: The risk difference between $\hat{p}_U$ and multiple shrinkage $\hat{p}_{H^*}$ when $\mu = (c, \cdots, c)', v_x = 1, v_y = 0.2, b_1 = (2, ..., 2)', b_2 = (-2, \ldots, -2)'$, and $w_1 = w_2 = 0.5$.