

# Improved Minimax Prediction Under Kullback-Leibler Loss

Edward I. GEORGE, Feng LIANG and Xinyi XU \*

July 2003

## Abstract

Let  $X | \mu \sim N_p(\mu, v_x I)$  and  $Y | \mu \sim N_p(\mu, v_y I)$  be independent  $p$ -dimensional multivariate normal vectors with common unknown mean  $\mu$ , and let  $p(x|\mu)$  and  $p(y|\mu)$  denote the conditional densities of  $X$  and  $Y$ . Based on only observing  $X = x$ , we consider the problem of obtaining a predictive distribution  $\hat{p}(y|x)$  for  $Y$  that is close to  $p(y|\mu)$  as measured by Kullback-Leibler loss. The natural straw man for this problem is the best invariant predictive distribution, the Bayes rule  $p_U(y|x)$  under the uniform prior  $\pi_U(\mu) \equiv 1$ , which is seen to be minimax. We show that  $p_U(y|x)$  is dominated by any Bayes rules for which the square root of the marginal distribution is superharmonic. This yields wide classes of dominating predictive distributions including Bayes rules under superharmonic priors. These dominating predictive shrinkage distributions can be constructed to adaptively shrink  $p_U(y|x)$  towards arbitrary points or subspaces. Those procedures corresponding to superharmonic priors can be further combined to obtain minimax multiple shrinkage predictive distributions that adaptively shrink  $p_U(y|x)$  towards an arbitrary number of points or subspaces. Fundamental similarities and differences with the parallel theory of estimating a multivariate normal mean under quadratic loss are described throughout.

*Keywords:* BAYESIAN PREDICTION; HEAT EQUATION; INADMISSIBILITY; MULTIPLE SHRINKAGE; PRIOR DISTRIBUTIONS; SHRINKAGE ESTIMATION; SUPERHARMONIC MARGINALS; UNBIASED ESTIMATE OF RISK.

---

\*Edward I. George is the Universal Furniture Professor and Xinyi Xu is a Doctoral Student, Statistics Department, The Wharton School, 3730 Walnut Street 400 JMH, Philadelphia, PA 19104-6340, edgeorge@wharton.upenn.edu and xinyi@wharton.upenn.edu. Feng Liang is Assistant Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, NC, 27708-0251, feng@ids.duke.edu. We would like to thank Andrew Barron, Larry Brown, Bill Strawderman and Cun-Hui Zhang for generous insights and suggestions. This work was supported by NSF grant DMS-0130819.

# 1 Introduction

Let  $X | \mu \sim N_p(\mu, v_x I)$  and  $Y | \mu \sim N_p(\mu, v_y I)$  be independent  $p$ -dimensional multivariate normal vectors with common unknown mean  $\mu$ , and let  $p(x | \mu)$  and  $p(y | \mu)$  denote the conditional densities of  $X$  and  $Y$ . We assume that  $v_x$  and  $v_y$  are known until Section 6.

Based on only observing  $X = x$ , we consider the problem of obtaining a predictive distribution  $\hat{p}(y | x)$  for  $Y$  that is close to  $p(y | \mu)$ . We measure this closeness by Kullback-Leibler (KL) loss,

$$L(\mu, \hat{p}(y | x)) = \int p(y | \mu) \log \frac{p(y | \mu)}{\hat{p}(y | x)} dy, \quad (1)$$

and evaluate a predictive procedure  $\hat{p}(y | x)$  by its expected loss or risk function

$$R_{KL}(\mu, \hat{p}) = \int p(x | \mu) L(\mu, \hat{p}(y | x)) dx. \quad (2)$$

For the comparison of two procedures, we say that  $\hat{p}_1$  dominates  $\hat{p}_2$  if  $R_{KL}(\mu, \hat{p}_1) \leq R_{KL}(\mu, \hat{p}_2)$  for all  $\mu$  and with strict inequality for some  $\mu$ . By a sufficiency and transformation reduction, this problem is seen to be equivalent to predicting  $Y_1, \dots, Y_n | \mu$  iid  $\sim N_p(\mu, v_y \Sigma)$  under KL loss (1) based on observing  $X_1, \dots, X_m | \mu$  iid  $\sim N_p(\mu, v_x \Sigma)$  which are independent of  $Y_1, \dots, Y_n | \mu$ .

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

$$p_\pi(y | x) = \frac{\int p(x | \mu) p(y | \mu) \pi(\mu) d\mu}{\int p(x | \mu) \pi(\mu) d\mu}, \quad (3)$$

see Aitchison (1975). Note that unless  $\pi$  is a trivial point prior, such  $p_\pi(y | x)$  will not be of the form of any  $p(y | \mu)$  so that such Bayes rules fall outside the parameter space of all  $p(y | \mu)$ .

The best invariant predictive density for this problem is the constant risk Bayes rule (3) under the uniform prior  $\pi_U(\mu) = 1$ , namely

$$p_U(y | x) = \frac{1}{\{2\pi(v_x + v_y)\}^{\frac{p}{2}}} \exp \left\{ -\frac{\|y - x\|^2}{2(v_x + v_y)} \right\}, \quad (4)$$

see Murray (1977) and Ng (1980). Indeed,  $p_U(y | x)$  dominates the naive ‘‘plug-in’’ predictive distribution  $p(y | \hat{\mu} = x)$  that simply substitutes the MLE  $\hat{\mu} = x$  for  $\mu$ , see Aitchison (1975). As will be seen in Section 2,  $p_U(y | x)$  is minimax for KL loss (1). That  $\pi_U$  is best invariant and minimax can also be seen as a special case of the more general recent results in Liang and Barron (2003), which also show that  $p_U(y | x)$  is admissible when  $p = 1$ .

However,  $p_U(y | x)$  is inadmissible when  $p \geq 3$ . In a breakthrough result, Komaki (2001) proved that when  $p \geq 3$ ,  $p_U(y | x)$  itself is dominated by the Bayes rule

$$p_H(y | x) = \frac{\int p(x | \mu) p(y | \mu) \pi_H(\mu) d\mu}{\int p(x | \mu) \pi_H(\mu) d\mu}, \quad (5)$$

where

$$\pi_H(\mu) \propto \|\mu\|^{-(p-2)} \quad (6)$$

is the harmonic prior recommended by Stein (1974), which we subscript by “ $H$ ” for harmonic. Although Komaki referred to  $\pi_H$  as harmonic, his proof did not directly exploit this property.

More recently, Liang (2002) showed that  $p_U(y|x)$  is dominated by the proper Bayes rule  $p_a(y|x)$  under the Strawderman prior  $\pi_a(\mu)$  for which

$$\mu | s \sim N_p(0, s v_0 I), \quad s \sim (1 + s)^{a-2}, \quad (7)$$

when  $v_x \leq v_0$ , and when  $p = 5$  and  $a \in [.5, 1)$  or  $p \geq 6$  and  $a \in [0, 1)$ . Note that  $\pi_a(\mu)$  depends on the constant  $v_0$  in (7), a dependence that will be maintained throughout this paper. The improper harmonic prior  $\pi_H(\mu)$  is well known to be the special case of  $\pi_a(\mu)$  when  $a = 2$ .

These results closely parallel some key developments concerning minimax estimation of a multivariate normal mean under quadratic loss. Based on observing  $X | \mu \sim N_p(\mu, I)$ , this problem is to estimate  $\mu$  under

$$R_Q(\mu, \hat{\mu}) = E_\mu \|\hat{\mu} - \mu\|^2. \quad (8)$$

The natural straw man here is the maximum likelihood estimator

$$\hat{\mu}_{MLE} = X,$$

which, under quadratic risk (8), is best invariant, minimax and admissible when  $p \leq 2$ . Note that  $\hat{\mu}_{MLE}$  plays the same role here that  $p_U(y|x)$  plays in our KL risk problem. As a precursor to Komaki’s result about the domination of  $p_H(y|x)$  over  $p_U(y|x)$ , Stein (1974) showed that  $\hat{\mu}_H = E_{\pi_H}(\mu|x)$ , the Bayes rule under  $\pi_H$ , dominates  $\hat{\mu}_{MLE}$  when  $p \geq 3$ . And as a precursor to Liang’s result about the domination of  $p_a(y|x)$  over  $p_U(y|x)$ , Strawderman (1971) showed that  $\hat{\mu}_a = E_{\pi_a}(\mu|x)$ , the proper Bayes rule under  $\pi_a$  when  $v_x = v_0 = 1$ , dominates  $\hat{\mu}_{MLE}$  when  $p \geq 5$  for exactly the same choices of  $a$ .

A unified theory for the quadratic risk estimation problem of these and other shrinkage domination results has been obtained by focusing on the marginal distribution of  $X$  under  $\pi$ , namely

$$m_\pi(x) = \int p(x|\mu)\pi(\mu)d\mu. \quad (9)$$

The key to this theory is the representation due to Brown (1971) that any Bayes rule  $\hat{\mu}_\pi = E_\pi(\mu|x)$  for quadratic risk is of the form

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

( $\nabla = (\partial/\partial x_1, \dots, \partial/\partial x_p)'$ ). To show that  $\hat{\mu}_H$  dominates  $\hat{\mu}_{MLE}$ , Stein (1974, 1981) used this representation to establish that  $R_Q(\mu, \hat{\mu}_{MLE}) - R_Q(\mu, \hat{\mu}_\pi) = E_\mu U(X)$ , where

$$U(X) = \|\nabla \log m_\pi(X)\|^2 - 2 \frac{\nabla^2 m_\pi(X)}{m_\pi(X)} \quad (11)$$

$$= -4 \frac{\nabla^2 \sqrt{m_\pi(X)}}{\sqrt{m_\pi(X)}} \quad (12)$$

is an unbiased estimate of the risk reduction of  $\hat{\mu}_\pi$  over  $\hat{\mu}_{MLE}$ , ( $\nabla^2 m_\pi(x) = \sum \frac{\partial^2}{\partial x_i^2} m_\pi(x)$ ). It follows immediately from (12) that the superharmonicity of  $\sqrt{m_\pi}$ , i.e.  $\nabla^2 \sqrt{m_\pi(x)} \leq 0$ , is sufficient (though not necessary) for  $\hat{\mu}_\pi$  to dominate  $\hat{\mu}_{MLE}$  and to be minimax. That the stronger condition of superharmonic  $m_\pi$ , i.e.  $\nabla^2 m_\pi(x) \leq 0$ , is sufficient follows immediately from (11). The fact that  $\hat{\mu}_H$  dominates  $\hat{\mu}_{MLE}$  when  $p \geq 3$ , now follows easily from the fact that superharmonic priors (of which the harmonic prior is a special case) yield superharmonic marginals  $m_\pi$  for  $X$ . Although the Strawderman priors  $\pi_a$  in (7) do not yield superharmonic  $m_\pi$  when  $p = 5$  and  $a \in [.5, 1)$  or when  $p \geq 6$  and  $a \in [0, 1)$ , Fourdrinier, Strawderman and Wells (1998) show that they do yield superharmonic  $\sqrt{m_\pi}$ , so that the dominance and minimaxity of the Strawderman estimator are established by (12). In fact, it follows from their results that  $\pi_a$  also yields superharmonic  $\sqrt{m_\pi}$  when  $a \in [1, 2)$  and  $p \geq 3$ , thereby broadening the class of minimax improper Bayes estimators.

A major thrust of the present paper is to establish an analogous unified theory for the KL risk prediction problem. Analogously to (10), we begin by showing how any Bayes rule  $p_\pi(y | x)$  can be explicitly represented in terms of the form of the corresponding marginal  $m_\pi(x)$ . This representation is seen to lead directly to an unbiased estimate of the KL risk reduction of Bayes rules  $p_\pi(y | x)$  over  $p_U(y | x)$ . Coupled with the heat equation, Stein's identity and the results of Fourdrinier et. al. (1998), the superharmonicity of  $\sqrt{m_\pi}$  is seen to be a sufficient condition for uniformly positive risk reduction over  $p_U(y | x)$ , thereby implying domination in the prediction problem. This general condition subsumes the specialized results of Komaki (2001) and Liang (2002), and can be used to obtain wide classes of improved minimax predictive distributions including the Bayes rules under  $\pi_H$  and  $\pi_a$ . It is further shown how the underlying priors and marginals can be adapted to obtain minimax shrinkage towards an arbitrary point or subspace, and how linear combinations of superharmonic priors and marginals can be constructed to obtain minimax multiple shrinkage predictive analogues of the minimax multiple shrinkage estimators of George (1986).

## 2 Improved Minimax Predictive Distributions

In this section, we develop and prove our main results concerning general conditions under which  $p_\pi(y | x)$  in (3) will dominate  $p_U(y | x)$  and will be minimax. We begin with three lemmas that may also be of independent interest.

The following general notation will be useful throughout. For  $Z | \mu \sim N_p(\mu, vI)$  and a prior  $\pi$  on  $\mu$ , we will denote the marginal distribution of  $Z$  by

$$m_\pi(z; v) = \int p(z | \mu) \pi(\mu) d\mu. \quad (13)$$

In terms of this notation, the marginal distributions of  $X | \mu \sim N_p(\mu, v_x I)$  and  $Y | \mu \sim N_p(\mu, v_y I)$  under  $\pi$  are then  $m_\pi(x; v_x)$  and  $m_\pi(y; v_y)$ , respectively. We will also make use of the weighted

mean

$$W = \frac{v_y X + v_x Y}{v_x + v_y}. \quad (14)$$

As  $X$  and  $Y$  are independent (conditionally on  $\mu$ ), it follows that  $W \mid \mu \sim N_p(\mu, v_w I)$  where

$$v_w = \frac{v_x v_y}{v_x + v_y}.$$

The marginal distribution of  $W$  is then  $m_\pi(w; v_w)$ .

**Lemma 1.** *If  $m_\pi(z; v)$  is finite for all  $z$ , then  $p_\pi(y \mid x)$  in (3) will be a probability distribution over  $y$ . Furthermore, the mean of  $p_\pi(y \mid x)$  is equal to  $E_\pi(\mu \mid x)$  if it exists.*

*Proof.* Both claims follows by integrating (3) wrt  $y$  and switching the order of integration. ‡

Lemma 1 is important because for our decision problem to be meaningful, it is necessary that a predictive estimate  $\hat{p}(y \mid x)$  be a proper probability distribution. By the laws of probability, the Bayes rule  $p_\pi(y \mid x)$  in (3) will be a proper probability distribution whenever  $\pi(\mu)$  is a proper prior. Lemma 1 shows that improper  $\pi(\mu)$  can still yield proper  $p_\pi(y \mid x)$  under a very weak condition. It further shows that the mean of  $p_\pi(y \mid x)$  is the Bayes rule for estimating  $\mu$  under quadratic loss, namely the posterior mean of  $\mu$ . In this sense,  $p_\pi(y \mid x)$  also carries the necessary information for that estimation problem. Our next lemma establishes an alternative representation of  $p_\pi(y \mid x)$ .

**Lemma 2.** *The Bayes rule  $p_\pi(y \mid x)$  in (3) can be expressed as*

$$p_\pi(y \mid x) = p_U(y \mid x) \frac{m_\pi(w; v_w)}{m_\pi(x; v_x)} \quad (15)$$

where  $p_U(y \mid x)$  is the Bayes rule under  $\pi_U(\mu) = 1$  given by (4).

*Proof.* The joint marginal distribution of  $X$  and  $Y$  under  $\pi$  is,

$$\begin{aligned} p_\pi(x, y) &= \int p(x \mid \mu) p(y \mid \mu) \pi(\mu) d\mu \\ &= \int \frac{1}{(2\pi v_x)^{\frac{p}{2}}} \exp\left\{-\frac{\|x - \mu\|^2}{2v_x}\right\} \frac{1}{(2\pi v_y)^{\frac{p}{2}}} \exp\left\{-\frac{\|y - \mu\|^2}{2v_y}\right\} \pi(\mu) d\mu \\ &= \int \frac{1}{\{2\pi(v_x + v_y)\}^{\frac{p}{2}}} \exp\left\{-\frac{\|y - x\|^2}{2(v_x + v_y)}\right\} \frac{1}{(2\pi v_w)^{\frac{p}{2}}} \exp\left\{-\frac{\|w - \mu\|^2}{2v_w}\right\} \pi(\mu) d\mu \\ &= p_U(y \mid x) m_\pi(w; v_w). \end{aligned}$$

The representation (15) now follows since  $p_\pi(x, y)$  and  $m_\pi(x; v_x)$  are the numerator and denominator of (3) respectively. ‡

Lemma 2 shows how  $p_\pi(y | x)$  is determined entirely by  $p_U(y | x)$  and the form of  $m_\pi(z; v)$ . Paralleling Brown's representation (10), this representation shows the explicit role played by the marginal distributions of the data under  $\pi$ . As will be seen in the proof of Theorem 1 below, this representation provides a key decomposition for establishing conditions for the dominance of  $p_\pi(y | x)$  over  $p_U(y | x)$ . But before moving on to this result, we provide one more result, an identity which plays a key role in establishing superharmonic conditions for dominance in Theorem 1. It may be of interest to note that this identity is in fact the well-known heat equation which has a long history in science and engineering, for example, see Steele (2001). Recently, Brown, DasGupta, Haff and Strawderman (2003) used identities derived from the heat equation, including one bearing a formal similarity to ours, in other contexts of inference and decision theory.

**Lemma 3.** *For  $Z | \mu \sim N_p(\mu, vI)$  and a given prior  $\pi$  on  $\mu$ , let  $m_\pi(z; v)$  be the marginal distribution of  $Z$  given by (13). Then*

$$\frac{\partial}{\partial v} m_\pi(z; v) = \frac{1}{2} \nabla^2 m_\pi(z; v). \quad (16)$$

*Proof.* The identity follows immediately by comparing

$$\begin{aligned} \frac{\partial}{\partial v} m_\pi(z; v) &= \frac{\partial}{\partial v} \int \frac{1}{(2\pi v)^{\frac{p}{2}}} \exp\left\{-\frac{\|z - \mu\|^2}{2v}\right\} \pi(\mu) d\mu \\ &= \int \left(-\frac{p}{2v} + \frac{\|z - \mu\|^2}{2v^2}\right) \frac{1}{(2\pi v)^{\frac{p}{2}}} \exp\left\{-\frac{\|z - \mu\|^2}{2v}\right\} \pi(\mu) d\mu, \end{aligned}$$

and

$$\begin{aligned} \nabla^2 m_\pi(z; v) &= \sum \frac{\partial^2}{\partial z_i^2} \int \frac{1}{(2\pi v)^{\frac{p}{2}}} \exp\left\{-\frac{\|z - \mu\|^2}{2v}\right\} \pi(\mu) d\mu \\ &= \int \left(-\frac{p}{v} + \frac{\|z - \mu\|^2}{v^2}\right) \frac{1}{(2\pi v)^{\frac{p}{2}}} \exp\left\{-\frac{\|z - \mu\|^2}{2v}\right\} \pi(\mu) d\mu. \quad \ddagger \end{aligned}$$

**Theorem 1.** *For  $Z | \mu \sim N_p(\mu, vI)$  and a given prior  $\pi$  on  $\mu$ , let  $m_\pi(z; v)$  be the marginal distribution of  $Z$  given by (13). If  $m_\pi(z; v)$  is finite for all  $z$ , then  $p_\pi(y | x)$  in (3) will dominate  $p_U(y | x)$  if any of the following hold for all  $v \leq v_x$ :*

- (i)  $\frac{\partial}{\partial v} E_{\mu, v} \log m_\pi(Z; v) < 0$ .
- (ii)  $\frac{\partial}{\partial v} m_\pi(\sqrt{v}z + \mu; v) \leq 0$  for all  $\mu$ , with strict inequality on some interval  $A$ .
- (iii)  $\sqrt{m_\pi(z; v)}$  is superharmonic with strict inequality on some interval  $A$ .
- (iv)  $m_\pi(z; v)$  is superharmonic with strict inequality on some interval  $A$ .
- (v)  $\pi(\mu)$  is superharmonic.

*Proof.* The difference between the risks of  $p_U(y | x)$  and  $p_\pi(y | x)$  can be expressed as

$$\begin{aligned}
R_{KL}(\mu, p_U) - R_{KL}(\mu, p_\pi) &= \int \int p(x | \mu) p(y | \mu) \log \frac{p_\pi(y | x)}{p_U(y | x)} dx dy \\
&= \int \int p(x | \mu) p(y | \mu) \log \frac{m_\pi(w; v_w)}{m_\pi(x; v_x)} dx dy \\
&= E_{\mu, v_w} \log m_\pi(W; v_w) - E_{\mu, v_x} \log m_\pi(X; v_x)
\end{aligned} \tag{17}$$

where (17) follows from the representation (15) of Lemma 1. Because  $v_w < v_x$ , (i) implies the dominance over  $p_U(y | x)$ .

Letting  $Z = \sqrt{v}Z^* + \mu \sim N_p(\mu, vI)$ , we obtain

$$\begin{aligned}
\frac{\partial}{\partial v} E_{\mu, v} \log m_\pi(Z; v) &= \frac{\partial}{\partial v} E \log m_\pi(\sqrt{v}Z^* + \mu; v) \\
&= E \frac{\frac{\partial}{\partial v} m_\pi(\sqrt{v}Z^* + \mu; v)}{m_\pi(\sqrt{v}Z^* + \mu; v)}
\end{aligned} \tag{19}$$

from which (ii) follows. Furthermore,

$$\begin{aligned}
\frac{\partial}{\partial v} m_\pi(\sqrt{v}z^* + \mu; v) &= \frac{\partial}{\partial v} \int \frac{1}{(2\pi v)^{\frac{p}{2}}} \exp \left\{ -\frac{\|\sqrt{v}z^* + \mu - \mu'\|^2}{2v} \right\} \pi(\mu') d\mu' \\
&= \int \left( -\frac{p}{2v} + \frac{\|z - \mu'\|^2}{2v^2} - \frac{\|z^*\|^2}{2v} - \frac{z^{*'}(\mu - \mu')}{2v^{3/2}} \right) p(z | \mu') \pi(\mu') d\mu' \\
&= \frac{\partial}{\partial v} m_\pi(z; v) - \int \frac{(z - \mu)'(z - \mu')}{2v^2} p(z | \mu') \pi(\mu') d\mu'.
\end{aligned}$$

Thus, by the heat equation (16) in Lemma 3 and Brown's representation  $E_\pi(\mu' | z) = z + \nabla \log m_\pi(z)$  from (10),

$$E \frac{\frac{\partial}{\partial v} m_\pi(\sqrt{v}Z^* + \mu; v)}{m_\pi(\sqrt{v}Z^* + \mu; v)} = E_{\mu, v} \left( \frac{1}{2} \frac{\nabla^2 m_\pi(Z; v)}{m_\pi(Z; v)} + \frac{(Z - \mu)' \nabla \log m_\pi(Z; v)}{2v} \right). \tag{20}$$

But by (2.3) of Stein (1981),

$$\begin{aligned}
E_{\mu, v} \frac{(Z - \mu)' \nabla \log m_\pi(Z; v)}{2v} &= E_{\mu, v} \frac{1}{2} \nabla^2 \log m_\pi(Z; v) \\
&= E_{\mu, v} \frac{1}{2} \nabla \frac{\nabla m_\pi(Z; v)}{m_\pi(Z; v)} \\
&= E_{\mu, v} \frac{1}{2} \left( \frac{\nabla^2 m_\pi(Z; v)}{m_\pi(Z; v)} - \|\nabla \log m_\pi(Z; v)\|^2 \right).
\end{aligned} \tag{21}$$

Combining (19), (20) and (21) yields

$$\frac{\partial}{\partial v} E_{\mu, v} \log m_\pi(Z; v) = E_{\mu, v} \left( \frac{\nabla^2 m_\pi(Z; v)}{m_\pi(Z; v)} - \frac{1}{2} \|\nabla \log m_\pi(Z; v)\|^2 \right) \tag{22}$$

$$= E_{\mu, v} 2 \frac{\nabla^2 \sqrt{m_\pi(Z; v)}}{\sqrt{m_\pi(Z; v)}} \tag{23}$$

where the last equality follows from the equality between (11) and (12). That (iii) implies (ii) follows from (23). That (iv) implies (iii) follows from (22) which demonstrates the well-known fact that  $\sqrt{m_\pi}$  will be superharmonic whenever  $m_\pi$  is superharmonic. Finally, it is straightforward to show that (v) implies (iv), (see problem 1.7.16 of Lehmann and Casella 1998). ‡

In the proof of Theorem 1,  $[\log m_\pi(w; v_w) - \log m_\pi(x; v_x)]$  appears in (18) as an unbiased estimate of the risk reduction of  $p_\pi(y | x)$  over  $p_U(y | x)$ . For the KL risk prediction problem, this estimate plays an analogous role to Stein's unbiased estimate of risk reduction  $U(x)$  in (11) and (12) for the quadratic risk estimation problem. However, the much stronger analogy is obtained by comparing (11) and (12) with (22) and (23). It follows directly from these that the risk reduction in the quadratic risk estimation problem can be expressed in terms of  $\log m_\pi$  as

$$R_Q(\mu, \hat{\mu}_{MLE}) - R_Q(\mu, \hat{\mu}_\pi) = -2 \left[ \frac{\partial}{\partial v} E_{\mu, v} \log m_\pi(Z; v) \right]_{v=1}.$$

We next turn to the issue of minimaxity. The following result is a special case of the general results of Barron and Liang (2003). We include it here with a direct proof for completeness.

**Lemma 4.**  $p_U(y | x)$ , the Bayes rule under  $\pi(\mu) = 1$ , is minimax for the risk  $R_{KL}(\mu, \hat{p})$  in (2).

*Proof.* By a transformation of variables,  $x \rightarrow (x - \mu)$  and  $y \rightarrow (y - \mu)$ , it is easy to see that  $R_{KL}(\mu, p_U) = R_{KL}(0, p_U) = r$  for all  $\mu$ , so that  $R_{KL}(\mu, p_U)$  is constant.

Next, we show that  $p_U(y | x)$  is a limit of Bayes rules  $p_{\pi_n}(y | x)$  with  $\pi_n(\mu) \sim N_p(0, \sigma_n^2 I)$ , where  $\sigma_n^2 \rightarrow \infty$  as  $n \rightarrow \infty$ . Using Lemma 1,

$$p_{\pi_n}(y | x) = p_U(y | x) \frac{m_{\pi_n}(w; v_w)}{m_{\pi_n}(x; v_x)} \rightarrow p_U(y | x)$$

since

$$m_{\pi_n}(z; v) = (2\pi(v + \sigma_n^2))^{-\frac{p}{2}} \exp \left\{ -\frac{\|z\|^2}{2(v + \sigma_n^2)} \right\}.$$

It follows that  $R_{KL}(\mu, p_{\pi_n}) \rightarrow R_{KL}(\mu, p_U) = r$  since  $R_{KL}(\mu, \hat{p})$  is a continuous function of  $\hat{p}$ , and then that  $r_{\pi_n} = \int R_{KL}(\mu, p_{\pi_n}) \pi_n(\mu) d\mu$ , the Bayes risk of  $p_{\pi_n}$  under  $\pi_n$ , goes to  $r$  since  $r_\pi$  is a continuous function of  $R_{KL}(\mu, p_\pi)$ . By Theorem 5.1.12 of Lehmann and Casella (1998), the minimaxity of  $p_U(y | x)$  now follows. ‡

If  $p_\pi$  satisfies any of the conditions of Theorem 1, it will dominate  $p_U$ . By Lemma 4, it will therefore be minimax. However, because all that is needed for minimaxity is  $R_{KL}(\mu, p_\pi) \leq R_{KL}(\mu, p_U)$  for all  $\mu$ , the conditions of Theorem 1 can be weakened. The proof of Theorem 2 below proceeds exactly as in the proof of Theorem 1.

**Theorem 2.** For  $Z | \mu \sim N_p(\mu, vI)$  and a given prior  $\pi$  on  $\mu$ , let  $m_\pi(z; v)$  be the marginal distribution of  $Z$  given by (13). If  $m_\pi(z; v)$  is finite for all  $z$ , then  $p_\pi(y | x)$  in (3) will be minimax if any of the following hold for all  $v \leq v_x$ :

- (i)  $\frac{\partial}{\partial v} E_{\mu, v} \log m_\pi(Z; v) < 0$ .
- (ii)  $\frac{\partial}{\partial v} m_\pi(\sqrt{v}z + \mu; v) \leq 0$  for all  $\mu$ .
- (iii)  $\sqrt{m_\pi(z; v)}$  is superharmonic.
- (iv)  $m_\pi(z; v)$  is superharmonic.
- (v)  $\pi(\mu)$  is superharmonic.

### 3 Minimax Shrinkage Towards 0

Let us now return to the Bayes rules  $p_H(y | x)$  and  $p_a(y | x)$ , the special cases of (3) under the harmonic prior  $\pi_H(\mu)$  in (6) and the Strawderman prior  $\pi_a(\mu)$  in (7). By Lemma 2, these Bayes rules can be expressed in terms of their respective marginals  $m_\pi$  as

$$p_\pi(y | x) = p_U(y | x) \frac{m_\pi(w; v_w)}{m_\pi(x; v_x)}. \quad (24)$$

Following Komaki (2001), the marginal of  $Z | \mu \sim N_p(\mu, vI)$  under  $\pi_H(\mu)$  can be expressed as

$$m_H(z; v) \propto v^{-(p-2)/2} \phi_p(\|z/\sqrt{v}\|) \quad (25)$$

where  $\phi_p(u) = u^{-p+2} \int_0^{\frac{1}{2}u^2} t^{\frac{p}{2}-2} \exp(-t) dt$  is the incomplete Gamma function. Because  $\pi_H$  is harmonic ( $\nabla^2 \pi_H(\mu) \equiv 0$ ), and hence superharmonic, for  $p \geq 3$ , Komaki's result that  $p_H(y|x)$  dominates  $p_U(y|x)$  follows immediately from (v) of Theorem 1, and the fact that  $p_H(y|x)$  is minimax from (v) of Theorem 2. This approach is both more direct and more general than Komaki's proof. For example, beyond  $p_H(y|x)$ , one might consider the class of Bayes rules  $p_\pi(y|x)$  corresponding to the (improper) multivariate  $t$  priors  $\pi(\mu) = (\|\mu\|^2 + 2/a_2)^{-(a_1+p/2)}$  considered by Faith (1978). Because these priors are superharmonic for  $a_1 \leq -1$  and  $p \geq 3$ , Bayes rules under such priors dominate  $p_U(y|x)$  by (v) of Theorem 1, and are minimax by (v) of Theorem 2.

Turning to  $p_a(y|x)$ , the marginal of  $Z | \mu \sim N_p(\mu, vI)$  under the Strawderman prior  $\pi_a$  can be expressed as

$$m_a(z; v) \propto \int_0^\infty \left\{ 2\pi v \left( \frac{v_0}{v} s + 1 \right) \right\}^{-p/2} \exp \left\{ -\frac{\|z/\sqrt{v}\|^2}{2 \left( \frac{v_0}{v} s + 1 \right)} \right\} (s+1)^{a-2} ds. \quad (26)$$

Because  $\pi_H(\mu)$  is the special case of  $\pi_a(\mu)$  when  $a = 2$ , it follows that  $m_H(z; v)$  is the special case of  $m_a(z; v)$  when  $a = 2$ . Liang (2002) proved that  $p_a(y|x)$  dominates  $p_U(y|x)$  by showing that

condition (i) of Theorem 1, namely  $\frac{\partial}{\partial v} E_{\mu, v} \log m_\pi(Z; v) < 0$ , holds for  $m_a(z; v)$  when  $v \leq v_x \leq v_0$ , and when  $p = 5$  and  $a \in [.5, 1)$  or  $p \geq 6$  and  $a \in [0, 1)$ . We now proceed to show that this result follows directly from (iii) of Theorem 1.

As Fourdrinier et. al. (1998) showed, the marginal for any proper prior cannot be superharmonic, so that conditions (iv) and (v) of Theorems 1 and 2 cannot hold for  $p_a(y|x)$  when  $a < 1$ . However, it does follow from Fourdrinier et. al. that  $\sqrt{m_a(z; v)}$  is superharmonic when  $v = v_0$  and when  $p = 5$  and  $a \in [.5, 1)$  or  $p \geq 6$  and  $a \in [0, 1)$ . However, to use Theorems 1 and 2, we will need the somewhat stronger result that  $\sqrt{m_a(z; v)}$  is superharmonic for any  $v \leq v_0$  when  $p = 5$  and  $a \in [.5, 1)$  or  $p \geq 6$  and  $a \in [0, 1)$ . Using Theorem 1 of Fourdrinier et. al. (1998), we obtain this result in Theorem 3 below.

For a nonnegative function  $h(s)$  over  $[0, \infty)$ , consider the scale mixture prior

$$\pi_h(\mu) = \int \pi(\mu | s v_0) h(s) ds, \quad (27)$$

where  $\pi(\mu | s v_0) = N_p(0, s v_0 I)$ . Note that  $\pi_a$  in (7) is the special case of  $\pi_h(\mu)$  when  $h(s) \propto (1+s)^{a-2}$ . For  $Z | \mu \sim N_p(\mu, vI)$ , the marginal distribution of  $Z$  under  $\pi_h(\mu)$  can be expressed as

$$m_h(z; v) \propto \int_0^\infty \{2\pi v(s+1)\}^{-p/2} \exp\left\{-\frac{\|z/\sqrt{v}\|^2}{2(s+1)}\right\} rh(rs) ds, \quad (28)$$

where  $r = v/v_0$ .

**Theorem 3.** *If  $h$  is a positive function such that*

(i)  $-(s+1)h'(s)/h(s)$  can be decomposed as  $l_1(s) + l_2(s)$  where  $l_1 \leq A$  is nondecreasing while  $0 < l_2 \leq B$  with  $\frac{1}{2}A + B \leq (p-2)/4$ ,

(ii)  $\lim_{s \rightarrow \infty} h(s)/(s+1)^{p/2} = 0$ .

Then

(a)  $\sqrt{m_h(z; v)}$  in (28) is superharmonic for all  $v \leq v_0$ .

(b) the Bayes rule  $p_h(y|x)$  under  $\pi_h(\mu)$  in (27) dominates  $p_U(y|x)$  and is minimax when  $v_x \leq v_0$ .

*Proof.* The proof of Theorem 1 of Fourdrinier et. al. (1998) shows that  $\sqrt{m_h(z; v)}$  in (28) is superharmonic when  $v = v_0 = 1$ , and it is straightforward to show that it is therefore superharmonic whenever  $v = v_0$ . From this fact and the expression for  $m_h(z; v)$  in (28), (a) will follow if  $h_r(s) := rh(rs)$  satisfies (i) and (ii) when  $r \in (0, 1]$ . First we show that  $h_r$  satisfies (i). By the assumptions on  $h$ , we have  $-(s+1)h'(s)/h(s)$  decomposed as  $\tilde{l}_1(s) + \tilde{l}_2(s)$ . Then

$$\begin{aligned} -(s+1) \frac{h'_r(s)}{h_r(s)} &= -\frac{r(s+1)}{rs+1} (rs+1) \frac{h'(rs)}{h(rs)} \\ &= \frac{r(s+1)}{rs+1} [\tilde{l}_1(s) + \tilde{l}_2(s)]. \end{aligned}$$

Choose  $l_i$  to be  $\tilde{l}_i$  multiplied by  $r(s+1)/(rs+1)$ . They can be checked to satisfy those conditions since the factor  $(rs+r)/(rs+1)$  is a nondecreasing function of  $s$  and less than or equal to 1 when  $0 < r \leq 1$ . To see that  $h_r$  satisfies (ii), note that

$$\frac{h_r(s)}{(s+1)^{\frac{p}{2}}} = \frac{h(rs)}{(rs+1)^{\frac{p}{2}}} r \left( \frac{rs+1}{s+1} \right)^{\frac{p}{2}}$$

goes to zero when  $s \rightarrow \infty$  since the first term goes to zero by the assumption on  $h$ .

By (a),  $\sqrt{m_h(z;v)}$  in (28) is superharmonic for all  $v \leq v_x$  when  $v_x \leq v_0$ . Thus, (b) follows from (iii) of Theorems 1 and 2. ‡

Fourdrinier et. al. (1998) showed that  $h(s) \propto (1+s)^{a-2}$ , which gives rise to the Strawderman prior  $\pi_a$  in (7), satisfies the conditions of Theorem 3 when  $3 - p/2 \leq a < 2$ . Thus, under the same conditions  $\sqrt{m_a(z;v)}$  will be superharmonic for any  $v \leq v_0$ , and the Bayes rule  $p_a(y|x)$  will dominate  $p_U(y|x)$  and be minimax when  $v_x \leq v_0$ . Since  $\pi_a(\mu)$  is proper when  $a \in [0,1)$ ,  $3 - p/2 \leq a < 2$  yields the same conditions for the domination and minimaxity of proper Bayes  $p_a(y|x)$  found by Liang (2002), namely  $p = 5$  and  $a \in [.5, 1)$  or when  $p \geq 6$  and  $a \in [0, 1)$ . Note that when  $a \in [1, 2)$  and  $\pi_a(\mu)$  is improper,  $3 - p/2 \leq a < 2$  is also satisfied for all  $p \geq 3$ , yielding another class of improper Bayes  $p_a(y|x)$  that dominate  $p_U(y|x)$  and are minimax when  $v_x \leq v_0$ . Going far beyond these results, Theorem 3 can be used to obtain wide classes of priors that yield proper Bayes minimax  $p_h(y|x)$  that dominate  $p_U(y|x)$ . Following the development in Section 4 of Fourdrinier et. al., such  $p_h(y|x)$  can be obtained with particular classes of shifted inverted gamma priors and classes of generalized t-priors.

Because  $\pi_H$  and  $\pi_a$  are unimodal about 0, it intuitively seems that the risk functions  $R_{KL}(\mu, p_H)$  and  $R_{KL}(\mu, p_a)$ , when  $p_a$  is minimax, should take on their minima at  $\mu = 0$ , and then asymptote up to  $R_{KL}(\mu, p_U)$  as  $\|\mu\| \rightarrow \infty$ . That this is exactly what happens is illustrated in Figures 1a and 1b which display the difference between the risk functions  $[R_{KL}(\mu, p_U) - R_{KL}(\mu, \hat{p})]$  for  $\hat{p} = p_H$  and  $\hat{p} = p_a$  with  $a = 0.5$ , at  $\mu = (c, \dots, c)'$ ,  $0 \leq c \leq 4$  when  $v_x = 1$  and  $v_y = 0.2$  for dimensions  $p = 3, 5, 7, 9$ . (These risk differences were calculated by simulating (17)). The largest risk reduction in all cases occurs close to  $\mu = 0$  and decreases rapidly to 0 as  $\|\mu\|$  increases. (Recall that  $R_{KL}(\mu, p_U)$  is constant as a function of  $\mu$ ). At the same time, risk reduction by  $p_H$  and  $p_a$  is larger for larger  $p$  at each fixed  $\|\mu\|$ . Note that  $p_a$  offers more risk reduction than  $p_H$ , apparently because it more sharply “shrinks  $p_U(y|x)$  towards 0” in the sense described below. Note also that when  $p = 3$ ,  $[R_{KL}(\mu, p_U) - R_{KL}(\mu, p_a)]$  is negative for larger  $\mu$ , a manifestation of the non minimaxity of  $p_a$  when  $a = 0.5$  and  $p = 3$ . All of these results parallel the risk relationships between minimax Stein estimators and the constant minimax risk of the MLE for quadratic loss estimation.

It is interesting to examine how these minimax Bayes rules  $p_\pi$  obtain the risk reduction when  $\mu$  is close 0. As is well-known, Bayes estimators of  $\mu$ , such as  $E_{\pi_H}(\mu|x)$  or  $E_{\pi_a}(\mu|x)$ , shrink  $x$  towards 0 by an adaptive multiplicative factor to obtain small risk (under quadratic loss) when  $\mu$

is close to 0. As can be seen from the representation (24),  $p_\pi(y | x)$  analogously “shrinks  $p_U(y | x)$  towards 0” by an adaptive multiplicative factor of the form

$$b_\pi(x, y) = \frac{m_\pi(w; v_w)}{m_\pi(x; v_x)}. \quad (29)$$

Figure 2 illustrates how this shrinkage occurs for  $p_H$  for various values of  $x$ . For  $x = (c_x, \dots, c_x)'$  and  $y = (c_y, \dots, c_y)'$ , Figure 2 plots  $p_H(y | x)$ ,  $p_U(y | x)$  and  $b_H(x, y) = b_\pi(x, y)$  when  $\pi = \pi_H$  as a function of  $c_y$  when  $c_x = 0, 1, 2$  and  $p = 5$ . Note first that  $p_U(y | x)$  is always the same shape centered at  $x$ . When  $c_x = 0$ ,  $b_H(x, y)$  modifies  $p_U(y | x)$  by increasing its concentration around 0. (The vertical scale for the  $b_H$  plots is several orders of magnitude larger than the scales for the  $p_H$  and  $p_U$  plots). When  $c_x = 1$  and 2,  $b_H(x, y)$  shrinks  $p_U(y | x)$  towards 0, leaving it skewed. As  $c_x$  increases, the spread of  $p_H(y | x)$  is increased. Note that although the height of the mode of  $b_H(x, y)$  increases as  $c_x$  increases, its effect on  $p_H(y | x)$  diminishes as it moves further from the mode of  $p_U(y | x)$ .

## 4 Minimax Shrinkage Towards Points or Subspaces

As seen in the previous section, shrinkage of  $p_U(y | x)$  towards zero by Bayes rules such as  $p_H(y | x)$  and  $p_a(y | x)$  offers most risk reduction when  $\mu$  is close to 0. Of course, one may prefer most reduction in other regions such as for  $\mu$  in a neighborhood of another  $b \in R^p$  or of a subspace  $B \subset R^p$ . Minimax Bayes rules  $p_\pi(y | x)$  yielding most reduction in such regions can easily be obtained by recentering the prior around such a point or subspace.

Recentering a prior  $\pi(\mu)$  around any  $b \in R^p$  is simply obtained as  $\pi^b(\mu) = \pi(\mu - b)$ . The marginal  $m_\pi^b$  corresponding to  $\pi^b$  can be directly obtained by recentering the marginal  $m_\pi$  corresponding to  $\pi$ , that is

$$m_\pi^b(z; v) = m_\pi(z - b; v). \quad (30)$$

Recentering priors such as  $\pi_H$  and  $\pi_a$ , which are unimodal around 0, yields priors  $\pi_H^b$  and  $\pi_a^b$  and hence marginals  $m_H^b$  and  $m_a^b$ , which are unimodal around  $b$ . Such recentered marginals yield predictive distributions

$$p_\pi^b(y | x) = p_U(y | x) \frac{m_\pi^b(w; v_w)}{m_\pi^b(x; v_x)} \quad (31)$$

that now shrink  $p_U(y | x)$  towards  $b$  rather than 0. It is straightforward to see that if  $m_\pi$  or  $\pi$  satisfy any of the conditions of Theorems 1, 2 or 3, then  $m_\pi^b$  or  $\pi^b$  will inherit those properties, so that dominance of  $p_U$  and minimaxity will be preserved by recentering.

More generally, to recenter a prior  $\pi(\mu)$  around a (possibly affine) subspace  $B \subset R^p$ , we restrict attention to spherically symmetric priors, namely priors that are functions of  $\mu$  only through  $\|\mu\|$ . Note that the harmonic prior  $\pi_H$ , the Strawderman prior  $\pi_a$  and all the other priors explicitly

mentioned in Section 3 are spherically symmetric. Recentering a spherically symmetric  $\pi(\mu)$  around  $B$  is obtained as

$$\pi^B(\mu) = \pi(\mu - P_B\mu), \quad (32)$$

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

$$\pi_H^B(\mu) = \|\mu - \bar{\mu}1_p\|^{-(p-3)}, \quad (33)$$

where  $\bar{\mu} = \mu'1_p/p$ . The exponent in (33) is  $(p - 1) - 1 = (p - 3)$  because  $(\mu - \bar{\mu}1_p)$  is  $(p - 1)$  dimensional.

The marginal  $m_\pi^B$  corresponding to the recentered  $\pi^B$  in (32) can be directly obtained by recentering the spherically symmetric marginal  $m_\pi$  corresponding to  $\pi$ , that is

$$m_\pi^B(z; v) = m_\pi(z - P_Bz; v), \quad (34)$$

where  $P_Bz$  is the projection of  $z$  onto  $B$ . Analogously to  $\pi^B(\mu)$ ,  $m_\pi^B(z; v)$  is uniform on  $P_Bz$  and applies  $m_\pi$  to  $(z - P_Bz)$ . Here too, the dimension of  $(z - P_Bz)$ , namely  $(p - \dim(B))$ , must be taken into account when evaluating  $m_\pi$ . For example, recentering the marginal  $m_H$  in around the subspace spanned by  $1_p = (1, \dots, 1)'$ , would entail replacing  $\|z\|$  by  $\|z - \bar{z}1_p\|$  where  $\bar{z} = z'1_p/p$ , and replacing  $p$  by  $(p - 1)$  in (25).

Applying the recentering (32) to priors such as  $\pi_H$  and  $\pi_a$ , which are unimodal around 0, yields priors  $\pi_H^B$  and  $\pi_a^B$  and hence marginals  $m_H^B$  and  $m_a^B$ , which are unimodal around  $B$ . Such recentered marginals yield predictive distributions

$$p_\pi^B(y | x) = p_U(y | x) \frac{m_\pi^B(w; v_w)}{m_\pi^B(x; v_x)} \quad (35)$$

that now shrink  $p_U(y | x)$  towards  $B$  by the multiplicative factor

$$b_\pi^B(x, y) = \frac{m_\pi^B(w; v_w)}{m_\pi^B(x; v_x)}. \quad (36)$$

For example, when  $m_\pi$  corresponds to the marginal  $m_H$ ,  $b_\pi^B(x, y)$  will behave much like  $b_H(x, y)$  in Figure 1. Shrinkage will be largest when  $x \in B$ , and will diminish as  $x$  moves away from  $B$ . Such  $p_\pi^B(y | x)$  obtain minimum risk when  $\mu \in B$ , but do not improve in any important way over  $p_U(y | x)$  when  $\mu$  is far from  $B$ .

The minimaxity and dominance of  $p_\pi^B(x, y)$  over  $p_U(y | x)$  can be established by applying Theorems 1, 2 or 3 directly to  $m_\pi^B$  or  $\pi^B$ . It will usually be the case that if  $m_\pi$  or  $\pi$  satisfy any of the conditions of Theorems 1, 2 or 3, then  $m_\pi^B$  or  $\pi^B$  will inherit those properties as long as  $(p - \dim(B))$  is large enough. For example, the recentered harmonic prior  $\pi_H^B(\mu)$  will only be superharmonic when  $(p - \dim(B)) \geq 3$ .

## 5 Minimax Multiple Shrinkage

To vastly enlarge the region of improved performance, we can further consider the following minimax multiple shrinkage modifications. For a spherically symmetric prior  $\pi(\mu)$ , a set of subspaces  $B_1, \dots, B_N$  of  $R^p$ , and a set of nonnegative weights  $w_1, \dots, w_N$  such that  $\sum_1^N w_i = 1$ , consider the mixture prior

$$\pi_*(\mu) = \sum_{i=1}^N w_i \pi^{B_i}(\mu), \quad (37)$$

where  $\pi^{B_i}$  are recentered priors given by (32). To simplify notation, we consider the case where each  $\pi^{B_i}$  is a recentering of the same  $\pi$ , although in principle such a construction can be applied with different priors. The marginal  $m_*$  corresponding to the mixture prior  $\pi_*$  in (37) is obtained as

$$m_*(z; v) = \sum_1^N w_i m_\pi^{B_i}(z; v) \quad (38)$$

where  $m_\pi^{B_i}$  are the recentered marginals corresponding to the  $\pi^{B_i}$  as given by (34).

Mixture marginals of the form  $m_*$  in (38) with superharmonic component  $m_\pi^{B_i}$  were used by George (1986abc) to construct minimax multiple shrinkage estimators of  $\mu$  under quadratic risk. These were obtained by applying the representation  $\hat{\mu}_\pi = x + \nabla \log m_\pi(x)$  in (10) to  $m_*$ . For the prediction problem, we instead apply the predictive construction (15) to  $m_*$  to obtain

$$p_*(y | x) = p_U(y | x) \frac{\sum_{i=1}^N w_i m_\pi^{B_i}(w; v_w)}{\sum_{i=1}^N w_i m_\pi^{B_i}(x; v_x)}. \quad (39)$$

It is easy to see that if every  $m_\pi^{B_i}(z; v)$  in (38) or  $\pi^{B_i}(\mu)$  in (37) satisfies conditions (ii), (iv) or (v) of Theorems 1 or 2, then  $m_*(z; v)$  in (38) will also satisfy those conditions, because those conditions are preserved under linear combination. Thus, we have the following which applies to recentered  $m_H$  and  $\pi_H$ , but not to recentered  $m_a$  and  $\pi_a$  which only satisfy (iii) of Theorems 1 and 2.

**Theorem 4.** *If each  $m_\pi^{B_i}(z; v)$  in (38) or  $\pi^{B_i}(\mu)$  in (37) satisfies conditions (ii), (iv) or (v) of Theorem 1, then  $p_*(y | x)$  in (39) will dominate  $p_U(y | x)$ . If each  $m_\pi^{B_i}(z; v)$  in (38) or  $\pi^{B_i}(\mu)$  in (37) satisfies conditions (ii), (iv) or (v) of Theorem 2, then  $p_*(y | x)$  in (39) will be minimax.*

To get some insight into the general behavior of  $p_*(y | x)$ , we can reexpress (39) as

$$p_*(y | x) = \sum_{i=1}^N p(B_i | x) p_\pi^{B_i}(y | x) \quad (40)$$

where  $p_\pi^{B_i}(y | x)$  is a single target predictive distribution of the form (35), and

$$p(B_i | x) = \frac{w_i m_\pi^{B_i}(x; v_x)}{\sum_{i=1}^N w_i m_\pi^{B_i}(x; v_x)}. \quad (41)$$

The form (40) reveals  $p_*(y | x)$  to be an adaptive convex combination of the individual shrinkage predictive distributions  $p_\pi^{B_i}(y | x)$ . The predictive distribution  $p_U(y | x)$  is doubly shrunk towards each  $B_i$  by  $p_*(y | x)$ ; it is first shrunk by  $p_\pi^{B_i}(y | x)$  and then by the component probability  $p(B_i | x)$ , the posterior weight on the  $i$ th prior component.

As a result of the shrinkage behavior of  $p_*$ , we would expect the risk reduction of  $R_{KL}(\mu, p_*)$  over  $R_{KL}(\mu, p_U)$  to be greatest when  $\mu$  is in or near any of  $B_1, \dots, B_N$ . This is precisely what happens with  $p_{H^*}$ , the multiple shrinkage version of  $p_H$ . Figure 3 illustrates the risk reduction  $[R_{KL}(\mu, p_U) - R_{KL}(\mu, p_{H^*})]$  at various  $\mu = (c, \dots, c)'$  obtained by  $p_{H^*}$  which adaptively shrinks  $p_U(y | x)$  towards the closer of the two points  $b_1 = (2, \dots, 2)$  and  $b_2 = (-2, \dots, -2)$  using equal weights  $w_1 = w_2 = 0.5$ . As in Figure 1a, 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  $\mu$  is close to  $b_1$  or  $b_2$ , and goes to 0 as  $\mu$  moves away from either of these points. At the same time, for each fixed  $\|\mu\|$ , risk reduction by  $p_{H^*}$  is larger for larger  $p$ . It is impressive that the size of the risk reduction offered by  $p_{H^*}$  is nearly the same as each of its single target counterparts. The cost of multiple shrinkage enhancement seems negligible, especially compared to the benefits.

## 6 The Case of Unknown Variance

In many realistic situations, the variances  $v_x$  and  $v_y$  will be unknown, but independent estimates of  $v_x$  may be available. If it can be assumed that  $v_y$  is proportional to  $v_x$ , so that  $v_y = r v_x$ , for a known constant  $r$ , then it is still possible to obtain improved prediction. We here consider the case where there exists an independent estimate of  $v_x$  of the form  $s/k$  where  $s$  is a realization of

$$S \sim v_x \chi_k^2, \quad (42)$$

$\chi_k^2$  representing the chi-square distribution with  $k$  degrees of freedom.

In this case, a straightforward solution is to simply substitute the estimates  $\hat{v}_x = s/k$ ,  $\hat{v}_y = rs/k$  and  $\hat{v}_w = \frac{r}{r+1}s/k$  for  $v_x$ ,  $v_y$  and  $v_w$  respectively, in  $p_U(y|x)$ ,  $m_\pi(x; v_x)$  and  $m_\pi(w; v_w)$ . The predictor

$$p_\pi^*(y | x) = p_U^*(y | x) \frac{m_\pi(w; \hat{v}_w)}{m_\pi(x; \hat{v}_x)} \quad (43)$$

will still dominate  $p_U^*(y | x)$  if any of the conditions in Theorem 1 is satisfied. This domination follows from the fact that

$$\begin{aligned} R_{KL}(\mu, p_U^*) - R_{KL}(\mu, p_\pi^*) &= \int \int \int p(s) p_{v_x}(x | \mu) p_{v_y}(y | \mu) \log \frac{p_\pi^*(y | x)}{p_U^*(y | x)} dx dy ds \\ &= \int \int \int p(s) p_{v_x}(x | \mu) p_{v_y}(y | \mu) \log \frac{m_\pi(w; \hat{v}_w)}{m_\pi(x; \hat{v}_x)} dx dy ds \\ &= E\{E_\mu[\log m_\pi(W; \hat{v}_w) - \log m_\pi(X; \hat{v}_x) | s]\}, \end{aligned}$$

and the fact that for any fixed  $s$ ,  $E_\mu[\log m_\pi(W; \hat{v}_w) - \log m_\pi(X; \hat{v}_x | s)]$  is positive given the conditions in Theorem 1.

Figure 4 illustrates the risk reduction  $[R_{KL}(\mu, p_U^*) - R_{KL}(\mu, p_H^*)]$  for  $\mu = (c, \dots, c)'$  obtained by  $p_{H^*}$  over  $p_U^*$ , the versions of  $p_H$  and  $p_U$  obtained with the substitutions  $\hat{v}_x = s/k$ ,  $\hat{v}_y = rs/k$  and  $\hat{v}_w = \frac{r}{r+1}s/k$  described above. As in Figure 1a, we considered the case  $v_x = 1, v_y = 0.2$  for  $p = 3, 5, 7, 9$  and so used  $r = 0.2$ . To obtain the variance estimates, we used  $s \sim \chi_k^2$  with  $k = 100$  degrees of freedom. Compared with Figure 1a, the plot reveals similar but slightly less risk reduction by  $p_{H^*}$  over  $p_U^*$ . The cost of estimating  $v_x$  is not substantial here. Of course, it should be emphasized that  $p_U^*(y | x)$  is no longer best invariant or minimax. Instead of plug-in variance estimates, a Bayesian treatment that integrates  $v_x$  out with respect to a prior would seem to be a more promising route for this unknown variance setup. We are currently investigating this approach and plan to report on it elsewhere.

## 7 Pseudo-Marginal Constructions?

Up to now, our results concerning Bayes rules for the prediction problem have paralleled the results for Bayes rules for the estimation problem as described in Section 1. However, as is well-known, Stein's unbiased estimate of risk reduction (11) for estimators of the form (10) can also be applied to certain pseudo-marginals, that is, functions  $m(x)$  that are not necessarily obtained as marginal distributions. For example, the positive-part James-Stein estimator

$$\hat{\mu}_S(x) = \left[ 1 - \frac{p-2}{\|x\|^2} \right]_+ x \quad (44)$$

can be expressed as

$$\mu_S(x) = x + \nabla \log m_S(x), \quad (45)$$

using the pseudo-marginal

$$m_S(x) = k_p \|x\|^{-(p-2)} \quad \text{if } \|x\|^2 \geq (p-2); \quad (46)$$

$$= \exp\{-\|x\|^2/2\} \quad \text{if } \|x\|^2 < (p-2), \quad (47)$$

where  $k_p = (e/(p-2))^{-(p-2)/2}$ , (see Stein 1974). Applying (11) with  $m_S$  in place of  $m_\pi$  reveals that the domination of  $\hat{\mu}_S$  over  $\hat{\mu}_{MLE}$  is a consequence of the fact that  $m_S(x)$  is superharmonic when  $p \geq 3$ .

Proceeding by analogy, it would seem that representation (15) from Lemma 2 could be formally extended to obtain similar results under KL loss. Letting  $m(z; v)$  be such a pseudo-marginal, we could formally define predictive estimates via

$$\hat{p}(y | x) = p_U(y | x) \frac{m(w; v_w)}{m(x; v_x)}. \quad (48)$$

If  $\hat{p}(y | x)$  so defined were a valid probability distribution and if  $m(z; v)$  were sufficiently smooth with respect to  $x$  and  $v$ , then condition (ii) in Theorems 1 and 2 would be sufficient to establish the domination and minimaxity of  $\hat{p}$ . If in addition,  $m(z; v)$  satisfied the heat equation (16), then conditions (iv) in Theorems 1 and 2 would also be sufficient. In this way, (48) could be used to formally construct minimax rules using functions  $m(z; v)$  that need not correspond to bona fide Bayesian marginals.

Unfortunately, we have so far been unable to find a satisfactory  $m(z; v)$  for such a minimax construction. An obvious pseudo-marginal candidate for  $m(z; v)$  would be the scale family elaboration of  $m_S(x)$ , namely

$$m_S(x; v) = \begin{cases} k_p \|x\|^{-(p-2)} & \text{if } \|x\|^2/v \geq (p-2); \end{cases} \quad (49)$$

$$= v^{-(p-2)/2} \exp\{-\|x\|^2/2v\} \quad \text{if } \|x\|^2/v < (p-2). \quad (50)$$

Although  $m_S(x; v)$  does not satisfy the heat equation (16), it does satisfy  $\frac{\partial}{\partial v} m_\pi(\sqrt{v}z + \mu; v) \leq 0$  for all  $\mu$ , and so by condition (ii) of Theorems 1 and 2, it would appear that

$$p_S(y | x) = p_U(y | x) \frac{m_S(w; v_w)}{m_S(x; v_x)} \quad (51)$$

dominates  $p_U(y | x)$  and be minimax. But this not true. The problem is that  $p_S(y | x)$  is not a bona fide predictive distribution,  $\int p_S(y | x) dy \neq 1$  and varies with  $x$ .

## 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., DasGupta, A., Haff, L.R., and Strawderman, W.E. (2003). Two Expectation Identities and Some Connections, *Jour. Statist. Planning and Inf., Special Issue in Memory of S.S. Gupta*, (to appear).
- 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.
- Komaki, F. (2001). A Shrinkage Predictive Distribution for Multivariate Normal Observations, *Biometrika*, 88, 859-864.
- 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.
- 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.
- Steele, J.M. (2001). *Stochastic Calculus and Financial Applications*. Springer, New York.
- 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.

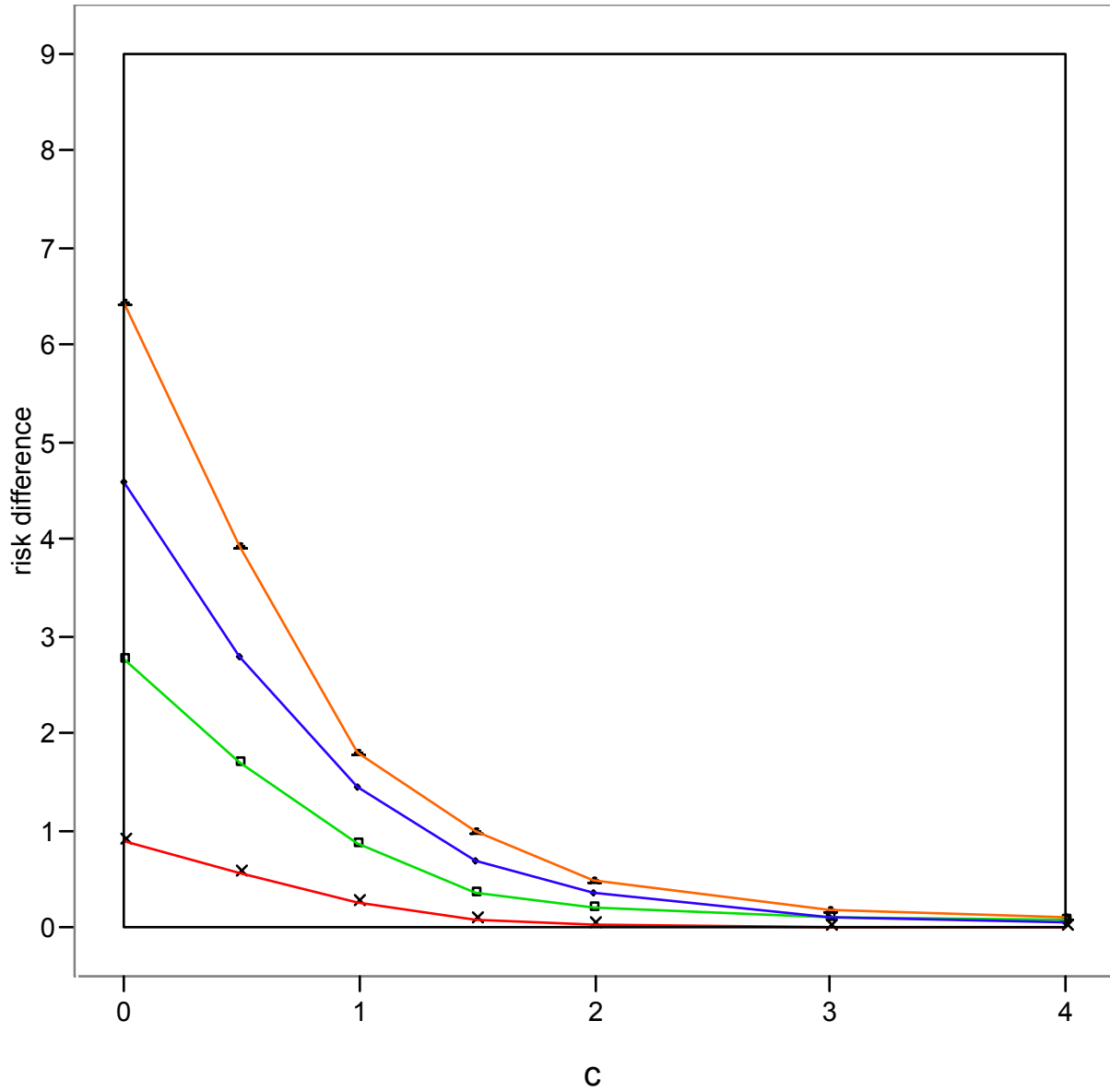


Figure 1a. The risk difference between  $p_U$  and  $p_H$ . Here  $\mu = (c, \dots, c)$ ,  $v_x = 1$ ,  $v_y = 0.2$ . The plots from lowest to highest correspond to  $p = 3, 5, 7, 9$ .

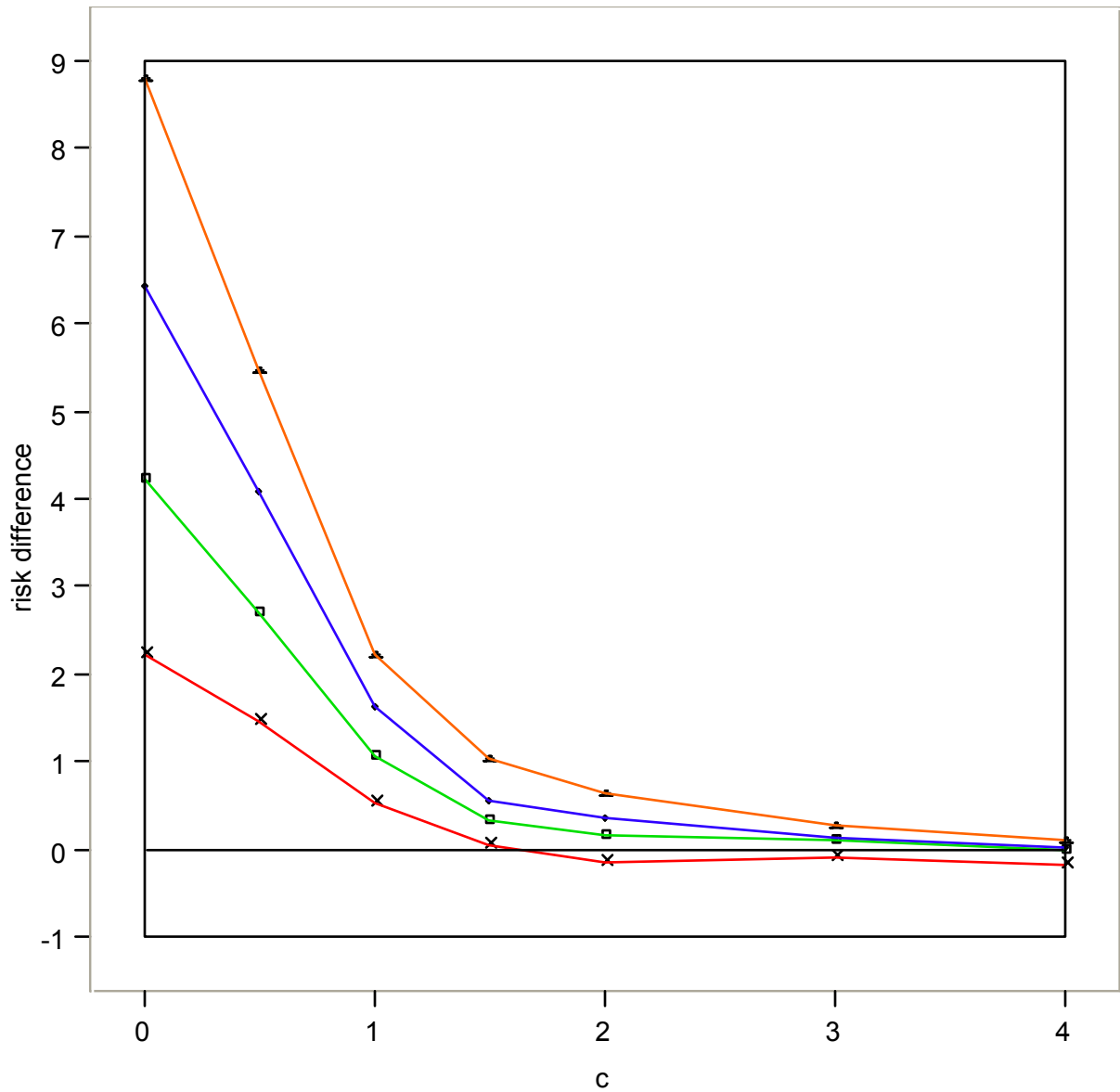


Figure 1b. The risk difference between  $p_U$  and  $p_a$  with  $a = 0.5$ . Here  $\mu = (c, \dots, c)$ ,  $v_x = 1$ ,  $v_y = 0.2$ . The plots from lowest to highest correspond to  $p = 3, 5, 7, 9$ .

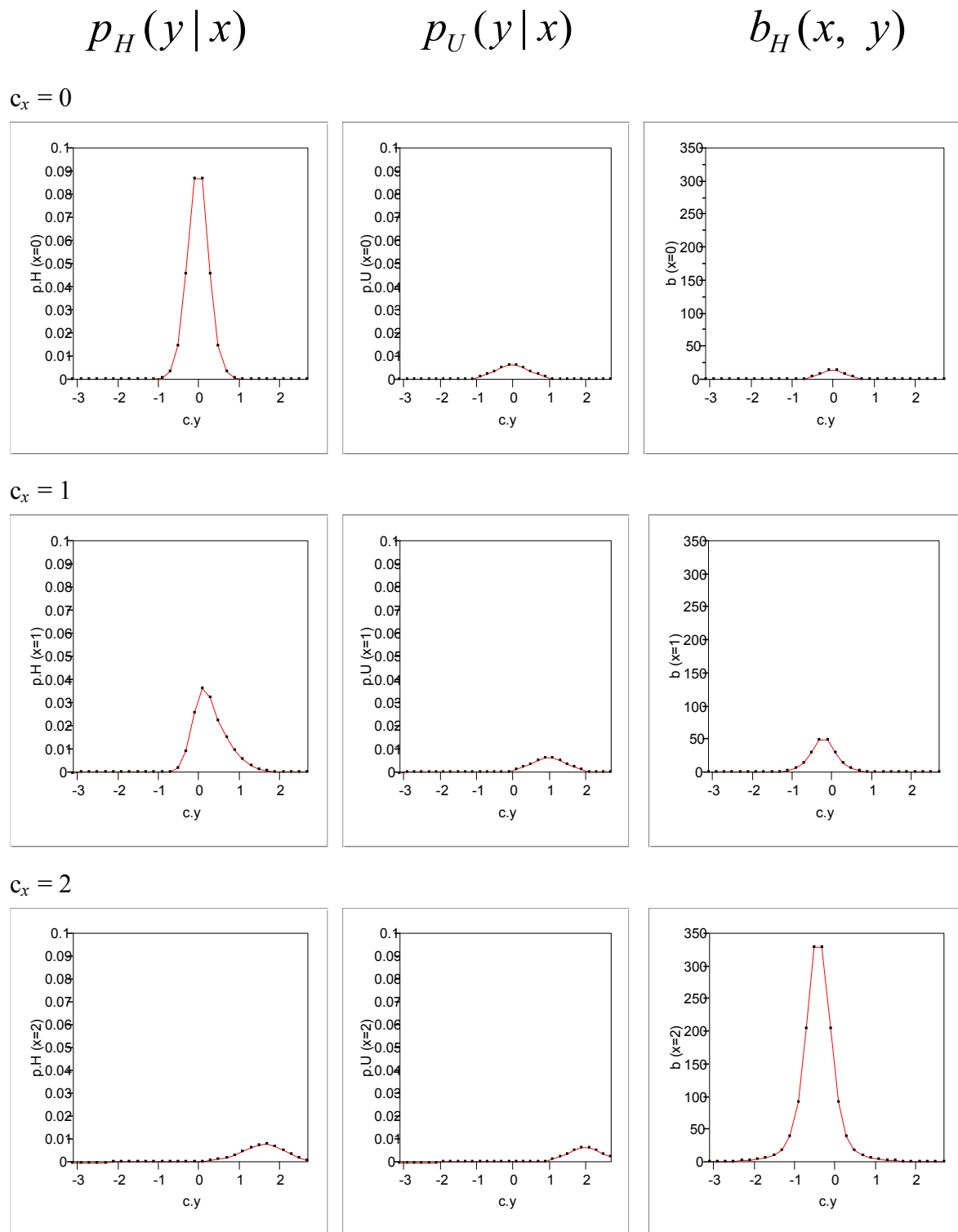


Figure 2. Shrinkage of  $p_U$  to obtain  $p_H$  when  $p = 5$ . For  $x = (c_x, \dots, c_x)$  and  $y = (c_y, \dots, c_y)$ , the plots show  $p_H(y|x)$ ,  $p_U(y|x)$ , and  $b_H(x,y)$  as functions of  $c_y$  when  $c_x = 0, 1, 2$ .

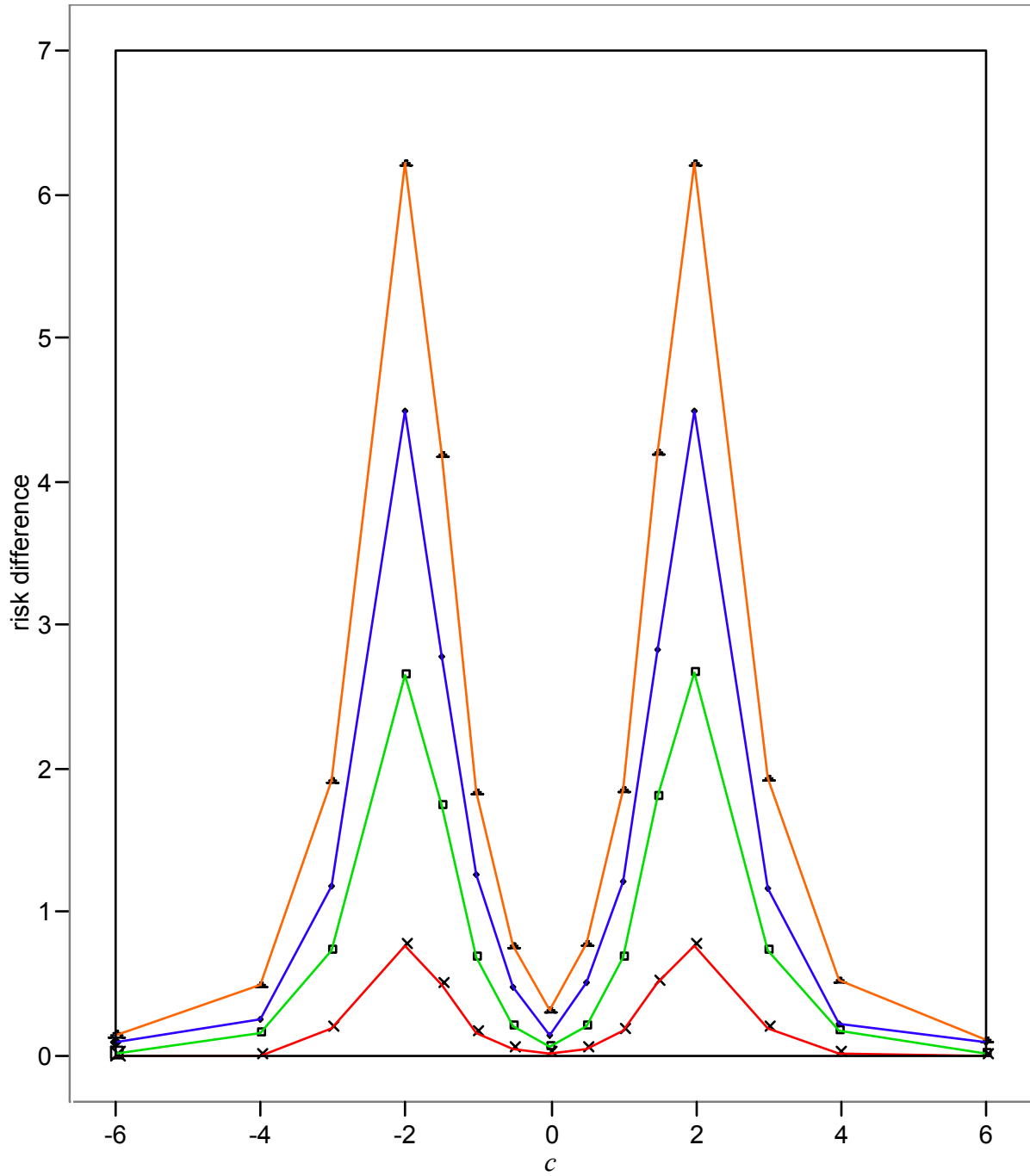


Figure 3. The risk difference between  $p_U$  and multiple shrinkage  $p_{H^*}$  with  $a = 0.5$ . Here  $\mu = (c, \dots, c)$ ,  $v_x = 1$ ,  $v_y = 0.2$ ,  $v_z = 0.2$ ,  $b_1 = -2$ ,  $b_2 = 2$ ,  $w_1 = w_2 = 0.5$ . The plots from lowest to highest correspond to  $p = 3, 5, 7, 9$ .

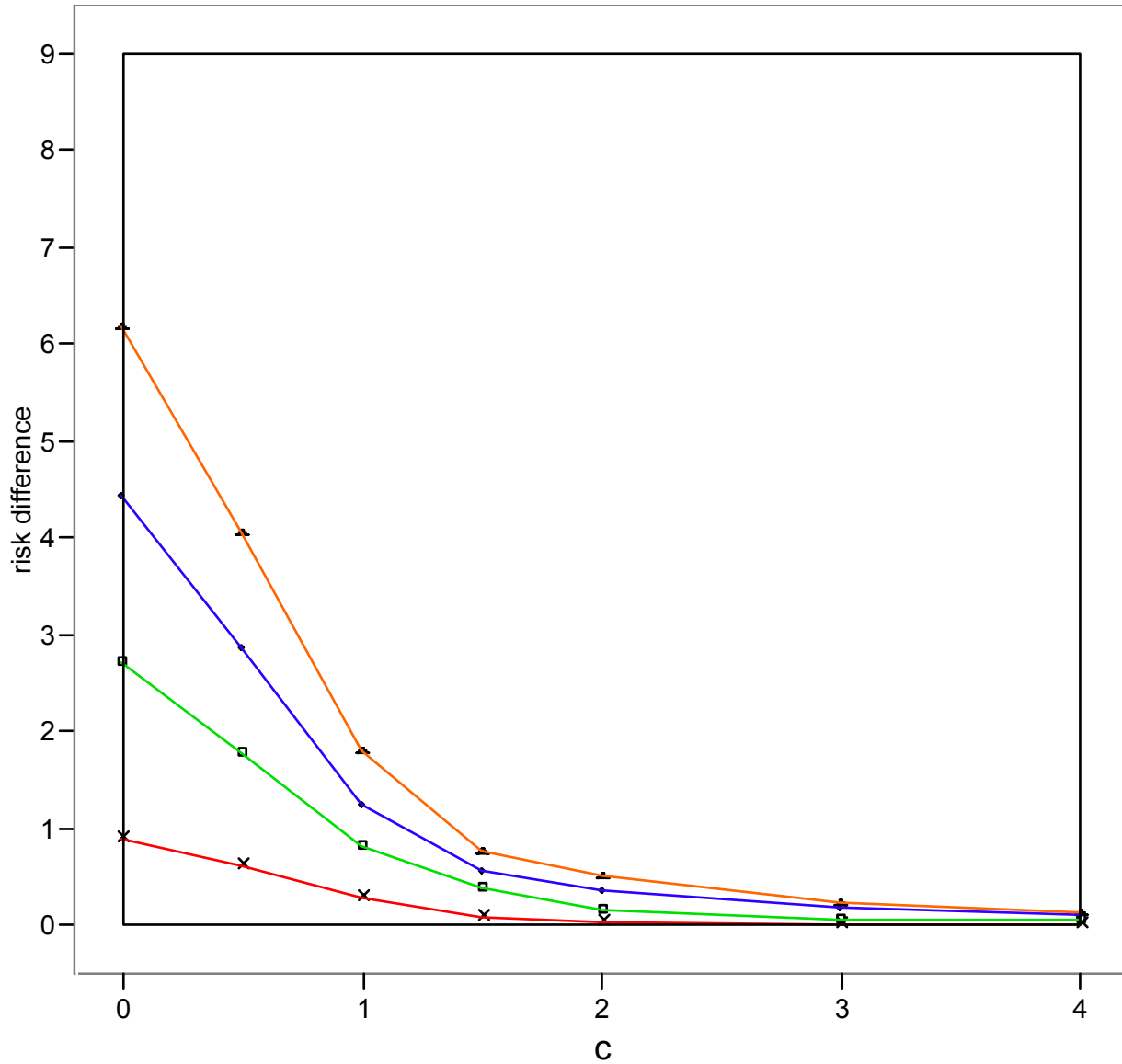


Figure 4. The risk difference between  $p_U$  and  $p_H$  in the unknown variance case. Here  $\mu = (c, \dots, c)$ ,  $v_x = 1$ ,  $v_y = 0.2$ . The plots from lowest to highest correspond to  $p = 3, 5, 7, 9$ .