High Dimensional Predictive Inference

Workshop on Current Trends and Challenges in Model Selection and Related Areas

Vienna, Austria
July 2008

Ed George
The Wharton School

(joint work with L. Brown, F. Liang, and X. Xu)

1. Estimating a Normal Mean: A Brief History

- Observe $X \mid \mu \sim N_p(\mu, I)$ and estimate μ by $\hat{\mu}$ under

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

- $\hat{\mu}_{MLE}(X) = X$ is the MLE, best invariant and minimax with constant risk

- Shocking Fact: $\hat{\mu}_{MLE}$ is inadmissible when $p \geq 3$. (Stein 1956)

- Bayes rules are a good place to look for improvements

- For a prior $\pi(\mu)$, the Bayes rule $\hat{\mu}_\pi(X) = E_\pi(\mu \mid X)$ minimizes $E_\pi R_Q(\mu, \hat{\mu})$

- Remark: The (formal) Bayes rule under $\pi_U(\mu) \equiv 1$ is

$$\hat{\mu}_U(X) \equiv \hat{\mu}_{MLE}(X) = X$$

- $\hat{\mu}_H(X)$, the Bayes rule under the Harmonic prior

$$\pi_H(\mu) = \|\mu\|^{-(p-2)},$$

dominates $\hat{\mu}_U$ when $p \geq 3$. (Stein 1974)

- $\hat{\mu}_a(X)$, the Bayes rule under $\pi_a(\mu)$ where

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

dominates $\hat{\mu}_U$ and is proper Bayes when $p = 5$ and $a \in [.5, 1)$ or when $p \geq 6$ and $a \in [0, 1)$. (Strawderman 1971)

- A Unifying Phenomenon: These domination results can be attributed to properties of the marginal distribution of X under π_H and π_a.

- The Bayes rule under $\pi(\mu)$ can be expressed as

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

where

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

is the marginal of X under $\pi(\mu)$. $(\nabla = (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_p})')$
(Brown 1971)

- The risk improvement of $\hat{\mu}_\pi(X)$ over $\hat{\mu}_U(X)$ can be expressed as

$$R_Q(\mu, \hat{\mu}_U) - R_Q(\mu, \hat{\mu}_\pi) = E_\mu \left[\|\nabla \log m_\pi(X)\|^2 - 2 \frac{\nabla^2 m_\pi(X)}{m_\pi(X)} \right]$$
$$= E_\mu \left[-4 \nabla^2 \sqrt{m_\pi(X)} / \sqrt{m_\pi(X)} \right]$$

$(\nabla^2 = \sum_i \frac{\partial^2}{\partial x_i^2})$ (Stein 1974, 1981)

- That $\hat{\mu}_H(X)$ dominates $\hat{\mu}_U$ when $p \geq 3$, follows from the fact that the marginal $m_\pi(X)$ under π_H is superharmonic, i.e.

$$\nabla^2 m_\pi(X) \leq 0$$

- That $\hat{\mu}_a(X)$ dominates $\hat{\mu}_U$ when $p \geq 5$ (and conditions on a), follows from the fact that the sqrt of the marginal under π_a is superharmonic, i.e.

$$\nabla^2 \sqrt{m_\pi(X)} \leq 0$$

(Fourdrinier, Strawderman and Wells 1998)

2. The Prediction Problem

- Observe $X \mid \mu \sim N_p(\mu, v_x I)$ and predict $Y \mid \mu \sim N_p(\mu, v_y I)$
  - Given μ, Y is independent of X
  - v_x and v_y are known (for now)

- The Problem: To estimate $p(y \mid \mu)$ by $q(y \mid x)$.

- Measure closeness by Kullback-Leibler loss,

$$L(\mu, q(y \mid x)) = \int p(y \mid \mu) \log \frac{p(y \mid \mu)}{q(y \mid x)} dy$$

- Risk function

$$R_{KL}(\mu, q) = \int L(\mu, q(y \mid x))\, p(x \mid \mu)\, dx = E_\mu[L(\mu, q(y \mid X)]$$

3. Bayes Rules for the Prediction Problem

- For a prior $\pi(\mu)$, the Bayes rule

$$p_\pi(y \mid x) = \int p(y \mid \mu)\pi(\mu \mid x)d\mu = E_\pi[p(y \mid \mu)|X]$$

minimizes $\int R_{KL}(\mu, q)\pi(\mu)d\mu$ (Aitchison 1975)

- Let $p_U(y \mid x)$ denote the Bayes rule under $\pi_U(\mu) \equiv 1$

- $p_U(y \mid x)$ dominates $p(y \mid \hat{\mu} = x)$, the naive "plug-in" predictive distribution (Aitchison 1975)

- $p_U(y \mid x)$ is best invariant and minimax with constant risk (Murray 1977, Ng 1980, Barron and Liang 2003)

- Shocking Fact: $p_U(y \mid x)$ is inadmissible when $p \geq 3$

- $p_H(y \mid x)$, the Bayes rule under the Harmonic prior

$$\pi_H(\mu) = \|\mu\|^{-(p-2)},$$

dominates $p_U(y \mid x)$ when $p \geq 3$. (Komaki 2001).

- $p_a(y \mid x)$, the Bayes rule under $\pi_a(\mu)$ where

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

dominates $p_U(y \mid x)$ and is proper Bayes when $v_x \leq v_0$ and when $p = 5$ and $a \in [.5, 1)$ or when $p \geq 6$ and $a \in [0, 1)$. (Liang 2002)

- Main Question: Are these domination results attributable to the properties of m_π?

4. A Key Representation for $p_\pi(y \mid x)$

- Let $m_\pi(x; v_x)$ denote the marginal of $X \mid \mu \sim N_p(\mu, v_x I)$ under $\pi(\mu)$.

- **Lemma**: The Bayes rule $p_\pi(y \mid x)$ can be expressed as

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

where

$$W = \frac{v_y X + v_x Y}{v_x + v_y} \sim N_p(\mu, v_w I)$$

- Using this, the risk improvement can be expressed as

$$R_{KL}(\mu, p_U) - R_{KL}(\mu, p_\pi) = \int\int p_{v_x}(x|\mu)\, p_{v_y}(y|\mu) \log \frac{p_\pi(y \mid x)}{p_U(y \mid x)} dx dy$$

$$= E_{\mu, v_w} \log m_\pi(W; v_w) - E_{\mu, v_x} \log m_\pi(X; v_x)$$

5. An Analogue of Stein's Unbiased Estimate of Risk

- **Theorem**:

$$\begin{aligned}\frac{\partial}{\partial v} E_{\mu,v} \log m_\pi(Z; v) &= E_{\mu,v}\left[\frac{\nabla^2 m_\pi(Z;v)}{m_\pi(Z;v)} - \frac{1}{2}\|\nabla \log m_\pi(Z;v)\|^2\right] \\ &= E_{\mu,v}\left[2\nabla^2 \sqrt{m_\pi(Z;v)}/\sqrt{m_\pi(Z;v)}\right]\end{aligned}$$

- Proof relies on using the heat equation

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

Brown's representation and Stein's Lemma.

6. General Conditions for Minimax Prediction

- Let $m_\pi(z; v)$ be the marginal distribution of $Z \mid \mu \sim N_p(\mu, vI)$ under $\pi(\mu)$.

- **Theorem**: If $m_\pi(z; v)$ is finite for all z, then $p_\pi(y \mid x)$ will be minimax if either of the following hold:

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

  (ii) $m_\pi(z; v)$ is superharmonic

- **Corollary**: If $m_\pi(z; v)$ is finite for all z, then $p_\pi(y \mid x)$ will be minimax if $\pi(\mu)$ is superharmonic

- $p_\pi(y \mid x)$ will dominate $p_U(y \mid x)$ in the above results if the superharmonicity is strict on some interval.

7. An Explicit Connection Between the Two Problems

- Comparing Stein's unbiased quadratic risk expression with our unbiased KL risk expression reveals

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

- Combined with our previous KL risk difference expression reveals a fascinating connection

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

- Ultimately it is this connection that yields the similar conditions for minimaxity and domination in both problems. Can we go further?

8. Sufficient Conditions for Admissibility

- Let $B_{KL}(\pi, q) \equiv E_\pi[R_{KL}(\mu, q)]$ be the average KL risk of $q(y \,|\, x)$ under π.

- **Theorem** (Blyth's Method): If there is a sequence of finite non-negative measures satisfying $\pi_n(\{\mu : \|\mu\| \leq 1\}) \geq 1$ such that

$$B_{KL}(\pi_n, q) - B_{KL}(\pi_n, p_{\pi_n}) \to 0$$

then q is admissible.

- **Theorem**: For any two Bayes rules p_π and p_{π_n}

$$B_{KL}(\pi_n, p_\pi) - B_{KL}(\pi_n, p_{\pi_n}) = \frac{1}{2} \int_{v_w}^{v_x} \frac{1}{v^2} \left[B_Q(\pi_n, \hat{\mu}_\pi) - B_Q(\pi_n, \hat{\mu}_{\pi_n})\right]_v dv$$

where $B_Q(\pi, \hat{\mu})$ is the average quadratic risk of $\hat{\mu}$ under π.

- Using the explicit construction of $\pi_n(\mu)$ from Brown and Hwang (1984), we obtain tail behavior conditions that prove admissibility of $p_U(y \,|\, x)$ when $p \leq 2$, and admissibility of $p_H(y \,|\, x)$ when $p \geq 3$.

9. A Complete Class Theorem

- **Theorem**: In the KL risk problem, all the admissible procedures are Bayes or formal Bayes procedures.

- Our proof uses the weak* topology from L^∞ to L^1 to define convergence on the action space which is the set of all proper densities on R^p.

- A Sletch of the Proof:

  (i) All the admissible procedures are non-randomized.

  (ii) For any admissible procedure $p(\cdot\,|\,x)$, there exists a sequence of priors $\pi_i(\mu)$ such that $p_{\pi_i}(\cdot\,|\,x) \to p(\cdot\,|\,x)$ weak* for a.e. x.

  (iii) We can find a subsequence $\{\pi_{i''}\}$ and a limit prior π such that $p_{\pi_{i''}}(\cdot\,|\,x) \to p_\pi(\cdot\,|\,x)$ weak* for almost every x. Therefore, $p(\cdot\,|\,x) = p_\pi(\cdot\,|\,x)$ for a.e. x, i.e. $p(\cdot\,|\,x)$ is a Bayes or a formal Bayes rule.

10. Predictive Estimation for Linear Regression

- Observe $\quad X_{m\times 1} = A_{m\times p}\, \beta_{p\times 1} + \varepsilon_{m\times 1}$
  and predict $\quad Y_{n\times 1} = B_{n\times p}\, \beta_{p\times 1} + \tau_{n\times 1}$
  - $\varepsilon \sim N_m(0, I_m)$ is independent of $\tau \sim N_n(0, I_n)$
  - $rank(A'A) = p$

- Given a prior π on β, the Bayes procedure $p^L_\pi(y \mid x)$ is

$$p^L_\pi(y \mid x) = \frac{\int p(x \mid A\beta) p(y \mid B\beta) \pi(\beta) d\beta}{\int p(x \mid A\beta) \pi(\beta) d\beta}$$

- The Bayes procedure $p^L_U(y \mid x)$ under the uniform prior $\pi_U \equiv 1$ is minimax with constant risk

11. The Key Marginal Representation

- For any prior π,

$$p_\pi^L(y \mid x) = \frac{m_\pi(\hat{\beta}_{x,y}, (C'C)^{-1})}{m_\pi(\hat{\beta}_x, (A'A)^{-1})}\, p_U^L(y \mid x)$$

where $C_{(m+n)\times p} = (A', B')'$ and

$$\hat{\beta}_x = (A'A)^{-1}A'x \sim N_p(\beta, (A'A)^{-1})$$

$$\hat{\beta}_{x,y} = (C'C)^{-1}C'(x', y')' \sim N_p(\beta, (C'C)^{-1})$$

12. Risk Improvement over $p_U^L(y \mid x)$

- Here the difference between the KL risks of $p_U^L(y \mid x)$ and $p_\pi^L(y \mid x)$ can be expressed as

$$R_{KL}(\beta, p_U^L) - R_{KL}(\beta, p_\pi^L) =$$

$$E_{\beta,(C'C)^{-1}} \log m_\pi(\hat{\beta}_{x,y}; (C'C)^{-1}) - E_{\beta,(A'A)^{-1}} \log m_\pi(\hat{\beta}_x; (A'A)^{-1})$$

- Minimaxity of $p_\pi^L(y \mid x)$ is here obtained when

$$\frac{\partial}{\partial \omega} E_{\mu,V_\omega} \log m_\pi(Z; V_\omega) < 0$$

where

$$V_\omega \equiv \omega(A'A)^{-1} + (1-\omega)(C'C)^{-1}$$

- This leads to weighted superharmonic conditions on m_π and π for minimaxity.

13. Minimax Shrinkage Towards 0

- Our Lemma representation

$$p_H(y \mid x) = \frac{m_H(w; v_w)}{m_H(x; v_x)}\, p_U(y \mid x)$$

  shows how $p_H(y \mid x)$ "shrinks $p_U(y \mid x)$ towards 0" by an adaptive multiplicative factor

- The following figure illustrates how this shrinkage occurs for various values of x.

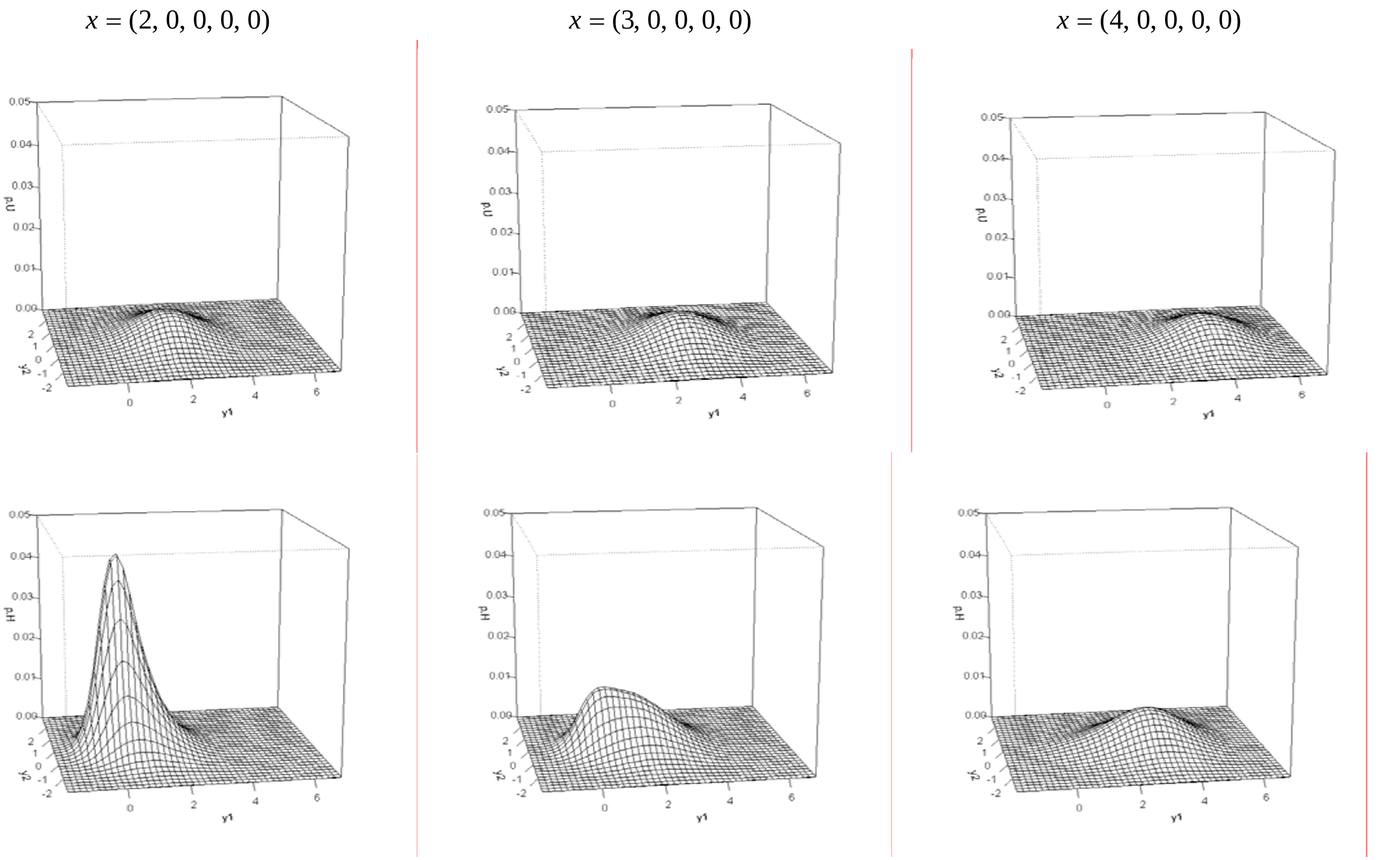


FIG 2. *Shrinkage of* $\hat{p}_U(y \mid x)$ *to obtain* $\hat{p}_H(y \mid x)$ *when* $v_x = 1$, $v_y = 0.2$ and $p = 5$. *Here* $y = (y_1, y_2, 0, 0, 0)$.

- Because π_H and $\sqrt{m_a}$ are superharmonic under suitable conditions, the result that $p_H(y \mid x)$ and $p_a(y \mid x)$ dominate $p_U(y \mid x)$ and are minimax follows immediately from our results.

- It also 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 Bayes rules that dominate $p_U(y \mid x)$ and are minimax.

- The following figures illustrate how the risk functions $R_{KL}(\mu, p_H)$ and $R_{KL}(\mu, p_a)$ take on their minima at $\mu = 0$, and then asymptote to $R_{KL}(\mu, p_U)$ as $\|\mu\| \to \infty$.

Figure 1a. The risk difference between q_U and q_H: $R(\mu, q_U) - R(\mu, q_H)$.
Here $\theta = (c, \cdots, c)$, $v_x = 1$, $v_y = 0.2$

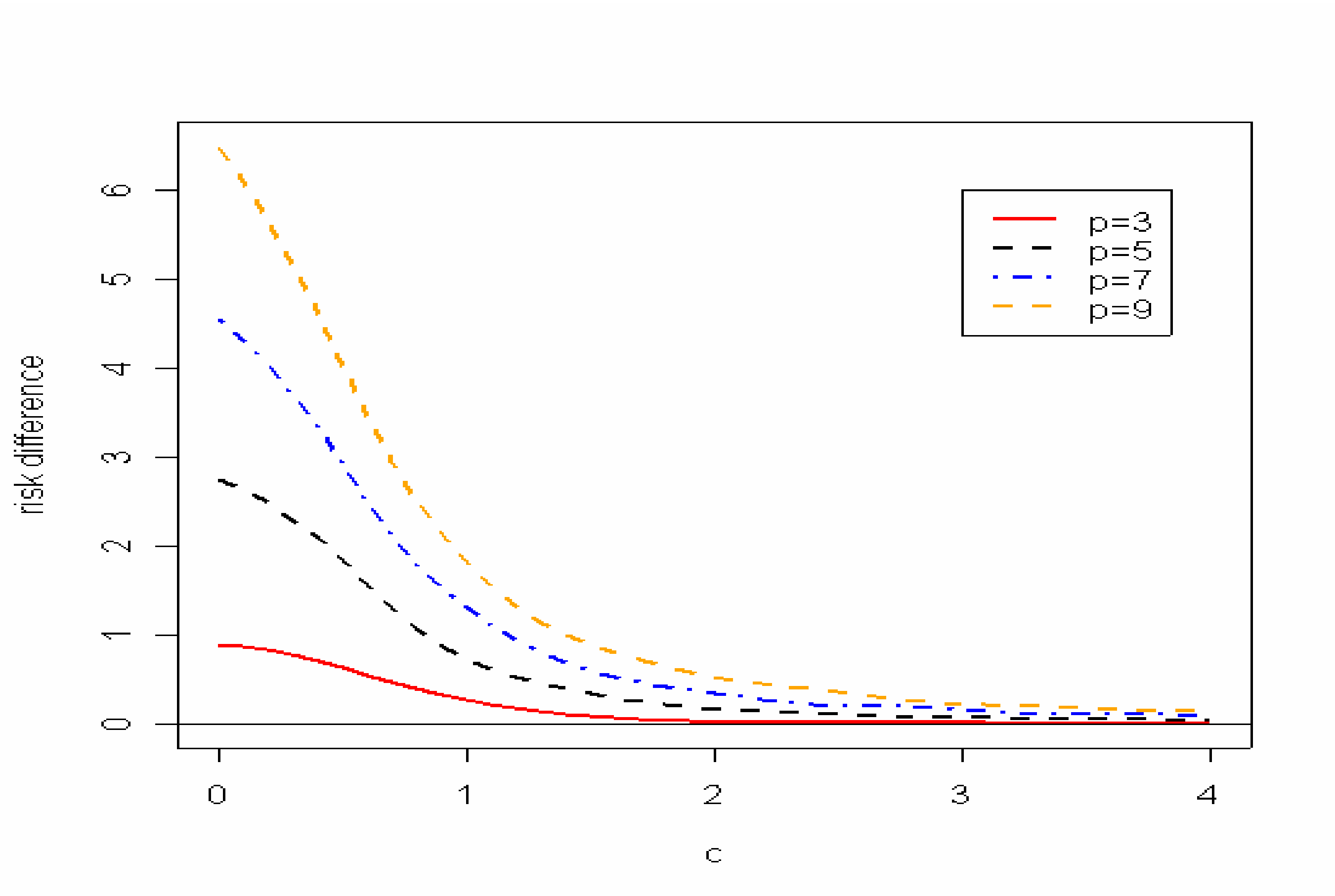
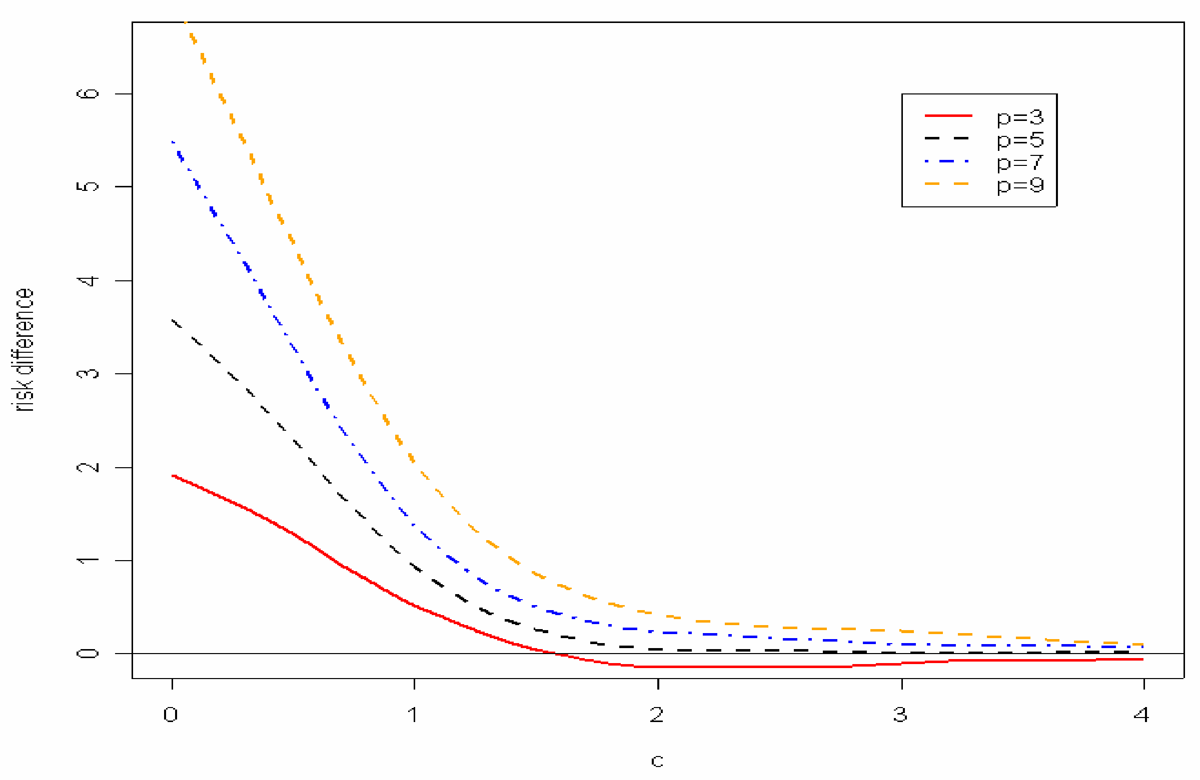


Figure 1b. The risk difference between q_U and q_a with $a = 0.5$: $R(\mu, q_U) - R(\mu, q_a)$.
Here $\theta = (c, \cdots, c)$, $v_x = 1$, $v_y = 0.2$



14. Shrinkage Towards Points or Subspaces

- We can trivially modify the previous priors and predictive distributions to shrink towards an arbitrary point $b \in R^p$.

- Consider the recentered prior

$$\pi^b(\mu) = \pi(\mu - b)$$

and corresponding recentered marginal

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

- This yields a predictive distribution

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

that now shrinks $p_U(y \mid x)$ towards b rather than 0.

- More generally, we can shrink $p_U(y \mid x)$ towards any subspace B of R^p whenever π, and hence m_π, is spherically symmetric.

- Letting $P_B z$ be the projection of z onto B, shrinkage towards B is obtained by using the recentered prior

$$\pi^B(\mu) = \pi(\mu - P_B \mu)$$

which yields the reecentered marginal

$$m_\pi^B(z; v) := m_\pi(z - P_B z; v).$$

- This modification yields a predictive distribution

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

that now shrinks $p_U(y \mid x)$ towards B.

- If $m_\pi^B(z; v)$ satisfies any of our superharmonic conditions for minimaxity, then $p_\pi^B(y \mid x)$ will dominate $p_U(y \mid x)$ and be minimax.

15. Minimax Multiple Shrinkage Prediction

- For any spherically symmetric prior, a set of subspaces $B_1, \ldots, B_N$, and corresponding probabilities $w_1, ..., w_N$, consider the recentered mixture prior

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

and corresponding recentered mixture marginal

$$m_*(z; v) = \sum_{1}^{N} w_i \, m_\pi^{B_i}(z; v).$$

- Applying the $\hat{\mu}_\pi(X) = X + \nabla \log m_\pi(X)$ construction with $m_*(X; v)$ yields minimax multiple shrinkage estimators of μ. (George 1986)

- Applying the predictive construction with $m_*(z; v)$ yields

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

where $p_\pi^{B_i}(y \mid x)$ is a single target predictive distribution and

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

is the posterior weight on the ith prior component.

- **Theorem**: If each $m_\pi^{B_i}(z; v)$ is superharmonic, then $p_*(y \mid x)$ will dominate $p_U(y \mid x)$ and will be minimax.

- The following final figure illustrates how the risk reduction obtained by the multiple shrinkage predictor p_{H^*} which adaptively shrinks $p_U(y|x)$ towards the closer of the two points $b_1 = (2, \ldots, 2)$ and $b_2 = (-2, \ldots, -2)$ using equal weights $w_1 = w_2 = 0.5$

Figure 3. The risk difference between p_U and multiple shrinkage p_{H^*}: $R(\mu,\, p_U) - R(\mu,\, p_{H^*})$.
Here $\theta = (c, \cdots, c)$, $v_x = 1$, $v_y = 0.2$, $a_1 = 2,\ a_2 = -2$, $w_1 = w_2 = 0.5$.