Statistics & Decisions 9, 81 - 105 (1991)

ON SHRINKAGE TOWARDS AN ARBITRARY ESTIMATOR

D. Sengupta

Received: Revised version: August 14, 1990

Abstract. *Suppose we want to estimate the unknown mean* θ *of a multi-normal distribution with independent components having a common variance* σ^2. *Several minimax estimators are known in this context. We consider a class of spherically symmetric estimators with random pivot of the following form*

$$d(X) = \hat{\theta}_0 + \left(1 - \frac{\hat{\sigma}^2 r(1/2\, \|X-\hat{\theta}_0\|^2)}{\|X-\hat{\theta}_0\|^2}\right)(X - \hat{\theta}_0),$$

where $\hat{\theta}_0$ *is another estimator of* θ *and* $\hat{\sigma}^2$ *is an estimator of* σ^2 *independent of* X. *We give conditions on* $\hat{\theta}_0$ *and* r *so that* d *is minimax under squared error loss. It is expected that* d *will improve upon* X *significantly in a region where* $\hat{\theta}_0$ *performs well. Now by choosing* $\hat{\theta}_0$ *a Bayes estimator under some prior* π, *we develop a concept of optimal shrinkage. The problem of selecting a good minimax estimator has already been emphasized by Berger (1982). Georde (1986a) gave a reasonable solution in some cases. Here we develop another way of looking at the above problem.*

AMS subject classifications: 62 F 10, 62 F 15

Key words: admissibility, minimaxity, Bayes estimators, exponential families, shrinkage estimators

1. Introduction

Beginning with the pioneering work of Stein (1956), a great deal of research has been done in minimax estimation of a multivariate normal mean. Suppose that

$$X \sim N_p(\theta, \sigma^2 I) \ , \tag{1.1}$$

with unknown mean $\theta = (\theta_1, \ldots, \theta_p)'$ and variance σ^2 (may be known). We consider the problem of estimating the mean vector θ under normalized quadratic loss. The risk function of an estimator $d\colon \mathbf{R}^p \to \mathbf{R}^p$ is given by

$$R(\theta, d) = \frac{1}{\sigma^2} E_\theta(d(X) - \theta)'(d(X) - \theta) = \frac{1}{\sigma^2} E_\theta\|d(X) - \theta\|^2 \ , \tag{1.2}$$

where E_θ denotes the expectation under (1.1).

Let $\mathbf{C}$ be the class of estimators with everywhere finite risk. An estimator $d^* \in \mathbf{C}$ is called minimax if the following condition holds

$$\sup_\theta R(\theta, d^*) = \inf_{d \in \mathbf{C}} \sup_\theta R(\theta, d) \ . \tag{1.3}$$

The classical example of a minimax estimator for the above problem is

$$d_0(X) = X \ .$$

The estimator d_0 is not only minimax, it has other statistically appealing features as well. It is known to be the maximum likelihood estimator and also the minimum variance unbiased estimator. It is now easy to see that

$$R(\theta,\ d_0) = p\ .$$

Because d_0 is minimax with constant risk, any other minimax estimator will have a risk not larger than d_0 at every $\theta \in \mathbf{R}^p$. Thus the existence of other minimax estimators of θ proves that d_0 is inadmissible under (1.2).

Next suppose that we have several minimax estimators of θ . The above fact asserts that each of them gives us some reduction in risk over d_0 . But, which one do we choose? This issue has to be taken care of before we decide to use some minimax estimator other than d_0 for estimating θ .

To fix idea, assume $\sigma^2 = 1$ and consider the family of estimators

$$d^{\mu}(X) = \left[1 - \frac{p-2}{\|X - \mu\|^2}\right]^{+}(X - \mu) + \mu\ ,\quad \mu \in \mathbf{R}^p\ . \tag{1.4}$$

For $p \geq 3$, it is known that each of these estimators is minimax. See Strawderman (1971). Also, it can be shown that for a fixed μ , d^{μ} has significantly smaller risk than d_0 only in a neighborhood of μ . The risk reduction is almost insignificant if the true θ is away from μ . Now, if we have any prior reason to believe that the true θ is close to some μ_0 , it is natural to use a d^{μ} with μ lying in a close vicinity of μ_0 .

Usually such a strong prior information is not available in practice. Thus, in general, the problem of selecting a good minimax estimator becomes quite involved.

The importance of this issue has been discussed in Berger (1982). Also, George (1986a, b) gives a reasonable solution to the problem of selecting a good minimax estimator in some cases of interest. Here we take the following approach.

Consider the class of estimators given by (1.4). Suppose that some kind of prior information is available for estimating θ (may be an order relation among various components of θ). Let $\hat{\theta}_0$ denote an estimator that uses the available prior information. In case we have a probabilistic prior π , $\hat{\theta}_0$ might be the Bayes estimator of θ .

Now as an alternative estimator of θ we consider the following adaptive linear combination of X and $\hat{\theta}_0$ for some $r \geq 0$:

$$d_*(X) = \hat{\theta}_0 + \left(1 - \frac{r(\frac{1}{2}\|X - \hat{\theta}_0\|^2)}{\|X - \hat{\theta}_0\|^2}\right)(X - \hat{\theta}_0) . \tag{1.5}$$

d_* has the property that it is closer to $\hat{\theta}_0$ than X when X is close to $\hat{\theta}_0$. Thus, provided d_* remains minimax it is obviously a reasonable minimax estimator.

George (1986a) solves the problem in a different manner. Suppose that we have a probabilistic prior π on $\mathbb{R}^p$. It was shown by George (1986a) that the following estimator

$$\delta_*(X) = \int_{\mathbb{R}^p} d^{\mu}(X) w_{\mu}(X) d\pi(\mu) \tag{1.6}$$

will be minimax under certain conditions for suitable choice of the weight function $w_\mu(X)$.

When π is a prior with support at finitely many points (1.6) reduces to the multiple shrinkage estimator introduced by George. The main result in this paper derives some conditions for which (1.5) is minimax.

The main result is given in Section 2. In Section 2.2, some applications are considered. In Section 3, we deal with the problem of shrinking towards a Bayes estimator and discuss an optimality criterion for optimal shrinkage. Finally, as concluding remarks, we discuss the problem of constructing Bayes minimax estimator. In section 2, we deal with the case where σ^2 is unknown. In the remaining sections σ^2 is assumed to be 1. However the results can be extended to the unknown σ^2 case under the assumptions of section 2 quite easily.

2. Shrinking towards an arbitrary estimator

In this section we treat the case where σ^2 is unknown. The observation vector X denotes the standard estimator of θ . Let $\hat{\theta}_0 = (\hat{\theta}_{01}, \ldots, \hat{\theta}_{0p})'$ be another estimator of θ with continuously differentiable components so that

$$E_\theta \frac{1}{\|X - \hat{\theta}_0\|^2} < \infty \quad \text{for all } \theta \in \mathbb{R}^p \text{ and } p \geq 3 . \tag{2.1}$$

Note that when $\hat{\theta}_0 = 0$, p must be at least 3 for (2.1) to be finite. We also need to assume that an estimator $\hat{\sigma}^2$ (> 0 with Probability 1) of σ^2 is available which is independent of X with the following property:

$$0 < \rho := \inf_{\sigma^2>0} \frac{\sigma^2 E\hat{\sigma}^2}{E\hat{\sigma}^4} . \tag{2.1}$$

Let $J_0(x) = [(j_{0,ki} = \hat{\theta}_{0k}^{(i)}, 1 \le i, k \le p)]$ where $\hat{\theta}_{0k}^{(i)} = \frac{\partial}{\partial x_i} \hat{\theta}_{0k}$ denote the Jacobian of $\hat{\theta}_0$. Now we state the main result of this section. A slight modification of the arguments given in Stein (1981) gives us the result. The proof is given in the appendix.

Theorem 2.1. Consider the estimator

$$d_*(X) = \hat{\theta}_0 + \left[1 - \frac{\hat{\sigma}^2 \, r(\tfrac{1}{2}\|X - \hat{\theta}_0\|^2)}{\|X - \hat{\theta}_0\|^2}\right](X - \hat{\theta}_0) \tag{2.2}$$

where $r: [0, \infty) \to \mathbf{R}$ is a bounded almost differentiable function with derivative r' with respect to the Lebesgue measure. Without loss of generality assume that $J_0(x)$ is symmetric and let $\lambda_{max}(J_0(x))$ and $\lambda_{min}(J_0(x))$ denote the maximum and the minimum eigenvalue of J_0 , respectively.

If $(I - J_0(x))$ is non-negative definite for all x satisfying (2.1) d_* (defined in (2.2)) has smaller risk than X for all θ provided either

a) $p > 2 + \text{trace}(J_0(x)) - 2\,\lambda_{min}(J_0(x)) + \epsilon$ for all x and some $\epsilon > 0$ with $r'(\cdot) \ge 0$ and $0 \le r(\cdot) < 2\epsilon\rho$, or

b) $\text{trace}(J_0(x)) > p-2 + 2\,\lambda_{max}(J_0(x)) + \epsilon$ for all x and some $\epsilon > 0$ with $r'(\cdot) \ge 0$ and $-2\epsilon\rho < r(\cdot) \le 0$.

If $(I - J_0(x))$ is non-positive definite for all x, the same conclusion as above holds if we replace $r' \geq 0$ by $r' \leq 0$.

Remark 1. It can be noticed that shrinking towards arbitrary estimators sometimes has an overall shrinkage effect and sometimes an overall expanding effect. The overall effect is determined by the Jacobian of the estimator $\hat{\theta}_0$. We discuss some examples below which can be derived using the above result.

Remark 2. When σ^2 is known we can use the estimator $\hat{\sigma}^2 = \sigma^2$ to estimate σ^2. In that case e [as defined by (2.1)] turns out to be 1. We make use of this fact in the next example.

2.1. Some examples

Example 1. Let us first consider the classical situation where $\hat{\theta}_0 = 0$. Therefore, J_0 is the zero matrix and the conditions of Theorem 2.1 can be easily verified to give us

$$d_*^0(X) = \left(1 - \frac{r(\frac{1}{2}\|X\|^2)}{\|X\|^2}\right)X \tag{2.9}$$

is minimax where $p \geq 3$ and $r(\cdot)$ is monotone nondecreasing with $0 < r(\cdot) \leq 2(p-2)$. [Hence $\epsilon < (p-2)$ works.] This has been obtained by Strawderman (1971). With the choice

$$r(t) = \begin{cases} 2t & \text{if } t < \frac{p-2}{2} \\ p-2 & \text{if } t \geq \frac{p-2}{2} . \end{cases}$$

Equation (2.9) coincides with the positive rule estimator.

For the choice $\hat{\theta}_0 = P_V X$, where P_V is the projection on some smaller dimensional (d) subspace V of $\mathbb{R}^p$, we can apply the Theorem 2.1 to obtain

$$d_*^V(X) = P_V X + \left(1 - \frac{r(\frac{1}{2}\|P_{V^\perp}X\|^2)}{\|P_{V^\perp}X\|^2}\right) P_{V^\perp}X \tag{2.10}$$

(where $P_{V^\perp}$ denote the projection onto the orthogonal complement of V) is minimax if $p > d + 2$ and $0 < r(\cdot) < 2(p-d-2)$ with $r' \geq 0$. See Lindley (1971) and Sclove, Morris and Radhakrishnan (1972) for these types of shrinkage estimators.

Example 2. Next we consider a situation which is not so common in the area of shrinkage estimation. Suppose we have a vague prior information on θ that each component is non-negative. Now from Katz (1961) it follows that there exists an estimator $\hat{\theta}_0$ which is given componentwise by

$$\hat{\theta}_{0i} = X_i + t(X_i) \ , \quad i=1,\, 2,\, \ldots,\, p \quad \text{for some} \quad t\colon \mathbb{R} \to \mathbb{R}_+ \tag{2.11}$$

and $\hat{\theta}_0$ is componentwise admissible when the i-th component θ_i is restricted to $\theta_i \geq 0$.

We now consider $\hat{\theta}_0$ of the form (2.11) and introduce a slightly different version of Katz' estimator. He considered

$$t(x) = \frac{\exp(-\frac{1}{2}x^2)}{\int_{-\infty}^{x} \exp(-\frac{1}{2}u^2)\,du} \ . \tag{2.12}$$

Unfortunately, $E_\theta\|X - \hat{\theta}_0\|^{-2}$ is not finite with this choice of t. We choose a truncated version of the above estimator of the following form

$$t_a(x) = t(x) \quad \text{for} \quad x < a$$
$$\qquad = t(a) \quad \text{for} \quad x \geq a \tag{2.13}$$

for some $a > 0$. By making a large enough, we can make t_a look like t on an arbitrary large interval. Another possibility is to choose a Bayes estimator where the prior puts all of its mass on the positive orthant. For the choice (2.13)

$$t'_a(x) = t'(x) \quad \text{for} \quad x < a$$
$$\qquad = 0 \quad \text{for} \quad x > a .$$

(Note that nondifferentiability at a single point does not change the almost differentiability.) From (2.12) it follows that

$$0 \geq t'_a(x) \geq -1 \quad \text{for all} \quad x .$$

Hence by the Theorem 2.1 it follows that for $p \geq 4$ the estimator

$$d^a_{*i}(X) = X_i + \frac{r(\frac{1}{2} \sum_{j=1}^{p} t_a^2(X_j))}{\sum_{j=1}^{p} t_a^2(X_j)} t_a(X_i) , \quad i=1, 2, \ldots, p \tag{2.14}$$

dominates X under quadratic loss. The performance of (2.14) is significantly better on some compact subset of the positive orthant and has finite risk for every $\theta \in \mathbb{R}^p$.

3. Shrinking towards Bayes estimators

When $\hat{\theta}_0$ is a Bayes estimator some extra simplification is possible. The classical James-Stein estimator shrinks towards the Bayes estimator under the point mass at 0 . Also the estimators suggested by Lindley (1962), Sclove, Morris, and Radhakrishnan (1972) are examples of shrinking towards Bayes estimators with lower dimensional generalized priors.

Consider a prior with density $\pi(\theta)$ with respect to some σ-finite measure ν on $\mathbb{R}^p$. The (formal) Bayes estimator with respect to π is given by

$$\hat{\theta}_\pi(X) = \frac{\int \theta \exp(-\tfrac{1}{2}\|\theta - X\|^2)\pi(\theta)d\nu(\theta)}{\int \exp(-\tfrac{1}{2}\|\theta - X\|^2)\pi(\theta)d\nu} =: \frac{I_X(\theta\pi(\theta))}{I_X(\pi(\theta))} \tag{3.1}$$

where for any integrable function $h: \mathbb{R}^p \to \mathbb{R}^p$

$$I_X(h) = \int h(\theta)\exp(-\tfrac{1}{2}\|\theta - X\|^2)d\nu(\theta) .$$

We assume $\sigma^2 = 1$ hereon. The Jacobian of $\hat{\theta}_\pi$ turns out to be

$$\hat{\theta}^{(i)}_{\pi,k}(X) = \frac{\partial}{\partial X_i} \frac{\int \theta_k \exp(\theta' X - \tfrac{1}{2}\|\theta\|^2)\pi(\theta)d\nu}{\int \exp(\theta' X - \tfrac{1}{2}\|\theta\|^2)\pi(\theta)d\nu}$$

$$= \mathrm{Cov}(\theta_k, \theta_i | X) . \tag{3.2}$$

Some more calculations along this line yields

$$\mathrm{Cov}(\theta_i,\ \theta_k|X) = \frac{\partial^2\psi}{\partial X_i \partial X_j} = \psi_{ij}(X) \quad ,\ \text{say}, \tag{3.5}$$

where

$$\psi(X) = \tfrac{1}{2}\|X\|^2 + \log E\pi(X + Z) \tag{3.6}$$

where Z is a $N(0,\ I)$ random variable under ν . We can interpret ψ as having two parts: first part coming from the likelihood and the latter from the curvature of the prior (vanishing if $\pi(\theta) = 1$).

Therefore we have

$$\psi_{ij}(X) = \delta_{ij} + \frac{\partial^2}{\partial X_i \partial X_j} \log E\pi(X + Z) \quad , \quad 1 \le i,\ j \le p\ . \tag{3.7}$$

Hence the Jacobian of $\hat{\theta}_\pi$ is given by

$$J_\pi(X) = I + H_\pi(X) = \Sigma(\theta|X) \tag{3.8}$$

where, $H_\pi(X) = [(h^\pi_{ij}(x))]$ with

$$h^\pi_{ij}(x) = \frac{\partial^2}{\partial X_i \partial X_j} \log E\pi(X + Z) \quad , \quad 1 \le i,\ j \le p\ .$$

Hence we obtain from (3.8) that

$$\mathrm{tr}(J_\pi) = p + \mathrm{tr}(H_\pi) = p + \nabla^2 \log E\pi(X + Z) \quad , \tag{3.9}$$

where ∇^2 denotes the Laplacian.

The above arguments can be considered as more formalized version of Stein (1981). We present it in the form of a theorem below.

Theorem 3.1. Let $\hat{\theta}_\pi$ be the Bayes estimator (possibly generalized) under the prior π and let

$$m_\pi(X) = \log E\pi(X + Z) \tag{3.10}$$

with Z as in (3.6). Then,

i) If m_π is concave, an equivalent condition for (a) and (b) to hold in the Theorem 2.1 are

$$\text{(a)}' \quad \nabla^2 m_\pi < 2\,\lambda_{\min}(H_\pi) - \epsilon \tag{3.11}$$

for all x and some $\epsilon > 0$ with $r'(\cdot) \geq 0$ and $0 \leq r(\cdot) < 2\epsilon$, and

$$\text{(b)}' \quad \nabla^2 m_\pi > 2\,\lambda_{\max}(H_\pi) + \epsilon \tag{3.12}$$

for all x and some $\epsilon > 0$ with $r'(\cdot) \geq 0$ and $-2\epsilon < r(\cdot) \leq 0$, respectively.

ii) If m_π is convex, an analogous conclusion holds after replacing $r'(\cdot) \geq 0$ by $r'(\cdot) \leq 0$ in (3.11) and (3.12), respectively.

iii) Sufficient conditions guaranteeing (3.11) and (3.12) are

$$\nabla^2 m_\pi < -2 - \epsilon \tag{3.11'}$$

$$\nabla^2 m_\pi > \epsilon \ , \tag{3.12)'}$$

respectively.

Proof. From (3.8) it follows that $(I - J_\pi)$ is non-negative or non-positive according as H_π is non-positive or non-negative. Since H_π is the hessian matrix of m_π it follows in turn that $(I - J_\pi)$ is non-negative if and only if m_π is concave. The remaining steps for proving (i) follow from (3.9). Similarly, we can prove (ii).

For (iii), notice that by (3.8) J_π is non-negative. So, it is now an obvious fact that

$$-1 \le \lambda_{\min}(H_\pi) \le \lambda_{\max}(H_\pi) \le 0 \ .$$

This completes the proof of the Theorem. []

Remarks. One implication of the above result is that it is not possible to shrink towards Bayes estimators which are too close to X . In the extreme case when $\pi(\theta) \equiv 1$, the generalized Bayes estimator turns out to be X . It is obvious that we can not shrink X towards itself.

Also it can be noticed that the above results can not be applied for priors of the form

$$\pi(\theta) \sim \|\theta\|^\alpha \text{ for large } \|\theta\| \ . \tag{3.14}$$

The difficulty arises due to the fact that H_π is neither positive nor negative. It turns out that

$$H_\pi(X) - \alpha\|X\|^{-2}(I - \frac{2}{\|X\|^2} X X') \tag{3.15}$$

for large $\|X\|$.

On the other hand for normal priors we can easily establish that (3.11) holds for $p \geq 3$. Consider a prior of the form

$$\pi(\theta) \sim \exp(- \frac{d}{2} \|\theta\|^2) \tag{3.16}$$

for some $d > 0$. Then it follows that

$$m_\pi(X) = - \tfrac{1}{2} \frac{d}{1+d} \|X\|^2 .$$

Thus, $m_\pi(X)$ is concave. Also,

$$\nabla^2 m_\pi - 2 \lambda_{\min}(H_\pi) = - \frac{(p-2)d}{1+d} < 0$$

if $p \geq 3$. Similarly, we can handle imaginary normal priors (i.e., $d < 0$) with (3.12). In the following proposition we develop a class of priors for which m_π is concave. Similar results can be obtained in other situations.

<u>Proposition 3.1</u>. If the prior π has the following form

$$\pi(\theta) \propto [\int_{\|\theta\|}^{\infty} s \exp(- \tfrac{1}{2} s^2 + m(s))ds] \exp(- \tfrac{1}{2} d\|\theta\|^2) \tag{3.17}$$

where $m(\cdot)$ is a nondecreasing function and $d > \frac{2}{p-2}$, d_* described in (1.5) is minimax.

The proof is left to the reader.

Remark. A more careful analysis will give us a larger class of priors. But, the class given by (3.17) is large enough for all practical purposes. Also, by using expanding priors (i.e., m_π is convex) we can obtain another class of priors admitting shrinkage.

3.1. A concept of optimal shrinkage

Let $\pi(\theta)$ be a prior density over $\mathbf{R}^p$ and let $\hat{\theta}_\pi$ $(= E(\theta|X))$ denote the corresponding Bayes estimator. For a class C of estimators we describe an optimality criterion below.

Definition 3.1. An estimator $d \in C$ is optimal minimax with respect to the prior π if

i) d is minimax.

ii) d has smallest Bayes risk among all other minimax estimators in C .

Next we define two classes of estimators for the multivariate normal mean problem. Let

$$B = \{d_{\rho,r}(X) = \rho(X) + \left(1 - \frac{r(\frac{1}{2}\|X - \rho\|^2)}{\|X - \rho\|^2}\right)[X - \rho(X)]: \text{ for some estimator } \rho(X) \text{ and almost differentiable } r(\cdot)\} .$$

B is the class of all spherically symmetric estimators with random center ρ. However, we shall deal with the following subclass of B because it is more manageable. Moreover, it contains the commonly used shrinkage estimators. Define

$$B_1 = \{d_{\rho,r} \in B: r'(\cdot) \geq 0 \text{ and } r(\cdot) \text{ is bounded by } \epsilon_0\},$$

where ϵ_0 is defined in the statement of theorem 3.2. We state the result below.

<u>Theorem 3.2</u>. Let π be a prior density such that m_π is concave and

$$\epsilon_0 = \inf_x \{2\lambda_{\min}[H_\pi(x)] - \nabla^2 m_\pi(x)\} > 0 .$$

Then the optimal minimax estimator in B_1 under π is given by $d_{\hat{\theta}_\pi, r_0}$ where $\hat{\theta}_\pi$ is the Bayes estimator under π and

$$r_0(t) = 2\min(t, \epsilon_0) .$$

<u>Proof</u>. Consider an arbitrary element $d_{\rho,r} \in B$. Let

$$\begin{aligned} I(r, \rho, \theta) &= R(X, \theta) - R(d_{\rho,r}, \theta) \qquad (3.18) \\ &= 2E_\theta\left\{ \frac{(X - \theta)'(X - \rho)}{\|X - \rho\|^2} r(\tfrac{1}{2}\|X - \rho\|^2) - \tfrac{1}{2}\frac{r^2(\tfrac{1}{2}\|X - \rho\|^2)}{\|X - \rho\|^2} \right\} . \end{aligned}$$

If $d_{\rho,r}$ is minimax, the right hand side of (3.18) is always non-negative. Therefore, in order to find out an optimal shrinkage we have to maximize $E_\pi I(r, \rho, \theta)$ subject to the constraint (3.18) is non-negative for all θ. In this context, note that the global optimal solution is given by:

$$\hat{\theta}_\pi = d_{\hat{\theta}_\pi, r*}$$

where $r*(t) = 2t$. Hence, the global solution cannot be minimax unless $\sup_X \|X - \hat{\theta}_\pi\| < \infty$.

Now the reduction in the Bayes risk for $d_{\rho,r}$ is given by

$$
\begin{aligned}
&E_\pi I(r, \rho, \theta) \\
&= 2 \int E_\theta \Big[\frac{(X - \theta)'(X - \rho)}{\|X - \rho\|^2} r(\tfrac{1}{2} \|X - \rho\|^2) \\
&\qquad - \tfrac{1}{2} \frac{r^2(\tfrac{1}{2}\|X - \rho\|^2)}{\|X - \rho\|^2} \Big] \pi(\theta) d\theta \qquad (3.19) \\
&= 2 \int E_\pi(Q|X)\, dm(X)
\end{aligned}
$$

(by Fubini's Theorem) where

$$Q = \frac{(X - \theta)'(X - \rho)}{\|X - \rho\|^2} r(\tfrac{1}{2}\|X - \rho\|^2) - \tfrac{1}{2} \frac{r^2(\tfrac{1}{2}\|X - \rho\|^2)}{\|X - \rho\|^2}$$

and m denotes the marginal. So, to maximize (3.19) it is sufficient to maximize the posterior expectation of Q, which is given by

$$E(Q|X) = \frac{(X - \hat{\theta}_\pi)'(X - \rho)}{\|X - \rho\|^2} r(½\|X - \rho\|^2) - ½\frac{r^2(\|X - \rho\|^2)}{\|X - \rho\|^2} . \tag{3.20}$$

Now denoting $(X - \hat{\theta}_\pi)$ by D and $(X - \rho)$ by U , respectively, we can rewrite (3.20) as

$$E(Q|X) = \frac{D'U}{\|U\|^2} r(½\|U\|^2) - ½\frac{r^2(½\|U\|^2)}{\|U\|^2} \tag{3.21}$$

$$= \frac{D'U}{\|U\|}\frac{r(½\|U\|^2)}{\|U\|} - ½\frac{r^2(½\|U\|^2)}{\|U\|^2} .$$

Next we maximize after fixing $\frac{r(½\|U\|^2)}{\|U\|}$ at a fixed level τ . It is easy to see that (3.21) is maximized if

$$\frac{U}{\|U\|} = \frac{D}{\|D\|} \tag{3.22}$$

Thus we put $\frac{U}{\|U\|} = \frac{D}{\|D\|}$ to obtain that for fixed τ ,the maximum value of $E(Q|X)$ is given by

$$E(Q|X) = \|D\|\tau - ½\tau^2 . \tag{3.23}$$

For any ρ , the relation (3.22) implies that

$$\rho = \rho_\alpha = X - \alpha D \tag{3.24}$$

for some random α . Hence, $d_{\rho,r}$ looks like

$$d_{\rho_\alpha,r} = X - \frac{r(\frac{1}{2}|\alpha|^2\|D\|^2)}{|\alpha|\,\|D\|^2} D . \tag{3.25}$$

Equation (3.25) follows from (3.22) and (3.24).

Since we restricted ourselves only to spherically symmetric estimators with random centers, it follows that (3.25) should be spherically symmetric about $\hat{\theta}_\pi$ in order to be a member of $\mathcal{B}$. In that case we can claim, in view of (3.22), that the optimal shrinkage lies within the subclass

$$B_\pi = \{d_{\hat{\theta}_\pi,r}:\ r \text{ is almost differentiable}\} . \tag{3.26}$$

Therefore (3.23) becomes

$$E(Q|X) = \|D\| \frac{r(\frac{1}{2}\|D\|^2)}{\|D\|} - \frac{1}{2} \frac{r^2(\frac{1}{2}\|D\|^2)}{\|D\|^2} . \tag{3.27}$$

Now by assumption of the Theorem m_π is concave. Thus by (3.11) we have the optimal choice of r to be

$$r(t) = 2 \min(t, \epsilon_0) \tag{3.28}$$

where

$$\epsilon_0 = 2 \inf_x [2 \lambda_{min}(H_\pi(x)) - \nabla^2 m_\pi(x)] .$$

This choice of r is optimal in the class of estimators

$$B_1 = \{d_{\rho,r}: \ r'(\cdot) \geq 0 \ \text{ and } \ 0 \leq r(\cdot) \leq \epsilon_0\} . \qquad []$$

3.3. Bayes minimax shrinkage estimators

In this section we shall investigate the Bayes property of the shrinkage estimators studied in earlier sections.

Suppose that ρ is the Bayes estimator with respect to a mixture prior π of the following form

$$\pi(\theta) = \int \phi\left(\frac{(\theta - \mu)\sqrt{1-a}}{\sqrt{a}} \right) h(1-a) g(\mu) da d\mu$$

where the mixing parameters $(\mu, a) \in \mathbb{R}^p \times [0, 1)$ have density $h(1-a)g(\mu)$ and ϕ denotes the $N_p(0, I)$ density. So, we have

$$\rho(X) = E(\theta|X) \tag{3.29}$$

$$= E[(1-a)\mu + aX|X]$$

$$= X - E[E((1-a)(X - \mu)|\mu, X)|X] .$$

Now it is well known that the conditional distribution of $(1-a)$ given (μ, X) is proportional to

$$\exp(-\tfrac{1}{2}(1-a)\|X - \mu\|^2) h_1(1-a) \tag{3.30}$$

for some h_1 . Hence we have

$$E[(1-a)|\mu, X] = \frac{r(\frac{1}{2}\|X - \mu\|^2)}{\|X - \mu\|^2} \tag{3.31}$$

for some r . Therefore, by (3.31) we have

$$\rho(X) = X - E\left[\frac{r(\frac{1}{2}\|X - \mu\|^2)}{\|X - \mu\|^2}(X - \mu)|X\right] \tag{3.32}$$

$$= E[d_{\mu,r}|X] \; .$$

This form of shrinkage estimators are very similar to ones suggested by George (1986a). One problem with these exact Bayes estimators is that they can be hard to compute. The proposed estimators in the earlier sections can be thought of as approximations to exact Bayes estimators where we substitute an estimate $E(\mu|X)$ to obtain $d_{E(\mu|X),r}$. The results in this article show that such approximations are quite reasonable from minimax point of view.

Acknowledgement: I would like to thank Professor Glen Meeden for his comments on an earlier version of this article and also the referees for their suggestions which improved the presentation of this article significantly.

APPENDIX

<u>Proof of Theorem 2.1</u>. Let $T = \|X - \hat{\theta}_0\|^2$. Then d_* can be written as

$$d_*(X) = X - \hat{\sigma}^2 \frac{r(\frac{1}{2}T)}{T} (X - \hat{\theta}_0) . \tag{2.3}$$

Next, by direct expansion of the loss function we obtain

$$R(d_*, \theta) - R(X, \theta) = -2\left[E \frac{r(\frac{1}{2}T)(X - \hat{\theta}_0)'(X - \theta)}{T\sigma^2}\right][E\hat{\sigma}^2]$$

$$+ \left[E_\theta \frac{r^2(\frac{1}{2}T)}{T\sigma^2}\right][E\hat{\sigma}^4] \tag{2.4}$$

using independence of X and $\hat{\sigma}^2$. Since r is bounded, Cauchy-Schwarz inequality asserts that

$$E \frac{r(\frac{1}{2}T)(X - \hat{\theta}_0)'(X - \theta)}{T} \le K\ E_\theta \frac{1}{T} < \infty \tag{2.5}$$

by assumption (provided $p \ge 3$).

Also, by assumption $r(\frac{1}{2}T)(X - \hat{\theta}_0)$ is almost differentiable, so using Stein's identity (viz. Stein (1973)) (2.4) can be rewritten as

$R(d_*, \theta) - R(X, \theta)$

$$= -2 \quad E\sigma^2 E \quad \frac{r'(\frac{1}{2}T) \sum_{i=1}^{p} (X_i - \hat{\theta}_{0i}) \sum_{k=1}^{p} (X_k - \hat{\theta}_{0k})(\delta_{ki} - \hat{\theta}_{0k}^{(i)})}{T}$$

$$+ r(\tfrac{1}{2}T) \sum_{i=1}^{p} \quad \frac{1 - \hat{\theta}_{0i}^{(i)}}{T} - \frac{2(X_i - \hat{\theta}_{0i}) \sum_{k=1}^{p} (X_k - \hat{\theta}_{0k})(\delta_{ki} - \hat{\theta}_{0k}^{(i)})}{T^2}$$

$$- \frac{1}{2\sigma^2} E\hat{\sigma}^4 \; E\{r^2(\tfrac{1}{2}T) \frac{1}{T}\} \tag{2.6}$$

(Here δ_{ki} denotes the Kronecker function ($\delta_{ki} = 1$ if $k = i$, 0 otherwise).) After some algebra, (2.6) simplifies to

$R(d_*, \theta) - R(X, \theta)$

$$= -2[E\hat{\sigma}^2 \; E\{r'(\tfrac{1}{2}T)U'(I - J_0)U + \frac{1}{T} r(\tfrac{1}{2}T)[p - tr(J_0)$$

$$- 2U'(I - J_0)U]\} - \tfrac{1}{2} E \; r^2(\tfrac{1}{2}T) \frac{1}{T} \frac{E\hat{\sigma}^4}{\sigma^2}] \tag{2.7}$$

where, $U = T^{-\frac{1}{2}}(X - \hat{\theta}_0)$ so that $\|U\| = 1$ and $tr(J_0)$ denotes the trace of J_0 .

Next consider the case where $(I - J_0)$ is nonnegative definite. Observe that $U'(I - J_0)U \le 1 - \lambda_{min}(J_0)$. Thus, if $p > 2 + tr(J_0) - 2\lambda_{min}(J_0) + \epsilon$ for all x and $0 \le r(\cdot) < 2\epsilon$ with $r' \ge 0$ (a.e. Lebesgue measure) the right hand side of (2.7) is always negative. (This follows after observing that $E\, r^2(\frac{1}{T}) \frac{1}{T} \le \sup_t r(t) \; Er(\frac{1}{2}T) \frac{1}{T}$.) Hence

$$R(d_*, \theta) < R(X, \theta) \text{ for all } \theta . \tag{2.8}$$

In a similar manner we can write

$$U'(I - J_0)U \geq 1 - \lambda_{max}(J_0) .$$

Therefore, if $tr(J_0) > p-2 + 2\lambda_{max}(J_0) + \epsilon$ for all x, (b) holds for some $r < 0$ with $r' \geq 0$ and $|r|$ bounded by $2\epsilon\rho$.

The other conclusions hold analogously. Hence the Theorem. []

References

Berger, J. O. (1982): Selecting a minimax estimator of a multivariate normal mean. Ann. Statist. 10, 81 - 92

Billingsley, P. (1979): Probability and Measure. John Wiliey, New York

George, E. (1986a): Minimax multiple shrinkage estimators. Ann. Statist. 14 188 - 205

George, E. (1986b): A formal Bayes multiple shrinkage estimator. Commun. Statist. - Theo. Meth. 15, 2099 - 2114

Katz, M. W. (1961): Admissible and minimax estimators of parameters in truncated spaces. Ann. Math. Statist. 32, 136 - 142

Lindley, D. V. (1971): The estimation of many parameters. Foundation of Statistical Inference, eds. V. P. Godambe and D. A. Sprott, Toronto

Sclove, S. L., Morris, C. and Radhakrishnan, R. (1972): Non-optimality of preliminary test estimators for the multi-normal mean. Ann. Math. Statist. 43, 1481 - 1490

Stein, C. (1956): Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proc. Third Berkeley Symp. Math. Statist. Probab., University of California Press, Berkeley

Stein, C. (1973): Estimation of the mean of a multivariate normal distribution. Proc. Prague Symp. Asymp. Statist.

Stein, C. (1981): Estimation of the mean of a multivariate normal distribution. Ann. Statist. 9, 1135 - 1151

Strawderman, W. E. (1971): Proper Bayes minimax estimators of the multivariate normal mean. Ann. Math. Statist. 42, 385 - 388

Strawderman, W. E. and Cohen, A. (1971): Admissibility of estimators of the mean vector of a multivariate normal distribution under quadratic loss. Ann. Math. Statist. 42, 270 - 296

D. Sengupta
Stat.-Math. Division
Indian Statistical Institute
203 B. T. Road
Calcutta 700 050
India