# Multivariate Estimation with Nonsymmetric Loss Functions

by

James Berger[1]
Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #517

November, 1977

[1]Research supported by the National Science Foundation under Grant #MCS76-06627A2, and by the John Simon Guggenheim Memorial Foundation.

Abstract

Multivariate estimation, in which the loss is a weighted sum of component losses, is considered. A method is given, whereby if "improved" estimators can be found in subproblems of the original problem, then an "improved" estimator can be found for the original problem. Several applications are given to problems of estimating, under weighted sum of squares error loss, the mean vectors of multivariate distributions from the exponential family. Of most importance is an application to the multivariate normal distribution. An estimator better than the vector of sample means (in 3 or more dimensions) and able to take significant advantage of (possibly vague) prior information is developed.

## Section 1. Introduction

Let X denote an arbitrary random variable, with a distribution depending on the vector $\theta = (\theta_1,\theta_2,\ldots,\theta_p)^t$ of interest and (possibly) upon a nuisance vector of parameters $\eta$. The $\theta_i$ could themselves be vectors or matrices. The loss in estimating $\theta$ by an estimator $\delta(x) = (\delta_1(x),\delta_2(x),\ldots,\delta_p(x))^t$ is assumed to be of the form

$$(1) \qquad L(\delta,\theta,\eta) = \sum_{i=1}^{p} q_i L_i(\delta_i,\theta_i,\eta),$$

where $q_i > 0$, $i=1,\ldots,p$. The loss can thus be decomposed into the component losses of estimating $\theta_i$ by $\delta_i$. Many problems have loss functions of this nature, or can be transformed so that they do.

As usual, an estimator will be evaluated in terms of its risk function (expected loss), given by

$$R(\delta,\theta,\eta) = E^{X}_{\theta,\eta}\,[L(\delta(X),\theta,\eta)].$$

(E stands for expectation, with subscripts denoting parameter values at which the expectation is to be taken, and superscripts denoting random variables over which the expectation is to be taken.)

The goal is to find an estimator offering significant improvement (in terms of risk) upon a given estimator, $\bar{\delta}(x)$, of $\theta$. $\bar{\delta}$ will usually be a "standard" estimator, such as the vector of sample means when estimating the mean of a p-variate normal distribution. Such standard estimators are typically inadmissible (estimators with smaller risk can be found) in high enough dimensions (usually 3 or more).

Frequently it is relatively easy to find estimators significantly better than $\bar{\delta}$ for losses which are certain specific linear combinations of the $L_i$.

For example, when $L_i(\delta_1,\theta_i,\eta) = (\delta_i - \theta_i)^2$, losses of the form $\sum_{i=1}^{p} (\delta_i - \theta_i)^2$ are often much easier to deal with than losses of the form $\sum_{i=1}^{p} q_i(\delta_i-\theta_i)^2$.

Another problem which can frequently arise is that certain $\theta_i$ may not be suitable for inclusion in the combined estimation problem. If an "extreme" $\theta_i$ (i.e. one likely to be considerably different from the others in some sense) is included in the combined estimation problem, great care must be taken in ensuring that its influence does not destroy the good effects of combined estimation. As an example, consider the situation in which $X = (X_1,\ldots,X_p)^t$ has a p-variate normal distribution with mean $\theta$ and covariance matrix $I_p$ (the $(p\times p)$ identity matrix). Assume the loss is

$$L(\delta,\theta) = \sum_{i=1}^{p} (\delta_i-\theta_i)^2.$$

The usual estimator of $\theta$ is, of course, $\bar{\delta}(x) = x$. James and Stein (1960) proposed the estimator

$$\delta(x) = \left(1 - \frac{(p-2)}{|x|^2}\right)x \tag{2}$$

($|x|^2$ is the Euclidean norm of x), showing it to have smaller risk than $\bar{\delta}$ for $p \geq 3$. Assume now that $p = 4$, with $\theta_4$ being an extreme coordinate in the sense that (apriori) nothing is known about $\theta_4$, while $\theta_1$, $\theta_2$ and $\theta_3$ are (apriori) believed to be near zero. It is then quite "likely" that $\theta_4$, and hence $|x|^2$, will be large. The estimate in (2) that would result would differ little from $\bar{\delta}(X) = X$, and no significant improvement in risk would be obtained. In such a situation the fourth coordinate should essentially be eliminated from the estimation problem. The estimator (2), used only with (and on) the first three coordinates, would perform quite well compared to $\bar{\delta}$, provided the $\theta_i$ are near zero as expected.

Similar problems are encountered with all "minimax" estimators of a multivariate normal mean so far discovered. The same type of difficulty also arises when the covariance matrix of X has "extreme" variances, or the loss is $\sum_{i=1}^{p} q_i(\delta_i-\theta_i)^2$ with "extreme" $q_i$. The need is for an estimator which automatically removes the influence of extreme coordinates.

In Section 2, a general method of dealing with the above difficulties is developed. The basic idea is to decompose the original problem into manageable subproblems. To save on notation when talking about subproblems, when z is a vector and B is a matrix, let

$$z^j = (z_1, z_2, \ldots, z_j)^t$$

and $B^j$ be the $(j\times j)$ upper left corner matrix of B. The subproblems which can be considered are those of estimating the $\theta^j$. In these subproblems it is possible to choose for loss functions linear combinations of the $L_i$ other than that given by (1). Indeed it is usually possible to choose a desired linear combination

$$L^{(j)}(\delta,\theta^j,\eta) = \sum_{i=1}^{j} t_i^j L_i(\delta_i,\theta_i,\eta) \tag{3}$$

$(t_i^j \geq 0)$, as the loss for the jth subproblem of estimating $\theta^j$. This choice can be made in a manner which makes the subproblem easy to analyze. If, for each subproblem, an estimator $\delta^{(j)}$ can be found which is better than $\bar{\delta}^j$ for estimating $\theta^j$ under $L^{(j)}$, then it will be seen how to construct an estimator better than $\bar{\delta}$ for the original problem. This idea of decomposition into subproblems was first used in a special case of the following theory by Bhattacharya (1966). See also Hudson (1974).

Section 2 contains the basic theory and a number of applications. Several of these applications deal with estimating the mean vector of an observation

from a distribution in the multivariate exponential family. Several papers have appeared (Clevenson and Zidek (1975), Peng (1976), Hudson (1977)) in which estimators better than the usual maximum liklihood estimator were found for losses of the form

$$\sum_{i=1}^{p} h(\theta_i)(\delta_i-\theta_i)^2.$$

Dealing with weighted losses

$$\sum_{i=1}^{p} q_i\, h(\theta_i)(\delta_i-\theta_i)^2$$

by the methods used in those papers appears very difficult. Such a generalization is easily obtained in Section 2, however.

The most important multivariate estimation problem is, of course, that of estimating a multivariate normal mean. Wide classes of estimators better than the usual estimator $\bar{\delta}(X) = X$ have been found. Unfortunately, they do not properly deal with the problem mentioned earlier of "extreme" coordinates.

The greatest difficulty is caused by the necessity to deal with prior information. The reason prior information must be considered is that the estimators better than $\bar{\delta}$ offer significant improvement in only a limited region of the parameter space, with risks being nearly equal to the risk of $\bar{\delta}$ elsewhere. For example, the estimator (2) has risk significantly smaller than $\bar{\delta}$ only for $|\theta|$ near zero. An improved estimator should thus be chosen whose region of significant improvement coincides with where $\theta$ is thought "likely" to be. If there is no such prior information about $\theta$, little can be gained by using estimators other than $\bar{\delta}$.

The above idea was developed in Berger (1977), wherein for quadratic loss a robust generalized Bayes estimator for $\theta$ was found. The estimator easily

allowed the incorporation of prior information, performing considerably better than $\bar{\delta}$ when the prior information was approximately correct, and yet was seldom worse than $\bar{\delta}$, even when the prior information was drastically wrong. Unfortunately, the estimator was uniformly better than $\bar{\delta}$(in terms of $R(\delta,\theta)$) only for certain quadratic losses.

In Section 3 an estimator is developed which incorporates prior information, is better than $\bar{\delta}$ for any given quadratic loss, and automatically handles the problem of extreme coordinates. The estimator is developed by the decomposition technique mentioned earlier. Subproblems are considered in which the losses (3) are chosen so that the appropriate generalized Bayes estimators of $\theta^j$ uniformly dominate $\bar{\delta}^j$. The theory then leads to the desired estimator in the original problem.

## Section 2. Decomposition to Subproblems

It is desired to find an estimator $\delta(x)$ such that

$$\Delta(\theta,\eta) = R(\bar{\delta},\theta,\eta) - R(\delta,\theta,\eta) \geq 0,$$

with strict inequality for some $\theta$ and $\eta$. To do this, consider the p subproblems of estimating $\theta^j$ under loss

$$(4) \qquad L^{(j)}(\delta,\theta^j,\eta) = \sum_{i=1}^{j} \alpha_i^j \, q_i L_i(\delta_i,\theta_i,\eta),$$

where the $\alpha_i^j$ satisfy the following condition.

Condition 1. $0 \leq \alpha_i^j \leq 1$, $\alpha_i^j = 0$ for $j < i$, and $\sum_{j=1}^{p} \alpha_i^j = 1$.

Assume that in the subproblems, estimators $\delta^{(j)}$ can be found which are as good as or better than the estimators $\bar{\delta}^j$. Thus

$$(5) \qquad \Delta_j(\theta^j,\eta) = R_j(\bar{\delta}^j,\theta^j,\eta) - R_j(\delta^{(j)},\theta^j,\eta) \geq 0,$$

where $R_j(\delta,\theta^j,\eta) = E_{\theta^j,\eta}[L^{(j)}(\delta(X),\theta^j,\eta)]$.

Two overall estimators of $\theta$ will be considered. First, the randomized estimator $\delta^*(x)$ defined componentwise by

$$\Pr(\delta_i^*(x) = \delta_i^{(j)}(x)) = \alpha_j^i, \tag{6}$$

and second the nonrandomized estimator $\delta'(x)$ whose ith component is given by

$$\delta_i'(x) = \sum_{j=i}^{p} \alpha_i^j \delta_i^{(j)}(x). \tag{7}$$

The following theorem gives the basic result.

Theorem 1.

(i) $\Delta^*(\theta,\eta) = R(\bar{\delta},\theta,\eta) - R(\delta^*,\theta,\eta) \geq 0$, with strict inequality for any $(\theta,\eta)$ for which $\Delta_j(\theta^j,\eta) > 0$ for some j. Thus $\delta^*$ is as good as or better than $\bar{\delta}$.

(ii) $\Delta'(\theta,\eta) = R(\bar{\delta},\theta,\eta) - R(\delta',\theta,\eta) \geq 0$ if $L(\cdot,\theta,\eta)$ is convex. The inequality is strict

(a) for any $(\theta,\eta)$ for which $\Delta_j(\theta^j,\eta) > 0$ for some j; or

(b) if $L(\cdot,\theta,\eta)$ is strictly convex and $\delta'(x)$ is not, with probability one, equal to $\delta^*(x)$.

Proof. Clearly

$$R(\delta^*,\theta,\eta) = E^X_{\theta,\eta} E^{\delta^*} \sum_{i=1}^{p} q_i L_i(\delta_i^*(X),\theta_i,\eta) \tag{8}$$

$$= E^X_{\theta,\eta} \sum_{i=1}^{p} q_i \sum_{j=i}^{p} \alpha_i^j L_i(\delta_i^{(j)}(X),\theta_i,\eta).$$

Since $\sum_{j=i}^{p} \alpha_i^j = 1$, it is also clear that

$$(9) \qquad R(\bar{\delta},\theta,\eta) = E^X_{\theta,\eta} \sum_{i=1}^{p} q_i \Big( \sum_{j=i}^{p} \alpha_i^j \Big) L_i(\bar{\delta}_i(X),\theta_i,\eta).$$

Combining (8) and (9) and performing a summation by parts gives

$$\begin{aligned}
\Delta^*(\theta,\eta) &= \sum_{i=1}^{p} \sum_{j=i}^{p} q_i \alpha_i^j \, E^X_{\theta,\eta} \, [L_i(\bar{\delta}_i(X),\theta_i,\eta) - L_i(\delta_i^{(j)}(X),\theta_i,\eta)] \\
&= \sum_{j=1}^{p} \sum_{i=1}^{j} q_i \alpha_i^j \, E^X_{\theta,\eta} [L_i(\bar{\delta}_i(X),\theta_i,\eta) - L_i(\delta_i^{(j)}(X),\theta_i,\eta)] \\
&= \sum_{j=1}^{p} E^X_{\theta,\eta} \, [L^{(j)}(\bar{\delta}^j(X),\theta^j,\eta) - L^{(j)}(\delta^{(j)}(X),\theta^j,\eta)] \\
&= \sum_{j=1}^{p} \Delta_j(\theta^j,\eta).
\end{aligned}$$

The conclusions in part (i) follow immediately from this and (5). Part (ii) is a direct application of Jensen's inequality, in that $\delta'(x) = E^{\delta^*}[\delta^*(x)]$.||

Application 1. Without loss of generality assume that $q_1 \geq q_2 \cdots \geq q_p$, and define

$$(10) \qquad \alpha_i^j = \begin{cases} 0 & \text{if } j < i \\ (q_j - q_{j+1})/q_i & \text{if } j \geq i \end{cases},$$

where $q_{p+1}$ is defined to be zero. Clearly Condition 1 is satisfied. The interest in this choice of $\alpha_i^j$ is that for $L^{(j)}$ as in (4),

$$(11) \qquad L^{(j)}(\delta,\theta^j,\eta) = (q_j - q_{j+1}) \sum_{i=1}^{j} L_i(\delta_i(x),\theta_i,\eta).$$

As an example of application, assume $X = (X_1,\ldots,X_p)^t$, where the $X_i$ are independently distributed with distributions $F_i$ from an exponential family. It is desired to estimate $\theta = E[X]$ under a loss

$$(12) \qquad L(\delta,\theta) = \sum_{i=1}^{p} q_i\, h(\theta_i)(\delta_i-\theta_i)^2.$$

For $q_i = 1$, the following improvements upon $\bar{\delta}(x) = x$ have been found:

1. James and Stein (1960):

$$\delta(x) = \left(1 - \frac{(p-2)}{|x|^2}\right)x,$$

when the $F_i$ are normal with unit variances, $h(\theta_i) = 1$, and $p \geq 3$.

2. Clevenson and Zidek (1975):

$$\delta(x) = \left(1 - \frac{p}{p + \sum_{i=1}^{p} x_i}\right)x_i,$$

when the $F_i$ are Poisson, $h(\theta_i) = 1/\theta_i$, and $p \geq 2$.

3. Peng (1976):

$$\delta(x) = x + (g_1(x),\ldots,g_p(x))^t$$

where $g_i(x) = -(p - N_0 - 2)^+ \ell(x_i)/S$, $\ell(x_i) = \sum_{k=1}^{x_i} \frac{1}{k}$, $S = \sum_{i=1}^{p} \ell(x_i)^2$, $N_0$ is the number of $x_i$ equal to zero, and "+" stands for the positive part, when the $F_i$ are Poisson, $h(\theta_i) = 1$, and $p \geq 3$.

4. Hudson (1977):

$$\delta(x) = x - \frac{(p-2)}{S}(B_1,\ldots,B_p)^t,$$

where $B_i = \log x_i$ and $S = \sum_{i=1}^{p} B_i^2$, when the $F_i$ are Gamma $(\theta_i,1)$ and $p \geq 3$.

To obtain improvements upon $\bar{\delta}$ in these cases for a loss of the form (12) with the $q_i$ unequal, simply choose $\alpha_i^j$ as in (10). The subproblem losses in

(11) are then constant multiples of the losses $\sum_{i=1}^{j} h(\theta_i)(\delta_i-\theta_i)^2$. In these subproblems, the estimators $\delta^{(j)}$ can be chosen to be the estimators in #1 through #4 above (depending on the problem), with p replaced by j and x replaced by $x^j$. (For $j = 1$, $\delta^{(1)}(x) = x_1$ must be chosen since $x_1$ is admissible for $\theta_1$, meaning no better estimator exists. Likewise for $j = 2$, $\delta^{(2)}(x) = (x_1,x_2)^t$ needs to be used, except for problem #2.) Theorem 1 (ii) then implies that the estimator $\delta'$ is uniformly better than $\bar{\delta}$ in the original problem. Bhattacharya (1966) proved this result for problem #1 and the $\alpha_i^j$ as in (10).

Application 2.

The choice of the $\alpha_i^j$ given in (10) is simple, but is not always suitable. Often, one would like to work with subproblems where

$$L^{(j)}(\delta,\theta^j,\eta) = \sum_{i=1}^{j} t_i^j L_i(\delta_i,\theta_i,\eta),$$

the $t_i^j$ being convenient nonidentical nonnegative numbers. Unfortunately, it rarely works to identify the $\alpha_i^j$ with the $t_i^j/q_i$, because of Condition 1. This difficulty can usually be resolved by noting that each $L^{(j)}$ can be multiplied by a nonnegative constant $\beta_j$, without affecting the subproblem. (An estimator as good as $\bar{\delta}^j$ under $L^{(j)}$ is also as good as $\bar{\delta}^j$ under $\beta_j L^{(j)}$.) Hence it is only necessary to find nonnegative constants $\beta_1,\ldots,\beta_p$ such that the $\alpha_i^j$, defined by

$$\alpha_i^j = \begin{cases} 0 & \text{if } j < i \\ \beta_j t_i^j/q_i & \text{if } j \geq i \end{cases}, \tag{13}$$

satisfy Condition 1. These $\alpha_i^j$ are clearly nonnegative, so Condition 1 will be satisfied if

$$(14) \qquad 1 = \sum_{j=i}^{p} \alpha_i^j = \frac{1}{q_i} \sum_{j=i}^{p} \beta_j t_i^j \ , \ i=1,\ldots,p.$$

This is a simple set of linear equations in $\beta_1,\ldots,\beta_p$. The solution is most easily found iteratively, starting with $\beta_p$. Indeed

$$\beta_p = q_p/t_p^p,$$

$$\beta_{p-i} = (q_{p-i} - \sum_{j=i+1}^{p} \beta_j t_{p-i}^j)/t_{p-i}^{p-i} \ .$$

If these solutions do not exist or are negative, we are out of luck and the $t_i^j$ must be altered. If they do exist, the $\alpha_i^j$ given in (13) satisfy Condition 1 and Theorem 1 can be used. The next section gives an important example of this type of application of Theorem 1.

## Section 3. Estimating a Multivariate Normal Mean.

Assume $X = (X_1,\ldots,X_p)^t$ has a p-variate normal distribution with mean $\theta$ and known positive definite covariance matrix $\Sigma$. (The situation of unknown $\Sigma$ will be discussed at the end of the section.) The loss function is assumed to be

$$L(\delta,\theta) = (\delta-\theta)^t Q(\delta-\theta),$$

where Q is a positive definite (p×p) matrix. The standard estimator of $\theta$ is, of course, $\bar{\delta}(x) = x$.

As mentioned in the introduction, it is necessary to make use of prior information in order to construct an estimator significantly better than $\bar{\delta}$. A reasonable way of summarizing prior information is in terms of a prior mean $\mu = (\mu_1,\ldots,\mu_p)^t$, and a prior covariance matrix A. The vector $\mu$ can be thought of as a best guess for $\theta$, and A as a measure of the believed accuracy

of this guess. Only rarely will additional prior information (such as knowledge of the tail of the prior) be available. See Berger (1977) for discussion of the development of $\mu$ and A in various standard situations.

In Berger (1977) the following generalized Bayes estimator for $\theta$ was proposed when $p \geq 3$:

$$\delta^B(x) = \left(I_p - \frac{r_p(||x-\mu||^2/\rho)\Sigma(\Sigma+A)^{-1}}{||x-\mu||^2}\right)(x-\mu)+\mu, \tag{15}$$

where $||x-\mu||^2 = (x-\mu)^t(\Sigma+A)^{-1}(x-\mu)$,

$$r_p(v) = \begin{cases} \dfrac{(v/2)^{p/2}}{(\frac{p}{2})!\left[\exp\{v/2\} - \sum_{i=0}^{(p-2)/2} \frac{(v/2)^i}{i!}\right]} & \text{if } p \text{ is even} \\ \dfrac{(v/2)^{p/2}}{\Gamma(\frac{p}{2})\left[\exp\{v/2\}\operatorname{erf}\{(v/2)^{1/2}\} - \sum_{i=0}^{(p-3)/2} \frac{(v/2)^{(i+1/2)}}{\Gamma(i+3/2)}\right]} & \text{if } p \text{ is odd} \end{cases}$$

where $\operatorname{erf}(z) = (2/\pi^{1/2}) \int_0^z \exp(-t^2)dt$, and $\rho = \min\{1, \max\{2\lambda, .6\}\}$

where $\lambda = ch_{max}\{\Sigma(\Sigma+A)^{-1}\}$, "$ch_{max}$" denoting maximum characteristic root. This estimator was shown to have a number of very attractive properties. Unfortunately, $R(\delta^B,\theta) \leq R(\bar{\delta},\theta)$ for all $\theta$ if and only if

$$(p+2) \leq \frac{2\ tr[Q\Sigma(\Sigma+A)^{-1}\Sigma]}{ch_{max}[Q\Sigma(\Sigma+A)^{-1}\Sigma]}, \tag{16}$$

where "tr" stands for trace. (This was established by Corollary 2.2.2 of Berger (1977).) The ideas discussed in Application 2 of Section 2 will be used to develop an estimator related to $\delta^B$ and better than $\bar{\delta}$ for all $\theta$.

As a first step, it is necessary to linearly transform the problem so that the loss is of the form (1). (Linearly transforming all elements of the

problem (X, $\theta$, $\mu$, Q, $\Sigma$, and A) gives an equivalent decision problem.) Indeed, it is convenient to consider the transformed problem of estimating $\xi = \Gamma Q^{1/2}\theta$ by an estimator based on the observation $Y = \Gamma Q^{1/2}X$, where $\Gamma$ is an orthogonal (p×p) matrix chosen so that

$$D = \Gamma\Sigma(\Sigma+A)^{-1}\Sigma\Gamma^{t}$$

is diagonal, with diagonal elements

$$d_1 \geq d_2 \cdots \geq d_p > 0.$$

It is easy to check that Q*, $\Sigma$* and A* in the transformed problem (corresponding to Q, $\Sigma$, and A in the original problem) satisfy $Q^* = I_p$ and $\Sigma^*(\Sigma^*+A^*)^{-1}\Sigma^* = D$. Rather than using the more cumbersome notation of the transformed problem, it will just be assumed that in the original problem

(17) $Q = I_p$ and $D = \Sigma(\Sigma+A)^{-1}\Sigma$ is diagonal with diagonal elements

$$d_1 \geq d_2 \cdots \geq d_p > 0.$$

The above transformation has imposed a particular ordering on the $\theta_i$ (corresponding to the order of the $d_i$). It is essential to achieve a proper ordering, in that the only subproblems which can be considered are those of estimating the ordered sequence $\theta^j = (\theta_1,\ldots,\theta_j)$. Essentially $\theta_1$ should be the "most important" coordinate, $\theta_2$ the next most important, etc. The meaning of "important coordinate" is here somewhat vague. It can best be interpreted as a reflection of the amount of improvement in risk that is likely to be obtained by including that coordinate in the combined estimation problem. If the estimator $\delta^B$ in (15) were used, it can be shown (see Berger (1977)) that the improvement in risk $\Delta_B(\theta) = R(\bar{\delta},\theta) - R(\delta^B,\theta)$ is given (when $Q = I_p$) by

$$(18)\qquad \Delta_B(\theta) = E\Big[\frac{r_p(||X-\mu||^2/\rho)}{||X-\mu||^2}\{2\,\mathrm{tr}[\Sigma(\Sigma+A)^{-1}\Sigma] - \Big[\frac{4+r_p(\ )}{||X-\mu||^2} - 4r_p'(\ )\Big](X-\mu)^t(\Sigma+A)^{-1}\Sigma^2(\Sigma+A)^{-1}(X-\mu)\}\Big]$$

where $r_p'(\cdot)$ is the derivative of $r_p(\cdot)$. The improvement in Bayes risk obtained by using $\delta^B$ is $E[\Delta_B(\theta)]$, where the expectation is taken with respect to the prior distribution of $\theta$. This is the same as the expectation of the integrand in (18) over both X and $\theta$, or equivalently over the marginal (unconditional) distribution of X. Marginally, X has mean $\mu$ and covariance matrix $(\Sigma+A)$. This makes plausible the rough approximations $||X-\mu||^2 \approx p$ and

$$(X-\mu)^t(\Sigma+A)^{-1}\Sigma^2(\Sigma+A)^{-1}(X-\mu) \approx \mathrm{tr}[\Sigma(\Sigma+A)^{-1}\Sigma].$$

Hence, roughly

$$E[\Delta_B(\theta)] = \frac{r_p(p/\rho)}{p}\{[2 - \frac{4+r_p(p/\rho)}{p} + 4r_p'(p/\rho)]\,\mathrm{tr}[\Sigma(\Sigma+A)^{-1}\Sigma]\}.$$

Since $r_p \le (p-2)$ and $r_p' \ge 0$ (see Berger (1977)), the expected improvement appears to be an increasing function of $\mathrm{tr}[\Sigma(\Sigma+A)^{-1}\Sigma] = \sum_{i=1}^{p} d_i$, with the larger $d_i$ giving rise to more improvement then the smaller $d_i$. This suggests measuring the importance of $\theta_i$ by the corresponding $d_i$.

In the subproblems of estimating the $\theta^j$, it is appealing to use the estimators (15) with p, $\Sigma$, A, x, and $\mu$ replaced by j, $\Sigma^j$, $A^j$, $x^j$, and $\mu^j$. For technical ease in the ensuing calculations, the following slightly different estimator will be used:

$$(19)\qquad \delta^{(j)}(x) = \Big(I_j - \frac{r_j(||x^j-\mu^j||_j^2/\rho_j)D^j(\Sigma^j)^{-1}}{||x^j-\mu^j||_j^2}\Big)(x^j-\mu^j) + \mu^j,$$

where $||x^j-\mu^j||_j^2 = (x^j-\mu^j)(\Sigma^j)^{-1}D^j(\Sigma^j)^{-1}(x^j-\mu^j)$, $r_j$ is given in (15) for $j \ge 3$

and equals zero otherwise, $\rho_j = \min\{1,\max\{2\lambda_j,.6\}\}$, and $\lambda_j = ch_{max}\{D^j(\Sigma^j)^{-1}\}$. This estimator is the subproblem version of (15) when $\Sigma$ and A are diagonal. Note that $r_j = 0$ when $j = 1$ or 2, so that $\delta^{(j)} = \bar{\delta}^j$ for $j = 1$ or 2. This is forced, since in one and two dimensions the usual estimator is admissible for any quadratic loss.

Having decided on the estimators to use in the subproblems, it is necessary to choose the losses,

$$L^{(j)}(\delta,\theta^j) = \sum_{i=1}^{j} t_i^j(\delta_i-\theta_i)^2,$$

so that $R_j(\delta^{(j)},\theta^j) \leq R_j(\bar{\delta}^j,\theta^j)$ for all $\theta^j$. (See Application 2 of Section 2.) Using Berger (1976), it can be shown that this will be true (for $j \geq 3$) if and only if

$$(20) \qquad (j+2) \leq \frac{2\sum_{i=1}^{j} t_i^j d_i}{\max_{1\leq i\leq j}\{t_i^j d_i\}} .$$

In choosing the $t_i^j$, it seems desirable to keep them as close as possible to one, the coefficients for the original loss. (Recall $Q = I_p$.) Noting that $t_i^j = d_1/d_i$ is always a solution to (20), a reasonable way of choosing the $t_i^j$ is, therefore,

$$(21) \qquad t_i^j = \tau_j + (1-\tau_j)d_1/d_i,$$

where if $j \geq 3$,

$$(22) \qquad \tau_j = \sup\{\tau:\ 0 \leq \tau \leq 1 \text{ and (20) is satisfied by (21)}\}.$$

In words, convex combinations of the "sure" choice $t_i^j = d_1/d_i$ and the "wishful" choice $t_i^j = 1$ are considered, the final choice being that convex combination

which satisfies (20) and is closest to $t_i^j = 1$. For $t_i^j$ as in (21), it is clear that

$$\underset{1\le i\le j}{ch_{max}}\{t_i^j d_i\} = t_1^j d_1 = d_1.$$

An easy calculation using (20), (21), and (22) then shows that (for $j \ge 3$)

$$\tau_j = \min\{\frac{(j-2)}{2(j - \sum_{i=1}^{j} \frac{d_i}{d_1})}, 1\}. \tag{23}$$

($\tau_j = 1$ if $d_i = d_1$ for $1 \le i \le j$.) For $j = 1$ or 2, it is reasonable to set $\tau_j = 1$, since $\delta^{(1)}$ and $\delta^{(2)}$ are the same as $\bar{\delta}^1$ and $\bar{\delta}^2$, and it is hence reasonable to use the original loss in these subproblems.

To apply Theorem 1, it remains only to find the $\alpha_i^j$ corresponding to the $t_i^j$. (See Application 2 of Section 2.) It is first necessary to find the solutions $\beta_1,\ldots,\beta_p$ of the equations in (14), which here are simply

$$\sum_{j=i}^{p} \beta_j t_i^j = 1, \quad 1 \le i \le p. \tag{24}$$

A tedious induction argument (starting with $i = p$ and working backwards) shows that for $t_i^j$ defined by (21) and (23), the solutions of (24) are

$$\beta_j = \frac{(\rho_j^* - \rho_{j+1}^*) \sum_{k=j+1}^{p} [1-\tau_k(1-\rho_{k+1}^*)]}{\sum_{k=j}^{p} [1-\tau_k(1-\rho_k^*)]}, \tag{25}$$

where

$$\rho_k^* = \begin{cases} d_k/d_1 & \text{if } k \le p \\ 0 & \text{if } k > p \end{cases}.$$

Since $0 < \rho_k^* \leq 1$, $\rho_k^* \geq \rho_{k+1}^*$, and $0 \leq \tau_k \leq 1$, it is clear from (25) that the $\beta_j$ are nonnegative, as was required of the solutions to (14).

Using (25), (23), and (13), the desired $\alpha_i^j$ are

$$(26)\quad \alpha_i^j = \begin{cases} 0 & \text{if } j < i \\ \dfrac{(\rho_j^*-\rho_{j+1}^*)[1-\tau_j(1-\rho_i^*)]\prod_{k=j+1}^{p}[1-\tau_k(1-\rho_{k+1}^*)]}{\rho_i^*\prod_{k=j}^{p}[1-\tau_k(1-\rho_k^*)]} & \text{if } j \geq i \end{cases}$$

where for $1 \leq j \leq p$,

$$\tau_j = \begin{cases} 1 & \text{if } j \leq 2, \text{ or } d_k = d_1 \text{ for } k \leq j \\ \min\left\{\dfrac{(j-2)}{2(j-\sum_{i=1}^{j}\rho_i^*)}, 1\right\} & \text{otherwise} \end{cases}$$

As in Application 2 of Section 2, it can be concluded from Theorem 1 that

$$(27)\qquad R(\delta',\theta) \leq R(\bar{\delta},\theta) \qquad \text{for all } \theta,$$

where

$$(28)\qquad \delta_i'(x) = \sum_{j=i}^{p} \alpha_i^j \delta_i^{(j)}(x),$$

$\alpha_i^j$ defined by (26) and $\delta^{(j)}$ by (19). Indeed if $p \geq 3$, it can be shown that $\Delta_p(\theta) > 0$ for all $\theta$. Hence by Theorem 1 the inequality in (27) is strict.

Of course, if the original problem had been transformed so that (17) was satisfied, the estimator in (28) must be transformed back to the original coordinate system.

Comments.

1. When $\tau_p = 1$, it is easy to check, using (26), that $\alpha_i^p = 1$ for $1 \leq i \leq p$,

and $\alpha_i^j = 0$ otherwise. Hence the estimator $\delta'$ is simply the estimator $\delta^B$ in (15). This is as would be desired, since by definition of $\tau_p = 1$ the estimator $\delta^{(p)}$ (which is the same as $\delta^B$) is uniformly better than $\bar{\delta}$ for the original loss.

2. The estimator is clearly somewhat messy. This poses no great calculational hardship, as most multivariate analyses are done on the computer anyway. More serious is the fact that the estimator, due to its complexity, is hard to examine for good or bad features. The fact that it uniformly dominates $\bar{\delta}$ in terms of risk helps greatly in eliminating fears that there might be some serious hidden fault. The main concern, therefore, is whether or not it makes good use of the prior information and eliminates the problems of "extreme" coordinates. The following special case indicates that it does so.

Assume $p = 5$, $Q = \Sigma = I$, $\mu = 0$, and $A$ is diagonal with diagonal elements $A_i = 1$ $(i = 1, 2, 3)$, and $A_i = c > 3$ $(i = 4, 5)$. Calculation using (19), (26), and (28) gives that

$$\delta_i'(x) = \begin{cases} \left(1 - \frac{1}{(c+1)(c+3)}\left[\frac{(c-1)(c-3)r_3(|x^3|^2)}{|x^3|^2} + \frac{8cr_5(||x||^2)}{||x||^2}\right]\right)x_3 & \text{if } i \leq 3 \\ \left(1 - \frac{r_5(||x||^2)}{||x||^2(1+c)}\right)x_i & \text{if } i = 4, 5, \end{cases}$$

where $||x||^2 = x^t(I+A)^{-1}x$. Note that the "bad" coordinates $\theta_4$ and $\theta_5$ get estimated as in (15) which is fine. For large c the "good" coordinates $\theta_1$, $\theta_2$, and $\theta_3$ get estimated essentially by

$$\delta_i(x) = \left(1 - \frac{r_3(|x^3|^2)}{|x^3|^2}\right)x_i.$$

In other words, when $\theta_4$ and $\theta_5$ get too extreme they essentially have no effect upon the estimation of $\theta^3$. This good behavior is in marked contrast to the behavior of the estimator (2), discussed in the introduction.

3. Usually the covariance matrix $\Sigma$ is not known. Frequently, however, it is assumed to be of the form $\Sigma = \sigma^2\Sigma_0$, where $\Sigma_0$ is known but $\sigma^2$ is unknown. In such a situation, assume a random variable $S^2$ can be observed, where $S^2/\sigma^2$ has a chi square distribution with m degrees of freedom (independent of X). It is then quite reasonable to use the estimator $\delta'$ for $\theta$, with $\Sigma$ replaced by $[S^2/(m+2)]\Sigma_0$ in the derivation. This estimator is probably still uniformly better than $\bar{\delta}$, but a general proof becomes very hard. The difficulty lies in the fact that the $d_i$ will, in general, be functions of $S^2$, so the decomposition to subproblems will vary as $S^2$ varies.

A result can be obtained in the special case when $A = c\Sigma_0$ for some constant c. It can then be checked that the $d_i$ are of the form $d_i(S^2) = h(S^2)d_i'$, the constants $d_i'$ not depending on $S^2$. Hence the decomposition to subproblems will be the same for all $S^2$. Likewise, the $\alpha_i^j$ given in (26) do not depend on $S^2$, as they are functions only of the $\rho_i^* = d_i(S^2)/d_1(S^2) = d_i'/d_1'$. Theorem 1 can thus be used ($\sigma^2$ is $\eta$) to show that $\delta'$ is better than $\bar{\delta}$ (for all $\theta$ and $\sigma^2$) provided the estimators $\delta^{(j)}$ are better than $\bar{\delta}^j$ in the subproblems. This last fact can be established for $\delta^{(j)}$ as in (19) with $\Sigma$ replaced by $(\frac{S^2}{m+2})\Sigma_0$, using Theorem 5.1 of Berger (1977).

4. It bears repeating that if one hopes to significantly improve upon $\bar{\delta}$, it is necessary to make use of prior information. One is then faced with an option, however. Either the uniformly better estimator $\delta'$ could be used, or the more truly Bayesian estimator $\delta^B$ in (15) could be used. When the two differ, $\delta^B$ will naturally perform better when the prior information is reasonably accurate, while $\delta'$ will be safer if the prior information is wrong. There is no clear way of deciding between the two estimators. Someone with Bayesian

inclinations will probably prefer $\delta^B$, while a more classical statistician might favor $\delta'$.

To aid in understanding the difference between the two estimators, the following special case is considered. Assume $p = 3$, $\xi = Q = I$, $\mu = 0$, and $A$ is diagonal with diagonal elements $A_1 = A_2 = 1$, $A_3 = c > 3$. Calculation using (19), (26), and (28) gives that

$$\delta_i^{\dagger}(x) = \begin{cases} [1 - \dfrac{2r_3(||x||^2)}{(1+c)||x||^2}]x_i & \text{if } i = 1, 2 \\ [1 - \dfrac{r_3(||x||^2)}{(1+c)||x||^2}]x_i & \text{if } i = 3, \end{cases}$$

where $||x||^2 = x^t(I+A)^{-1}x$. If $c$ is large, $\delta'(x) \cong x, (r_3(||x||^2)/||x||^2 < 1)$, so little improvement upon $\bar{\delta}$ is obtained. Indeed no estimator uniformly better than $\bar{\delta}$ could offer much improvement in this rather nasty situation. In contrast, the estimator

$$\delta^B(x) = (I - \frac{r_3(||x||^2)(I+A)^{-1}}{||x||^2})x$$

will be significantly better than $\bar{\delta}$ if the prior information is reasonably accurate. The penalty (compared to $R(\bar{\delta},\theta) = 3$) if the prior information is inaccurate will never be more than 5% in this example. In higher dimensional examples, however, $R(\delta^B,\theta)$ can be considerably more than 5% worse than $R(\bar{\delta},\theta)$ if the prior information is quite wrong. See Berger (1977) for further discussion of this.

An interesting possibility exists for compromising between $\delta^B$ and $\delta'$. The idea is to choose the $\alpha_i^j$ as in (26), with the $\tau_j$ replaced by

$$\tau_j^* = (1 - \gamma) + \gamma\tau_j,$$

where $\gamma$ is a number between zero and one. $\gamma = 1$ would give the estimator $\delta'$, while $\gamma = 0$ would give the estimator $\delta^B$ (see comment 1). Ideally, one could perhaps decide on the largest acceptable value of $\sup_0 R(\delta,\theta)$ and then use the smallest $\gamma$ for which the compromise estimator has risk within this bound. Unfortunately, the risks would have to be calculated numerically, making the approach perhaps unfeasible.

## References

[1] Berger, J. (1976). Minimax estimation of a multivariate normal mean under arbitrary quadratic loss. J. Multivariate Anal. 6, 256-264.

[2] Berger, J. (1977). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. Technical Report #480, Department of Statistics, Purdue University.

[3] Bhattacharya, P. K. (1966). Estimating the mean of a multivariate normal population with general quadratic loss function. Ann. Math. Statist. 37, 1819-1927.

[4] Clevenson, M. L. and Zidek, J. V. (1975). Simultaneous estimation of the mean of independent Poisson laws. J. Amer. Statist. Assoc. 70, 698-705.

[5] Efron, B. and Morris, C. (1973). Combining possibly related estimation problems. J. Roy. Statist. Soc., B, 35, 379-421.

[6] Hudson, H. M. (1974). Empirical Bayes estimation. Stanford Univ. Technical Report No. 58.

[7] Hudson, H. M. (1977). A natural identity for exponential families with applications in multiparameter estimation. Maquarie Univ. Research Paper No. 138.

[8] James, W. and Stein, C. (1960). Estimation with quadratic loss. Proc. Fourth Berkeley Symposium Math. Stat. Prob. 1, 361-379. University of California Press.

[9] Morris, C. (1977). Interval estimation for empirical Bayes generalizations of Stein's estimator. The Rand Paper Series. Rand Corporation, California.

[10] Peng, J. C. (1976). Simultaneous estimation of the parameters of independent Poisson distributions. To appear in Ann. Statist.

[11] Stein, C. (1974). Estimation of the parameters of a multivariate normal distribution. I. Estimation of the means. Stanford Univ. Technical Report No. 63.