The Annals of Statistics
1982, Vol. 10, No. 1, 81–92

SELECTING A MINIMAX ESTIMATOR OF A MULTIVARIATE NORMAL MEAN[1]

BY JAMES O. BERGER

Purdue University

The problem of estimating a p-variate normal mean under arbitrary quadratic loss when $p \geq 3$ is considered. Any estimator having uniformly smaller risk than the maximum likelihood estimator δ^0 will have significantly smaller risk only in a fairly small region of the parameter space. A relatively simple minimax estimator is developed which allows the user to select the region in which significant improvement over δ^0 is to be achieved. Since the desired region of improvement should probably be chosen to coincide with prior beliefs concerning the whereabouts of the normal mean, the estimator is also analyzed from a Bayesian viewpoint.

1. Introduction. Let $X = (X_1, \cdots, X_p)'$ have a p-variate normal distribution with mean vector $\theta = (\theta_1, \cdots, \theta_p)'$ and known covariance matrix $\mathbf{\Sigma}$. It is desired to estimate θ, using an estimator $\delta(x) = (\delta_1(x), \cdots, \delta_p(x))'$, under a quadratic loss

$$L(\theta, \delta) = (\theta - \delta)'Q(\theta - \delta), \tag{1.1}$$

Q being a known positive definite matrix. An estimator will be evaluated by its risk function

$$R(\theta, \delta) = E_\theta\{L(\theta, \delta(X))\},$$

i.e. the expected loss, or mean squared error if $Q = I_p$.

The usual estimator $\delta^0(x) = x$ is inadmissible if $p \geq 3$, a surprising fact discovered by Stein (1955) for $Q = \mathbf{\Sigma} = I_p$. Estimators having uniformly smaller risk than δ^0 for the general situation above have been found by Berger (1976a, 1976b, and 1979), Bhattacharya (1966), Bock (1975), Casella (1977), Hudson (1974), Judge and Bock (1977), Kariya (1977), Shinozaki (1974), Strawderman (1978), Thisted and Morris (1980), and many others in spherically symmetric or other special cases. Indeed, the class of estimators better than δ^0 is almost embarrassingly large, in that it is very difficult to choose one among them for actual use.

The basic difficulty is discussed in Berger (1980a), namely that any estimator better than δ^0 is, unfortunately, *significantly* better only for θ in a fairly small region, or subspace, of the parameter space. Hence selecting an alternative to δ^0 corresponds to selecting the region of the parameter space in which significant improvement is desired. It is convenient to think of such a region as, say, an ellipse of the form

$$\{\theta : (\theta - \mu)'A^{-1}(\theta - \mu) \leq p\}, \tag{1.2}$$

where μ specifies the center of the region and A determines the axes and orientation of the ellipse. It is natural here to think in Bayesian terms. Indeed, it seems inescapable that one would want to choose the region of significant improvement to be that region in which it is suspected that θ resides. Hence the point μ can be thought of as a prior guess for θ, and A as a prior covariance matrix for θ. We will interchangeably view the problem as a

Received August 1980; revised July 1981.

[1] Research supported by the National Science Foundation grant MCS78-02300, and by the Alfred P. Sloan Foundation.

AMS 1970 *subject classifications.* Primary 62C99; secondary 62F15,62F10, 62H99.

Key words and phrases. Minimax, normal mean, quadratic loss, risk function, prior information, Bayes risk.

Bayesian would, where μ and A represent the prior mean and covariance matrix, and as a non-Bayesian would, where μ and A determine the region of desired improvement in risk.

In this paper, a relatively simple minimax estimator, uniformly better than δ^0, will be developed; an estimator which allows the direct incorporation of μ and A and achieves its major improvement in risk over δ^0 in the region specified by μ and A. This estimator thus provides a reasonable solution to the selection problem.

Before proceeding, several explanatory comments are in order, first for Bayesians. The problem, as formulated here, is decidedly non-Bayesian in outlook. A Bayesian, after determining μ and A, would tend to construct around them a particular prior, and then calculate the Bayes estimator. The concern that an estimator has risk uniformly less than that of δ^0 is irrelevant, at first sight, to a Bayesian. Of considerable importance to a Bayesian (or at least to many Bayesians), however, is the robustness of the estimator with respect to misspecification of the prior chosen. Any prior that is written down is, after all, only an approximation to one's true prior knowledge. Consideration of possible misspecification of μ and A and of the chosen functional form for the prior leads one to worry about risk properties of the Bayes rule; see Berger (1980a, 1980b) for discussion of this and for earlier references. A Bayesian could consider the desire to have risk uniformly smaller than that of δ^0 to be a desire for the greatest degree of robustness possible, in that δ^0 is minimax and is the "noninformative prior" Bayes estimator, implying that even complete misspecification of the prior (here, complete misspecification of μ and A) will not have disastrous consequences.

A non-Bayesian may question the need to specify μ and A. The need to specify μ, the point towards which one "shrinks," is fairly well accepted, since it is easy to see that the region of significant improvement of any minimax estimator is concentrated about μ. The need to specify A is less clearcut, however, and so an example seems in order. The following example deals with a minimax estimator developed independently in Berger (1976a) and Hudson (1974). This estimator will also be needed in the later development. Define

$$\delta^{BH}(x) = \left\{ I_p - \frac{r(\|x-\mu\|^{2)}}{\|x-\mu\|^2} Q^{-1} \Sigma^{-1} \right\} (x-\mu) + \mu, \tag{1.3}$$

where $\|x-\mu\|^2 = (x-\mu)' \Sigma^{-}1Q-1\, \Sigma^{-}1(x-\mu)$ and r is any positive nondecreasing function less than or equal to $2(p-2)$. This estimator has risk less than that of δ^0. Note that δ^{BH} shrinks towards μ, but does not incorporate A; it can be shown to correspond to a particular, rather unrealistic, choice of A.

EXAMPLE 1. To see what can go wrong with δ^{BH}, assume $p = 5$, $Q = I_5$, Σ is diagonal with diagonal elements $\{10, 1, 1, 1, 0.1\}$, and the θ_i are thought likely to lie between 0 and 2. Thus we would like an estimator which does well in, say, the ellipse (1.2) with $\mu = (1, 1, 1, 1, 1)'$ and $A = I_5$; or, in Bayesian terms, we feel that μ and A are reasonable as the prior mean and covariance matrix. Sticking with the Bayesian language for clarity, we roughly expect the unconditional distribution of X, averaged over the prior, to have mean μ and covariance matrix $(\Sigma + A) = (\Sigma + I_5)$. Hence, very crudely, we expect to observe

$$\|x-\mu\|^2 \cong \operatorname{tr}\,(\Sigma + A)\Sigma^{-1}Q^{-1}\Sigma^{-1} = 116.11.$$

But then $\delta^{BH}(x)$ will tend to be very close to x itself, which indicates that the estimator will improve very little upon δ^0 in the desired region. Indeed this is the difficulty that usually befalls anyone trying to automatically apply a "Stein estimator" in a nonsymmetric situation: the result is usually very close to x itself, indicating that the estimator has not been appropriately centered and scaled by μ and A.

The above example indicates that A can be an important component of the analysis. Even more striking difficulties are encountered if certain of the θ_i are much less accurately specified (a priori) than others, i.e., if A has a wide spectrum of eigenvalues. Then $\|x-\mu\|^2$ will almost certainly be very large, leading to very little improvement. Note that A and Σ with wide spectrums tend to be the rule in multivariate analysis.

Before leaving the discussion of A, it is important to note that the amount of improvement over δ^0 that can be obtained in the specified region is inversely related to the size of the region. When A is very large, there will be only a small amount of improvement in the specified region. Therefore, unless one has reasonably accurate prior beliefs about at least some of the θ_i, it is a waste of time to use anything but δ^0.

An important qualification and limitation of this work should be mentioned, namely that it is designed for situations in which the data are of no or little use in determining A. As an example in which the data can be used to estimate A, suppose it is felt a priori that μ is of the form $\mu_0(1, \cdots, 1)'$, i.e. the θ_i have a common prior mean μ_0, and that A is of the form $\text{diag}(A_1 I_k, A_2 I_{p-k})$, where both k and $p - k$ are moderately large. Then A_1 can be estimated by $(1/k) \sum_{i=1}^{k} (x_i - \bar{x}_k)^2$, where $\bar{x}_k = (1/k) \sum_{i=1}^{k} x_i$, and A_2 can be similarly estimated. Reasonable estimators for this situation, estimators probably better than those suggested in this paper, are developed in Efron and Morris (1973). Another such example, discussed in Efron and Morris (1972), is the so-called matrix of means problem. This arises in, say, trying to simultaneously estimate several "exchangeable" regression equations. It is often possible, in such a situation, to completely estimate A from the data. Stein (1966) and Stein (1981) also deal with situations that can be interpreted as involving estimation of A (and μ).

The difficulties in the theory dealing with estimation of A, or certain features of A, from the data are that (i) the theory has only been developed for a few very special symmetric cases; (ii) the answers in the special cases seem drastically different, so no easy "recipe" can be given; and (iii) it is not clear how and when to combine data estimates of facets of A with subjective estimates, a necessity when p is not very large. The problems in developing a comprehensive general theory thus seem to be considerable. Actually, the majority of multivariate estimation problems probably lack the necessary symmetries or a large enough dimension to make estimation of features of A feasible. Hence the results in this paper, which are based mainly on subjective determination of μ and A, should be widely applicable. As will be seen later, the estimators proposed here can be adapted easily to deal with situations in which μ can be estimated from the data, such as when μ is thought to be of the form $\mu_0(1, \cdots, 1)'$. Thus only estimation of A itself is really precluded.

In Berger (1980a), a robust generalized Bayes estimator using μ and A was developed. The estimator is given by

$$(1.4) \qquad \delta^{RB}(x) = \left\{ I_p - \frac{r((x-\mu)'(\boldsymbol{\Sigma} + A)^{-1}(x-\mu))}{(x-\mu)'(\boldsymbol{\Sigma} + A)^{-1}(x-\mu)} \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + A)^{-1} \right\} (x-\mu) + \mu,$$

where r is a certain increasing function which can be reasonably approximated by $r(z) = \min(p - 2, z)$. This estimator was shown to provide significant improvement in risk over δ^0 in the region specified by μ and A, but unfortunately it is not necessarily minimax (i.e., uniformly better than δ^0). It is thus of interest to develop usable minimax estimators incorporating μ and A. (Many statisticians, the author included, do not feel that minimaxity is an absolutely essential criterion for an estimator, particularly since minimaxity is dependent on the subjective loss structure assumed. Nevertheless, there are those who would feel more comfortable knowing that their alternative to δ^0 was a uniform improvement, and in any case a good minimax estimator is needed as a base of comparison for nonminimax estimators.)

A very complicated minimax estimator incorporating μ and A was developed in Berger (1979). In our paper, a much simpler minimax estimator is developed, one which can be easily and usefully applied. This estimator is constructed in the next section, and an important special case is given. The third section of the paper compares this minimax estimator with δ^{RB} and more common Bayes estimators, indicating the effect of insisting on minimaxity. Generalizations and concluding remarks are presented in Section 4.

2. The Minimax Estimator. The estimator that will be proposed is a combination of δ^{BH} and a minimax estimator developed in Bhattacharya (1966). The estimator of Bhattacharya, as generalized by Berger (1979), can be described as follows. Suppose that it is desired to estimate θ under loss

$$L(\theta, \delta) = \textstyle\sum_{i=1}^{p} q_i^*(\theta_i - \delta_i)^2, \quad \text{where } q_1^* \geq q_2^* \geq \cdots \geq q_p^*, \tag{2.1}$$

and that it is known that, in estimating $\theta^j = (\theta_1, \theta_2, \cdots, \theta_j)'$ under sum of squares error loss, the estimator $\delta^{(j)}(x) = (\delta_1^{(j)}(x), \cdots, \delta_j^{(j)}(x))'$ is minimax. Define the estimator δ^{MB} componentwise by

$$\delta_i^{MB}(x) = q_i^{*^{-1}} \textstyle\sum_{j=i}^{p} (q_j^* - q_{j+1}^*)\, \delta_i^{(j)}(x), \tag{2.2}$$

where q_{p+1}^* is defined to be zero. Then

Lemma 1. *δ^{MB} is a minimax estimator of θ under the loss* (2.1).

The above lemma is Application 1 of Berger (1979), although the idea of the proof can be found in Bhattacharya (1966). After a suitable transformation of the orginal problem, it will be seen that the estimator we seek is δ^{MB} with the $\delta^{(j)}$ chosen as in (1.3).

To properly transform the original problem, let Λ be the $p \times p$ orthogonal matrix such that

$$Q^* = \Lambda(\Sigma + A)^{-1/2}\Sigma Q \Sigma(\Sigma + A)^{-1/2}\, \Lambda = \operatorname{diag}(q_1^*, \cdots, q_p^*), \tag{2.3}$$
$$\text{where } q_1^* \geq q_2^* \geq \cdots \geq q_p^*,$$

and define

$$B = \Lambda(\Sigma + A)^{1/2}\Sigma^{-1}, \quad X^* = BX, \quad \theta^* = B\theta, \tag{2.4}$$
$$\Sigma^* = B\Sigma B', \quad \mu^* = B\mu, \quad \text{and } A^* = BAB'.$$

The problem of estimating θ^* under loss $\sum_{i=1}^{p} q_i^*(\theta_i^* - \delta_i^*)^2$, based on X^*, Σ^*, μ^*, and A^*, can be easily seen to be equivalent to the original problem. The reason for considering this transformed problem is that, in Bayesian terms, the improvement that can be expected by shrinking a particular coordinate in a Bayesian fashion is roughly (in the diagonal case) $q_i\sigma_i^4/(\sigma_i^2 + A_i)$, where q_i, σ_i^2, and A_i are the corresponding diagonal elements of Q, Σ, and A; this will be indicated in the next section. Hence the above transformation rescales and rotates the problem so that the true "importance" of the ith coordinate is indeed accurately measured by q_i^*. This is crucial, in that δ^{MB} forces minimaxity, in a sense, by ignoring coordinates with larger indices to the extent necessary.

In the following, it will be convenient to denote the first j components of a vector y by $y^j = (y_1, \cdots, y_j)^t$, and to let C_j denote the $j \times j$ upper left corner matrix of a matrix C.

To employ Lemma 1 in the transformed problem, minimax estimators of θ^{*j} must be found for sum of squares error loss. We also desire these "subproblem" estimators to incorporate μ^* and A^*. An extremely fortunate fact can be observed in the transformed problem, namely that $(\Sigma^* + A^*) = \Sigma^{*2}$. If, furthermore, we are in the diagnoal case where Σ^* and A^* are diagonal matrices, then $\Sigma_j^* + A_j^* = (\Sigma_j^*)^2$. Hence the robust generalized Bayes estimator for θ^{*j}, as defined in (1.4), can be written

$$\delta^{(j)}(x^{*j}) = \left\{ I_j - \frac{r(\|x^{*j} - \mu^{*j}\|^2)}{\|x^{*j} - \mu^{*j}\|^2} \Sigma_j^{*-1} \right\} (x^{*j} - \mu^{*j}) + \mu^{*j},$$

where $\|x^{*j} - \mu^{*j}\|^2 = (x^{*j} - \mu^{*j})'\, \Sigma_j^{*-2}\, (x^{*j} - \mu^{*j})$. But this is an estimator of the form (1.3) (recall that the loss is assumed here to be the sum of squares error loss), and is hence minimax if $j > 2$ (only $\delta^0(x^{*j}) = x^{*j}$ is minimax if $j \leq 2$) and if r is any positive nondecreasing function less than or equal to $2(j - 2)$. Thus the robust Bayes estimator, which is known to be good in the region specified by μ^{*j} and A_j^*, is minimax for the transformed subproblems. Note that $\delta^{(j)}$ is minimax regardless of whether or not Σ^* and A^* are diagonal. When Σ^* and A^* are not diagonal, $\delta^{(j)}$ need not correspond exactly with the robust generalized Bayes estimator, but it still utilizes μ^* and A^* in a reasonable fashion and should perform well.

The only remaining concern is how best to choose the function r above. From the viewpoint of maximizing improvement in the region corresponding to μ and A, the choice $r^*(z) = \min\{2(j-2), z\}$ is reasonable. The conjugate prior Bayes rule, known to perform very well in the specified region, corresponds to $r(z) = z$. But to guarantee minimaxity, r must be bounded by $2(j-2)$. Hence the use of r^* allows $\delta^{(j)}$ to perform like the Bayes rule as much as possible, consistent with minimaxity. Thus the recommended subproblem estimator is

$$\delta^{(j)}(x^{*j}) = \left[I_j - \frac{\min\{2(j-2)^+, \|x^{*j} - \mu^{*j}\|^2\}}{\|x^{*j} - \mu^{*j}\|^2} \Sigma_j^{*-1} \right] (x^{*j} - \mu^{*j}) + \mu^{*j}, \tag{2.5}$$

where $(j-2)+$ equals $(j-2)$ if $j \geq 2$ and equals 0 if $j = 1$.

The search is now ended. The $\delta^{(j)}$ in (2.5) can be used in (2.2), the resulting estimator being guaranteed to be minimax by Lemma 1. Transforming back to the original coordinates establishes the following summarizing theorem.

THEOREM 1. *For the original problem of estimating θ under the loss* (1.1), *the estimator δ^{MB}, defined as follows, is minimax*:

$$\{B^{-1}\delta^{MB}(x)\}_i = q_i^{*-1} \textstyle\sum_{j=i}^p (q_j^* - q_{j+1}^*)\, \delta_i^{(j)}\, ((Bx)^j), \tag{2.6}$$

where the q_i^*, B, and $\delta^{(j)}$ are defined in (2.3), (2.4), *and* (2.5); *recall that $q_{p+1}^* = 0$.*

The above estimator is easier to evaluate in the diagonal case, so we state for convenience

COROLLARY 1. *Assume that Q, Σ, and A are diagonal with diagonal elements q_i, σ_i^2, and A_i, respectively, and that the X_i are indexed so that $q_1^* \geq q_2^* \geq \cdots \geq q_p^*$, where $q_i^* = q_i\sigma_i^4/(\sigma_i^2 + A_i)$. Then a minimax estimator of θ is given, coordinatewise, by*

$$\delta_i^{MB}(x) = x_i - \frac{\sigma_i^2}{(\sigma_i^2 + A_i)} (x_i - \mu_i) \left[\frac{1}{q_i^*} \textstyle\sum_{j=i}^p (q_j^* - q_{j+1}^*) \min\left\{1, \frac{2(j-2)^+}{\|x^j - \mu^j\|^2}\right\}\right], \tag{2.7}$$

where

$$\|x^j - \mu^j\|^2 = \textstyle\sum_{\ell=1}^j (x_\ell - \mu_\ell)^2/(\sigma_\ell^2 + A_\ell) \quad \text{and } q_{p+1}^* \equiv 0.$$

COMMENT. When $A = c\Sigma$ and Q is a multiple of $\Sigma-1$, δ^{MB} reduces to

$$\delta^{MB}(x) = \left[1 - \min\left\{ \frac{1}{1+c}, \frac{2(p-2)}{(x-\mu)'\Sigma^{-1}(x-\mu)} \right\} \right] (x-\mu) + \mu,$$

which is a version of the James-Stein estimator. Observe also, that if the expression in square brackets in (2.7) were 1, then the estimator would be the conjugate prior Bayes estimator.

3. Evaluation of the Minimax Estimator. In this section, the improvement offered by δ^{MB} over δ^0 will be evaluated. Also, δ^{MB} will be compared with δ^{RB} and the standard Bayes rule for the problem. Attention will be restricted to the diagonal situation, and, for ease of calculation, the following simpler versions of δ^{MB} and δ^{RB} will be considered:

$$\delta^{RB*}(x) = \left\{ I_p - \frac{(p-2)}{(x-\mu)'(\Sigma + A)^{-1}(x-\mu)} \Sigma(\Sigma + A)^{-1} \right\} (x-\mu) + \mu,$$

and (see Corollary 1 for definitions)

$$\delta_i^{MB*}(x) = x_i - \frac{\sigma_i^2}{(\sigma_i^2 + A_i)} (x_i - \mu_i) \left\{ \frac{1}{q_i^*} \textstyle\sum_{j=i}^p (q_j^* - q_{j+1}^*) \frac{(j-2)^+}{\|x^j - \mu^j\|^2} \right\}.$$

These estimators are undoubtedly worse than δ^{RB} and δ^{MB}; the use of $(p-2)$ instead of

$r(\|x-\mu\|^2)$ introduces an undesirable singularity as $x \to \mu$ in δ^{RB*}, and the use of $(j-2)^+$ instead of $\min\{2(j-2)^+, \|x^j - \mu^j\|^2\}$, besides introducing such singularities, causes δ^{MB*} to generally be farther from the conjugate prior Bayes rule than δ^{MB}. The improvements actually obtained using δ^{RB} and δ^{MB} will thus probably even be greater than the improvements calculated here.

We are interested in measuring, in some fashion, the improvement in risk over δ^0, in the region specified by μ and A, that can be obtained through use of δ^{RB*} and δ^{MB*}. A natural way to do this is to average the risk over some prior distribution for θ, a prior concentrated near the specified region. Since μ and A can be interpreted as a prior mean and covariance matrix, it is reasonable to choose as the averaging prior the $\mathcal{N}_p(\mu, A)$ distribution, henceforth denoted by π. We will thus compare the Bayes risks, i.e. $r(\pi, \delta) = E^\pi\{R(\theta, \delta)\}$, of δ^{RB*} and δ^{MB*} to that of δ^0. Note that

$$r(\pi, \delta^0) = E^\pi\{\mathrm{tr}(Q\Sigma)\} = \mathrm{tr}(Q\Sigma).$$

Note also that if π really were the true prior distribution for the problem, then one would want to use the Bayes estimator

$$\delta^B(x) = \{I_p - \Sigma(\Sigma + A)^{-1}\}(x-\mu) + \mu = x - \Sigma(\Sigma + A)^{-1}(x-\mu),$$

which has risk

$$R(\theta, \delta^B) = \mathrm{tr}[\Sigma\{I - \Sigma(\Sigma + A)^{-1}\}'Q\{I - \Sigma(\Sigma + A)^{-1}\}] + \theta'(\Sigma + A)^{-1}\Sigma Q \Sigma(\Sigma + A)^{-1}\,\theta, \tag{3.1}$$

and Bayes risk

$$r(\pi, \delta^B) = \mathrm{tr}(Q\Sigma) - \mathrm{tr}\{(\Sigma + A)^{-1}\Sigma Q \Sigma\} = r(\pi, \delta^0) - \textstyle\sum_{i=1}^p q_i^*. \tag{3.2}$$

Since $r(\pi, \delta)$ is minimized by δ^B, the maximum possible improvement over δ^0 for this measure of average improvement is clearly $\sum_{i=1}^p q_i^*$. It is also clear that q_i^* does indeed reflect the amount of improvement obtainable in estimating θ_i, a fact used in ordering the coordinates in the derivation of δ^{MB}.

The calculation of $r(\pi, \delta^{RB*})$ was carried out in Dey (1980), the conclusion being

$$r(\pi, \delta^{RB*}) = r(\pi, \delta^0) - \left(1 - \frac{2}{p}\right)\textstyle\sum_{i=1}^p q_i^*. \tag{3.3}$$

Finally, the Bayes risk of δ^{MB*} is given in the following lemma.

LEMMA 2. *If π is the $\mathcal{N}_p(\mu, A)$ prior distribution, then*

$$\begin{aligned} r(\pi, \delta^{MB*}) &= r(\pi, \delta^0) - \textstyle\sum_{i=3}^p q_i^* - 2\sum_{i=3}^p \frac{q_i^*}{i}\left\{1 - \frac{q_i^*}{(i-1)}\sum_{j=1}^{i-1}\frac{1}{q_j^*}\right\} \\ &\leq \; r(\pi, \delta^0) - \textstyle\sum_{i=3}^p q_i^*. \end{aligned} \tag{3.4}$$

PROOF. See Appendix. □

Using (3.2), (3.3), and (3.4), it is easy to compare the potential improvements over δ^0 of δ^B, δ^{RB*}, and δ^{MB*}, namely $\sum_{i=1}^p q_i^*$, $\left(1 - \frac{2}{p}\right)\sum_{i=1}^p q_i^*$, and

$$\textstyle\sum_{i=3}^p q_i^* + 2\sum_{i=3}^p \frac{q_i^*}{i}\left\{1 - \frac{q_i^*}{(i-1)}\sum_{j=1}^{i-1}\frac{1}{q_j^*}\right\},$$

respectively. Thus, taking a Bayesian approach but using the robust generalized Bayes estimator costs about $(2/p)\sum_{i=1}^p q_i^*$ in potential improvement, while insisting on a minimax estimator costs at most $(q_1^* + q_2^*)$ in potential improvement.

These "costs" in being robust or minimax are often not excessive, particularly for larger p; see, however, Remarks 2 and 3 in Section 4, which are concerned with certain other problems arising when p is large. Although for small p these "costs" can be large, recall that δ^{RB} and δ^{MB} are probably considerably better than δ^{RB*} and δ^{MB*}, especially for small p where the singularities have greater impact. Hence the true "cost" should be substantially less than indicated above for small p.

It is important to recall why δ^{RB} and δ^{MB} are usually preferable to δ^{B}. The risk of δ^{B}, given in (3.1), is a quadratic function of θ, and can hence be terrible outside of the region specified by μ and A. The estimator δ^{RB}, on the other hand, rarely has risk much worse than that of δ^{0} (Berger, 1980a), and hence is considerably safer than δ^{B}. The estimator δ^{MB}, being minimax, is, of course, the ultimate in safety. Again, to a Bayesian, the "safety" here can be interpreted as safety with respect to misspecification of the prior.

It should be stressed that, when q_1^* and q_2^* are large compared with the remaining q_i^*, then δ^{MB} can be considerably worse from the Bayesian viewpoint than δ^{B} or δ^{RB}. Insisting on minimaxity can thus eliminate most of the potential gains available from prior information when the coordinates are quite disparate in terms of the q_i^*. The robust Bayesian estimator δ^{RB} does not suffer this inadequacy, and hence is generally preferred by the author when significant prior information is deemed to be available.

To conclude this section, it is worthwhile to revisit Example 1, and to see how well δ^{MB} performs.

EXAMPLE 1 (*continued*). A calculation gives that $q_1 = 100/11$, $q_2 = q_3 = q_4 = 1/2$, and $q_5 = 1/110$. Thus, defining

$$r_j(x) = \min\left\{1, \frac{2(j-2)^+}{\|x^j - \mu^j\|^2}\right\},$$

it follows from (2.7) that

$$\delta_i^{MB}(x) = \begin{cases} x_i - \dfrac{\sigma_i^2}{(\sigma_i^2 + A_i)}(x_i - \mu_i)\left[q_i^{*-1}\left\{\left(\dfrac{1}{2} - \dfrac{1}{110}\right)r_4(x) + \dfrac{1}{110}r_5(x)\right\}\right] & \text{for } i \leq 4 \\ x_i - \dfrac{\sigma_i^2}{(\sigma_i^2 + A_i)}(x_i - \mu_i)r_5(x) \quad \text{for } i = 5. & \end{cases}$$

Observe that, if the prior information is correct, then $\|x^j - \mu^j\|-2$ should be roughly $(j-2)-1$ so that r_4 and r_5 should be close to 1. Hence δ_i^{MB} should be, for $i \geq 2$, roughly like the conjugate prior Bayes estimator

$$\delta_i^B(x) = x_i - \frac{\sigma_i^2}{(\sigma_i^2 + A_i)}(x_i - \mu_i).$$

Unfortunately, q_1^* is large compared to $q_2^* = 1/2$, so δ_1^{MB} is much closer to $\delta_1^0(x) = x_1$ than to δ_1^B. Again, this is unavoidable if one insists on minimaxity; coordinates with exceptionally large q_i^* cannot be shrunk as much as a Bayesian would desire. Note, however, that x_5 is not allowed to mess up the other x_i, as it did in the δ^{BH} estimator. Thus, for example, if the problem were to estimate $(\theta_2, \theta_3, \theta_4, \theta_5)$, δ^{MB} would perform exceptionally well, while δ^{BH} would still collapse back to δ^0. Incidentally, δ^{RB} will behave like

$$x_i - \frac{\sigma_i^2}{(\sigma_i^2 + A_i)}(x_i - \mu_i)\, r_5(x)$$

for all coordinates, and hence will have good Bayesian performance. Finally, note that if the prior information assumed does not accurately reflect the location of θ, then $\|x^j - \mu^j\|^2$ will tend to be large, and $r_i(x)$ small. Thus δ^{MB} will collapse back to δ^0 automatically when the data seems to contradict the prior. Such behavior, of course, is necessary for minimaxity.

4. Concluding Remarks.

REMARK 1. In developing μ and A, it is often important to incorporate believed a priori relationships among the θ_i. Certain relationships, such as a belief in the equality of the θ_i, or, more generally, a belief that certain linear restrictions hold, can easily be handled in the framework of this paper. Indeed, assume that it is felt that $B\theta = \theta_0$ with "accuracy matrix" C; in Bayesian terms, say, $C = E^{\pi}\{(B\theta - \theta_0)(B\theta - \theta_0)'\}$, π now simply denoting the prior information. Then if B is a nonsingular $p \times p$ matrix, it is natural to choose $\mu = B-1\theta_0$ and $A = B-1C(B-1)'$ and proceed. If, on the other hand, B is not of full rank, say B is an $m \times p$ matrix $(3 \le m < p)$ of rank m, and simply define $Y = BX - \theta_0$, $\eta = B\theta - \theta_0$, $\tilde{\Sigma} = B \Sigma B'$, $\tilde{Q} = (BQ-1B')-1$, $\mu = 0$, and $A = C$. Calculate $\delta^{MB}(y)$ for this transformed m-dimensional problem, and then in the original problem use

$$\delta(x) = x + Q^{-1}B'\tilde{Q}\{\delta^{MB}(Bx - \theta_0) - (Bx - \theta_0)\}. \tag{4.1}$$

This estimator is minimax by Theorem 2 of Berger and Bock (1977), and is tailored to offer significant improvement over δ^0 in, say, the region $\{\theta\colon (B\theta - \theta_0)'C-1(B\theta - \theta_0) \le m\}$.

As an example of the above ideas, suppose it is felt that the θ_i arise independently from a common prior distribution with unknown mean θ^*. This could be modeled, as above, by choosing B to be the $(p - 1) \times p$ matrix

$$\begin{pmatrix} 1 & -1 & 0 & \cdots & & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & & & & & \\ 0 & \cdots & & 0 & 1 & -1 \end{pmatrix}$$

choosing $\theta_0 = (0, \cdots, 0)'$, and letting C be the matrix

$$\begin{pmatrix} a & b & 0 & \cdots & & 0 \\ b & a & b & 0 & \cdots & 0 \\ \vdots & & & & & \\ 0 & \cdots & & 0 & b & a & b \end{pmatrix}$$

where

$$a = E^{\pi}(\theta_i - \theta_{i+1})^2 = E^{\pi}\{(\theta_i - \theta^*) - (\theta_{i+1} - \theta^*)\}^2 = 2\tau^2,$$

τ^2 being the believed prior variance, and

$$b = E^{\pi}\{(\theta_i - \theta_{i+1})(\theta_{i+1} - \theta_{i+2})\} = -E^{\pi}(\theta_{i+1} - \theta^*)^2 = -\tau^2.$$

Note that it is necessary to (subjectively) estimate τ^2. The estimator (4.1) can be seen, in this situation, to be an estimator which shrinks towards a point (the grand mean in symmetric situations) determined by the data.

It is also possible, in the above example, to use prior information about the value of θ^* if available. To do this, simply add the row $(0, \cdots, 0, 1)$ to the matrix B above, define θ_0 to be $(0, \cdots, 0, \theta_0^*)$, where θ_0^* is the prior guess for θ^*, and let C be as above with the addition of a pth row and a column of zeroes, except for the (p, p) element $C_{p,p} = E^{\pi}(\theta^* - \theta_0^*)^2$ representing the believed accuracy of the guess for the common mean. The result will be an estimator which shrinks towards a point determined by both the data and subjective knowledge.

REMARK 2. The estimators considered in this paper "group" all coordinates together, in the sense that all of the x_j are involved in the estimation of each θ_i. There are several situations in which this is probably not desirable. One such situation is when, say, q_1 (and maybe also q_2) are much larger than the other q_i. Then one might well want to sacrifice minimaxity, using separate though still robust Bayesian estimates of θ_1 (and maybe θ_2). The point is that the relatively unimportant coordinates should not be given the chance to foul up the important coordinates by, say, misspecification of the corresponding μ_i and A_i. Of course, if minimaxity is insisted upon, at least the third coordinate must be given the chance to foul things up.

Another situation involving a "grouping" problem is when it is felt that the μ_i and A_i are less accurately specified for certain coordinates than for others. It then may be desirable to group the coordinates according to the accuracy of the prior specification; see Dey (1980) for results on this problem.

REMARK 3. For large p, another modification of the estimators suggested here is probably desirable, namely "truncation" of large coordinates as suggested in Stein (1981). The problem is that, when p is large, some of the θ_i can be expected to be "outliers" with respect to the prior beliefs; or, equivalently, the prior distribution may well have fat tails. Such outliers could result in an excessively large value of $(x - u)'(\Sigma + A)-1(x - \mu)$, and hence cause the estimators considered here to be essentially equivalent to δ^0, as in Example 1. To correct this, Stein (1981) proposed, for the symmetric estimation problem, truncating large values of the x_i.

To apply this idea to our situation, consider estimators of the form (1.4), and define

$$v = (\Sigma + A)^{-1/2}(x - \mu).$$

The estimator (1.4) can then be written

$$\delta^{RB}(x) = x - \frac{r(|v|^2)}{|v|^2}\Sigma(\Sigma + A)^{-1/2}v.$$

Next, let $w_i = |v_i|$, and define $w_{(1)} \le w_{(2)} \le \cdots \le w_{(p)}$ to be the ordered w_i. Finally, for some integer $k < p$, define

$$z_\iota = (\operatorname{sgn} v_i)(\min w_i, w_{(k)}), \quad \text{and } z = (z_1, \cdots, z_p)'.$$

Clearly z_i is simply v_i, truncated so that its absolute value is no larger than the kth order statistic of the $|v_i|$. If the prior inputs μ and A are "correct," then the v_i should be small, so that truncating the v_i as above will indeed protect against outliers. Using these truncated variables, the estimator of interest can be written

$$\delta^{RBT}(x) = x - \frac{r_T(|z|^2)}{|z|^2}\Sigma(\Sigma + A)^{-1/2} z,$$

where r_T should now be calculated as if the dimension were k (the truncation effectively reduces the dimension to k).

The integer k should be chosen to be some appreciable fraction of p, say $k = \langle \alpha p \rangle$, where $0 < \alpha < 1$ and $\langle y \rangle$ denotes the smallest integer greater than or equal to y. From Stein (1981) and from more detailed studies in Dey (1980) for a variety of possible situations, the choice $\alpha = .7$ seems quite reasonable when p is very large. For smaller p, one must be more careful, and an overall choice of k such as

$$k = 3 + \langle .7(p - 3) \rangle$$

might be reasonable. Another possibility, which we do not pursue here, is to choose that k which maximizes $(k - 2)/|z|^2$, the idea being (assuming $r_T \equiv k - 2$) to let the data choose the truncation point.

To develop a truncated version of δ^{MB} and to prove that it is minimax, let $\mathcal{O}_j$ be an orthogonal matrix such that $\mathcal{O}_j \Sigma_j^* \mathcal{O}_j'$ is diagonal, and then define

$$v^{(j)} = \mathcal{O}_j \Sigma_j^{*-1} (x^{*j} - \mu^{*j}),\ w_i^{(j)} = |v_i^{(j)}|,\ k(j) = 3 + \langle .7(j-3) \rangle,$$
$$z_i^{(j)} = (\operatorname{sgn} v_i^{(j)}) \min\{w_i^{(j)}, w_{(k(j))}^{(j)}\}, \quad \text{and } z^{(j)} = (z_1^{(j)}, \cdots, z_j^{(j)})'.$$

Finally, choose

$$\delta^{(1)}(x^{*1}) = x^{*1},\ \delta^{(2)}(x^{*2}) = x^{*2},$$
$$\delta^{(j)}(x^{*j}) = x^{*j} - \frac{\min\{2(k(j)-2)^+, |z^{(j)}|^2\}}{|z^{(j)}|^2} \mathcal{O}_j' z^{(j)}, \quad j \geq 3, \tag{4.2}$$

and define δ^{MBT} as in (2.6), with (4.2) used for $\delta^{(j)}$.

Using the subproblem argument of Section 2, it is clear that to establish minimaxity of δ^{MBT} it is only necessary to verify that $\delta^{(j)}$ is minimax under sum of squares error loss. This can be done using integration by parts, after first diagonalizing Σ^* (the reason for the presence of the $\mathcal{O}_j$ in the definition of $\delta^{(j)}$), along the lines of the proof for the symmetric case in Stein (1981). For brevity, the proof is omitted.

REMARK 4. When Σ is unknown (or partially unknown) and a reasonable estimate of it is available, simply plugging the estimate into δ^{MB} or δ^{RB} seems to work quite well. For example, with even a moderate number of degrees of freedom for the estimate of Σ, δ^{MB} will probably still be minimax and offer substantial improvement over δ^0. Proving this in general is enormously difficult, however, since the ordering of the q_i^* will depend on the estimate of Σ. For certain special situations, minimax results can be established along the lines of Comment 3 of Berger (1979).

REMARK 5. A variety of previous investigations (see Brandwein and Strawderman (1980) and Berger (1976c) for certain results and other references) indicate that the improvement of the estimators discussed here will not be particularly dependent on the exact functional form of the loss and density. For most losses that are symmetric in the errors of estimation and for many densities such as Student's t, the estimators suggested here should significantly improve upon δ^0. Of course, for non-normal densities, the sample mean or least squares estimator may not be appropriate at all. No matter what the loss or density, however, the matrices Q and Σ are still very relevant. See Berger (1976c) for appropriate definitions of Q and Σ when dealing with general location losses and densities.

It is interesting to observe that δ^{RB} does not depend on Q. This is because δ^{RB} is a generalized Bayes estimator, and in Bayesian estimation with quadratic loss the generalized Bayes estimator is the posterior mean, regardless of Q. Whether this freedom from Q is considered a liability or an advantage of δ^{RB} probably depends on whether one is a Bayesian or a non-Bayesian. Note that any minimax estimator must depend heavily on Q.

REMARK 6. A continuum of estimators between δ^B and δ^{MB} can be defined, with δ^{RB} somewhere in the middle, allowing one to compromise between the minimax and Bayesian viewpoints to any extent desired. One could, for example, determine the maximum risk one is willing to suffer, and then find the "most Bayesian" estimator in this continuum subject to it having a maximum risk less than or equal to this predetermined maximum. The implementation of such a program would be a difficult numerical problem, however, and usually either δ^B, δ^{RB}, or δ^{MB} will be adequate.

Appendix. Proof of Lemma 2. In Dey (1980), it is shown that if $\sum_{j=\iota}^p \alpha_i^j = 1$ for $i = 1, \cdots, p$, and δ^* is given coordinatewise by

$$\delta_i^*(x) = \sum_{j=1}^p \alpha_i^j \left[\left\{ 1 - \frac{(j-2)^+}{\|x^j - \mu^j\|^2} \frac{\sigma_i^2}{(\sigma_i^2 + A_i)} \right\} (x_i - \mu_i) + \mu_i \right],$$

then

$$r(\pi, \delta^*) = r(\pi, \delta^0) - \sum_{i=1}^{p} q_i^* \sum_{j=i}^{p} \frac{(j-2)^+}{j} \alpha_i^j \left(2 \sum_{\ell=i}^{j-1} \alpha_i^\ell + \alpha_i^j\right).$$

Clearly δ^{MB*} is just δ^* with $\alpha_i^j = q_i^{*-1}(q_j^* - q_{j+1}^*)$. Observing that

$$\sum_{\ell=i}^{j-1} \alpha_i^\ell = \sum_{\ell=i}^{j-1} q_i^{*-1}(q_j^* - q_{j+1}^*) = q_i^{*-1}(q_i^* - q_j^*),$$

it follows that

(A1) $r(\pi, \delta^{MB*}) - r(\pi, \delta^0) =$

$$-2 \sum_{i=1}^{p} \sum_{j=i}^{p} \frac{(j-2)^+}{j} (q_j - q_{j+1}) + \sum_{i=1}^{p} \frac{1}{q_i} \sum_{j=i}^{p} \frac{(j-2)^+}{j} (q_j^2 - q_{j+1}^2),$$

where for simplicity of notation we now drop the superscript * from the q_i^*. Summing by parts shows that

(A2) $\sum_{i=1}^{p} \sum_{j=i}^{p} \frac{(j-2)^+}{j} (q_j - q_{j+1}) = \sum_{j=1}^{p} \sum_{i=1}^{j} \frac{(j-2)^+}{j} (q_j - q_{j+1})$

$= \sum_{j=1}^{p-2} \sum_{i=1}^{j} (q_{j+2} - q_{j+3}) = \sum_{i=1}^{p-2} \sum_{j=i}^{p-2} (q_{j+2} - q_{j+3}) = \sum_{i=3}^{p} q_i,$ recalling that $q_{p+1} \equiv 0$.

Next, defining $i^* = \max(3, i)$, it is clear that

(A3) $$\sum_{j=i}^{p} \frac{(j-2)^+}{j} (q_j^2 - q_{j+1}^2) = q_{i^*}^2 - 2 \sum_{j=i^*}^{p} \frac{1}{j} (q_j^2 - q_{j+1}^2).$$

Combining (A1), (A2), and (A3), it follows that

$$r(\pi, \delta^{MB*}) - r(\pi, \delta^0) = - \sum_{i=3}^{p} q_i - \left\{ - \left(\frac{1}{q_1} + \frac{1}{q_2}\right) q_3^2 + 2 \sum_{i=1}^{p} \frac{i}{q_i} \sum_{j=i^*}^{p} \frac{1}{j} (q_j^2 - q_{j+1}^2) \right\} .$$

It thus remains only to show that

(A4) $$- \left(\frac{1}{q_1} + \frac{1}{q_2}\right) q_3^2 + 2 \sum_{i=1}^{p} \frac{1}{q_i} \sum_{j=i^*}^{p} \frac{1}{j} (q_j^2 - q_{j+1}^2) = 2 \sum_{i=3}^{p} \frac{q_i}{i} \left\{ 1 - \frac{q_i}{(i-1)} \sum_{j=1}^{i-1} \frac{1}{q_j} \right\} .$$

To establish (A4), induction on p will be used. For $p = 3$, (A4) can be verified by direct calculation. Thus assume (A4) holds for $p = n$, and we must show that it then holds for $p = n + 1$. Now, recalling that $q_{p+1} \equiv 0$,

$$-\left(\frac{1}{q_1} + \frac{1}{q_2}\right) q_3^2 + 2 \sum_{i=1}^{n+1} \frac{1}{q_i} \sum_{j=i^*}^{n+1} \frac{1}{j} (q_j^2 - q_{j+1}^2)$$

$$= -\left(\frac{1}{q_1} + \frac{1}{q_2}\right) q_3^2 + 2 \sum_{i=1}^{n} \frac{1}{q_i} \left\{ \sum_{j=i^*}^{n-1} \frac{1}{j} (q_j^2 - q_{j+1}^2) + \frac{1}{n} q_n^2 \right\} + \frac{2q_{n+1}}{(n+1)} \left(1 - \frac{q_{n+1}}{n} \sum_{j=1}^{n} \frac{1}{q_j}\right)$$

$$= 2 \sum_{i=3}^{n} \frac{q_i}{i} \left\{ 1 - \frac{q_i}{(i-1)} \sum_{j=1}^{i-1} \frac{1}{q_j} \right\} + \frac{2q_{n+1}}{(n+1)} \left(1 - \frac{q_{n+1}}{n} \sum_{j=1}^{n} \frac{1}{q_j}\right) .$$

The last step follows from the induction hypothesis, i.e. (A4) applied with $p = n$, so that q_{n+1}^2 in (A4) is zero. But this last expression is precisely the right hand side of (A4) for $p = n + 1$, completing the induction argument.

The bound on $r(\pi, \delta^{MB*})$ in (3.4) follows from the observation that

$$\sum_{j=1}^{i-1} \frac{1}{q_{j^*}} \le \frac{i-1}{q_{i^*}},$$

since $q_1^* \ge q_2^* \ge \cdots \ge q_p^* > 0$. This completes the proof of Lemma 2. □

REFERENCES

Berger, J. (1976a). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Ann. Statist.* 4 223-226.

BERGER, J. (1976b). Minimax estimation of a multivariate normal mean under arbitrary quadratic loss. *J. Multivariate Anal.* **6** 256–264.

BERGER, J. (1976c). Tail minimaxity in location vector problems and its applications. *Ann. Statist.* **4** 33–50.

BERGER, J. (1979). Multivariate estimation with nonsymmetric loss functions. In *Optimizing Methods in Statistics*, J. S. Rustagi (Ed.). Academic, New York.

BERGER, J. (1980a). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist.* **8** 716–761.

BERGER, J. (1980b). *Statistical Decision Theory: Foundations, Concepts and Methods.* Springer-Verlag, New York.

BERGER, J. and BOCK, M. E. (1977). Improved minimax estimators of normal mean vectors for certain types of covariance matrices. *Statistical Decision Theory and Related Topics, II*, S. Gupta and D. Moore, eds. Academic, New York.

BERGER, J. and BOCK, M. E. (1976). Eliminating singularities of Stein-type estimators of location vectors. *J. Roy. Statist. Soc. B* **38** 166–170.

BHATTACHARYA, P. K. (1966). Estimating the mean of a multivariate normal population with general quadratic loss function. *Ann. Math. Statist.* **37** 1819–1824.

BOCK, M. E. (1975). Minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.* **3** 209–218.

BRANDWEIN, A. and STRAWDERMAN, W. (1980). Minimax estimators of location parameters for spherically symmetric distributions with concave loss. *Ann. Statist.* **8** 279–284.

CASELLA, G. (1977). Minimax ridge regression estimation. *Ann. Statist.* **8** 1036–1056.

DEY, D. (1980). On the choice of coordinates in simultaneous estimation of normal means. Mimeograph Series #80-23, Department of Statistics, Purdue University.

EFRON, B. and MORRIS, C. (1972). Empirical Bayes on vector observations: an extension of Stein's method. *Biometrika* **59** 335–347.

EFRON, B. and MORRIS, C. (1973). Combining possibly related estimation problems (with discussion). *J. Roy. Statist. Soc. B* **35** 379–421.

HUDSON, H. M. (1974). Empirical Bayes estimation. Ph.D. thesis, Department of Statistics, Stanford University.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. Proc. Fourth Berkeley Symp. Math. Statist. Prob. 1. 361–379. University of California Press.

JUDGE, G. and BOCK, M. E. (1977). *Implications of Pretest and Stein Rule Estimators in Econometrics.* North Holland, Amsterdam.

KARIYA, T. (1977). A class of minimax estimators in a regression model with arbitrary quadratic loss. *J. Japan Statist. Soc.* **7** 67–73.

SHINOZAKI, N. (1974). A note on estimating the mean vector of a multivariate normal distribution with general quadratic loss function. Keio Engineering Reports 27, 105–112.

STEIN, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proc. Third Berkeley Symp. Math. Statist. Prob. 1, 197–206. University of California Press.

STEIN, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. In Festschrift for J. Neyman (F. N. David, ed.), 351–366. Wiley, New York.

STEIN, C. (1981). Estimation of the parameters of a multivariate normal distribution-I. Estimation of the means. *Ann. Statist.* **9** 1135–1151.

STRAWDERMAN, W. E. (1978). Minimax adaptive generalized ridge regression estimators. *J. Amer. Statist. Assoc.* **73** 623–627.

THISTED, R. and MORRIS, C. (1980). Theoretical results for adaptive ordinary ridge regression estimators. Technical Report No. 94, University of Chicago.

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, IN 47907