Truncation of Shrinkage Estimators of Normal Means in the Nonsymmetric Case*

by

James O. Berger
Purdue University

and

Dipak K. Dey
Texas Tech University

Technical Report #83-11

April 1983

Purdue University
West Lafayette, IN

*Research supported by the National Science Foundation under Grants MCS 78-02300 and MCS 81-01670.

ABSTRACT

The problem of estimating a multivariate normal mean under quadratic loss is considered for the nonsymmetric situation. Shrinkage estimators have been developed which allow incorporation of prior information, performing very well when the prior information is correct while being quite satisfactory when the prior information is misspecified. In this paper versions of these estimators are developed which perform well even when the prior information is misspecified for some of the coordinates of the normal mean. This is done by utilizing a truncation technique developed by Stein for the symmetric situation.

1. INTRODUCTION

Let $X = (X_1,\ldots,X_k)^t$ have a k-variate normal distribution with mean vector $\theta = (\theta_1,\ldots,\theta_k)^t$ and known positive definite covariance matrix Σ. It is desired to estimate θ, using an estimator $\delta(X) = (\delta_1(X),\ldots,\delta_k(X))^t$, under a quadratic loss

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

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

$$R(\theta,\delta) = E_\theta[L(\theta,\delta(X))],$$

that is the expected loss. For the sum of squares error loss function, that is $Q = I$, and for $\Sigma = \sigma^2 I$ (σ^2 known), James and Stein (1961) showed that the usual estimator $\delta^0(X) = X$ is inadmissible when $k \geq 3$ and that the estimator

$$\delta(X) = \left(1 - \frac{(k-2)\sigma^2}{X^tX}\right)X \tag{1.1}$$

has uniformly smaller risk than δ^0. Estimators having uniformly smaller risk than δ^0 for the general situation above have been found by many authors. See Berger (1982a) for references.

A key feature of any Stein type estimator is that it has risk significantly better than that of δ^0 only in a relatively small region (or subspace) of the parameter space. For Stein estimation to result in significant improvement, therefore, one must carefully select an estimator designed to do well in the region in which θ is thought likely to lie. This is essentially done by finding a Stein type estimator which shrinks towards the desired region. Often, regions in which θ is thought to lie can be represented as ellipses such as

$$C = \{\theta: (\theta-\mu)^t A^{-1}(\theta-\mu) \leq k\}.$$

A more convenient way to think of this is to use Bayesian ideas, specifying a prior mean μ and covariance matrix A for θ. Such simple features of prior information are relatively easy to specify, as opposed to more involved features such as the functional form of the prior. See Berger (1980, 1982a) for more discussion.

In Berger (1980) an estimator was developed which utilized μ and A, was significantly better than δ^0 if this prior information accurately reflected the location of θ, and yet was little, if any, worse than δ^0 for θ in conflict with the prior information. A simpler and in some ways better (see Berger (1982b)) version of this estimator is

$$\delta^{RB}(X) = X - \frac{r_k((X-\mu)^t(\Sigma+A)^{-1}(X-\mu))}{(X-\mu)^t(\Sigma+A)^{-1}(X-\mu)}\,\Sigma(\Sigma+A)^{-1}(X-\mu), \tag{1.2}$$

where $r_k(v) = \min\{k-2, v\}$. For a general discussion of the properties of this estimator see Berger (1980 and 1982b). Its desirable behavior is at least indicated by noting that when

$$(X-\mu)^t(\Sigma+A)^{-1}(X-\mu) \leq k-2, \tag{1.3}$$

δ^{RB} is the usual conjugate prior Bayes estimator, while otherwise δ^{RB} is a Stein type estimator with bounded risk $R(\theta,\delta)$ (typically smaller than $R(\theta,\delta^0)$). Since (1.3) can be roughly interpreted as implying that the data supports the prior assumptions (the marginal mean and covariance matrix of X are μ and $\Sigma+A$, respectively) the claimed desirable frequentist and Bayesian properties of δ^{RB} seem plausible.

The estimator δ^{RB} is not always minimax (i.e. uniformly better than δ^0 in terms of risk). While it can be argued that this is not a serious concern, uniform domination of δ^0 may be demanded by some (although this can entail a substantial decrease in average improvement). An estimator allowing incorporation of μ and A and yet guaranteeing such dominance was developed in Berger (1982a).

For notational simplicity, only the case in which Q, $\sharp$, and A are diagonal with diagonal elements q_i, σ_i^2, and A_i, respectively, will be considered. (The general case can be dealt with along the lines of Berger (1982a).) If, without loss of generality, the X_i are indexed so that $q_1^* \geq q_2^* \geq \ldots \geq q_k^*$, where $q_i^* = q_i\sigma_i^2/(\sigma_i^2+A_i)$, the minimax estimator is given coordinatewise by

$$\delta_i^{MB}(X) = X_i - \frac{\sigma_i^2}{(\sigma_i^2+A_i)}(X_i-\mu_i)\left[\frac{1}{q_i^*}\sum_{i=1}^{k}(q_j^*-q_{j+1}^*)\min\left\{1, \frac{2(j-2)^+}{||X^j-\mu^j||_j^2}\right\}\right], \tag{1.4}$$

where

$$||X^j-\mu^j||_j^2 = \sum_{n=1}^{j}(X_n-\mu_n)^2/(\sigma_n^2+A_n) \text{ and } q_{k+1}^* \equiv 0.$$

This estimator will also often act like the conjugate prior Bayes estimator when the data "supports" the prior, yet it is <u>always</u> better than δ^0 in terms of <u>frequentist</u> risk.

An undesirable feature of δ^{RB} and δ^{MB} is that, if the prior information is misspecified for a few of the coordinates θ_i, then $(X-\mu)^t(\sharp+A)^{-1}(X-\mu)$ or the $||X^j-\mu^j||_j^2$ will tend to be large and the estimators will collapse back to $\delta^0(X) = X$, even when the prior information specified for the other coordinates is fine and could result in substantial improvement in estimation of those coordinates. One could, of course, informally drop the offending coordinates from the simultaneous estimation problem, dealing with them separately. While satisfactory to a Bayesian, this would result in a procedure with uncertain frequentist risk, an unappealing consequence for non-Bayesians. Also, it would be desirable to have an automated procedure to deal with the problem.

The technique we adopt is one employed by Stein (1981) in the symmetric case. In place of (1.1) he proposed use of

$$\delta_i^{(\ell)}(X) = \left(1 - \frac{(\ell-2)\sigma^2 \min\{1,\ Z_{(\ell)}/|X_i|\}}{\sum_{j=1}^{k} X_j^2 \wedge Z_{(\ell)}^2}\right) X_i, \tag{1.5}$$

where ℓ is a large fraction of k, $a \wedge b$ denotes the minimum of a and b,

$$Z_i = |X_i| \quad \text{and} \quad Z_{(1)} < Z_{(2)} < \ldots < Z_{(k)} \tag{1.6}$$

are the order statistics of $Z_1,\ldots,Z_k$. Stein proposed this to (i) eliminate the detrimental influence on (1.1) of a few very large θ_i, and (ii) reduce the maximum component risk of the estimator. This estimator was further developed and analyzed in Dey and Berger (1983). In Sections 2 and 3 we apply this truncation technique to δ^{RB} and δ^{MB} to reduce the detrimental effect of misspecified prior information for a few coordinates. A beneficial side effect, as with (1.6), will be a reduction in maximum component risk of the estimators (see also Efron and Morris (1972)).

It should be noted that we are <u>not</u> considering an empirical Bayes situation here, in which the θ_i are felt to be related in some fashion and the other X_j can be of use in estimating features of this relationship. The prior inputs μ and A are considered to be solely subjective inputs and are not estimable in any way from the data. To a Bayesian, if there is no suspected relationship among the θ_i, there seems (at first sight) to be no reason to combine the X_i in a simultaneous shrinkage estimator. A justification can be given, however, in terms of Bayesian robustness, i.e. robustness with respect to possible misspecification of the prior information. For discussion of this issue see Berger (1980, 1982b, 1983). At the very least, a Bayesian can be somewhat satisfied with δ^{RB} since it will frequently be equal to the conjugate prior Bayes estimator with inputs μ and A.

As a final comment, the results of Berger and Dey (1983) should be mentioned. In that paper it was shown, somewhat surprisingly, that coordinates should <u>not</u>

be dropped from the simultaneous estimation problem if there is no fear of prior misspecification. (Intuitively, one might have thought that large $\sigma_i^2+A_i$ would make elimination of θ_i from the simultaneous estimation problem desirable.)

2. TRUNCATION FOR δ^{RB}

The following series of transformations exhibits δ^{RB} in a form where truncation can be easily implemented. Let Γ be the (kxk) orthogonal matrix such that

$$Q^* = \Gamma(\Sigma+A)^{-1/2}\Sigma Q\Sigma(\Sigma+A)^{-1/2}\Gamma^t \tag{2.1}$$

is diagonal with diagonal elements $q_1^* \geq q_2^* \geq \ldots \geq q_k^*$, and define

$$B = \Gamma(\Sigma+A)^{1/2}\Sigma^{-1}, \quad X^* = BX, \quad \theta^* = B\theta, \tag{2.2}$$
$$\Sigma^* = B\Sigma B^t, \quad \mu^* = B\mu, \text{ and } A^* = BAB^t.$$

The problem of estimating θ^* under loss $\sum_{i=1}^{k} q_i^*(\theta_i^*-\delta_i^*)^2$, based on X^*, Σ^*, μ^*, and A^*, can be easily seen to be equivalent to the original problem of estimating θ under loss $(\theta-\delta)^tQ(\theta-\delta)$.

It can be observed, as in Berger (1982a), that in the transformed problem $(\Sigma^*+A^*) = \Sigma^{*2}$. Thus the robust generalized Bayes estimator for θ^*, as defined in (1.2), becomes

$$\delta^{RB*}(X^*) = X^* - \frac{r_k((X^*-\mu^*)^t\Sigma^{*-2}(X^*-\mu^*))}{(X^*-\mu^*)^t\Sigma^{*-2}(X^*-\mu^*)}\,\Sigma^{*-1}(X^*-\mu^*). \tag{2.3}$$

Defining $Y = \Sigma^{*-1}(X^*-\mu^*)$ this can be written

$$\delta^{RB*}(X^*) = X^* - \frac{r_k(|Y|^2)}{|Y|^2}Y. \tag{2.4}$$

This last form is very convenient, since an easy calculation shows that Y has marginal mean 0 and marginal covariance matrix I_k. Thus, unusually large

values of $|Y_i|$ indicate that the prior information about the corresponding (transformed) coordinates seems to be in error, and will have the effect of collapsing δ^{RB*} back to X*, even if the other Y_j are reasonably small. Furthermore, Stein's truncation procedure can be easily applied to (2.4). Thus we propose, as the truncated estimator for θ^* (which can easily be transformed back into an estimator for θ),

$$\delta^{\ell}(X^*) = X^* - \frac{r_{\ell}(|Z|^2)Z}{|Z|^2} \quad , \tag{2.5}$$

where

$$Z_i = (\text{sgn } Y_i)\{|Y_i| \wedge |Y|_{(\ell)}\}, \; i=1,2,\ldots,k, \tag{2.6}$$

and $|Y|_{(1)} < |Y|_{(2)} < \ldots < |Y|_{(k)}$ are the ordered $|Y_i|$. This estimator clearly limits the influence of large $|Y_i|$.

The remaining issue that must be addressed is that of choosing ℓ. The approach we will take is to try to choose ℓ so as to optimize the overall performance of δ^{ℓ} with respect to plausible prior distributions π. Thus we will investigate

$$r(\pi,\delta^{\ell}) = E^{\pi}R(\theta^*,\delta^{\ell}).$$

The reason for looking at this overall measure is that δ^{ℓ} will usually have satisfactory frequentist risk for reasonable ℓ (in terms of $\sup_{\theta^*} R(\theta^*,\delta^{\ell})$), so of concern in choosing ℓ is its overall average performance.

In choosing plausible π, recall that θ was determined to have prior mean μ and covariance matrix A, but that further information was presumed not to be available. Note also that, for the marginal distribution $m^*(y)$ (obtained by transformation from the marginal distribution $m(x) = E^{\pi}f(x|\theta)$, $f(x|\theta)$ being the normal density of X), the Y_i have marginal means 0 and variances 1. To make

further progress we will make the somewhat restrictive assumption that

$$m^*(y) = \prod_{i=1}^{k} p(y_i), \tag{2.7}$$

i.e., that the Y_i are independent with a common marginal density p. This, of course, will be the case if the original prior π is taken to be normal with mean μ and covariance matrix A, and it can be shown to hold in a number of other cases, such as when π is an appropriate mixture of normals. Indeed mixtures of normals can be found which result in m^* as in (2.7) with p having tails as thick as desired. This is important, in that it is precisely thick tails for the prior, and hence m^*, that are feared. (Thick tails for the prior are a convenient way to represent the feeling that some of the prior specifications for the θ_i might not accurately reflect the location of the θ_i. To a Bayesian actually believing in a prior distribution, thick tails are very reasonable since "surprising" θ_i are a fairly common experience.) The assumption that all Y_i have a common marginal is, of course, unrealistic, but we are only seeking rough guidelines as to how to choose ℓ and will virtually never have detailed information about the functional form of π or m^*; thus proceding on the basis of (2.7) is not unreasonable. (In certain of the following results (2.7) could be relaxed somewhat, but the relaxation leads to the same methodological conclusions.) The following theorem gives a useful expression for $r(\pi,\delta)$.

Theorem 1. Suppose m^* is exchangeable (clearly true if (2.7) is satisfied) and

$$\delta(X^*) = X^* - g(|Z|^2)Z, \tag{2.8}$$

where g is a bounded, continuous, piecewise differentiable function. Then

$$r(\pi,\delta) = \mathrm{tr}(Q^*\Psi^*) - \frac{1}{k}(\mathrm{tr}Q^*)E^{m^*}[2g(|Z|^2)\ell + |Z|^2\{4g'(|Z|^2) - g^2(|Z|^2)\}]. \tag{2.9}$$

Proof. Clearly Y is (conditionally) normally distributed with mean $\eta = \Sigma^{*-1}(\theta^*-\mu^*)$ and covariance matrix Σ^{*-1}. Defining $h(Y) = (h_1(Y),\ldots,h_k(Y))^t = g(|Z|^2)Z$, an expansion gives

$$R(\theta^*,\delta) = E^{X^*}_{\theta^*}[(X^*-\theta^*-h(Y))^t Q^*(X^*-\theta^*-h(Y))]$$
$$= tr(Q^*\Sigma^*) + E^{X^*}_{\theta^*}[h(Y)^t Q^* h(Y)] - 2E^{X^*}_{\theta^*}[(X^*-\theta^*)^t Q^* h(Y)]. \qquad (2.10)$$

Also,

$$E^{X^*}_{\theta^*}[(X^*-\theta^*)^t Q^* h(Y)] = E^{Y}_{\eta}[(\Sigma^*(Y-\eta))^t Q^* h(Y)]$$
$$= E^{Y}_{\eta}[\sum_{i=1}^{k} q_i^* \frac{\partial}{\partial Y_i} h_i(Y)],$$

the last step following from an integration by parts (i.e. use of Stein's unbiased estimator of risk). Taking expectations over θ^* (or η) in (2.10) thus gives

$$r(\pi,\delta) = tr(Q^*\Sigma^*) - \sum_{i=1}^{k} q_i^* E^{m^*}[g^2(|Z|^2)Z_i^2]$$
$$- 2\sum_{i=1}^{k} q_i^* E^{m^*}[\frac{\partial}{\partial Y_i} h_i(Y)]. \qquad (2.11)$$

By the assumption of exchangeability of m^*, it must be true that, for all i,

$$E^{m^*}[g^2(|Z|^2)Z_i^2] = \frac{1}{k}\sum_{i=1}^{k} E^{m^*}[g^2(|Z|^2)Z_i^2]$$
$$= \frac{1}{k} E^{m^*}[g^2(|Z|^2)|Z|^2], \qquad (2.12)$$

and

$$E^{m^*}[\frac{\partial}{\partial Y_i} h_i(Y)] = \frac{1}{k} E^{m^*}[\sum_{i=1}^{k} \frac{\partial}{\partial Y_i} h_i(Y)]. \qquad (2.13)$$

Letting I_B denote the usual indicator function of an event B, it is clear that

$$\frac{\partial}{\partial Y_i} Z_j = \begin{cases} I_{\{|Y_i| \le |Y|_{(\ell)}\}} & \text{if } j=i \\ (\text{sgn}Y_i)(\text{sgn}Y_j) I_{\{|Y_j| > |Y_i| = |Y|_{(\ell)}\}} & \text{if } j \neq i, \end{cases}$$

and hence

$$\frac{\partial}{\partial Y_i} h_i(Y) = g(|Z|^2) I_{\{|Y_i| \le |Y|_{(\ell)}\}} + 2g'(|Z|^2) Z_i$$

$$x\{Z_i I_{\{|Y_i| \le |Y|_{(\ell)}\}} + \sum_{j \ne i} Z_j (\mathrm{sgn} Y_i)(\mathrm{sgn} Y_j) I_{\{|Y_j| > |Y_i| = |Y|_{(\ell)}\}}\}.$$

Using this in (2.13) yields

$$E^{m^*}[\frac{\partial}{\partial Y_i} h_i(Y)] = \frac{1}{k} E^{m^*}[g(|Z|^2)\ell + 2g'(|Z|^2)\{\sum_{i=1}^{\ell} Z^2_{(i)} + (k-\ell)Z^2_{(\ell)}\}]$$

$$= \frac{1}{k} E^{m^*}[g(|Z|^2)\ell + 2g'(|Z|^2)|Z|^2].$$

Using this and (2.12) in (2.11) yields the desired conclusion. ||

Corollary 1.1. If m^* is exchangeable, then

$$r(\pi, \delta^{\ell}) = \mathrm{tr}(Q^* \ddagger^*) - \frac{1}{k}(\mathrm{tr}Q^*)[\int_{|z^2| \le \ell-2} (2\ell - |z|^2) m^*(y)dy$$

$$+ \int_{|z|^2 > \ell-2} \{(\ell-2)^2/|z|^2\} m^*(y)dy]. \qquad (2.14)$$

Proof. Simple calculation. ||

Theorem 2. When $Q^* = cI$ and δ is as in (2.8), then

$$R(\theta^*, \delta) = c\mathrm{tr}\ddagger^* - cE_{\theta^*}[2g(|Z|^2)\ell$$

$$+ |Z|^2\{4g'(|Z|^2) - g^2(|Z|^2)\}]. \qquad (2.15)$$

Proof. Almost identical to that of Theorem 1. ||

Corollary 2.1. When $Q^* = cI$, (2.9) and (2.14) hold for all m^*.

Corollary 2.2. When $Q^* = cI$, $R(\theta^*,\delta^{\ell}) < R(\theta^*,\delta^{0^*})$.

Proof. The same calculation as in (2.14) shows that $R(\theta^*,\delta^{\ell}) < \text{ctr}\ddagger^* = R(\theta^*,\delta^{0^*})$. ||

For particular m^*, the expression (2.14) for $r(\pi,\delta^{\ell})$ could be minimized numerically over ℓ to find the optimal choice of ℓ. Since m^* will rarely be known, however, we seek general guidelines by looking at the asymptotic (large k) situation as in Dey and Berger (1983) (which dealt with the symmetric case). We consider the choice

$$\ell = [\lambda k], \tag{2.16}$$

where $0 < \lambda \leq 1$ and $[n]$ denotes the nearest integer to n. Thus we will be considering truncating the fraction $1-\lambda$ of the largest $|Y_i|$. Of interest will be

$$r(\lambda) = \lim_{k\to\infty} \frac{1}{k}\left[\int_{|z|^2\leq \ell-2} \{2\ell-|z|^2\}m^*(y)dy + \int_{|z|^2>\ell-2} \{(\ell-2)^2/|z|^2\}m^*(y)dy\right], \tag{2.17}$$

since an easy calculation (using (2.14)) shows that

$$r(\lambda) = \lim_{k\to\infty} \frac{r(\pi,\delta^{0^*}) - r(\pi,\delta^{\ell})}{r(\pi,\delta^{0^*}) - r(\pi)},$$

where $r(\pi)$ is the Bayes risk of the optimal Bayes rule. (Of course, π is the prior corresponding to m^*.) Thus $r(\lambda)$ can be interpreted as the asymptotic proportional improvement of δ^{ℓ} over δ^{0^*} compared with the maximum possible improvement. We will seek a value of λ which seems to provide good $r(\lambda)$ for a wide range of m^*. To this end, the following theorem is needed.

Theorem 3. For ℓ as in (2.16) and m^* as in (2.7),

$$r(\lambda) = \begin{cases} \lambda^2/V_{\lambda} & \text{if } \lambda < V_{\lambda} \\ 2\lambda - V_{\lambda} & \text{if } \lambda \geq V_{\lambda}, \end{cases} \tag{2.18}$$

where

$$V_\lambda = E^{m^*}[Y_i^2 \wedge \alpha^2(\lambda)]$$
$$= \int_{-\alpha(\lambda)}^{\alpha(\lambda)} y^2 m^*(y)dy + \alpha^2(\lambda)(1-\lambda), \tag{2.19}$$

and $\alpha(\lambda)$ is the λth fractile of $|Y_i|$ defined by

$$\int_{-\alpha(\lambda)}^{\alpha(\lambda)} m^*(y)dy = \lambda.$$

Proof. Thd idea of the proof is to note that $|Y|_{(\ell)} \to \alpha(\lambda)$ and $k^{-1}|Z|^2 \to V_\lambda$, and that the condition $|Z|^2 \leq \ell-2$ becomes in the limit (after dividing by k) $V_\lambda \leq \lambda$. The rest follows in a very straightforward way from (2.17). The details are very similar to a related proof in Dey and Berger (1983) for the symmetric case, and will be omitted. ||

It remains only to try various m* in (2.19) and calculate the λ maximizing (2.18). Ideally, one would want to use m* that clearly correspond to actual priors π, but for calculational simplicity attention was restricted to m* of the form (2.7) with the $p(y_i)$ being t-densities given by

$$p(y_i) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)\sigma}\left(1 + \frac{y_i^2}{n\sigma^2}\right)^{-(n+1)/2}, \quad n \geq 1. \tag{2.20}$$

This class provides a wide range of tail behaviors for m*, and should be sufficiently representative of realizable marginals for our purposes.

One final choice that was needed was that of appropriate scale factors σ for the $p(y_i)$. Here we reasoned as follows. In talking about a prior mean μ and covariance matrix A, one is implicitly thinking in terms of a conjugate normal prior. For more realistic thicker-tailed priors, however, means and variances may not even exist. Thus it is more realistic to think of μ and A as indicating prior

medians and fractiles. For the conjugate prior, the Y_i will be marginally normal with mean zero and variance one, which can be thought of instead as normal with median zero and quartiles ±.68. Thus we choose the scale factors σ in (2.20) to give densities with the same quartiles ±.68. (Again, there is some inaccuracy in matching up the marginals instead of the priors themselves, but we are only seeking rough guidelines.)

The five cases considered were n=1, (Cauchy density), 2, 3, 4, and ∞ (normal density). The corresponding matched σ were .68, .82, .89, .91, and 1, respectively. Table 1 gives values of $r(\lambda)$, calculated from (2.18) and (2.19). (Formulas for V_λ in these cases can be found in Dey and Berger (1983).)

Table 1

Values of $r(\lambda)$

	n				
λ	1	2	3	4	∞
0	0	0	0	0	0
.1	.19	.19	.19	.19	.19
.2	.36	.35	.35	.35	.35
.3	.50	.49	.49	.49	.48
.4	.62	.61	.61	.61	.60
.5	.70	.70	.70	.70	.70
.6	.63	.77	.77	.77	.79
.7	.63	.73	.80	.83	.86
.8	.45	.73	.78	.83	.92
.9	.24	.59	.71	.81	.96
1	0	0	.42	.60	1

As can be seen from Table 1, the optimal values of λ seem to be about .5, .6, .7, .8, and 1 for the five cases studied. A good compromise value seems to be λ = .8, which costs about 8% in the normal (n=∞) case, is optimal for n=4, costs 3% when n=3, and costs 5% when n=2. (The cost for the Cauchy case is 36%, but the Cauchy case may be a bit extreme.) Of course, these results are asymptotic results as $k \to \infty$, and must be modified for smaller k. A choice of ℓ such as

$$\ell = 3 + [.8(k-3)]$$

seems reasonable for general use.

3. TRUNCATION FOR δ^{MB}

Consider the estimator δ^{MB} given in (1.4), where recall we are assuming (for simplicity) that Q, Σ, and A are diagonal. Paralleling (2.1) and (2.2), define $q_i^* = q_i\sigma_i^2/(\sigma_i^2+A_i)$ (assumed to be decreasingly ordered),

$$X_i^* = (\sigma_i^2+A_i)^{1/2}X_i/\sigma_i^2,\ \theta_i^* = (\sigma_i^2+A_i)^{1/2}\theta_i/\sigma_i^2,$$

$$\sigma_i^{2*} = (\sigma_i^2+A_i)/\sigma_i^2,\ \mu_i^* = (\sigma_i^2+A_i)^{1/2}\mu_i/\sigma_i^2.$$

The estimator δ^{MB} was derived (see Berger (1982a)) from consideration of the "subproblems" of estimating the first j coordinates of θ^*, namely $\theta^{*j} = (\theta_1^*,\ldots,\theta_j^*)^t$, based on $X^{*j} = (X_1^*,\ldots,X_j^*)^t$. In each subproblem one can use the truncation methods of Section 2. Thus define

$$Y_i = (X_i^* - \mu_i^*)/\sigma_i^{2*},$$

and order $|Y_1|,\ldots,|Y_j|$; denote the resulting order statistics

$$|Y|_{(1)}^{(j)} < |Y|_{(2)}^{(j)} < \ldots < |Y|_{(j)}^{(j)} .$$

(The superscript j is included because the ordering will typically depend on the subproblem, j, considered.) Then let $0 \le \ell_j \le j$ denote the desired truncation point in the subproblem, and define

$$Z_i^{(j)} = (\text{sgn}Y_i)(|Y_i| \wedge |Y|_{(i)}^{(j)})$$

and $Z^{(j)} = (Z_1^{(j)},\ldots,Z_j^{(j)})$. The truncated robust Bayes estimator in (2.5) for the subproblem (with a slightly altered choice of r_ℓ) is

$$\delta^{(j)}(X^{*j}) = X^{*j} - \min\{1, \frac{2(\ell_j-2)^+}{|Z^{(j)}|^2}\}Z^{(j)} , \tag{3.1}$$

where a^+ denotes the positive part of a. Note that $\delta^{(1)} = X^{*1}$ and $\delta^{(2)} = X^{*2}$, since $\ell_j \leq j$ implies that $(\ell_j-2)^+ = 0$ for j=1 or 2. The use of the constants $2(\ell_j-2)^+$ instead of $(\ell_j-2)^+$ in the estimators is because $X^{*j}-(Y_1,\ldots,Y_j)$ can be seen to be the conjugate prior Bayes estimator for the subproblem, and it seems desirable from a Bayesian viewpoint to use an estimate as close to $X^{*j}-Z^{(j)}$ (and hence as close to $X^{*j}-(Y_1,\ldots,Y_j)$) as possible. The choice $2(\ell_j-2)^+$ is optimal from this viewpoint, in that it is the largest possible constant for which the resulting estimator is still minimax.

As in Berger (1982a), the appropriate estimator for the entire parameter vector θ^* is constructed from the $\delta^{(j)}$ (based on an idea in Bhattacharya (1966)), and is given componentwise by

$$\begin{aligned}\delta_i^*(X^*) &= q_i^{*-1} \sum_{j=i}^{k} (q_j^*-q_{j+1}^*)\delta_i^{(j)}(X^{*j}) \\ &= X_i^* - Z_i^{(j)}[q_i^{*-1} \sum_{j=i}^{k} (q_j^*-q_{j+1}^*)\min\{1, \frac{2(\ell_j-2)^+}{|Z^{(j)}|^2}\}],\end{aligned} \tag{3.2}$$

where $q_{k+1}^* \equiv 0$. (Of course, multiplying by $\sigma_i^2/(\sigma_i^2+A_i)^{1/2}$ will convert this back to an estimate of θ_i.)

Theorem 4. The estimator δ^* defined by (3.2) is minimax (i.e., $R(\theta^*,\delta^*) \leq R(\theta^*,\delta^{0*})$ for all θ^*).

Proof. Following the argument in Berger (1982a), it is only necessary to show that the subproblem estimators $\delta^{(j)}$ in (3.1) are minimax under sum of squares error loss. But using Theorem 2, a calculation gives

$$\begin{aligned}R(\theta^*,\delta^{(j)}) &= \sum_{i=1}^{j} \sigma_i^{2*} - E_{\theta^*}[(2\ell_j - |Z^{(j)}|^2)I_{\{|Z^{(j)}|^2 \leq 2(\ell_j-2)^+\}}] \\ &< \sum_{i=1}^{j} \sigma_i^{2*} = R(\theta^*,\delta^{0*}). \quad ||\end{aligned}$$

No attempt will be made to formally investigate the effect of truncation in δ^*, but the nature of the construction of the estimator from the subproblem estimators suggests that the ℓ_j should be chosen to be approximately equal to the "optimum" truncation values in the subproblems. The analysis in Section 2 thus suggests the choices

$$\ell_j = 3 + [.8(j-3)].$$

References

[1] Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. Ann. Statist. 8, 716-761.

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

[3] Berger, J. (1982b). Bayesian robustness and the Stein effect. J. Amer. Statist. Assoc. 77, 358-368.

[4] Berger, J. (1983). The robust Bayesian viewpoint. In Robustness in Bayesian Statistics (J. Kadane, ed.), North-Holland, Amsterdam.

[5] Berger, J. and Dey, D.K. (1983). Combining coordinates in simultaneous estimation of normal means. To appear in J. Statist. Planning and Inference.

[6] Bhattacharya, P.K. (1966). Estimating the mean of a multivariate normal population with general quadratic loss function. Ann. Math. Statist. 37, 1819-1824.

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

[8] Dey, D.K. and Berger, J. (1983). On truncation of shrinkage estimators in simultaneous estimation of normal means. To appear in J. Amer. Statist. Assoc.

[9] Efron, B. and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators - Part II: The empirical Bayes case. J. Amer. Statist. Assoc. 67, 130-139.

[10] Efron, B. and Morris, C. (1973a). Stein's estimation rule and its competitors - an empirical Bayes approach. J. Amer. Statist. Assoc. 68, 117-130.

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

[12] James, W. and Stein, C. (1961). Estimation with quadratic loss. Proc. Fourth Berkeley Symposium Math. Stat. Prob. 1, 316-379. University of California Press.

[13] Stein, C. (1981). Estimation of the parameters of a multivariate normal distribution. I. Estimation of the mean. Ann. Statist. 9, 1135-1151.