Bayesian Input in Stein Estimation and A New Minimax Empirical Bayes Estimator*

by

James Berger
Purdue University

L. Mark Berliner
Ohio State University

Technical Report #83-36

Department of Statistics
Purdue University

October 1983

*This research was supported by the National Science Foundation under Grant MCS-8101670A1.

BAYESIAN INPUT IN STEIN ESTIMATION AND A NEW MINIMAX EMPIRICAL BAYES ESTIMATOR*

James BERGER
Purdue University, West Lafayette, IN 47907, USA
L. Mark BERLINER
Ohio State University, Columbus, OH 43210, USA

The relationship between Stein estimation of a multivariate normal mean and Bayesian analysis is considered. The necessity to involve prior information is discussed, and the various methods of so doing are reviewed. These include direct Bayesian analyses, adhoc utilization of prior information, restricted class Bayesian and Γ-minimax analyses, and Type II maximum likelihood (empirical Bayes) methods. A new empirical Bayes Stein-type estimator is developed, via the latter method, for an interesting ε-contamination class of priors, and is shown to be minimax under reasonable conditions. The minimax proof contains some novel theoretical features.

*This research was supported by the National Science Foundation under Grant MCS-8101670A1.

Send proofs to James Berger, Department of Statistics, Purdue University, West Lafayette, IN 47907

1. Introduction

1.1 Goals

The major goal of this paper is to discuss the intimate relationship between Stein estimation of a multivariate normal mean and Bayesian analysis. Indeed Stein estimation can not be sensibly carried out without at least informal consideration of Bayesian aspects of the problem (see section 1.3). The paper is organized around the four major methods (see Section 1.4) in which prior information can be used in concert with, or related to, Stein estimation. Although the paper will be formally self-contained, references to in-depth studies of these methods will be given. Many of these ideas apply also to non-normal multivariate estimation problems, but for simplicity we consider only the normal case in this paper.

A second goal of the paper is to present some interesting new Stein-type procedures and evaluations of them. Foremost of these new results is the development of a new empirical Bayes estimator, and proof of its minimaxity, in section 5. The novelty of the estimator lies in its being empirical Bayes with respect to a "sensible" class of prior distributions, while the minimax proof contains several novel theoretical ideas that should be useful elsewhere.

1.2 Notation

A huge variety of multivariate estimation problems can be reduced to the canonical form where $X = (X_1,\ldots,X_p)^t$ is a random vector with a $\mathcal{N}_p(\theta,\Sigma)$ distribution, the mean $\theta = (\theta_1,\ldots,\theta_p)^t$ being unknown and the $(p\times p)$ positive definite covariance matrix Σ being assumed known. (It is typical of this subject that, when Σ is unknown, insertion of suitable estimates of Σ into all expressions leads to qualitatively similar conclusions.) It is desired to estimate θ by an estimator $\delta(x) = (\delta_1(x),\ldots,\delta_p(x))^t$, under the sum of squares error loss

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

(The more general situation of a quadratic loss, $(\theta-\delta)^t Q(\theta-\delta)$, can be reduced to the above case by a linear transformation. Also, losses having different functional dependencies on $(\theta-\delta)$ are known to give qualitatively the same results.) The usual frequentist evaluation of an estimator δ is via its frequentist risk function (or mean squared error)

$$R(\theta,\delta) = E_\theta^X L(\theta,\delta(X)) = E_\theta^X |\theta-\delta(X)|^2.$$

(In this paper E will denote expectation, with superscripts giving the random quantity or distribution over which the expectation is to be taken and subscripts standing for fixed parameters.)

Since Bayesian considerations will be important, we will use $\pi(\theta)$ to denote a prior density for θ. Of interest will be the posterior density,

$$\pi(\theta|x) = \pi(\theta)f(x|\theta)/m(x|\pi),$$

where $f(x|\theta)$ is the $\mathcal{N}_p(\theta,\Phi)$ density and

$$m(x|\pi) = E^\pi f(x|\theta) = \int_{R^p} f(x|\theta)\ \pi(\theta)\ d\theta \tag{1.1}$$

is the marginal or predictive distribution of X for the prior π. A posterior Bayesian would be primarily interested in the posterior expected loss of $\delta(x)$ (for given data x), namely

$$\rho(x,\delta) = E^{\pi(\theta|x)}|\theta-\delta(x)|^2 = \int |\theta-\delta(x)|^2 \pi(\theta|x)\ d\theta, \tag{1.2}$$

while a frequentist (or global) Bayesian would be concerned with the Bayes risk

$$r(\pi,\delta) = E^m \rho(X,\delta(X)) = E^\pi R(\theta,\delta), \tag{1.3}$$

which is just the average loss over both θ and X. Via either measure, the optimal estimator is the posterior mean

$$\delta^\pi(x) = E^{\pi(\theta|x)}[\theta].$$

A common Bayesian analysis of this situation is to use a conjugate $\mathcal{N}_p(\mu,A)$ prior, where $\mu = (\mu_1,\ldots,\mu_p)^t$ is the prior mean and A is the $(p\times p)$ prior covariance matrix. Denoting this prior π_N, standard calculations (c.f., Berger (1980a)) give

that the posterior mean is

$$\delta^{\pi_N}(x) = x - \Sigma(\Sigma+A)^{-1}(x-\mu), \tag{1.4}$$

and that

$$\rho(x,\delta^{\pi_N}(x)) = tr\Sigma - tr\Sigma(\Sigma+A)^{-1}\Sigma = r(\pi_N,\delta^{\pi_N}), \tag{1.5}$$

$$R(\theta,\delta^{\pi_N}) = tr[\Sigma^{-1}+2A^{-1}+A^{-1}\Sigma A^{-1}]^{-1} + |\Sigma(\Sigma+A)^{-1}(\theta-\mu)|^2, \tag{1.6}$$

and

$$m(x|\pi_N) \text{ is } \mathcal{N}_p(\mu,\Sigma+A). \tag{1.7}$$

1.3 The Stein effect and prior information

The classical estimator of θ is, of course, $\delta^0(x) = x$, and it is minimax with constant risk

$$R(\theta,\delta^0) \equiv tr\Sigma.$$

The Stein effect concerns the fact that, when $p \geq 3$, estimators δ^* can be found that dominate δ^0, in the sense that, for all θ,

$$R(\theta,\delta^*) < R(\theta,\delta^0). \tag{1.8}$$

(Any such δ^* is also clearly minimax.) Study of this phemonenon began with Stein (1956) and James and Stein (1960). General discussions and other references can be found in Judge and Bock (1979) and in Berger (1983a).

The necessity for the involvement of prior information follows from the basic fact that $R(\theta,\delta^*)$ is substantially less than $R(\theta,\delta^0)$ only in a relatively small region or subspace of the parameter space. Hence, unless θ is thought to lie in this region, use of δ^* will gain only a negligible amount. (The same can be said for all other approaches to multiparameter estimation, such as ridge regression.) Clearly, therefore, among the great variety of Stein effect estimators that are available, the one selected for use should be one whose region of significant improvement coincides with where θ is thought likely to be (apriori).

The simplest situation to conceptualize is that in which θ is thought to lie in some ellipse, say

$$C = \{\theta: (\theta-\mu)^t A^{-1}(\theta-\mu) \leq p\}. \tag{1.9}$$

(Here μ could be thought of as the "best guess" for θ, and A as corresponding to specifications of accuracies (variances) and dependencies (covariances) for these opinions.) It has long been recognized that specification of μ is necessary in Stein estimation (essentially being the point to which one shrinks), but it is shown in Berger (1982a) that the "shape feature" A is also usually crucial. These appear to be the key needed inputs (see Berger (1980b, 1982a)), and luckily correspond to aspects of prior knowledge which are fairly easy to specify. (For details about how prior information concerning regression parameters can often be reduced to this form, see Berger (1980b, 1982a).)

A different important possibility is that θ may be thought to lie near some subspace of R^p, say the line where all θ_i are equal. Versions of Stein estimators are then appropriate which shrink towards this subspace; substantial risk improvement can often be guaranteed along the entire subspace. Or θ may be thought to lie near the surface of some convex set, in which case the Stein estimator should shrink towards this surface (c.f., Bock (1983)). Or various combinations of the above types of information may be available.

Given a region in which θ is thought to lie, the question arises as to how to select an estimator good for this region. The most natural solution is to find the Stein effect estimator, δ^*, which does best "on the average" over this region. And the most direct way to implement this is to choose a prior distribution π_0, reflecting where θ is thought likely to lie, and then measure the average performance of δ^* by

$$r(\pi_0,\delta^*) = \int R(\theta,\delta^*)\, \pi_0(\theta)\, d\theta.$$

In the situation of (1.9), for instance, it would frequently be reasonable to model the prior information as being that π_0 is $\mathcal{N}_p(\mu,A)$. If, on the other hand, it is thought that the θ_i are similar, it would be reasonable to model this in the usual empirical Bayes fashion of assuming that π_0 is $\mathcal{N}_p(\mu,A)$, where $\mu = (1,\ldots,1)^t\mu_0$, $A = \tau^2 I$, and μ_0 and τ^2 are unknown or partially unknown. Indeed, it will virtually always be most efficient in Stein estimation to begin by attempting a (possibly crude)

quantification of a prior π_0. The goal, of course, will then be to find a good Stein effect estimator which also has good Bayesian performance with respect to π_0.

To non-Bayesians, this approach may be somewhat disturbing. Note, however, that the usual fears about Bayesian analysis are not applicable here; even if the prior π_0 is completely inappropriate, the selected δ^* (if it satisfies (1.8)) will still not be worse than δ^0 (though it will be essentially equivalent). One could argue that sometimes no prior information is available (although this would be rare in economic situations). The answer then is quite simple: just use δ^0, since no Stein effect estimator would have much of a chance of offering significant improvement.

For the above reasons, even frequentists working in multiparameter estimation should focus substantial attention on prior specification and its use in estimator selection. The "common" practice of proposing an estimator and comparing it with others via simulations is inappropriate. Different estimators will do well in different regions (or, equivalently, for different priors), so prior information is the only way to differentiate among estimators. A number of Stein-like approaches, such as ridge regression, offer the allure of avoidance of prior input. The allure is a specious one, however; any particular ridge regression estimator will do well only in a particular region of the parameter space. Any suggestion that the estimator will magically shrink towards the correct region is simply erroneous: even in empirical Bayes situations, where the data helps select the direction and amount of shrinkage, prior information concerning the relationships of the θ_i must be specified to direct the shrinkage. Among the many good articles concerning this issue is Smith and Campbell (1980).

Finally, it should be remarked that one can often do amazingly well from both the Bayesian and frequentist perspectives, simultaneously. This has long been known in empirical Bayes settings (c.f. Efron and Morris (1973)). Other examples of this will be seen throughout the paper, especially in section 4.

1.4 Using π_0 in Stein estimation

Essentially four methods have been proposed for the needed incorporation of prior information in Stein estimation. The first is to simply derive the Bayes estimator for an appropriate π_0, and then to check its frequentist properties (such as dominance or near dominance of δ^0). This is discussed in section 2.

The second method is the "adhoc" method, which consists of taking existing Stein type estimators and attempting to modify them to incorporate prior information. This is discussed in section 3.

The third method is that of finding the Bayes estimator for π_0 within a restricted class of estimators, a class usually chosen for its good frequentist properties. A very interesting example is the Hodges and Lehmann (1952) "restricted risk Bayes" method, which is as follows:

$$\text{Among all } \delta^* \text{ for which } R(\theta,\delta^*) \leq R(\theta,\delta^0) + C,$$
$$\text{select that } \delta^* \text{ which minimizes } r(\pi_0,\delta^*). \tag{1.10}$$

When $C = 0$ in (1.10), one is finding the minimax estimator which is best for π_0. This restricted risk approach is closely related to the Γ-minimax problem of considering the class of priors

$$\Gamma = \{\pi = (1-\varepsilon)\pi_0 + \varepsilon Q;\ Q \in \mathcal{Q}\}, \tag{1.11}$$

where ε is a specified constant reflecting uncertainty in π_0, and $\mathcal{Q}$ is a class of "contaminations," and then choosing δ^* to minimize

$$\sup_{\pi \in \Gamma} r(\pi,\delta^*). \tag{1.12}$$

The idea here is, of course, that, providing π_0, ε, and $\mathcal{Q}$ are chosen reasonably, Γ should contain all "plausible" prior distributions. If δ^* does well for all π in Γ, it should thus be quite satisfactory for use (at least from a frequentist Bayes viewpoint). These approaches are discussed in Section 4, where some quite surprising results are presented.

The final method of approaching the problem is the empirical Bayes or, more

precisely, Type-II maximum likelihood method (see Good (1980)). This method is related to the Γ-minimax approach, in that one considers a class of priors such as (1.11), but one then chooses the "most likely" prior in Γ, according to the data. Since the "likelihood" of a prior is simply $m(x|\pi)$, this leads to the following definition.

Definition. For a class Γ of priors and given data x, the ML-II (type II maximum likelihood) prior in Γ, $\hat{\pi}$, is defined (assuming it exists and is unique) as that $\pi \in \Gamma$ maximizing $m(x|\pi)$ over all $\pi \in \Gamma$.

Most empirical Bayes analyses proceed by letting Γ be all conjugate priors, choosing the ML-II $\hat{\pi}$, and then doing a Bayesian analysis pretending that $\hat{\pi}$ is the true prior. Although this smacks of adhocery (the "prior" $\hat{\pi}$ being chosen in a data dependent fashion), a number of justifications for the approach can be given. (See Berger and Berliner (1983) for discussion and references.) Also, the approach seems to work remarkably well. In section 5 we apply the approach to realistic classes of priors such as (1.11). Indeed for Γ as in (1.11), with π_0 being a conjugate prior and $\mathcal{Q}$ being a reasonable class of contaminations, a quite attractive estimator is developed (via this approach) which is a data adaptive compromise between δ^{π_0} and a standard empirical Bayes estimator. Furthermore, the estimator is shown to be minimax (for large enough ε) by a minimax proof incorporating some novel features.

2. Bayesian Stein-type estimators

2.1 Introduction

There are several advantages to approaching Stein estimation through Bayesian development of estimators. The first is that one can be certain that the needed prior information is used correctly. The second is that the resulting estimator is often admissible (c.f., Strawderman (1971) and Berger (1976)). Related to this is the fact that Bayesian estimators are guaranteed to be fine conditionally: the original James-Stein (1960) estimator (for $\Sigma = I$),

$$\delta^{J\text{-}S}(x) = (1 - (p-2)/|x|^2)x,$$

has $R(\theta,\delta^{J\text{-}S}) < R(\theta,\delta^0)$, but if $p = 3$ and $x = (.01, -.01, .01)^t$ then $\delta^{J\text{-}S}(x) \cong (-33, 33, -33)^t$, a ridiculous result. Of course, the positive part version corrects this obvious conditional deficiency, but only Bayesian development of procedures can generally guarantee that the procedures are sensible for each x, and not just on the average (c.f., Berger and Wolpert (1984)).

The final and most important practical reason to adopt a Bayesian approach to Stein estimation is that one can easily obtain confidence sets, perform tests, etc.. The posterior covariance matrix is usually not much harder to calculate than the posterior mean, and can be used to indicate the variability of the estimates (c.f., Box and Tiao (1973), Berger (1980b), Morris (1983a, 1983b) and Van der Merwe et. al. (1981)). Trying to develop estimates of variability from the frequentist perspective has proved enormously difficult, even in the simplest cases (c.f., Hwang and Casella (1982)).

The disadvantage with Bayesian development is that the estimators are sometimes expressible only as numerical integrals, and that frequentist risk properties can be hard to verify. Situations where good frequentist risks are known to result are discussed in this section. Note that the conjugate priors, π_N, discussed in section 1.2, do <u>not</u> have good frequentist risk properties, in that (see (1.6))

$$\sup_\theta R(\theta,\delta^{\pi_N}) = \infty.$$

2.2 Flat-tailed prior distributions

There is substantial evidence (c.f., Rubin (1977) and Berger (1983b)) that estimators developed from priors with flat (polynomial like) tails are Stein-type and have good risk properties. One example of such a development is Berger (1980b), in which such a prior is used as an alternative to π_N, and results in the Stein-type estimator (posterior mean)

$$\delta^{RB}(x) = x - \frac{r(||x-\mu||^2)}{||x-\mu||^2}\,\tfrac{1}{2}(\tfrac{1}{2}+A)^{-1}(x-\mu),$$

where $||x-\mu||^2 = (x-\mu)^t(\Sigma+A)^{-1}(x-\mu)$ and $r(z)$ is approximately $\min\{z,p-2\}$. Note the similarity of this estimator with $\delta^{\pi N}$ in (1.4). Indeed, if μ and A are accurate reflections of the location of θ, then $||x-\mu||^2$ will be approximately p (see (1.7)), and (for moderately large p) δ^{RB} will be essentially $\delta^{\pi N}$. If, however, θ is far from μ, then $r(||x-\mu||^2)/||x-\mu||^2$ will be small, and δ^{RB} will be essentially δ^0. This is the general behavior of estimators developed from flat tailed priors, and is the source of their good frequentist properties. (Indeed δ^{RB} is sometimes minimax and always has bounded risk - see Berger (1980b).) Confidence regions for θ, centered about δ^{RB}, are also given in Berger (1980b).

2.3 Hierarchical priors

When the prior information consists of structural knowledge about similarities or relationships among the θ_i, this can often be conveniently modeled by hierarchical priors. For instance, the usual empirical Bayes situation, in which the θ_i are felt to be similar, can be modeled by supposing that the θ_i are (independently $\mathcal{N}(\nu_0,\tau_0^2)$, and then putting a second stage (perhaps diffuse) prior on ν_0 and τ_0^2. The resulting estimator (posterior mean) will again be Stein-type and have good frequentist risk, providing the second stage prior on τ_0^2 has flat tails. There is a huge literature on this approach, with a wealth of excellent statistical estimators (and associated confidence sets). References can be found in Lindley and Smith (1972), Good (1980), Morris (1983b), and Berger (1983b).

2.4 Posterior robustness

An exciting possibility exists for bypassing the (often difficult) verification of good frequentist risk behavior of a Bayesianly derived Stein-type estimator. The idea is to look at the range of posterior means for all π in a class Γ of plausible priors, such as (1.11). If this range is small, then posterior robustness obtains, and a Bayesian would be completely satisfied with use of $\delta^{\pi 0}(x)$. It seems likely (see Brown (1983)) that the estimate is also then satisfactory from a frequentist viewpoint. The great appeal of this approach lies in the fact that posterior

robustness needs to be investigated only for the actual observed x. (Of course, to a Bayesian, this is not just a convenient technique, but is the only fundamentally sound statistical analysis.) See Berger (1983b) and Berger and Berliner (1983) for discussion of some of these issues.

3. Adhoc incorporation of prior information

One can simply incorporate the needed prior information into estimators in an "intuitive" fashion. One example is Berger (1982a), where a minimax version of δ^{RB} was developed, incorporating μ and A. This was done by simply taking an existing class of estimators due to Bhattacharya (1966) and Berger (1979), and incorporating μ and A. The resulting estimator, when Σ and A are diagonal, can be written (coordinatewise) as

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

where $\{\sigma_i^2\}$ and $\{A_i\}$ are the diagonal elements of Σ and A, $q_i = \sigma_i^4/(\sigma_i^2+A_i)$ (and a relabelling has been done, if necessary, so that $q_1 \geq q_2 \geq \cdots \geq q_p > 0 \equiv q_{p+1}$), and $||x^j-\mu^j||^2 = \sum_{\ell=1}^{j}(x_\ell-\mu_\ell)^2/(\sigma_\ell^2+A_\ell)$. This estimator whas shown to <u>always</u> be minimax and to have good Bayesian properties.

A second example of such an adhoc approach is that of Bock (1983), in which minimax estimators are developed which shrink towards the surface of convex sets (such as a sphere, the positive orthant, or a wedge). These estimators would be of Bayesian value when the prior information is that θ is likely to be near such a surface. (As an example, one might have vague prior knowledge concerning the length, $|\theta|$, of θ, and hence want to shrink towards the surface of the appropriate sphere.)

The chief usefulness of such adhoc estimator developments is that they allow (usually by design) proof of minimaxity (or some other desirable property) of the final estimator, while bearing some relation to the desired prior input. Their

weaknessess are those that have already been mentioned: they may utilize the prior information in an inferior way; they are often not admissible; and they do not lead to error estimates.

4. Restricted class Bayes and Γ-minimax estimators

A reasonable approach to incorporating prior information into Stein estimation is to restrict oneself to a class of estimators known to have desirable frequentist risk properties, and within this class seek an estimator good with respect to π_0. Although there are several examples of this in the literature, we will restrict discussion here to the restricted risk Bayes problem posed by (1.10). There is also the closely related Γ-minimax problem discussed in (1.11) and (1.12). These two problems are actually equivalent in a wide variety of situations, as the following lemma shows. For use in this lemma define the "orthant at θ'" as the set

$$\Lambda(\theta') = \{\theta: \theta_i > \theta_i' \text{ if } \theta_i' > 0 \text{ and } \theta_i < \theta_i' \text{ if } \theta_i' < 0\}.$$

Lemma 1. Suppose that Γ is as in (1.11) and that, for each θ', there exists $Q_{\theta'} \in \mathcal{Q}$ such that $P^{Q_{\theta'}}(\Lambda(\theta')) = 1$. Suppose also that δ is an estimator such that $R(\theta,\delta)$ is nondecreasing in $|\theta_i| > K$ for some K and all i. Then

$$\sup_{\pi \in \Gamma} r(\pi,\delta) = (1-\varepsilon) r(\pi_0,\delta) + \varepsilon \sup_{\theta} R(\theta,\delta). \tag{4.1}$$

Proof. Clearly

$$\sup_{\pi \in \Gamma} r(\pi_0,\delta) = (1-\varepsilon)\, r(\pi_0,\delta) + \varepsilon \sup_{Q \in \mathcal{Q}} r(Q,\delta).$$

Define $M = \sup_{\theta} R(\theta,\delta)$, and observe that, for any $\lambda > 0$, one can find a point θ^λ such that $|\theta_i^\lambda| > K$ for all i and $R(\theta^\lambda,\delta) > M - \lambda$. Since $R(\theta,\delta)$ is nondecreasing for $|\theta_i| > K$, it follows that

$$r(Q_{\theta^\lambda},\delta) = \int_{\Lambda(\theta^\lambda)} R(\theta,\delta)\, dQ_{\theta^\lambda}(\theta) > M - \lambda.$$

But λ is arbitrary, so that

$$\sup_{Q \in \mathcal{Q}} r(Q,\delta) \geq M.$$

On the other hand,

$$r(Q,\delta) = \int R(\theta,\delta)\, dQ(\theta) \leq \int M\, dQ(\theta) = M,$$

establishing the conclusion.||

Virtually all classes of contaminations Ω that are considered (c.f., Berger and Berliner (1983)) satisfy the mild condition of the lemma, and furthermore, it can frequently be shown (for, e.g., unimodal π_0) that any δ minimizing $\sup_{\pi \in \Gamma} r(\pi_0,\delta)$ must also satisfy the condition of the lemma. But it follows from a standard game-theoretic argument that any δ minimizing the right hand side of (4.1) must also be the solution to (1.10), where ε and C are related in a monotonic fashion.

As a specific example, consider the situation where $\Sigma = \sigma^2 I$ and π_0 is $\mathcal{N}_p(\mu,\tau^2 I)$. In Berger (1982b, 1982c) it is shown, when $p \geq 3$ and C=0 in (1.10) (or equivalently ε=1 in the Γ-minimax case), that the approximate optimal restricted risk Bayes rule is the Stein type estimator

$$\delta^R(x) = \begin{cases} \delta^{\pi_0}(x) = x - \dfrac{\sigma^2}{(\sigma^2+\tau^2)}(x-\mu) & \text{if } |x-\mu|^2 < 2(p-2)(\sigma^2+\tau^2) \\ x - \dfrac{2(p-2)\sigma^2}{|x-\mu|^2}(x-\mu) & \text{if } |x-\mu|^2 > 2(p-2)(\sigma^2+\tau^2). \end{cases}$$

Since C=0, δ^R is minimax, and hence <u>always</u> better than δ^0 in terms of frequentist risk. It is convenient to measure the Bayesian performance of δ^R by the "relative savings risk" (see Efron and Morris (1972))

$$RSR(\pi_0,\delta) = \frac{r(\pi_0,\delta) - r(\pi_0,\delta^{\pi_0})}{r(\pi_0,\delta^0) - r(\pi_0,\delta^{\pi_0})}.$$

When RSR is near zero, δ is essentially as good as the optimal Bayes rule with respect to π_0 (namely, δ^{π_0}) while, when RSR is near one, δ^{π_0} is no better than the standard estimator δ^0. Table 1 gives $RSR(\pi_0,\delta^R)$ for various values of p. These values are startling, in that δ^R has essentially optimal Bayesian performance (at least for $p \geq 5$) while maintaining minimaxity. That one could have such optimal

frequentist and Bayesian performance simultaneously is astonishing.

Results in this situation for $C>0$ (or $\varepsilon<1$) are also given in Berger (1982b, 1982c) when $p\geq 2$, and when $p=1$ in Efron and Morris (1971). The (approximate) solutions depend in general on modified Bessel functions. For $p=1$ and $p=3$, however, they have the fairly simple form (respectively)

$$\delta^{1,M} = \begin{cases} x - \dfrac{\sigma^2}{(\sigma^2+\tau^2)}(x-\mu) & \text{if } (x-\mu)^2 < M(\sigma^2+\tau^2) \\ x - (\text{sgn}[x-\mu])[M\sigma^4/(\sigma^2+\tau^2)]^{1/2} & \text{otherwise,} \end{cases}$$

$$\delta^{3,M} = \begin{cases} x - \dfrac{\sigma^2}{(\sigma^2+\tau^2)}(x-\mu) & \text{if } |x-\mu|^2 \leq (\sigma^2+\tau^2)d \\ x - \left[\dfrac{2\sigma^2}{|x-\mu|^2} + \dfrac{\sqrt{3M}\,\sigma^2}{\sqrt{\sigma^2+\tau^2}\;|x-\mu|}\right](x-\mu) & \text{otherwise,} \end{cases}$$

where $d = \frac{3}{2}M + \frac{4}{3} + M\sqrt{1+8/(3M)}$, and

$$M = \frac{C}{r(\pi_0,\delta^0) - r(\pi_0,\delta^{\pi_0})} = \frac{C(\sigma^2+\tau^2)}{p\sigma^4} .$$

Here M indicates the amount that $\delta^{p,M}$ is worse than a minimax rule (for which $M=C=0$) in terms of $\sup_\theta R(\theta,\delta)$, normalized to be on the same scale as $RSR(\pi_0,\delta)$. Table 2 presents M, $RSR(\pi_0,\delta^{p,M})$, and also values (see (4.1)) of

$$r(\varepsilon) = \frac{\sup_{\pi\in\Gamma} r(\pi,\delta^{p,M}) - r(\pi_0,\delta^{\pi_0})}{r(\pi_0,\delta^0) - r(\pi_0,\delta^{\pi_0})} \tag{4.2}$$

i.e., the suitably normalized Γ-minimax risk of $\delta^{p,M}$. It can be shown that

$$r(\varepsilon) = (1-\varepsilon)RSR(\pi_0,\delta^{p,M}) + \varepsilon(1+M). \tag{4.3}$$

Actually, it was more convenient to just give the results for $\varepsilon = .1, .2, .3,$ and $.4$, with M being determined as the corresponding value giving the optimal (approximate)

Γ-minimax rule (see the discussion after Lemma 1). The results are also given for p=2. (See Berger (1982b or 1982c) for the appropriate estimator in this case.)

The tradeoff between increased minimax risk (M) and Bayesian performance (RSR) is well illustrated by Table 2. When p=1, for instance, one can have reasonable Bayesian performance, sacrificing only 32% of the possible Bayesian gains, by being willing to accept an increase of 40% in the minimax risk. For p=3 the situation is very pleasant; by allowing an increase of 10% in minimax risk one surrenders only 13% of the possible Bayesian gains. The data for p=2 are included because they show that the "Stein effect" is operating even in two dimensions: the values of M and RSR are much better than when p=1.

The Γ-minimax data from Table 2 are also worth perusing. For instance, in p=3, if one elicits μ and τ^2 and feels that a normal shape for the prior is reasonable, but feels that the prior specification could be off (in terms of misspecified prior probabilities) by, say 20% (i.e., ε=.2), then using $\delta^{3,.2}$ would guarantee a Bayes risk no worse than 31% from "optimal" by the standardized measure. (Using (4.2) and (4.3) it is easy to convert this into actual Γ-minimax risk if desired.) Thus, for <u>any</u> prior deemed reasonable, $\delta^{3,.2}$ would have excellent overall risk, which should be satisfactory even to frequentists.

It should be mentioned that, for conditional Bayesians, the estimators discussed in this section are very sensible, being simply the conjugate prior Bayes estimator when x is near μ (and so compatible with the prior), while being similar to Bayes estimators with flat tails otherwise. Also, although the problem becomes much more difficult when Σ and A are not multiples of the identity, good restricted risk Bayes procedures and Γ-minimax procedures have been developed for such situations in Chen (1983).

5. Type-II maximum likelihood and empirical Bayes estimators

5.1 The assumed prior structure

Consider the "empirical Bayes" situation in which the θ_i are thought to be

independent realizations from a common $\mathcal{N}(\nu,\tau^2)$ prior density, but ν and τ^2 are partially unknown. Model this uncertainty in a robust hierarchical Bayesian fashion, i.e., assume that ν and τ^2 have a prior density

$$h(\nu,\tau^2) = (1-\varepsilon)\, h_0(\nu,\tau^2) + \varepsilon Q(\nu,\tau^2),$$

where h_0 is an elicited prior, ε is the possible error in elicitation, and Q is a member of a class $\mathcal{Q}$ of possible contaminations. The overall prior for θ is then

$$\begin{aligned}\pi(\theta) &= \int \Big[\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\tau} \exp\{-(\theta_i-\nu)^2/2\tau^2\}\Big] h(\nu,\tau^2)\, d\nu d\tau^2 \\ &= (1-\varepsilon)\pi_0(\theta) + \varepsilon Q^*(\theta), \end{aligned} \tag{5.1}$$

where

$$\pi_0(\theta) = \int \Big[\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\tau} \exp\{-(\theta_i-\nu)^2/2\tau^2\}\Big] h_0(\nu,\tau^2)\, d\nu d\tau^2$$

and

$$Q^*(\theta) = \int \Big[\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\tau} \exp\{-(\theta_i-\nu)^2/2\tau^2\}\Big] Q(\nu,\tau^2)\, d\nu d\tau^2. \tag{5.2}$$

We are thus faced with the class of possible priors

$$\Gamma = \{\pi = (1-\varepsilon)\,\pi_0 + \varepsilon Q^*;\ \ Q \in \mathcal{Q}\}. \tag{5.3}$$

Notice that this is considerably more refined than the usual empirical Bayes setup, which effectively assumes that $\varepsilon=1$ and $\mathcal{Q}$ = {all unit point masses}, so that Γ is just the set of all $\mathcal{N}_p(\nu \underset{\sim}{1},\tau^2 I)$ priors, where $\underset{\sim}{1} = (1,1,\ldots,1)^t$. Usually prior knowledge about ν and τ^2 is available and, as briefly discussed in Berger (1982b) and Berger (1983c), can provide valuable gains (unless p is quite large).

5.2 The ML-II prior and posterior mean

An adhoc, but intuitively reasonable method of dealing with Γ is to choose the "most likely" prior in Γ. Recall, from section 1.4, that this is called the ML-II prior, $\hat{\pi}$, and is that prior in Γ (if it exists) maximizing $m(x|\pi)$ for the given data x. (Recall that $m(x|\pi)$ (see (1.1)) is the predictive density of x given π, so that $\hat{\pi}$ is the "maximum likelihood" prior in the usual sense. Good (1965)

calls this "Type-II maximum likelihood," and we will stick with his nomenclature.)

This technique is extensively discussed in Berger and Berliner (1983), to which the reader is referred for further justification. Note, at least, that this would correspond to the usual empirical Bayes technique (when Γ consists of all $\mathcal{N}_p(\nu\underset{\sim}{1},\tau^2 I)$ distributions) of estimating ν and τ^2 via the maximum likelihood method from the predictive density

$$m(x|\nu,\tau^2) = \mathcal{N}_p(\nu\underset{\sim}{1},\ \Sigma+\tau^2 I). \tag{5.4}$$

When π is as in (5.1), it is clear that

$$m(x|\pi) = (1-\varepsilon)m(x|\pi_0) + \varepsilon m(x|Q^*). \tag{5.5}$$

Hence

$$\sup_{\pi\in\Gamma} m(x|\pi) = (1-\varepsilon)m(x|\pi_0) + \varepsilon \sup_{Q\in\mathcal{Q}} m(x|Q^*). \tag{5.6}$$

Furthermore, if $\hat{Q}^*$ maximizes this last term, then $\hat{\pi} = (1-\varepsilon)\ \pi_0 + \varepsilon\hat{Q}^*$, and the posterior mean with respect to $\hat{\pi}$ is

$$\begin{aligned}\delta^{\hat{\pi}}(x) &= \frac{\int \theta\ f(x|\theta)\ \hat{\pi}(\theta)\ d\theta}{m(x|\hat{\pi})} \\ &= \hat{\lambda}(x)\ \delta^{\pi_0}(x) + (1-\hat{\lambda}(x))\ \delta^{\hat{Q}}(x),\end{aligned} \tag{5.7}$$

where

$$\hat{\lambda}(x) = (1-\varepsilon)\ m(x|\pi_0)/[(1-\varepsilon)m(x|\pi_0) + \varepsilon m(x|\hat{Q})]. \tag{5.8}$$

Note that, when the data x "agrees with" π_0, $m(x|\pi_0)$ will be reasonably large and $\hat{\lambda}(x)$ will be close to one. If, on the other hand, x gives considerably more support to $\hat{Q}$, then $\hat{\lambda}(x)$ will be close to zero. This adaptive behavior of $\delta^{\hat{\pi}}$ is what makes it so attractive.

5.3 A special case

The simplest case to deal with is that in which $\Sigma = \sigma^2 I$, $h_0(\nu,\tau^2)$ is a point mass at (ν_0,τ_0^2) and

$$\mathcal{Q} = \{\text{all distributions concentrated on } \tau^2 \geq \tau_0^2\}.$$

The idea here is that (ν_0,τ_0^2) is simply the best guess as to the "hyperparameters," ε is a measure of the strength of belief in this guess, and there is enough uncertainty to want to allow all contaminations in $\mathcal{Q}$, subject to the constraint that $\tau^2 \geq \tau_0^2$. (The reasons for this constraint are partly technical, so that the analysis goes smoothly, and partly to prevent "spurious" precision from creeping in.) It shouldbe mentioned that, while choice of such a large $\mathcal{Q}$ seems to work well (in a conservative sense) in estimation, other uses of the resulting $\hat{\pi}$ (such as for confidence sets) are suspect (see Berger and Berliner (1983)), although the problems of so using $\hat{\pi}$ here are not too severe.

For this situation,

π_0 is $\mathcal{N}_p(\nu_0\underset{\sim}{1},\tau_0^2 I)$, $m(x|\pi_0)$ is $\mathcal{N}_p(\nu_0\underset{\sim}{1},\ (\sigma^2+\tau_0^2)I)$,

$$\delta^{\pi_0}(x) = x - \frac{\sigma^2}{(\sigma^2+\tau_0^2)}\,(x-\nu_0\underset{\sim}{1}), \tag{5.9}$$

and (see (5.2) and (5.4))

$$m(x|Q^*) = \int m(x|\nu,\tau^2)\ Q(\nu,\tau^2)\ d\nu d\tau^2.$$

Clearly $m(x|Q^*)$ is maximized over $Q \in \mathcal{Q}$ by choosing Q to be a point mass at the maximum likelihood estimate of (ν,τ^2) (subject to the constraint $\tau^2 \geq \tau_0^2$, of course). But, from (5.4), it is clear that the maximum likelihood estimate is simply

$$\hat{\nu} = \bar{x},\ \hat{\tau}^2 = \max\{\tau_0^2,\ |s|^2/p-\sigma^2\}, \tag{5.10}$$

where $s^2 = \sum_{i=1}^{p} (x_i-\bar{x})^2$. Thus

$$m(x|\hat{Q}) \text{ is } \mathcal{N}_p(\hat{\nu}\underset{\sim}{1},\ (\sigma^2+\hat{\tau}^2)\underline{I}),$$

and

$$\delta^{\hat{Q}}(x) = x - \frac{\sigma^2}{\max\{\tau_0^2+\sigma^2,|s|^2/p\}}\ (x-\bar{x}\underset{\sim}{1}), \tag{5.11}$$

which is more or less the "standard" empirical Bayes estimate of θ in the exchangeable

case. Also, $\hat{\lambda}(x)$ can be written (after some algebra and letting max denote $\max\{\tau_0^2+\sigma^2, |s|^2/p\}$) as

$$\hat{\lambda}(x) = \left[1+\frac{\varepsilon}{(1-\varepsilon)}\cdot\left(\frac{\sigma^2+\tau_0^2}{\max}\right)^{p/2}\exp\left\{\frac{s^2}{2}\left[\frac{1}{(\sigma^2+\tau_0^2)}-\frac{1}{\max}\right]+\frac{p(\bar{x}-\nu_0)^2}{2(\sigma^2+\tau_0^2)}\right\}\right]^{-1}. \qquad (5.12)$$

Using this with (5.9) and (5.11) in (5.7) gives the ML-II posterior mean, which will be a very appealing data adaptive compromise between the Bayes estimator for the specified (ν_0,τ_0^2) and the empirical Bayes estimator which assumes that these hyperparameters are unknown. More discussion of this approach, along with examples of more realistic or richer structures that can be assumed, is given in Berger and Berliner (1983).

5.4 Minimaxity of $\hat{\delta}^{\pi}$

The intuitive rationale and justification for $\hat{\delta}^{\pi}$, together with the fact that the estimator behaves similarly to more familiar Stein-type estimators in the limits, lends support to the feeling that the estimator will have good frequentist properties (as well as good Bayesian properties). It is of interest, however, to attempt direct verification of this, by attempting to establish minimaxity of $\hat{\delta}^{\pi}$. Unfortunately, we were unsuccessful in this attempt, due to technical difficulties arising from the introduction of $\hat{\nu}$. Workers in this area are, however, familiar with the fact that such estimation of the supposed common mean rarely affects minimaxity by more than a needed slight alteration of constants, so we considered the simpler problem of known ν. Thus suppose that $X \sim \mathcal{N}_p(\theta,\sigma^2 I)$, π_0 is $\mathcal{N}_p(\nu_0,\tau_0^2 I)$, and

$$\Gamma = \{\text{all distributions concentrated on } \{\nu=\nu_0, \tau^2 \geq \tau_0^2\}\}.$$

An analysis identical to that of the previous section gives, as the ML-II posterior mean,

$$\hat{\delta}^{\pi} = \begin{cases} x - \dfrac{\sigma^2}{(\sigma^2+\tau_0^2)}(x-\nu_0\underset{\sim}{1}) & \text{if } \sum_{i=1}^{p}(x_i-\nu_0)^2 < p(\tau_0^2+\sigma^2) \\ x - \dfrac{\sigma^2 r(v)}{v}(x-\nu_0\underset{\sim}{1}) & \text{otherwise,} \end{cases} \qquad (5.13)$$

where $v = \sum_{i=1}^{p} (x_i-\nu_0)^2$,

$$r(v) = p[1 + (\frac{v}{p(\sigma^2+\tau_0^2)} - 1)\,\hat{\lambda}(v)], \tag{5.14}$$

$$\hat{\lambda}(v) = \left\{1 + \frac{\varepsilon}{(1-\varepsilon)} \cdot \left[\frac{(\sigma^2+\tau_0^2)p}{v}\right]^{p/2} e^{-p/2}\ e^{v/2(\sigma^2+\tau_0^2)}\right\}^{-1}.$$

Theorem. For $p\geq 5$, $\hat{\delta}^{\pi}$ in (5.13) is minimax (i.e., satisfies (1.8)) providing $\varepsilon \geq \hat{\varepsilon}_p$, ε_p being given in Table 3 for $5\leq p\leq 26$. (For $p>26$, ε_p is less than .009, so $\hat{\delta}^{\pi}$ will be minimax for any sensible ε.)

Proof. By making the appropriate linear transformation, it is sufficient to prove the theorem for $\nu_0=0$ and $\sigma^2+\tau_0^2=1$, which we henceforth assume. Using the familiar unbiased estimate of risk of Stein (1981), to show that an estimator,

$$\delta^*(x) = (1-\sigma^2 r^*(v)/v)x, \tag{5.15}$$

is minimax (where $X \sim \mathcal{N}_p(\theta,\sigma^2 I)$ and $v=|x|^2$), it suffices to show that

$$4\,\frac{d}{dv}\,r^*(v) + r^*(v)\,v^{-1}[2(p-2) - r^*(v)] \geq 0. \tag{5.16}$$

For $v<p$, $\hat{\delta}^{\pi}$ is of the form (5.15) with $r^*(v) = v$ (recall $\sigma^2+\tau_0^2=1$, $\nu_0=0$), for which verification of (5.16) is trivial. For $v>p$, we must verify (5.16) for $r^*=r$ as in (5.14). Substituting (5.14) into (5.16), and after some algebra, the problem reduces to showing that (for $v>p$)

$$K_1(v) + K_2(v)\,c + K_3(v)\,c^2 \geq 0, \tag{5.17}$$

where

$$c = \varepsilon(1-\varepsilon)^{-1}\,p^{p/2}e^{-p/2},$$

$$K_1(v) = 2vp-v^2,$$

$$K_2(v) = v^{-p/2}e^{v/2}[4p(v-1)-2v^2],$$

$$K_3(v) = p(p-4)v^{-p}e^{v}. \tag{5.18}$$

Since $K_3(v)$ is positive (for $p\geq 5$), in establishing (5.17) it is only necessary to prove that the left hand side has no roots in c. Clearly it is only necessary to consider the case where

$$K_2^2(v) - 4K_1(v)\,K_3(v) > 0, \tag{5.19}$$

and then to show that

$$c > [-K_2 + (K_2^2 - 4K_1K_3)^{1/2}]/2K_3. \tag{5.20}$$

Calculation shows that (5.19) is satisfied (for $v>p$) only for $v \in B_1 \cup B_2$, where

$$B_1=(p,p + [\frac{p^2}{2} - \frac{p}{2}\{p^2-16\}^{1/2}]^{1/2}),\ B_2 = (p + [\frac{p^2}{2} + \frac{p}{2}\{p^2-16\}^{1/2}]^{1/2},\infty).$$

Algebra also gives that (5.20) is equivalent to

$$\frac{\varepsilon}{1-\varepsilon} > \frac{p^{-p/2}e^{p/2}v^{p/2}e^{-v/2}}{p(p-4)}\,[(2p-a) + (a^2-p^2a+4p^2)^{1/2}], \tag{5.21}$$

where $a = -v^2+2pv$. For $v \in B_1$, it is straightforward to check that the right hand side of (5.21) is negative, establishing the result in this range.

Consider, finally, the functions

$$\psi(v) = v^{p/2}e^{-v/2}[(2p-a) + (a^2-p^2a+4p^2)^{1/2}],$$

$$H(v) = \log\{v^{p/2}[(2p-a) + (a^2-2pa+4p^2)^{1/2}]\}.$$

It is easy to check that $\psi(v)$ is positive on B_2. Also, it can be shown that $H'(v)$ is a positive continuous function, decreasing from ∞ (at the left end point of B_2) to 0 (at $v=\infty$). Thus

$$\frac{d}{dv}\log\psi(v) = -\frac{1}{2} + H'(v) = 0$$

has a unique solution v_0, and hence

$$\sup_{v \in B_2} \psi(v) = \psi(v_0).$$

(The equation $H'(v) = \frac{1}{2}$ turns out to be a fifth degree polynomial equation, so the computer was used to find the root v_0 in B_2. These roots are also given in Table 3 for the various p.)

Finally, it is clear that (5.21) is satisfied on B_2 if

$$\frac{\varepsilon}{(1-\varepsilon)} > \frac{p^{-p/2}e^{p/2}}{p(p-4)}\ \psi(v_0),$$

or if

$$\varepsilon > 1/\{1 + p(p-4)\, p^{p/2}e^{-p/2}/\psi(v_0)\} \equiv \varepsilon_p.$$

This completes the proof of the theorem.||

Comments.

1. All previous minimax proofs we have seen, that use Stein's technique, depend on the $r^*(v)$ in (5.16) being increasing, so that the derivative of r^* can be ignored. A substantial part of the difficulty of the above proof was due to r not being monotonically increasing.
2. A substantial simplification was achieved through the approach of analyzing (5.17) as a quadratic in c and showing that there can be no roots. As pointed out in Gleser (1983), this can often be a useful technique.
3. The estimator cannot be minimax if $p \leq 4$. This is mainly because, as $v \to \infty$, $r(v) \to p$ (see (5.14)), and not to (p-2) as with more familiar Stein type estimators. An adhoc adjustment of $\hat{\delta}^{\pi}$ could probably be effected to achieve minimaxity for p=3 and 4, but the major point of the theorem was to indicate that the estimator does have reasonable frequentist properties.
4. The Bayesian performance, with respect to π_0, of $\hat{\delta}^{\pi}$ will not be as good as the Bayesian performance of δ^R in section 4. However, δ^R will fare poorly (though never worse than δ^0) when θ is not near μ, while $\hat{\delta}^{\pi}$ will continue to perform well as long as the θ_i are similar (i.e., as long as the exchangeability assumption is valid). The strength of the Type-II maximum likelihood approach, with ε-contamination classes of priors, is that a number of different types of prior information can be built into Γ, and the data will shift $\hat{\delta}^{\pi}$ towards the posterior mean corresponding to the most plausible prior input (in light of the data). Further discussion and examples can be found in Berger and Berliner (1983).

References

Berger, J., 1976, Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss, Annals of Statistics 4, 223-226.

Berger, J., 1979, Multivariate estimation with nonsymmetric loss functions, in: J.S. Rustagi, ed., Optimizing methods in statistics (Academic Press, New York) 5-26.

Berger, J., 1980a, Statistical decision theory: foundations, concepts, and methods (Springer-Verlag, New York).

Berger, J., 1980b, A robust generalized Bayes estimator and confidence region for a multivariate normal mean, Annals of Statistics 8, 716-761.

Berger, J., 1982a, Selecting a minimax estimator of a multivariate normal mean, Annals of Statistics 10, 81-92.

Berger, J., 1982b, Bayesian robustness and the Stein effect, Journal of the American Statistical Association 77, 358-368.

Berger, J., 1982c, Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results, in: S.S. Gupta and J. Berger, eds., Statistical decision theory and related topics III (Academic Press, New York).

Berger, J., 1983a, The Stein effect, in: S. Kotz and N. L. Johnson, eds., Encyclopedia of statistical sciences (Wiley, New York).

Berger, J., 1983b, The robust Bayesian viewpoint, in: J. Kadane, ed., Robustness in Bayesian statistics (North Holland, Amsterdam).

Berger, J., 1983c, Discussion of: C. Morris, Parametric empirical Bayes inference: theory and applications, Journal of the American Statistical Association 78, 55-57.

Berger, J. and Berliner, L. M., 1983, Robust Bayes and empirical Bayes analysis with ε-contaminated priors, Technical Report No. 83-35, Statistics Department, Purdue University, West Lafayette.

Berger, J. and Wolpert, R., 1984, The likelihood principle: a review and generalizations (to appear in the Institute of Mathematical Statistics Monograph Series).

Berliner, L. M., 1983, A decision theoretic structure for robust Bayesian analysis with applications to the estimation of a multivariate normal mean, Technical Report, Statistics Department, Ohio State University, Columbus.

Bhattacharya, P. K., 1966, Estimating the mean of a multivariate normal population with general quadratic loss function, Annals of Mathematical Statistics 37, 1819-1824.

Bock, M. E., 1983, Minimax estimators that shift towards a hypersphere for location vectors of spherically symmetric distributions, Journal of Multivariate Analysis.

Box, G. E. P. and Tiao, G. C., 1973, Bayesian inference in statistical analysis (Addison-Wesley, Reading).

Brown, L. D., 1983, Discussion of: J. Berger, The robust Bayesian viewpoint, in: J. Kadane, ed., Robustness in Bayesian statistics (North Holland, Amsterdam).

Chen, S. Y., 1983, Restricted risk Bayes estimation, Technical Report #83-33, Statistics Department, Purdue University, West Lafayette.

Efron, B. and Morris, C., 1981, Limiting the risk of Bayes and empirical Bayes estimators - part I: the Bayes case, Journal of the American Statistical Association 66, 807-815.

Efron, B. and Morris, C., 1972, Limiting the risk of Bayes and empirical Bayes estimators - part 2: the empirical Bayes case, Journal of the American Statistical Association 67, 130-139.

Efron, B. and Morris, C., 1973, Stein's estimation rule and its competitors - an empirical Bayes approach, Journal of the American Statistical Association 68, 117-130.

Gleser, L. J., 1983, Improving inadmissible estimators under quadratic loss, Technical Report No. 83-19, Department of Statistics, Purdue University, West Lafayette.

Good, I. J., 1965, The estimation of probabilities (M.I.T. Press, Cambridge).

Good, I. J., 1980, Some history of the hierarchical Bayesian methodology, in: J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds., Bayesian statistics (University Press, Valencia).

Hodges, J. L., Jr. and Lehmann, E. L., 1952, The use of previous experience in reaching statistical decisions, Annals of Mathematical Statistics 23, 392-407.

Hwang, J. T. and Casella, G., 1982, Minimax confidence sets for the mean of a multivariate normal distribution, Annals of Statistics 10, 868-881.

James, W. and Stein, C., 1961, Estimation with quadratic loss, in: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability 1 (University of California Press, Berkeley) 361-379.

Judge, G. and Bock, M. E., 1978, Statistical implications of pre-test and Stein rule estimators in econometrics (North Holland, Amsterdam).

Lindley, D. V. and Smith, A. F. M., 1972, Bayes estimates for the linear model, Journal of the Royal Statistical Society B 34, 1-41.

Morris, C., 1983a, Parametric empirical Bayes confidence sets, in: G. E. P. Box, T. Leonard, and C. F. Wu, eds., Scientific inference, data analysis, and robustness (Academic Press, New York).

Morris, C., 1983b, Parametric empirical Bayes inference: theory and applications, Journal of the American Statistical Association 78, 47-65.

Rubin, H., 1977, Robust Bayesian estimators, in: S. S. Gupta and D. S. Moore, eds., Statistical decision theory and related topics II (Academic Press, New York).

Smith, G. and Campbell, F., 1980, A critique of some ridge regression methods, Journal of the American Statistical Association 75, 74-103.

Stein, C., 1956, Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, in: Proceedings of the third Berkeley symposium on mathematical statistics and probability 1 (University of California Press, Berkeley) 197-206.

Stein, C., 1981, Estimation of the mean of a multivariate normal distribution, Annals of Statistics 9, 1135-1151.

Strawderman, W. E., 1971, Proper Bayes minimax estimators of the multivariate normal mean, Annals of Mathematical Statistics 42, 385-388.

Van Der Merwe, A. J., Groenewald, P. C. N., Nel, D. G. and Van Der Merwe, C. A., 1981, Confidence intervals for a multivariate normal mean in the case of empirical Bayes estimation using Pearson curves and normal approximations, Technical Report No. 70, Department of Mathematical Statistics, University of the Orange Free State, Bloemfontein.

Zellner, A. and Vandaele, W., 1971, Bayes-Stein estimators for K-means, regression and simultaneous equation models, in: S. Fienberg and A. Zellner, eds., Studies in Bayesian econometrics and statistics (North Holland, Amsterdam).

Table 1. $RSR(\pi_0, \delta^R)$

p	3	4	5	6	7	8	9	10	15
RSR	.296	.135	.073	.043	.027	.017	.012	.008	.002

Table 2. M, ε, $r(\varepsilon)$, and $RSR(\pi_0, \delta^{p,M})$

p	1				2				3			
ε	.1	.2	.3	.4	.1	.2	.3	.4	.1	.2	.3	.4
$r(\varepsilon)$	.33	.50	.64	.75	.22	.37	.49	.59	.18	.31	.42	.52
M	1.4	.8	.4	.4	.6	.5	.2	.2	.4	.2	.14	.1
RSR	.10	.18	.32	.32	.07	.09	.19	.19	.05	.09	.11	.13

Table 3. ε_p and v_0

p	ε_p	v_0	p	ε_p	v_0
5	.777	11.50	16	.061	33.65
6	.604	13.52	17	.050	35.66
7	.471	15.55	18	.041	37.67
8	.369	17.56	19	.033	39.67
9	.290	19.58	20	.027	41.68
10	.229	21.59	21	.022	43.68
11	.182	23.60	22	.018	45.69
12	.145	25.61	23	.015	47.69
13	.116	27.63	24	.013	49.70
14	.094	29.64	25	.010	51.70
15	.076	31.64	26	.009	53.71

83-36

Journal of Econometrics 25 (1984) 87–108. North-Holland

BAYESIAN INPUT IN STEIN ESTIMATION AND A NEW MINIMAX EMPIRICAL BAYES ESTIMATOR*

J. BERGER

Purdue University, West Lafayette, IN 47907, USA

L.M. BERLINER

Ohio State University, Columbus, OH 43210, USA

The relationship between Stein estimation of a multivariate normal mean and Bayesian analysis is considered. The necessity to involve prior information is discussed, and the various methods of so doing are reviewed. These include direct Bayesian analyses, ad hoc utilization of prior information, restricted class Bayesian and Γ-minimax analyses, and Type II maximum likelihood (empirical Bayes) methods. A new empirical Bayes Stein-type estimator is developed, via the latter method, for an interesting ϵ-contamination class of priors, and is shown to be minimax under reasonable conditions. The minimax proof contains some novel theoretical features.

1. Introduction

1.1. Goals

The major goal of this paper is to discuss the intimate relationship between Stein estimation of a multivariate normal mean and Bayesian analysis. Indeed Stein estimation can not be sensibly carried out without at least informal consideration of Bayesian aspects of the problem (see section 1.3). The paper is organized around the four major methods (see section 1.4) in which prior information can be used in concert with, or related to, Stein estimation. Although the paper will be formally self-contained, references to in-depth studies of these methods will be given. Many of these ideas apply also to non-normal multivariate estimation problems, but for simplicity we consider only the normal case in this paper.

A second goal of the paper is to present some interesting new Stein-type procedures and evaluations of them. Foremost of these new results is the development of a new empirical Bayes estimator, and proof of its minimaxity, in section 5. The novelty of the estimator lies in its being empirical Bayes with respect to a 'sensible' class of prior distributions, while the minimax proof contains several novel theoretical ideas that should be useful elsewhere.

* This research was supported by the National Science Foundation under Grant MCS-8101670A1.

1.2. Notation

A huge variety of multivariate estimation problems can be reduced to the canonical form where $X=(X_1,\ldots,X_p)^t$ is a random vector with a $\mathcal{N}_p(\theta,\Sigma)$ distribution, the mean $\theta=(\theta_1,\ldots,\theta_p)^t$ being unknown and the $(p\times p)$ positive definite covariance matrix Σ being assumed known. (It is typical of this subject that, when Σ is unknown, insertion of suitable estimates of Σ into all expressions leads to qualitatively similar conclusions.) It is desired to estimate θ by an estimator $\delta(x)=(\delta_1(x),\ldots,\delta_p(x))^t$, under the sum of squares error loss

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

[The more general situation of a quadratic loss, $(\theta-\delta)^tQ(\theta-\delta)$, can be reduced to the above case by a linear transformation. Also, losses having different functional dependencies on $(\theta-\delta)$ are known to give qualitatively the same results.] The usual frequentist evaluation of an estimator δ is via its *frequentist risk* function (or mean squared error)

$$R(\theta,\delta)=\mathrm{E}_\theta^X L(\theta,\delta(X))=\mathrm{E}_\theta^X|\theta-\delta(X)|^2.$$

(In this paper E will denote expectation, with superscripts giving the random quantity or distribution over which the expectation is to be taken and subscripts standing for fixed parameters.)

Since Bayesian considerations will be important, we will use $\pi(\theta)$ to denote a *prior density* for θ. Of interest will be the posterior density,

$$\pi(\theta|x)=\pi(\theta)f(x|\theta)/m(x|\pi),$$

where $f(x|\theta)$ is the $\mathcal{N}_p(\theta,\Sigma)$ density and

$$m(x|\pi)=\mathrm{E}^\pi f(x|\theta)=\int_{R^p} f(x|\theta)\pi(\theta)\,\mathrm{d}\theta \tag{1.1}$$

is the marginal or *predictive* distribution of X for the prior π. A posterior Bayesian would be primarily interested in the *posterior expected loss* of $\delta(x)$ (for given data x), namely,

$$\rho(x,\delta)=\mathrm{E}^{\pi(\theta|x)}|\theta-\delta(x)|^2=\int|\theta-\delta(x)|^2\pi(\theta|x)\,\mathrm{d}\theta, \tag{1.2}$$

while a frequentist (or global) Bayesian would be concerned with the *Bayes risk*

$$r(\pi,\delta)=\mathrm{E}^m\rho(X,\delta(X))=\mathrm{E}^\pi R(\theta,\delta), \tag{1.3}$$

which is just the average loss over both θ and X. Via either measure, the optimal estimator is the posterior mean

$$\delta^{\pi}(x) = \mathrm{E}^{\pi(\theta|x)}[\theta].$$

A common Bayesian analysis of this situation is to use a conjugate $\mathcal{N}_p(\mu, A)$ prior, where $\mu = (\mu_1, \ldots, \mu_p)'$ is the prior mean and A is the $(p \times p)$ prior covariance matrix. Denoting this prior π_N, standard calculations [cf., Berger (1980a)] give that the posterior mean is

$$\delta^{\pi_N}(x) = x - \Sigma(\Sigma + A)^{-1}(x - \mu), \tag{1.4}$$

and that

$$\rho(x, \delta^{\pi_N}(x)) = \operatorname{tr}\Sigma - \operatorname{tr}\Sigma(\Sigma + A)^{-1}\Sigma = r(\pi_N, \delta^{\pi_N}), \tag{1.5}$$

$$R(\theta, \delta^{\pi_N}) = \operatorname{tr}[\Sigma^{-1} + 2A^{-1} + A^{-1}\Sigma A^{-1}]^{-1} + |\Sigma(\Sigma + A)^{-1}(\theta - \mu)|^2, \tag{1.6}$$

and

$$m(x|\pi_N) \text{ is } \mathcal{N}_p(\mu, \Sigma + A). \tag{1.7}$$

1.3. The Stein effect and prior information

The classical estimator of θ is, of course, $\delta^0(x) = x$, and it is minimax with constant risk

$$R(\theta, \delta^0) \equiv \operatorname{tr}\Sigma.$$

The Stein effect concerns the fact that, when $p \geq 3$, estimators δ^* can be found that dominate δ^0, in the sense that, for all θ,

$$R(\theta, \delta^*) < R(\theta, \delta^0). \tag{1.8}$$

(Any such δ^* is also clearly minimax.) Study of this phenomenon began with Stein (1956) and James and Stein (1961). General discussions and other references can be found in Judge and Bock (1978) and in Berger (1983a).

The necessity for the involvement of prior information follows from the basic fact that $R(\theta, \delta^*)$ is *substantially* less than $R(\theta, \delta^0)$ only in a relatively small region or subspace of the parameter space. Hence, unless θ is thought to lie in this region, use of δ^* will gain only a negligible amount. (The same can be said for all other approaches to multiparameter estimation, such as ridge

regression.) Clearly, therefore, among the great variety of Stein effect estimators that are available, the one selected for use should be one whose region of significant improvement coincides with where θ is thought likely to be (a priori).

The simplest situation to conceptualize is that in which θ is thought to lie in some ellipse, say

$$C = \{\theta : (\theta - \mu)^t A^{-1} (\theta - \mu) \leq p\}. \tag{1.9}$$

[Here μ could be thought of as the 'best guess' for θ, and A as corresponding to specifications of accuracies (variances) and dependencies (covariances) for these opinions.] It has long been recognized that specification of μ is necessary in Stein estimation (essentially being the point to which one shrinks), but it is shown in Berger (1982a) that the 'shape feature' A is also usually crucial. These appear to be the key needed inputs [see Berger (1980b, 1982a)], and luckily correspond to aspects of prior knowledge which are fairly easy to specify. [For details about how prior information concerning regression parameters can often be reduced to this form, see Berger (1980b, 1982a).]

A different important possibility is that θ may be thought to lie near some subspace of R^p, say the line where all θ_i are equal. Versions of Stein estimators are then appropriate which shrink towards this subspace; substantial risk improvement can often be guaranteed along the entire subspace. Or θ may be thought to lie near the surface of some convex set, in which case the Stein estimator should shrink towards this surface [cf., Bock (1983)]. Or various combinations of the above types of information may be available.

Given a region in which θ is thought to lie, the question arises as to how to select an estimator good for this region. The most natural solution is to find the Stein effect estimator, δ^*, which does best 'on the average' over this region. And the most direct way to implement this is to choose a prior distribution π_0, reflecting where θ is thought likely to lie, and then measure the average performance of δ^* by

$$r(\pi_0, \delta^*) = \int R(\theta, \delta^*) \pi_0(\theta)\, d\theta.$$

In the situation of (1.9), for instance, it would frequently be reasonable to model the prior information as being that π_0 is $\mathcal{N}_p(\mu, A)$. If, on the other hand, it is thought that the θ_i are similar, it would be reasonable to model this in the usual empirical Bayes fashion of assuming that π_0 is $\mathcal{N}_p(\mu, A)$, where $\mu = (1, \ldots, 1)^t \mu_0$, $A = \tau^2 I$, and μ_0 and τ^2 are unknown or partially unknown. Indeed, it will virtually always be most efficient in Stein estimation to begin by attempting a (possibly crude) quantification of a prior π_0. The goal, of course,

will then be to find a good Stein effect estimator which also has good Bayesian performance with respect to π_0.

To non-Bayesians, this approach may be somewhat disturbing. Note, however, that the usual fears about Bayesian analysis are not applicable here; even if the prior π_0 is *completely* inappropriate, the selected δ^* [if it satisfies (1.8)] will still not be worse than δ^0 (though it will be essentially equivalent). One could argue that sometimes *no* prior information is available (although this would be rare in economic situations). The answer then is quite simple: just use δ^0, since no Stein effect estimator would have much of a chance of offering significant improvement.

For the above reasons, even frequentists working in multiparameter estimation should focus substantial attention on prior specification and its use in estimator selection. The 'common' practice of proposing an estimator and comparing it with others via simulations is inappropriate. Different estimators will do well in different regions (or, equivalently, for different priors), so prior information is the only way to differentiate among estimators. A number of Stein-like approaches, such as ridge regression, offer the allure of avoidance of prior input. The allure is a specious one, however; any particular ridge regression estimator will do well only in a particular region of the parameter space. Any suggestion that the estimator will magically shrink towards the correct region is simply erroneous: even in empirical Bayes situations, where the data helps select the direction and amount of shrinkage, prior information concerning the relationships of the θ_i must be specified to direct the shrinkage. Among the many good articles concerning this issue is Smith and Campbell (1980).

Finally, it should be remarked that one can often do amazingly well from *both* the Bayesian and frequentist perspectives, simultaneously. This has long been known in empirical Bayes settings [cf. Efron and Morris (1973)]. Other examples of this will be seen throughout the paper, especially in section 4.

1.4. Using π_0 in Stein estimation

Essentially four methods have been proposed for the needed incorporation of prior information in Stein estimation. The first is to simply derive the Bayes estimator for an appropriate π_0, and then to check its frequentist properties (such as dominance or near dominance of δ^0). This is discussed in section 2.

The second method is the 'ad hoc' method, which consists of taking existing Stein-type estimators and attempting to modify them to incorporate prior information. This is discussed in section 3.

The third method is that of finding the Bayes estimator for π_0 within a restricted class of estimators, a class usually chosen for its good frequentist properties. A very interesting example is the Hodges and Lehmann (1952)

'restricted risk Bayes' method, which is as follows:

$$\text{Among all } \delta^* \text{ for which } R(\theta, \delta^*) \leq R(\theta, \delta^0) + C,$$

$$\text{select that } \delta^* \text{ which minimizes } r(\pi_0, \delta^*). \tag{1.10}$$

When $C = 0$ in (1.10), one is finding the minimax estimator which is best for π_0. This restricted risk approach is closely related to the Γ-minimax problem of considering the class of priors

$$\Gamma = \{ \pi = (1 - \varepsilon)\pi_0 + \varepsilon Q; Q \in \mathcal{Q} \}, \tag{1.11}$$

where ε is a specified constant reflecting uncertainty in π_0, and $\mathcal{Q}$ is a class of 'contaminations', and then choosing δ^* to minimize

$$\sup_{\pi \in \Gamma} r(\pi, \delta^*). \tag{1.12}$$

The idea here is, of course, that, providing π_0, ε, and $\mathcal{Q}$ are chosen reasonably, Γ should contain all 'plausible' prior distributions. If δ^* does well for all π in Γ, it should thus be quite satisfactory for use (at least from a frequentist Bayes viewpoint). These approaches are discussed in section 4, where some quite surprising results are presented.

The final method of approaching the problem is the empirical Bayes or, more precisely, Type II maximum likelihood method [see Good (1980)]. This method is related to the Γ-minimax approach, in that one considers a class of priors such as (1.11), but one then chooses the 'most likely' prior in Γ, according to the data. Since the 'likelihood' of a prior is simply $m(x|\pi)$, this leads to the following definition.

Definition. For a class Γ of priors and given data x, the ML-II (Type II maximum likelihood) prior in Γ, $\hat{\pi}$, is defined (assuming it exists and is unique) as that $\pi \in \Gamma$ maximizing $m(x|\pi)$ over all $\pi \in \Gamma$.

Most empirical Bayes analyses proceed by letting Γ be all conjugate priors, choosing the ML-II $\hat{\pi}$, and then doing a Bayesian analysis pretending that $\hat{\pi}$ is the true prior. Although this smacks of adhocery (the 'prior' $\hat{\pi}$ being chosen in a data dependent fashion), a number of justifications for the approach can be given. [See Berger and Berliner (1983) for discussion and references.] Also, the approach seems to work remarkably well. In section 5 we apply the approach to realistic classes of priors such as (1.11). Indeed for Γ as in (1.11), with π_0 being a conjugate prior and $\mathcal{Q}$ being a reasonable class of contaminations, a quite attractive estimator is developed (via this approach) which is a data

adaptive compromise between δ^{π_0} and a standard empirical Bayes estimator. Furthermore, the estimator is shown to be minimax (for large enough ε) by a minimax proof incorporating some novel features.

2. Bayesian Stein-type estimators

2.1. Introduction

There are several advantages to approaching Stein estimation through Bayesian development of estimators. The first is that one can be certain that the needed prior information is used correctly. The second is that the resulting estimator is often admissible [cf., Strawderman (1971) and Berger (1976)]. Related to this is the fact that Bayesian estimators are guaranteed to be fine conditionally: the original James–Stein (1961) estimator (for $\Sigma = I$),

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

has $R(\theta, \delta^{JS}) < R(\theta, \delta^0)$, but if $p = 3$ and $x = (0.01, -0.01, 0.01)^t$ then $\delta^{JS}(x) \cong (-33, 33, -33)^t$, a ridiculous result. Of course, the positive part version corrects this obvious conditional deficiency, but only Bayesian development of procedures can generally guarantee that the procedures are sensible for each x, and not just on the average [cf., Berger and Wolpert (1984)].

The final and most important practical reason to adopt a Bayesian approach to Stein estimation is that one can easily obtain confidence sets, perform tests, etc. The posterior covariance matrix is usually not much harder to calculate than the posterior mean, and can be used to indicate the variability of the estimates [cf., Box and Tiao (1973), Berger (1980b), Morris (1983a, b) and Van der Merwe et al. (1981)]. Trying to develop estimates of variability from the frequentist perspective has proved enormously difficult, even in the simplest cases [cf., Hwang and Casella (1982)].

The disadvantage with Bayesian development is that the estimators are sometimes expressible only as numerical integrals, and that frequentist risk properties can be hard to verify. Situations where good frequentist risks are known to result are discussed in this section. Note that the conjugate priors, π_N, discussed in section 1.2, do *not* have good frequentist risk properties, in that [see (1.6)]

$$\sup_{\theta} R(\theta, \delta^{\pi_N}) = \infty.$$

2.2. Flat-tailed prior distributions

There is substantial evidence [cf., Rubin (1977) and Berger (1983b)] that estimators developed from priors with flat (polynomial like) tails are Stein-type

and have good risk properties. One example of such a development is Berger (1980b), in which such a prior is used as an alternative to π_N, and results in the Stein-type estimator (posterior mean)

$$\delta^{\mathrm{RB}}(x) = x - \frac{r(\|x-\mu\|^2)}{\|x-\mu\|^2}\Sigma(\Sigma+A)^{-1}(x-\mu),$$

where $\|x-\mu\|^2 = (x-\mu)'(\Sigma+A)^{-1}(x-\mu)$ and $r(z)$ is approximately $\min\{z, p-2\}$. Note the similarity of this estimator with δ^{π_N} in (1.4). Indeed, if μ and A are accurate reflections of the location of θ, then $\|x-\mu\|^2$ will be approximately p [see (1.7)], and (for moderately large p) δ^{RB} will be essentially δ^{π_N}. If, however, θ is far from μ, then $r(\|x-\mu\|^2)/\|x-\mu\|^2$ will be small, and δ^{RB} will be essentially δ^0. This is the general behavior of estimators developed from flat tailed priors, and is the source of their good frequentist properties. [Indeed δ^{RB} is sometimes minimax and always has bounded risk – see Berger (1980b).] Confidence regions for θ, centered about δ^{RB}, are also given in Berger (1980b).

2.3. *Hierarchical priors*

When the prior information consists of structural knowledge about similarities or relationships among the θ_i, this can often be conveniently modeled by hierarchical priors. For instance, the usual empirical Bayes situation, in which the θ_i are felt to be similar, can be modeled by supposing that the θ_i are (independently) $\mathcal{N}(\nu_0, \tau_0^2)$, and then putting a second-stage (perhaps diffuse) prior on ν_0 and τ_0^2. The resulting estimator (posterior mean) will again be Stein-type and have good frequentist risk, providing the second-stage prior on τ_0^2 has flat tails. There is a huge literature on this approach, with a wealth of excellent statistical estimators (and associated confidence sets). References can be found in Lindley and Smith (1972), Good (1980), Morris (1983b), and Berger (1983b).

2.4. *Posterior robustness*

An exciting possibility exists for bypassing the (often difficult) verification of good frequentist risk behavior of a Bayesianly derived Stein-type estimator. The idea is to look at the range of posterior means for all π in a class Γ of plausible priors, such as (1.11). If this range is small, then posterior robustness obtains, and a Bayesian would be completely satisfied with use of $\delta^{\pi_0}(x)$. It seems likely [see Brown (1983)] that the estimate is also then satisfactory from a frequentist viewpoint. The great appeal of this approach lies in the fact that posterior robustness needs to be investigated only for the actual observed x.

(Of course, to a Bayesian, this is not just a convenient technique, but is the only fundamentally sound statistical analysis.) See Berger (1983b) and Berger and Berliner (1983) for discussion of some of these issues.

3. Ad hoc incorporation of prior information

One can simply incorporate the needed prior information into estimators in an 'intuitive' fashion. One example is Berger (1982a), where a minimax version of δ^{RB} was developed, incorporating μ and A. This was done by simply taking an existing class of estimators due to Bhattacharya (1966) and Berger (1979), and incorporating μ and A. The resulting estimator, when Σ and A are diagonal, can be written (coordinatewise) as

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

where $\{\sigma_i^2\}$ and $\{A_i\}$ are the diagonal elements of Σ and A, $q_i = \sigma_i^4/(\sigma_i^2 + A_i)$ (and a relabelling has been done, if necessary, so that $q_1 \geq q_2 \geq \cdots \geq q_p > 0 \equiv q_{p+1}$), and $\|x^j - \mu^j\|^2 = \sum_{l=1}^{j}(x_l - \mu_l)^2/(\sigma_l^2 + A_l)$. This estimator was shown to *always* be minimax and to have good Bayesian properties.

A second example of such an ad hoc approach is that of Bock (1983), in which minimax estimators are developed which shrink towards the surface of convex sets (such as a sphere, the positive orthant, or a wedge). These estimators would be of Bayesian value when the prior information is that θ is likely to be near such a surface. (As an example, one might have vague prior knowledge concerning the length, $|\theta|$, of θ, and hence want to shrink towards the surface of the appropriate sphere.)

The chief usefulness of such ad hoc estimator developments is that they allow (usually by design) proof of minimaxity (or some other desirable property) of the final estimator, while bearing some relation to the desired prior input. Their weaknesses are those that have already been mentioned: they may utilize the prior information in an inferior way; they are often not admissible; and they do not lead to error estimates.

4. Restricted class Bayes and Γ-minimax estimators

A reasonable approach to incorporating prior information into Stein estimation is to restrict analysis to a class of estimators known to have desirable frequentist risk properties, and within this class seek an estimator good with

respect to π_0. Although there are several examples of this in the literature, we will restrict discussion here to the restricted risk Bayes problem posed by (1.10). There is also the closely related Γ-minimax problem discussed in (1.11) and (1.12). These two problems are actually equivalent in a wide variety of situations, as the following lemma shows. For use in this lemma define the 'orthant at θ'' as the set

$$\Lambda(\theta') = \{\theta : \theta_i > \theta_i' \text{ if } \theta_i' > 0 \text{ and } \theta_i < \theta_i' \text{ if } \theta_i' < 0\}.$$

Lemma 1. Suppose that Γ is as in (1.11) and that, for each θ', there exists $Q_{\theta'} \in \mathcal{Q}$ such that $P^{Q_{\theta'}}(\Lambda(\theta')) = 1$. Suppose also that δ is an estimator such that $R(\theta, \delta)$ is non-decreasing in $|\theta_i| > K$ for some K and all i. Then

$$\sup_{\pi \in \Gamma} r(\pi, \delta) = (1-\varepsilon) r(\pi_0, \delta) + \varepsilon \sup_{\theta} R(\theta, \delta). \tag{4.1}$$

Proof. Clearly

$$\sup_{\pi \in \Gamma} r(\pi_0, \delta) = (1-\varepsilon) r(\pi_0, \delta) + \varepsilon \sup_{Q \in \mathcal{Q}} r(Q, \delta).$$

Define $M = \sup_\theta R(\theta, \delta)$, and observe that, for any $\lambda > 0$, one can find a point θ^λ such that $|\theta_i^\lambda| > K$ for all i and $R(\theta^\lambda, \delta) > M - \lambda$. Since $R(\theta, \delta)$ is non-decreasing for $|\theta_i| > K$, it follows that

$$r(Q_{\theta^\lambda}, \delta) = \int_{\Lambda(\theta^\lambda)} R(\theta, \delta)\, dQ_{\theta^\lambda}(\theta) > M - \lambda.$$

But λ is arbitrary, so that

$$\sup_{Q \in \mathcal{Q}} r(Q, \delta) \geq M.$$

On the other hand,

$$r(Q, \delta) = \int R(\theta, \delta)\, dQ(\theta) \leq \int M\, dQ(\theta) = M,$$

establishing the conclusion.

Virtually all classes of contaminations $\mathcal{Q}$ that are considered [cf., Berger and Berliner (1983)] satisfy the mild condition of the lemma, and furthermore, it

can frequently be shown (for, e.g., unimodal π_0) that any δ minimizing $\sup_{\pi\in\Gamma} r(\pi_0, \delta)$ must also satisfy the condition of the lemma. But it follows from a standard game-theoretic argument that any δ minimizing the right-hand side of (4.1) must also be the solution to (1.10), where ε and C are related in a monotonic fashion.

As a specific example, consider the situation where $\Sigma = \sigma^2 I$ and π_0 is $\mathcal{N}_p(\mu, \tau^2 I)$. In Berger (1982b, c) it is shown, when $p \geq 3$ and $C = 0$ in (1.10) (or equivalently $\varepsilon = 1$ in the Γ-minimax case), that the approximate optimal restricted risk Bayes rule is the Stein-type estimator

$$\delta^R(x) = \delta^{\pi_0}(x) = x - \frac{\sigma^2}{(\sigma^2+\tau^2)}(x-\mu) \quad \text{if} \quad |x-\mu|^2 < 2(p-2)(\sigma^2+\tau^2),$$

$$= x - \frac{2(p-2)\sigma^2}{|x-\mu|^2}(x-\mu) \quad \text{if} \quad |x-\mu|^2 > 2(p-2)(\sigma^2+\tau^2).$$

Since $C = 0$, δ^R is minimax, and hence *always* better than δ^0 in terms of frequentist risk. It is convenient to measure the Bayesian performance of δ^R by the 'relative savings risk' [see Efron and Morris (1972)],

$$RSR(\pi_0, \delta) = \frac{r(\pi_0, \delta) - r(\pi_0, \delta^{\pi_0})}{r(\pi_0, \delta^0) - r(\pi_0, \delta^{\pi_0})}.$$

When RSR is near zero, δ is essentially as good as the optimal Bayes rule with respect to π_0 (namely, δ^{π_0}), while, when RSR is near one, δ^{π_0} is no better than the standard estimator δ^0. Table 1 gives $RSR(\pi_0, \delta^R)$ for various values of p. These values are startling, in that δ^R has essentially optimal Bayesian performance (at least for $p \geq 5$) while maintaining minimaxity. That one could have such optimal frequentist and Bayesian performance simultaneously is astonishing.

Results in this situation for $C > 0$ (or $\varepsilon < 1$) are also given in Berger (1982b, c) when $p \geq 2$, and when $p = 1$ in Efron and Morris (1971). The (approximate) solutions depend in general on modified Bessel functions. For

Table 1
$RSR(\pi_0, \delta^R)$.

p	3	4	5	6	7	8	9	10	15
RSR	0.296	0.135	0.073	0.043	0.027	0.017	0.012	0.008	0.002

$p = 1$ and $p = 3$, however, they have the fairly simple form (respectively)

$$\delta^{1,M} = x - \frac{\sigma^2}{(\sigma^2+\tau^2)}(x-\mu) \quad \text{if} \quad (x-\mu)^2 < M(\sigma^2+\tau^2),$$

$$= x - (\operatorname{sgn}[x-\mu])\left[M\sigma^4/(\sigma^2+\tau^2)\right]^{\frac{1}{2}} \quad \text{otherwise},$$

$$\delta^{3,M} = x - \frac{\sigma^2}{(\sigma^2+\tau^2)}(x-\mu) \quad \text{if} \quad |x-\mu|^2 \le (\sigma^2+\tau^2)d,$$

$$= x - \left[\frac{2\sigma^2}{|x-\mu|^2} + \frac{\sqrt{3M}\,\sigma^2}{\sqrt{\sigma^2+\tau^2}\,|x-\mu|}\right](x-\mu) \quad \text{otherwise},$$

where

$$d = \tfrac{3}{2}M + \tfrac{4}{3} + M\sqrt{1+8/(3M)}\,,$$

and

$$M = \frac{C}{r(\pi_0,\delta^0) - r(\pi_0,\delta^{\pi_0})} = \frac{C(\sigma^2+\tau^2)}{p\sigma^4}.$$

Here M indicates the amount that $\delta^{p,M}$ is worse than a minimax rule (for which $M = C = 0$) in terms of $\sup_\theta R(\theta,\delta)$, normalized to be on the same scale as $RSR(\pi_0,\delta)$. Table 2 presents M, $RSR(\pi_0,\delta^{p,M})$, and also values [see (4.1)] of

$$r(\varepsilon) = \frac{\sup_{\pi\in\Gamma} r(\pi,\delta^{p,M}) - r(\pi_0,\delta^{\pi_0})}{r(\pi_0,\delta^0) - r(\pi_0,\delta^{\pi_0})}, \tag{4.2}$$

i.e., the suitably normalized Γ-minimax risk of $\delta^{p,M}$. It can be shown that

$$r(\varepsilon) = (1-\varepsilon)RSR(\pi_0,\delta^{p,M}) + \varepsilon(1+M). \tag{4.3}$$

Table 2

M, ε, $r(\varepsilon)$, and $RSR(\pi_0,\delta^{p,M})$.

p	1				2				3			
ε	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
$r(\varepsilon)$	0.33	0.50	0.64	0.75	0.22	0.37	0.49	0.59	0.18	0.31	0.42	0.52
M	1.4	0.8	0.4	0.4	0.6	0.5	0.2	0.2	0.4	0.2	0.14	0.1
RSR	0.10	0.18	0.32	0.32	0.07	0.09	0.19	0.19	0.05	0.09	0.11	0.13

Actually, it was more convenient to just give the results for $\varepsilon = 0.1$, 0.2, 0.3, and 0.4, with M being determined as the corresponding value giving the optimal (approximate) Γ-minimax rule (see the discussion after Lemma 1). The results are also given for $p = 2$. [See Berger (1982b or 1982c) for the appropriate estimator in this case.]

The tradeoff between increased minimax risk (M) and Bayesian performance (RSR) is well illustrated by table 2. When $p = 1$, for instance, one can have reasonable Bayesian performance, sacrificing only 32% of the possible Bayesian gains, by being willing to accept an increase of 40% in the minimax risk. For $p = 3$ the situation is very pleasant; by allowing an increase of 10% in minimax risk one surrenders only 13% of the possible Bayesian gains. The data for $p = 2$ are included because they show that the 'Stein effect' is operating even in two dimensions: the values of M and RSR are much better than when $p = 1$.

The Γ-minimax data from table 2 are also worth perusing. For instance, in $p = 3$, if one elicits μ and τ^2 and feels that a normal shape for the prior is reasonable, but feels that the prior specification could be off (in terms of misspecified prior probabilities) by, say 20% (i.e., $\varepsilon = 0.2$), then using $\delta^{3,0.2}$ would guarantee a Bayes risk no worse than 31% from 'optimal' by the standardized measure. [Using (4.2) and (4.3) it is easy to convert this into actual Γ-minimax risk if desired.] Thus, for *any* prior deemed reasonable, $\delta^{3,0.2}$ would have excellent overall risk, which should be satisfactory even to frequentists.

It should be mentioned that, for conditional Bayesians, the estimators discussed in this section are very sensible, being simply the conjugate prior Bayes estimator when x is near μ (and so compatible with the prior), while being similar to Bayes estimators with flat tails otherwise. Also, although the problem becomes much more difficult when Σ and A are not multiples of the identity, good restricted risk Bayes procedures and Γ-minimax procedures have been developed for such situations in Chen (1983).

5. Type II maximum likelihood and empirical Bayes estimators

5.1. *The assumed prior structure*

Consider the 'empirical Bayes' situation in which the θ_i are thought to be independent realizations from a common $\mathcal{N}(\nu, \tau^2)$ prior density, but ν and τ^2 are partially unknown. Model this uncertainty in a *robust* hierarchical Bayesian fashion, i.e., assume that ν and τ^2 have a prior density

$$h(\nu, \tau^2) = (1 - \varepsilon)h_0(\nu, \tau^2) + \varepsilon Q(\nu, \tau^2),$$

where h_0 is an elicited prior, ε is the possible error in elicitation, and Q is a

member of a class $\mathcal{Q}$ of possible contaminations. The overall prior for θ is then

$$\pi(\theta) = \int \left[\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\tau} \exp\left\{ -(\theta_i - \nu)^2 / 2\tau^2 \right\} \right] h(\nu, \tau^2)\, d\nu\, d\tau^2 \\ = (1-\varepsilon)\pi_0(\theta) + \varepsilon Q^*(\theta), \tag{5.1}$$

where

$$\pi_0(\theta) = \int \left[\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\tau} \exp\left\{ -(\theta_i - \nu)^2 / 2\tau^2 \right\} \right] h_0(\nu, \tau^2)\, d\nu\, d\tau^2,$$

and

$$Q^*(\theta) = \int \left[\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\tau} \exp\left\{ -(\theta_i - \nu)^2 / 2\tau^2 \right\} \right] Q(\nu, \tau^2)\, d\nu\, d\tau^2. \tag{5.2}$$

We are thus faced with the class of possible priors

$$\Gamma = \left\{ \pi = (1-\varepsilon)\pi_0 + \varepsilon Q^*;\ Q \in \mathcal{Q} \right\}. \tag{5.3}$$

Notice that this is considerably more refined than the usual empirical Bayes setup, which effectively assumes that $\varepsilon = 1$ and $\mathcal{Q} = \{\text{all unit point masses}\}$, so that Γ is just the set of all $\mathcal{N}_p(\nu \boldsymbol{1}, \tau^2 I)$ priors, where $\boldsymbol{1} = (1, 1, \ldots, 1)^t$. Usually prior knowledge about ν and τ^2 *is* available and, as briefly discussed in Berger (1982b, 1983c), can provide valuable gains (unless p is quite large).

5.2. *The ML-II prior and posterior mean*

An ad hoc, but intuitively reasonable method of dealing with Γ is to choose the 'most likely' prior in Γ. Recall, from section 1.4, that this is called the ML-II prior, $\hat{\pi}$, and is that prior in Γ (if it exists) maximizing $m(x|\pi)$ for the given data x. (Recall that $m(x|\pi)$ [see (1.1)] is the predictive density of x given π, so that $\hat{\pi}$ is the 'maximum likelihood' prior in the usual sense. Good (1965) calls this 'Type II maximum likelihood', and we will stick with his nomenclature.)

This technique is extensively discussed in Berger and Berliner (1983), to which the reader is referred for further justification. Note, at least, that this would correspond to the usual empirical Bayes technique [when Γ consists of all $\mathcal{N}_p(\nu \boldsymbol{1}, \tau^2 I)$ distributions] of estimating ν and τ^2 via the maximum likelihood method from the predictive density

$$m(x|\nu, \tau^2) = \mathcal{N}_p(\nu \boldsymbol{1}, \Sigma + \tau^2 I). \tag{5.4}$$

When π is as in (5.1), it is clear that

$$m(x|\pi) = (1-\varepsilon)m(x|\pi_0) + \varepsilon m(x|Q^*). \tag{5.5}$$

Hence

$$\sup_{\pi\in\Gamma} m(x|\pi) = (1-\varepsilon)m(x|\pi_0) + \varepsilon \sup_{Q\in\mathcal{Q}} m(x|Q^*). \tag{5.6}$$

Furthermore, if $\hat{Q}^*$ maximizes this last term, then $\hat{\pi} = (1-\varepsilon)\pi_0 + \varepsilon\hat{Q}^*$, and the posterior mean with respect to $\hat{\pi}$ is

$$\begin{aligned}\delta^{\hat{\pi}}(x) &= \int \frac{\theta f(x|\theta)\hat{\pi}(\theta)\,d\theta}{m(x|\hat{\pi})} \\ &= \hat{\lambda}(x)\delta^{\pi_0}(x) + (1-\hat{\lambda}(x))\delta^{\hat{Q}}(x),\end{aligned} \tag{5.7}$$

where

$$\hat{\lambda}(x) = (1-\varepsilon)m(x|\pi_0)/\left[(1-\varepsilon)m(x|\pi_0) + \varepsilon m(x|\hat{Q})\right]. \tag{5.8}$$

Note that, when the data x 'agrees with' π_0, $m(x|\pi_0)$ will be reasonably large and $\hat{\lambda}(x)$ will be close to one. If, on the other hand, x gives considerably more support to $\hat{Q}$, then $\hat{\lambda}(x)$ will be close to zero. This adaptive behavior of $\delta^{\hat{\pi}}$ is what makes it so attractive.

5.3. *A special case*

The simplest case to deal with is that in which $\Sigma = \sigma^2 I$, $h_0(\nu, \tau^2)$ is a point mass at (ν_0, τ_0^2) and

$$\mathcal{Q} = \left\{\text{all distributions concentrated on } \tau^2 \geq \tau_0^2\right\}.$$

The idea here is that (ν_0, τ_0^2) is simply the best guess as to the 'hyperparameters', ε is a measure of the strength of belief in this guess, and there is enough uncertainty to want to allow all contaminations in $\mathcal{Q}$, subject to the constraint that $\tau^2 \geq \tau_0^2$. (The reasons for this constraint are partly technical, so that the analysis goes smoothly, and partly to prevent 'spurious' precision from creeping in.) It should be mentioned that, while choice of such a large $\mathcal{Q}$ seems to work well (in a conservative sense) in estimation, other uses of the resulting $\hat{\pi}$ (such as for confidence sets) are suspect [see Berger and Berliner (1983)], although the problems of so using $\hat{\pi}$ here are not too severe.

For this situation,

$$\pi_0 = \mathcal{N}_p\left(\nu_0 \mathbf{1}, \tau_0^2 I\right),$$

$$m(x|\pi_0) = \mathcal{N}_p\left(\nu_0 \mathbf{1}, \left(\sigma^2 + \tau_0^2\right) I\right),$$

$$\delta^{\pi_0}(x) = x - \frac{\sigma^2}{\left(\sigma^2 + \tau_0^2\right)}(x - \nu_0 \mathbf{1}), \tag{5.9}$$

and [see (5.2) and (5.4)]

$$m(x|Q^*) = \int m(x|\nu, \tau^2) Q(\nu, \tau^2)\, \mathrm{d}\nu\, \mathrm{d}\tau^2.$$

Clearly $m(x|Q^*)$ is maximized over $Q \in \mathcal{Q}$ by choosing Q to be a point mass at the maximum likelihood estimate of (ν, τ^2) (subject to the constraint $\tau^2 \geq \tau_0^2$, of course). But, from (5.4), it is clear that the maximum likelihood estimate is simply

$$\hat{\nu} = \bar{x}, \qquad \hat{\tau}^2 = \max\left\{\tau_0^2, |s|^2/p - \sigma^2\right\}, \tag{5.10}$$

where

$$s^2 = \sum_{i=1}^{p} (x_i - \bar{x})^2.$$

Thus

$$m(x|\hat{Q}) = \mathcal{N}_p\left(\hat{\nu}\mathbf{1}, \left(\sigma^2 + \hat{\tau}^2\right) I\right),$$

and

$$\delta^{\hat{Q}}(x) = x - \frac{\sigma^2}{\max\left\{\tau_0^2 + \sigma^2, |s|^2/p\right\}}(x - \bar{x}\mathbf{1}), \tag{5.11}$$

which is more or less the 'standard' empirical Bayes estimate of θ in the exchangeable case. Also, $\hat{\lambda}(x)$ can be written (after some algebra and letting max denote $\max\{\tau_0^2 + \sigma^2, |s|^2/p\}$) as

$$\hat{\lambda}(x) = \left[1 + \frac{\varepsilon}{(1-\varepsilon)} \cdot \left(\frac{\sigma^2 + \tau_0^2}{\max}\right)^{p/2} \times \exp\left\{\frac{s^2}{2}\left[\frac{1}{\left(\sigma^2 + \tau_0^2\right)} - \frac{1}{\max}\right] + \frac{p(\bar{x} - \nu_0)^2}{2\left(\sigma^2 + \tau_0^2\right)}\right\}\right]^{-1}. \tag{5.12}$$

Using this with (5.9) and (5.11) in (5.7) gives the ML-II posterior mean, which will be a very appealing data adaptive compromise between the Bayes estimator for the specified (ν_0, τ_0^2) and the empirical Bayes estimator which assumes that these hyperparameters are unknown. More discussion of this approach, along with examples of more realistic or richer structures that can be assumed, is given in Berger and Berliner (1983).

5.4. *Minimaxity of* $\delta^{\hat{\pi}}$

The intuitive rationale and justification for $\delta^{\hat{\pi}}$, together with the fact that the estimator behaves similarly to more familiar Stein-type estimators in the limits, lends support to the feeling that the estimator will have good frequentist properties (as well as good Bayesian properties). It is of interest, however, to attempt direct verification of this, by attempting to establish minimaxity of $\delta^{\hat{\pi}}$. Unfortunately, we were unsuccessful in this attempt, due to technical difficulties arising from the introduction of $\hat{\nu}$. Workers in this area are, however, familiar with the fact that such estimation of the supposed common mean rarely affects minimaxity by more than a needed slight alteration of constants, so we considered the simpler problem of known ν. Thus suppose that $X \sim \mathcal{N}_p(\theta, \sigma^2 I)$, $\pi_0 = \mathcal{N}_p(\nu_0, \tau_0^2 I)$, and

$$\mathcal{Q} = \left\{\text{all distributions concentrated on } \left\{\nu = \nu_0, \tau^2 \geq \tau_0^2\right\}\right\}.$$

An analysis identical to that of the previous section gives, as the ML-II posterior mean,

$$\begin{aligned} \delta^{\hat{\pi}} &= x - \frac{\sigma^2}{\left(\sigma^2 + \tau_0^2\right)}(x - \nu_0 \mathbf{1}) \quad \text{if} \quad \sum_{i=1}^{p} (x_i - \nu_0)^2 < p\left(\tau_0^2 + \sigma^2\right), \\ &= x - \frac{\sigma^2 r(v)}{v}(x - \nu_0 \mathbf{1}) \qquad \text{otherwise,} \end{aligned} \tag{5.13}$$

where

$$\begin{aligned} v &= \sum_{i=1}^{p} (x_i - \nu_0)^2, \\ r(v) &= p\left[1 + \left(\frac{v}{p\left(\sigma^2 + \tau_0^2\right)} - 1\right)\hat{\lambda}(v)\right], \\ \hat{\lambda}(v) &= \left\{1 + \frac{\varepsilon}{(1-\varepsilon)} \cdot \left[\frac{\left(\sigma^2 + \tau_0^2\right)p}{v}\right]^{p/2} e^{-p/2} e^{v/2\left(\sigma^2 + \tau_0^2\right)}\right\}^{-1}. \end{aligned} \tag{5.14}$$

Theorem. *For $p \geq 5$, $\delta^{\hat{\pi}}$ in (5.13) is minimax [i.e., satisfies (1.8)] providing $\varepsilon \geq \varepsilon_p$, ε_p being given in table 3 for $5 \leq p \leq 26$. (For $p > 26$, ε_p is less than 0.009, so $\delta^{\hat{\pi}}$ will be minimax for any sensible ε.)*

Proof. By making the appropriate linear transformation, it is sufficient to prove the theorem for $\nu_0 = 0$ and $\sigma^2 + \tau_0^2 = 1$, which we henceforth assume. Using the familiar unbiased estimate of risk of Stein (1981), to show that an estimator,

$$\delta^*(x) = \left(1 - \sigma^2 r^*(v)/v\right)x, \tag{5.15}$$

is minimax [where $X \sim \mathcal{N}_p(\theta, \sigma^2 I)$ and $v = |x|^2$], it suffices to show that

$$4\frac{\mathrm{d}}{\mathrm{d}v}r^*(v) + r^*(v)v^{-1}\left[2(p-2) - r^*(v)\right] \geq 0. \tag{5.16}$$

For $v < p$, $\delta^{\hat{\pi}}$ is of the form (5.15) with $r^*(v) = v$ (recall $\sigma^2 + \tau_0^2 = 1$, $\nu_0 = 0$), for which verification of (5.16) is trivial. For $v > p$, we must verify (5.16) for $r^* = r$ as in (5.14). Substituting (5.14) into (5.16), and after some algebra, the problem reduces to showing that (for $v > p$)

$$K_1(v) + K_2(v)c + K_3(v)c^2 \geq 0, \tag{5.17}$$

where

$$\begin{aligned} c &= \varepsilon(1-\varepsilon)^{-1}p^{p/2}\mathrm{e}^{-p/2}, \\ K_1(v) &= 2vp - v^2, \\ K_2(v) &= v^{-p/2}\mathrm{e}^{v/2}\left[4p(v-1) - 2v^2\right], \\ K_3(v) &= p(p-4)v^{-p}\mathrm{e}^{v}. \end{aligned} \tag{5.18}$$

Table 3

ε_p and v_0.

p	ε_p	v_0	p	ε_p	v_0
5	0.777	11.50	16	0.061	33.65
6	0.604	13.52	17	0.050	35.66
7	0.471	15.55	18	0.041	37.67
8	0.369	17.56	19	0.033	39.67
9	0.290	19.58	20	0.027	41.68
10	0.229	21.59	21	0.022	43.68
11	0.182	23.60	22	0.018	45.69
12	0.145	25.61	23	0.015	47.69
13	0.116	27.63	24	0.013	49.70
14	0.094	29.64	25	0.010	51.70
15	0.076	31.64	26	0.009	53.71

Since $K_3(v)$ is positive (for $p \geq 5$), in establishing (5.17) it is only necessary to prove that the left-hand side has no roots in c. Clearly it is only necessary to consider the case where

$$K_2^2(v) - 4K_1(v)K_3(v) > 0, \tag{5.19}$$

and then to show that

$$c > \left[-K_2 + \left(K_2^2 - 4K_1K_3\right)^{\frac{1}{2}}\right] \Big/ 2K_3. \tag{5.20}$$

Calculation shows that (5.19) is satisfied (for $v > p$) only for $v \in B_1 \cup B_2$, where

$$B_1 = \left(p, p + \left[\frac{p^2}{2} - \frac{p}{2}\left\{p^2 - 16\right\}^{\frac{1}{2}}\right]^{\frac{1}{2}}\right),$$

$$B_2 = \left(p + \left[\frac{p^2}{2} + \frac{p}{2}\left\{p^2 - 16\right\}^{\frac{1}{2}}\right]^{\frac{1}{2}}, \infty\right).$$

Algebra also gives that (5.20) is equivalent to

$$\frac{\varepsilon}{1-\varepsilon} > \frac{p^{-p/2}e^{p/2}v^{p/2}e^{-v/2}}{p(p-4)}\left[(2p-a) + \left(a^2 - p^2a + 4p^2\right)^{\frac{1}{2}}\right], \tag{5.21}$$

where $a = -v^2 + 2pv$. For $v \in B_1$, it is straightforward to check that the right-hand side of (5.21) is negative, establishing the result in this range.

Consider, finally, the functions

$$\psi(v) = v^{p/2}e^{-v/2}\left[(2p-a) + \left(a^2 - p^2a + 4p^2\right)^{\frac{1}{2}}\right],$$

$$H(v) = \log\left\{v^{p/2}\left[(2p-a) + \left(a^2 - 2pa + 4p^2\right)^{\frac{1}{2}}\right]\right\}.$$

It is easy to check that $\psi(v)$ is positive on B_2. Also, it can be shown that $H'(v)$ is a positive continuous function, decreasing from ∞ (at the left end point of B_2) to 0 (at $v = \infty$). Thus

$$\frac{d}{dv}\log\psi(v) = -\frac{1}{2} + H'(v) = 0$$

has a unique solution v_0, and hence

$$\sup_{v \in B_2} \psi(v) = \psi(v_0).$$

[The equation $H'(v) = \frac{1}{2}$ turns out to be a fifth-degree polynomial equation, so the computer was used to find the root v_0 in B_2. These roots are also given in table 3 for the various p.]

Finally, it is clear that (5.21) is satisfied on B_2 if

$$\frac{\varepsilon}{(1-\varepsilon)} > \frac{p^{-p/2}e^{p/2}}{p(p-4)}\psi(v_0),$$

or if

$$\varepsilon > 1/\{1 + p(p-4)p^{p/2}e^{-p/2}/\psi(v_0)\} \equiv \varepsilon_p.$$

This completes the proof of the theorem.

Comments

1. All previous minimax proofs we have seen, that use Stein's technique, depend on the $r^*(v)$ in (5.16) being increasing, so that the derivative of r^* can be ignored. A substantial part of the difficulty of the above proof was due to r *not* being monotonically increasing.
2. A substantial simplification was achieved through the approach of analyzing (5.17) as a quadratic in c and showing that there can be no roots. As pointed out in Gleser (1983), this can often be a useful technique.
3. The estimator cannot be minimax if $p \leq 4$. This is mainly because, as $v \to \infty$, $r(v) \to p$ [see (5.14)], and not to $(p-2)$ as with more familiar Stein-type estimators. An ad hoc adjustment of $\delta^{\hat{\pi}}$ could probably be effected to achieve minimaxity for $p = 3$ and 4, but the major point of the theorem was to indicate that the estimator does have reasonable frequentist properties.
4. The Bayesian performance, with respect to π_0, of $\delta^{\hat{\pi}}$ will not be as good as the Bayesian performance of δ^R in section 4. However, δ^R will fare poorly (though never worse than δ^0) when θ is not near μ, while $\delta^{\hat{\pi}}$ will continue to perform well as long as the θ_i are similar (i.e., as long as the exchangeability assumption is valid). The strength of the Type II maximum likelihood approach, with ε-contamination classes of priors, is that a number of different types of prior information can be built into Γ, and the data will shift $\delta^{\hat{\pi}}$ towards the posterior mean corresponding to the most plausible prior input (in light of the data). Further discussion and examples can be found in Berger and Berliner (1983).

References

Berger, J., 1976, Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss, Annals of Statistics 4, 223–226.

Berger, J., 1979, Multivariate estimation with nonsymmetric loss functions, in: J.S. Rustagi, ed., Optimizing methods in statistics (Academic Press, New York) 5–26.

Berger, J., 1980a, Statistical decision theory: Foundations, concepts, and methods (Springer-Verlag, New York).
Berger, J., 1980b, A robust generalized Bayes estimator and confidence region for a multivariate normal mean, Annals of Statistics 8, 716–761.
Berger, J., 1982a, Selecting a minimax estimator of a multivariate normal mean, Annals of Statistics 10, 81–92.
Berger, J., 1982b, Bayesian robustness and the Stein effect, Journal of the American Statistical Association 77, 358–368.
Berger, J., 1982c, Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results, in: S.S. Gupta and J. Berger, eds., Statistical decision theory and related topics III (Academic Press, New York).
Berger, J., 1983a, The Stein effect, in: S. Kotz and N.L. Johnson, eds., Encyclopedia of statistical sciences (Wiley, New York).
Berger, J., 1983b, The robust Bayesian viewpoint, in: J. Kadane, ed., Robustness in Bayesian statistics (North-Holland, Amsterdam).
Berger, J., 1983c, Discussion of: C. Morris, Parametric empirical Bayes inference: Theory and applications, Journal of the American Statistical Association 78, 55–57.
Berger, J. and L.M. Berliner, 1983, Robust Bayes and empirical Bayes analysis with ε-contaminated priors, Technical report no. 83-35 (Statistics Department, Purdue University, West Lafayette, IN).
Berger, J. and R. Wolpert, 1984, The likelihood principle: A review and generalizations (to appear in the Institute of Mathematical Statistics Monograph Series).
Berliner, L.M., 1983, A decision theoretic structure for robust Bayesian analysis with applications to the estimation of a multivariate normal mean, Technical report (Statistics Department, Ohio State University, Columbus, OH).
Bhattacharya, P.K., 1966, Estimating the mean of a multivariate normal population with general quadratic loss function, Annals of Mathematical Statistics 37, 1819–1824.
Bock, M.E., 1983, Minimax estimators that shift towards a hypersphere for location vectors of spherically symmetric distributions, Journal of Multivariate Analysis.
Box, G.E.P. and G.C. Tiao, 1973, Bayesian inference in statistical analysis (Addison-Wesley, Reading, MA).
Brown, L.D., 1983, Discussion of: J. Berger, The robust Bayesian viewpoint, in: J. Kadane, ed., Robustness in Bayesian statistics (North-Holland, Amsterdam).
Chen, S.Y., 1983, Restricted risk Bayes estimation, Technical report no. 83-33 (Statistics Department, Purdue University, West Lafayette, IN).
Efron, B. and C. Morris, 1971, Limiting the risk of Bayes and empirical Bayes estimators – Part 1: The Bayes case, Journal of the American Statistical Association 66, 807–815.
Efron, B. and C. Morris, 1972, Limiting the risk of Bayes and empirical Bayes estimators – Part 2: The empirical Bayes case, Journal of the American Statistical Association 67, 130–139.
Efron, B. and C. Morris, 1973, Stein's estimation rule and its competitors – An empirical Bayes approach, Journal of the American Statistical Association 68, 117–130.
Gleser, L.J., 1983, Improving inadmissible estimators under quadratic loss, Technical report no. 83-19 (Department of Statistics, Purdue University, West Lafayette, IN).
Good, I.J., 1965, The estimation of probabilities (M.I.T. Press, Cambridge, MA).
Good, I.J., 1980, Some history of the hierarchical Bayesian methodology, in: J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds., Bayesian statistics (University Press, Valencia).
Hodges, J.L., Jr. and E.L. Lehmann, 1952, The use of previous experience in reaching statistical decisions, Annals of Mathematical Statistics 23, 392–407.
Hwang, J.T. and G. Casella, 1982, Minimax confidence sets for the mean of a multivariate normal distribution, Annals of Statistics 10, 868–881.
James, W. and C. Stein, 1961, Estimation with quadratic loss, in: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, Vol. 1 (University of California Press, Berkeley, CA) 361–379.
Judge, G. and M.E. Bock, 1978, Statistical implications of pre-test and Stein rule estimators in econometrics (North-Holland, Amsterdam).
Lindley, D.V. and A.F.M. Smith, 1972, Bayes estimates for the linear model, Journal of the Royal Statistical Society B 34, 1–41.

Morris, C., 1983a, Parametric empirical Bayes confidence sets, in: G.E.P. Box, T. Leonard and C.F. Wu, eds., Scientific inference, data analysis, and robustness (Academic Press, New York).
Morris, C., 1983b, Parametric empirical Bayes inference: Theory and applications, Journal of the American Statistical Association 78, 47–65.
Rubin, H., 1977, Robust Bayesian estimators, in: S.S. Gupta and D.S. Moore, eds., Statistical decision theory and related topics II (Academic Press, New York).
Smith, G. and F. Campbell, 1980, A critique of some ridge regression methods, Journal of the American Statistical Association 75, 74–103.
Stein, C., 1956, Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, in: Proceedings of the third Berkeley symposium on mathematical statistics and probability, Vol. 1 (University of California Press, Berkeley, CA) 197–206.
Stein, C., 1981, Estimation of the mean of a multivariate normal distribution, Annals of Statistics 9, 1135–1151.
Strawderman, W.E., 1971, Proper Bayes minimax estimators of the multivariate normal mean, Annals of Mathematical Statistics 42, 385–388.
Van Der Merwe, A.J., P.C.N. Groenewald, D.G. Nel and C.A. Van Der Merwe, 1981, Confidence intervals for a multivariate normal mean in the case of empirical Bayes estimation using Pearson curves and normal approximations, Technical report no. 70 (Department of Mathematical Statistics, University of the Orange Free State, Bloemfontein).
Zellner, A. and W. Vandaele, 1971, Bayes–Stein estimators for K-means, regression and simultaneous equation models, in: S. Fienberg and A. Zellner, eds., Studies in Bayesian econometrics and statistics (North-Holland, Amsterdam).