81-1

BAYESIAN ROBUSTNESS AND THE STEIN EFFECT[1]

by

James Berger
Purdue University

Mimeograph Series #81-1

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #81-1

[1]Research supported by the Alfred P. Sloan Foundation and by the National Science Foundation under Grant #MCS 78-02300.

Abstract

In simultaneous estimation of normal means, it is shown that, through use of the Stein effect, surprisingly large gains of a Bayesian nature can be achieved, at little or no cost if the prior information is misspecified. This provides a justification, in terms of robustness with respect to mis-specification of the prior, for employing the Stein effect, even when combining apriori independent problems (i.e., problems in which no empirical Bayes effects are obtainable). To study this issue, a class of minimax estimators which closely mimic the conjugate prior Bayes estimators is introduced.

Key Words and Phrases: Bayesian robustness, Stein effect, simultaneous estimation, risk, Bayes risk.

1. Introduction

Suppose that it is desired to simultaneously estimate $\theta_1,\ldots,\theta_p$ on the basis of independent observations $X_i \sim \mathcal{N}(\theta_i,\sigma_i^2)$, where for simplicity we assume the σ_i^2 are known. The loss in estimation will be assumed to be the sum of squared errors, i.e., in estimating $\theta = (\theta_1,\ldots,\theta_p)^t$ by $\delta(x) = (\delta_1(x),\ldots,\delta_p(x))^t$, $L(\theta,\delta) = |\theta-\delta|^2 = \sum_{i=1}^{p}(\theta_i-\delta_i)^2$. (If $L(\theta,\delta)$ is a weighted sum of squared errors, the problem can be easily transformed into the above situation.) James and Stein (1961) showed (for $\sigma_i^2 = \sigma^2$) that the usual maximum likelihood estimator for θ, namely $\delta^0(x)=x=(x_1,\ldots,x_p)^t$ has larger expected loss or risk (risk of $\delta \equiv R(\theta,\delta) \equiv E_\theta|\theta-\delta(X)|^2$) than the estimator

$$\delta^{JS}(x) = \left(1 - \frac{(p-2)\sigma^2}{|x|^2}\right)x \tag{1.1}$$

when $p \geq 3$. This effect, that one can improve upon the standard estimators of "independent" problems through combined estimation, will be called the Stein effect.

Though at first treated with skepticism, the Stein effect gained considerable acceptance when explanations of it were given in Bayesian and empirical Bayesian terms. Thus if the θ_i are thought to be related in some fashion, not completely known, then information about θ_j could conceivably be obtained from the other X_i. For example, a common situation considered is that in which the θ_i are thought to arise independently from a prior distribution G, but that G is not completely known. A complete subjective Bayesian treatment would put a prior μ on G, the result being an overall prior density (we will talk in terms of densities for convenience)

$$\pi(\theta) = \int \prod_{i=1}^{p} G(\theta_i)\mu(dG) .$$

Alternatively, an empirical Bayes approach could be taken, with uncertainties in G being estimated from the data. In either case, all of X becomes involved in the estimation of each θ_i, and the Stein effect is intuitively plausible. (See Efron and Morris (1973a) for analysis of the above situation when G is assumed to be a normal distribution.)

Suppose, instead, that the θ_i are (apriori) completely unrelated. A Bayesian could model this, in terms of a prior density for θ, as

$$\pi^T(\theta) = \prod_{i=1}^{p} \pi_i^T(\theta_i) ,$$

where the superscript "T" is used to indicate that this is the "True" prior density, which to a Bayesian could be obtained by infinite reflection on the problem. Saying that the θ_i are apriori unrelated also means that there are no suspected relationships among the π_i^T; so in determining π_i^T knowledge of θ_j or π_j^T for $j \neq i$ would be of no use. It follows that the Bayes estimator of θ_i(with respect to π^T) will only depend on X_i. Thus formal Bayesian reasoning supports the intuition that if the θ_i are unrelated, then it does not make sense to use an estimator for θ_i depending on all of X. This point was made quite vigorously by several of the discussants in Efron and Morris (1973b) and also in Copas (1969). These comments piqued our curiosity, in essence posing the claim that the Stein effect is of use only in empirical Bayes types of situations, and not when dealing with apriori independent problems. Considerable evidence will be presented that this is not the case.

The weak point in the above Bayesian argument is, of course, the assumption that π^T is obtainable. Even a total Bayesian will acknowledge that, in a finite amount of time, only subjective approximations to π^T can be constructed. It is thus natural to desire an analysis which is robust with respect to possible misspecification of π^T. It will be seen in this paper that the Stein effect can be of almost astonishing benefit in the attaining of such robustness (even when p=2). Indeed the Stein effect can be used to achieve most of the potential Bayesian gains available, at no (or little) cost in terms of robustness.

In the process of examining the above issue, a new class of Stein-like estimators is proposed. These estimators seem better able to take advantage of prior information than the usual James-Stein estimators.

Section 2 discusses methods of measuring Bayesian robustness and Bayesian gains of estimators. Section 3 indicates the degree of robustness obtainable if the Stein effect is not employed. In Sections 4 and 5, the new estimators are presented, and their robustness and Bayes risk improvement analyzed and compared with other estimators. Section 6 presents conclusions.

2. Measures of Bayesian Robustness

Bayesian robustness has been considered by a variety of statisticians. References and an extensive discussion can be found in Berger (1980a). The usual approach is to consider, instead of a single selected prior, a class Γ of possible priors which is felt to contain the true prior π^T. For example, suppose it is felt, in the situation of this paper, that 68% (subjective) confidence intervals for the θ_i are $(\mu_i-\tau_i,\ \mu_i+\tau_i)$, and that each π_i^T is symmetric and unimodal about μ_i. Further subjective features of π^T may be deemed too difficult to determine, in which case it would be reasonable to consider

$$\Gamma = \{\pi(\theta)\colon \pi(\theta) = \prod_{i=1}^{p} \pi_i(\theta_i), \text{ where (for } i=1,\ldots,p)\pi_i \text{ is symmetric and unimodal about } \mu_i \text{ and } \int_{|\theta_i-\mu_i|\le\tau_i} \pi_i(\theta_i)d\theta_i = .68\}. \qquad (2.1)$$

We will consider this situation throughout the paper.

Ideally, a procedure should be sought with good performance for all $\pi \in \Gamma$. How to measure performance, here, is a matter of some controversy. Although posterior measures (such as posterior expected loss) are natural to a Bayesian, it is argued in Berger (1980a) that they are incapable of sufficiently distinguishing among procedures. Instead, measures based on the frequentist risk $R(\theta,\delta)$ or overall Bayes risk $r(\pi,\delta) = E^{\pi}[R(\theta,\delta)]$ are advocated. Although these involve averages over the sample space, they can be of meaning to a Bayesian in a megaproblem sense, indicating the overall performance of their Bayesian methodology. In the situation of this paper and (2.1), for example, the "standard" Bayesian methodology is to pretend that π_i^T is $\mathcal{N}(\theta_i,\tau_i^2)$ (denote the resulting prior π^N) and then use the Bayes rule (the posterior mean) given coordinatewise by

$$\delta_i^N(x) = \left(1 - \frac{\sigma_i^2}{\sigma_i^2+\tau_i^2}\right)(x_i-\mu_i) + \mu_i \ . \tag{2.2}$$

If one faced this situation repeatedly and always assumed a normal prior, but encountered θ_i occurring according to, say, the Cauchy prior (π^C) in Γ (not at all unreasonable), then an easy calculation shows that the overall average loss would converge to

$$r(\pi^C,\delta^N) = E^{\pi^C}[R(\theta,\delta^N)] = \infty \ .$$

(Of course, real losses are bounded, but significant harm could occur for many reasonable losses.) Thus the Bayesian methodology of assuming normal priors in this situation is contraindicated, at least as the standard method to be employed automatically in, say, computer packages. (Ideas similar to the above are presented in Hill (1974).)

A very strict robustness requirement is to insist that, for appropriately small C,

$$R(\theta,\delta) \leq C, \tag{2.3}$$

in that this is equivalent to insisting that $r(\pi,\delta) \leq C$ for <u>all</u> π. Although this is stronger than usually needed, we will consider it here, partly because the resulting problem is tractable and partly because (2.3) is a condition that many frequentist decision theorists would demand.

While desiring robustness in the sense of (2.3) to protect against prior misspecification, we want an estimator which performs well if the prior specification is accurate. To measure this, we will for calculational ease consider $r(\pi^N,\delta)$, since estimators satisfying (2.3) will have similar Bayes risks for all $\pi \in \Gamma$ when C is small (i.e., close to $R(\theta,\delta^0) = \sum_{i=1}^{p} \sigma_i^2$). In interpreting numerical results it is easier to use a normalized version of $r(\pi^N,\delta)$, namely the relative savings risk of Efron and Morris (1972), defined by

$$RSR(\pi,\delta) = \frac{[r(\pi,\delta^0)-r(\pi)]-[r(\pi,\delta^0)-r(\pi,\delta)]}{r(\pi,\delta^0)-r(\pi)} = \frac{r(\pi,\delta)-r(\pi)}{r(\pi,\delta^0)-r(\pi)}, \tag{2.4}$$

where $r(\pi)$ is the Bayes risk of the Bayes estimator. This is the proportion of the possible improvement over δ^0 that is sacrificed by use of δ instead of the (theoretically) optimal Bayes rule.

It is similarly convenient to evaluate the robustness of δ by a scaled version of $\sup_\theta R(\theta,\delta)$, namely

$$\rho(\pi,\delta) = \frac{[\sup_\theta R(\theta,\delta)]-[\sup_\theta R(\theta,\delta^0)]}{r(\pi,\delta^0)-r(\pi)} = \frac{[\sup_\theta R(\theta,\delta)]-\sum_{i=1}^{p}\sigma_i^2}{r(\pi,\delta^0)-r(\pi)}. \tag{2.5}$$

This is the maximum possible <u>harm</u> that could be encountered by using δ instead of the "most robust" (i.e., minimax) estimator δ^0, relative to the potential Bayes risk improvement over δ^0. The idea here is that one might be willing to be worse than δ^0 only in proportion to the potential gain obtainable in using the prior information.

As mentioned earlier, all computations will be done with respect to π^N, the product of $\mathcal{N}(\mu_i,\tau_i^2)$ densities. Ideally, we hope to find estimators δ with small values of both $RSR(\pi^N,\delta)$ and $\rho(\pi^N,\delta)$.

3. Robustness for Coordinatewise Independent Estimators

Since the θ_i are, apriori, thought to be independent, it is natural to see if Bayesian robustness can be achieved with coordinatewise independent estimators, i.e. estimators of the form

$$\delta(x) = (\delta_1(x_1),\ldots,\delta_p(x_p))^t .$$

Since sum of squared errors loss is being considered and we are assuming that the prior beliefs are approximated by the normal density

$$\pi^N(\theta) = \prod_{i=1}^{p} \pi_i^N(\theta_i) \quad (\pi_i^N(\theta_i) \text{ is } \mathcal{N}(\mu_i,\tau_i^2)) ,$$

$R(\theta,\delta)$ and $r(\pi^N,\delta)$ will be simply the sum of component risks for such estimators. Hence it suffices to consider a single component, say, θ_i.

One possible formulation of the robustness problem stated in section 2 is that of minimizing (over the choice of estimator δ_i) the quantity $RSR(\pi_i^N,\delta_i)$, subject to the constraint $\rho(\pi_i^N,\delta_i) \leq M$. Such a δ_i would be the "optimal" Bayesian estimator having the desired degree of robustness. This problem was introduced by Hodges and Lehmann (1952). An exact mathematical solution is very difficult, but Efron and Morris (1971) show that a very close approximate solution is given by the "limited translation rule"

$$\delta_i^M(x_i) = \begin{cases} x_i+[M/(\sigma_i^2+\tau_i^2)]^{\frac{1}{2}} & \text{if } (x_i-\mu_1)<-[M(\sigma_i^2+\tau_i^2)]^{\frac{1}{2}}/\sigma_i^2 \\ \delta_i^N(x_i) & \text{if } |x_i-\mu_i|<[M(\sigma_i^2+\tau_i^2)]^{\frac{1}{2}}/\sigma_i^2 , \\ x_i-[M/(\sigma_i^2+\tau_i^2)]^{\frac{1}{2}} & \text{if } (x_i-\mu_i)>[M(\sigma_i^2+\tau_i^2)]^{\frac{1}{2}}/\sigma_i^2 \end{cases} \qquad (3.1)$$

where $\delta_i^N(x_i)$ is the Bayes estimator given in (2.2). Efron and Morris (1971) prove that

$$\rho(\pi_i^N,\delta_i^M) = \frac{\sup_{\theta_i} R(\theta_i,\delta_i^M)-\sigma_i^2}{r(\pi_i^N,\delta_i^0)-r(\pi_i^N,\delta_i^N)} = \frac{[\sigma_i^2+M^2\sigma_i^2/(\sigma_i^2+\tau_i^2)]-\sigma_i^2}{\sigma_i^2-\sigma_i^2\tau_i^2/(\sigma_i^2+\tau_i^2)} = M \quad ,$$

and that

$$RSR(\pi_i^N,\delta_i^M) = 2[(M+1)(1-\phi(\sqrt{M}))-\sqrt{M}\varphi(\sqrt{M})] \quad ,$$

where ϕ and φ are the standard normal c.d.f. and density function. (It is interesting that neither RSR nor ρ depend on σ_i^2 and τ_i^2.) The following table gives some typical values of RSR.

Table 1. RSR and ρ for Limited Translation Rules

M(i.e. $\rho(\pi_i^N,\delta_i^M)$)	0	.002	.02	.10	.2	.4	.6	.8	1.0	1.4	4	5	6	∞
$RSR(\pi_i^N,\delta_i^M)$	1	.93	.80	.58	.46	.32	.24	.18	.16	.10	.0115	.006	.0035	0

The numbers in Table 1 are not extremely encouraging. For example, if $M = .2$ (i.e., one is unwilling to "risk" being more than 20% worse than δ^0, as measured by ρ), then 46% of the possible Bayes risk improvement over δ^0 must be sacrificed.

4. Robustness Using the Stein Effect: Symmetric Case

In this section it will be assumed that $\sigma_i^2 = \sigma^2$ and $\tau_i^2 = \tau^2$ for i=1, ...,p. This is not the usual empirical Bayes situation in which the θ_i are thought to have the same τ^2, which can then be estimated from the data, in that the τ_i^2 are apriori thought to be unrelated. We are simply imagining that independent subjective analyses for each of the θ_i just happen to result in equal τ_i^2. This, of course, is rather unrealistic, unless a large number of independent θ_i are being considered and can be grouped according to their prior variances. (Note that one cannot simply rescale to make the τ_i^2 equal, since rescaling would alter the σ_i^2 and the weights in the loss function.) Realism aside, it is likely that the symmetric case is most favorable to the Stein effect, so the results of this section can be viewed as indicating whether or not the Stein effect is of potential usefulness in attaining Bayesian robustness. Some partial results for the nonsymmetric situation will be given in the next section.

For this symmetric situation and $p \geq 3$, Efron and Morris (1973a) show that the relative savings risk of the James-Stein estimator,

$$\delta^{JS}(x) = \left(1 - \frac{(p-2)\sigma^2}{|x-\mu|^2}\right)(x-\mu) + \mu \ , \tag{4.1}$$

(which shrinks towards $\mu = (\mu_1,\ldots,\mu_p)^t$) is

$$RSR(\pi^N,\delta^{JS}) = 2/p \ .$$

This is of great interest, since δ^{JS} is also minimax, implying that

$$\rho(\pi^N,\delta^{JS}) = 0 \ .$$

Thus, if $p = 5$, only 40% of the Bayes risk improvement is sacrificed, while total robustness is achieved.

The Stein effect can be used to even greater benefit, however, in more sophisticated estimators. Consider first the positive part versions of the James-Stein estimator, namely

$$\delta^{c+}(x) = \left(1 - \frac{c\sigma^2}{|x-\mu|^2}\right)^+(x-\mu) + \mu \ , \tag{4.2}$$

which is known to be better than δ^{JS} for $c = (p-2)$. The following theorem, whose proof is given in the appendix, provides a formula for $RSR(\pi^N,\delta^{c+})$.

<u>Theorem 1.</u> If $p \geq 3$, $X \sim \mathcal{N}_p(\theta,\sigma^2 I)$, and π^N is $\mathcal{N}_p(\mu,\tau^2 I)$, then

$$RSR(\pi^N,\delta^{c+}) = A^2 + \frac{c^{(1+p/2)}e^{-c/2(1+A)}}{\Gamma(1+p/2)[2(1+a)]^{p/2}} \{ \frac{2}{p} - \frac{(1+A)}{(p-2)} - \frac{c}{p(p-2)} \}$$

$$+ \{1-A^2 - \frac{2c}{p} + \frac{c^2}{p(p-2)} \}[1-\psi_{(p+2)}(\frac{c}{1+A})] , \tag{4.3}$$

where $A = \tau^2/\sigma^2$ and $\psi_\nu(\cdot)$ is the c.d.f. of the χ^2 distribution with ν degrees of freedom.

<u>Corollary 1.</u> If p=4 in Theorem 1, then

$$RSR(\pi^N,\delta^{c+}) = A^2 + (1-A^2 - \frac{cA}{2})e^{-c/2(1+A)} . \tag{4.4}$$

This is minimized when $c = 2(1+A^{-1})$.

<u>Proof</u>. Simple calculation shows that

$$1 - \psi_6(b) = (1+b/2+b^2/8)e^{-b/2} ,$$

which, together with (4.3), immediately yields (4.4). Differentiating in (4.4) with respect to c demonstrates the optimality of $c = 2(1+A^{-1})$. ||

As an indication of the improvement obtainable through use of the positive part rules, Table 2 gives RSR for p=4, various values of A, the usual choice $c = (p-2) = 2$ in (4.2), and the optimal choice $c^* = \min\{4,2(1+A^{-1})\}$. (Restricting c to be no larger than $4 = 2(p-2)$ ensures that δ^{c+} is minimax, and hence that $\rho(\pi^N,\delta^{c+}) = 0$.)

Table 2. $RSR(\pi^N,\delta^{c+})$, p=4

A	0	.1	.5	.8	1.0	1.5	2.0	4.0	10	50	∞
RSR for c=2	.368	.369	.378	.388	.394	.407	.417	.444	.472	.494	.500
RSR for c=c*	.135	.138	.184	.232	.264	.325	.361	.424	.468	.493	.500

Since δ^{JS} has RSR = 2/p = .5 in this situation, it is clear that the positive part rules can be significantly better. Furthermore, choosing c according to the prior information can be of substantial benefit. (These improvements are less significant when p is larger.) Compare, also, Table 2 with Table 1 (recalling that $\rho(\pi^N,\delta^{c+}) = 0$), to see the benefits of the Stein effect.

To obtain the full value of the Stein effect in the Bayesian framework, it seems plausible to use estimators which mimick the Bayes rule δ^N (as did the limited translation estimators). Thus consider estimators of the form (again letting $A = \tau^2/\sigma^2$)

$$\delta^{c;A}(x) = \begin{cases} \delta^N(x) = (1-\frac{1}{(1+A)})(x-\mu) + \mu & \text{if} \quad |x-\mu|^2 \le c(1+A) \\ , & \\ (1-\frac{c}{|x-\mu|^2})(x-\mu) + \mu & \text{if} \quad |x-\mu|^2 > c(1+A). \end{cases} \tag{4.5}$$

This estimator has the intuitive justification of being the (normal prior) Bayes estimator when the prior information is supported by the data ($|x-\mu|^2$ small), and being a James-Stein estimator otherwise. Using Baranchik (1970), it is easy to verify that $\delta^{c,A}$ is minimax if $c \le 2(p-2)$, and hence that

$$\rho(\pi^N,\delta^{c,A}) = 0 \quad \text{if } c \le 2(p-2). \tag{4.6}$$

Of course, for these estimators and the positive part estimators values of c greater than 2(p-2) could be used. The value of $\rho(\pi^N,\delta)$ would then be greater than zero, however, and would have to be determined numerically.

To calculate the relative savings risk of $\delta^{c,A}$, we will use the following theorem, the generality of which will be useful later.

<u>Theorem 2.</u> Suppose $p \ge 3$, $X \sim \mathcal{N}_p(\theta,\sigma^2 I)$, and π^N is $\mathcal{N}_p(\mu,\tau^2 I)$. Define $A_0 = \tau_0^2/\sigma^2$, $A = \tau^2/\sigma^2$, and

$$\lambda = (1+A_0)/(1+A) .$$

Then

$$RSR(\pi^N,\delta^{c,A_0}) = [1-\lambda^{-1}]^2+\{1-\frac{2c}{p}+\frac{c^2}{p(p-2)}-[1-\lambda^{-1}]^2\}\{1-\psi_{(p+2)}(c\lambda)\}$$

$$+\ 2^{-p/2}[\Gamma(\frac{p}{2}+1)]^{-1}\{\frac{2}{p}-\frac{1}{(p-2)\lambda}-\frac{c}{p(p-2)}\}c^{(p+2)/2}\lambda^{p/2}e^{-c\lambda/2}. \quad (4.7)$$

Proof. Given in the Appendix. ||

Corollary 2. If $p \geq 3$, $X \sim \mathcal{N}_p(\theta,\sigma^2 I)$, π^N is $\mathcal{N}_p(\mu,\tau^2 I)$, and $A = \tau^2/\sigma^2$, then

$$RSR(\pi^N,\delta^{c,A}) = \{1-\frac{2c}{p}+\frac{c^2}{p(p-2)}\}\{1-\psi_{(p+2)}(c)\}$$

$$+\ 2^{-p/2}[\Gamma(\frac{p}{2}+1)]^{-1}\frac{(p-4-c)}{p(p-2)}c^{(p+2)/2}e^{-c/2} \quad . \quad (4.8)$$

Proof. Follows directly from Theorem 2, since $\lambda=1$. ||

Corollary 3. If $p=4$ in the situation of Theorem 2, then

$$RSR(\pi^N,\delta^{c,A}) = e^{-c\lambda/2}[\frac{c}{2}+\frac{(4-c)}{2\lambda}-\frac{1}{\lambda^2}] \quad .$$

Proof. Direct calculation, as in Corollary 1. ||

The choice of c which minimizes RSR in (4.8), subject to (4.6), can be seen to be $c^* = 2(p-2)$. Table 3 presents $RSR(\pi^N,\delta^{c^*,A})$ for various values of p. (Observe that, as for the limited translation rules, $RSR(\pi^N,\delta^{c,A})$ does not depend on A.)

Table 3. $RSR(\pi^N,\delta^{c^*,A})$

p	3	4	5	6	7	8	9	10	15	20
RSR	.296	.135	.0727	.0427	.0267	.0174	.0117	.0080	.0016	.0004

The values of RSR in the above table are startling. When p=5, for example, one can achieve total robustness (i.e., minimaxity) with only a 7% decrease in possible Bayes risk improvement over δ^0. This is overwhelmingly superior to what can be achieved using coordinatewise independent rules, and lends strong support to the claim that the Stein effect can be of great benefit in obtaining Bayesian robustness .

The estimator $\delta^{c^*,A}$ also seems to be considerably better than the usual James-Stein or positive part estimators, as can be seen by comparing Table 2 with Table 3 for p=4. In a sense, δ^{JS} and the usual positive part estimator tend to pull in too much for small $|x-\mu|$, but not enough for large $|x-\mu|$. (See also Leonard (1976).) To a Bayesian, $\delta^{c^*,A}$ would also be appealing from the viewpoint of posterior expected loss, since it coincides exactly with the normal Bayes estimator over a large range of x.

We conclude this section with a discussion of the situation when p=2. Of course when p=2 it is no longer possible to require that $\rho(\pi^N,\delta) = 0$, since only $\delta^0(x) = x$ is minimax in two dimensions. Nevertheless, the Stein effect can still be useful in achieving increased robustness. Again considering the estimators $\delta^{c,A}$ in (4.5), a calculation similar to that in Theorem 2 verifies the following result.

Theorem 3. If p=2, $X \sim \mathcal{N}_2(\theta,\sigma^2 I)$, π^N is $\mathcal{N}_2(\mu,\tau^2 I)$, and $A = \tau^2/\sigma^2$, then

$$RSR(\pi^N,\delta^{c,A}) = (1-\frac{c}{2})e^{-c/2} + \frac{c^2}{4}\int_{c/2}^{\infty} y^{-1}e^{-y}dy \ . \tag{4.9}$$

The following is a brief table of RSR for various c. (Again, RSR is independent of A.)

Table 4. $RSR(\pi^N,\delta^{c,A})$, p=2

c	0	.2	.4	.6	.8	1.0	1.2	1.4	1.6	1.8	2.0	2.5	3.0	4.0	5.0	6.0	∞
RSR	1	.833	.704	.600	.515	.443	.383	.332	.289	.251	.219	.157	.114	.061	.033	.018	0

Values of $\rho(\pi^N,\delta^{c,A})$ for various c and A are given in Table 5. (These values were found by simulation. The standard error of an entry is less than or equal to the unit of the last digit; e.g., for A = 1.0 and c = .8, the standard error is less than .001.)

Table 5. $\rho(\pi^N,\delta^{c,A})$, p=2

A \ c	0	.2	.4	.6	.8	1.0	1.2	1.4	1.6	1.8	2.0	2.5	3.0	4.0	5.0	6.0	∞
.1	0	.002	.007	.018	.035	.054	.07	.10	.13	.16	.20	.30	.41	.68	.97	1.28	∞
.5	0	.003	.011	.024	.042	.065	.09	.12	.16	.19	.24	.34	.46	.74	1.04	1.38	∞
1.0	0	.003	.012	.026	.047	.072	.10	.14	.18	.22	.26	.39	.53	.83	1.16	1.50	∞
2.0	0	.004	.018	.038	.063	.096	.13	.17	.22	.27	.32	.45	.61	.93	1.29	1.65	∞
4.0	0	.008	.028	.055	.090	.13	.18	.23	.28	.37	.39	.53	.70	1.04	1.41	1.81	∞
10.0	0	.01	.04	.07	.12	.18	.23	.29	.35	.42	.48	.65	.84	1.22	1.62	2.04	∞

A comparison of Tables 4 and 5 with Table 1 shows the value of the Stein effect in two dimensions. For example, if A=1 and c=1, $\delta^{c,A}$ loses 44% in possible Bayes risk improvement over δ^0, but is only 7% worse than δ^0 in terms of robustness. In contrast, the limited translation estimator must sacrifice 20% in robustness to achieve a loss in Bayes risk improvement of 46%. For smaller A the advantage of $\delta^{c,A}$ is even more pronounced. A careful examination of Tables 4 and 5 leaves a strong feeling that one should use Stein effect estimators even in 2 dimensions. Using the estimator with c=1, for example, can provide significant Bayesian gains over δ^0 at negligible cost in terms of robustness.

5. Robustness Using the Stein Effect: Nonsymmetric Case

In this section we will consider the more general situation where $X \sim \mathcal{N}_p(\theta,\ddagger)$, π^N is $\mathcal{N}_p(\mu,A)$, and the loss is $L(\theta,\delta) = (\theta-\delta)^t Q(\theta-\delta)$. The vector μ and the positive definite matrices Σ, A, and Q are all assumed to be known or specified. Again it is desired to make both $\rho(\pi^N,\delta)$ and $RSR(\pi^N,\delta)$ small, through use of the Stein effect.

A natural modification of the estimator in (4.5) is

$$\delta^{c,A}(x) = \begin{cases} \delta^N(x) = (I-\Sigma(\Sigma+A)^{-1})(x-\mu) + \mu & \text{if } ||x-\mu||^2 \le c \\ \left(I - \dfrac{c}{||x-\mu||^2}\Sigma(\Sigma+A)^{-1}\right)(x-\mu) + \mu & \text{if } ||x-\mu||^2 > c , \end{cases} \tag{5.1}$$

where $||x-\mu||^2 = (x-\mu)^t(\Sigma+A)^{-1}(x-\mu)$. (This is closely related to the estimator discussed in Berger (1980b).) Using results in Berger (1976), the following lemma is easily established.

Lemma 1. The estimator $\delta^{c,A}$ is minimax (so $\rho(\pi^N,\delta^{c,A}) = 0$) if $p \ge 3$ and

$$c \le \frac{2\,\mathrm{tr}\{\Sigma Q\Sigma(\Sigma+A)^{-1}\}}{ch_{max}\{\Sigma Q\Sigma(\Sigma+A)^{-1}\}} - 4 \overset{(\text{defn.})}{=} c^* , \tag{5.2}$$

where "tr" and "ch_{max}" denote trace and maximum characteristic root, respectively.

Theorem 4. For the above situation and $p \ge 3$, $RSR(\pi^N,\delta^{c,A})$ is given by (4.8) of Corollary 2. For p=2, $RSR(\pi^N,\delta^{c,A})$ is given by (4.9) of Theorem 3.

Proof. Given in the Appendix. ||

Example 1. Suppose p=4. Then from Theorem 4 and Corollary 3 (with λ=1) it follows that $RSR(\pi^N,\delta^{c,A}) = e^{-c/2}$. If, for instance, Σ=Q=I and A is the diagonal matrix with diagonal elements {2,3,4,5}, then c^* = 1.70 and $RSR(\pi^N,\delta^{c^*,A})$ = .427.

Example 2. Suppose p=8, Σ=Q=I, and A is the diagonal matrix with diagonal elements {2,2,2,4,4,4,5,10}. Then calculation give c^* = 7.145 and (from (4.8)) $RSR(\pi^N,\delta^{c^*,A})$ = .108.

It is clear that the Stein effect can be of considerable benefit in nonsymmetric problems. Actually, the estimators $\delta^{c,A}$ are not necessarily even close to optimum in the nonsymmetric situation, so the Stein effect is potentially of even greater benefit than indicated here. There are many reasonable alternatives to $\delta^{c,A}$, such as the minimax Bayesian estimator in Berger (1980c). Calculation of RSR for these estimators is difficult, however.

6. Conclusions and Comments

Comment 1. For the situation considered in this paper, the improvement in Bayesian robustness that can be obtained by use of the Stein effect is startling. The same should be true for more general or other simultaneous estimation problems, and also for different measures of Bayesian robustness. In particular, for simultaneous estimation of normal means it is well known that estimation of the σ_i^2 or changes in the loss function do not significantly affect the benefits of the Stein effect. Thus the Stein effect can be an important general tool to the Bayesian seeking robustness. It appears that combining unrelated estimation problems can definitely be of benefit.

Comment 2. It should be emphasized that the risk improvements obtained through use of the Stein effect are improvements in total risk, and not necessarily improvements in each of the coordinatewise risks. Thus the results are formally applicable only when it is reasonable to add the losses from the component problems, such as when a business must simultaneously make estimates in p problems. If it is also desired to ensure that component risks are not excessive, componentwise limited translation Stein effect estimators could be employed. (See, for example, Efron and Morris (1972, 1973a) and Shapiro (1972, 1975).) Of course, the estimators considered in this paper are, coordinatewise, considerably more robust than is the usual conjugate prior Bayes estimator, δ^N.

Comment 3. The purpose of this article was mainly theoretical; to establish the importance of the Stein effect in nonempirical Bayes situations. Nevertheless, the estimators (4.5) and (5.1) are very attractive in their own right, especially when substantial subjective prior information is available. Even in standard empirical Bayes situations these estimators can be superior to the usual empirical Bayes estimators, particularly if p is fairly small. Consider, for example, the simple empirical Bayes situation in which $X_i \sim \mathcal{N}(\theta_i,\sigma^2)$, $i=1,\ldots,p$, are independently observed, and it is thought that the θ_i arise independently from a common $\mathcal{N}(0,\tau^2)$ distribution. (It is common to also suppose a common prior mean μ_0 for the θ_i, and then estimate μ_0 from the data. This can also be done for the estimators (4.5), however, and hence does not qualitatively affect the results below.) The usual empirical Bayes estimator in this situation is the positive part James-Stein estimator in (4.2) with $c = (p-2)$ (and $\mu = (0,\ldots,0)^t$). Note that this estimator does not require a subjective specification of τ^2, as does the new estimator (4.5). It is nevertheless somewhat surprising that, if p is fairly small and τ^2 is specified correctly, then the new estimator does much better than the positive part estimator. (Compare Table 2 for c=2 with Table 3 for p=4.) Of course, the obvious question is - how badly does the new estimator do if τ^2 is misspecified?

Theorem 2 can be used to answer this question. Indeed, suppose that the subjective estimate for τ^2 is τ_0^2, so that the estimator δ^{c,A_0} in (4.5) would be used ($A_0 = \tau_0^2/\sigma^2$). Theorem 2 gives RSR for this estimator and the "true" prior π^N, which is $\mathcal{N}_p(0,\tau^2 I)$. RSR depends only on c and $\lambda = (1+A_0)/(1+A) = (\sigma^2+\tau_0^2)/(\sigma^2+\tau^2)$. Table 6 gives some typical values of RSR when p=4 (see Corollary 3) and c = 2(p-2) = 4.

Table 6. $RSR(\pi^N, \delta^{4,A_0})$, Misspecified Prior Variance

λ	0	.1	.3	.5	.7	.9	1.0	1.1	1.5	2	3	5	6	10	100	∞
RSR	1	.764	.444	.264	.174	.139	.135	.138	.189	.282	.450	.640	.694	.810	.980	1

Comparing Table 6 with Table 2 (for c=2) shows that δ^{4,A_0} is superior to the positive part estimator for a surprisingly wide range of λ. For example, if $A = \tau^2/\sigma^2 = 4$, then δ^{4,A_0} will be superior provided (roughly) that $.3 \leq \lambda \leq 3$, or (equivalently) $.5 \leq A_0 \leq 14$. Thus the subjective estimate, τ_0^2, need only be in the ballpark for δ^{4,A_0} to outperform the usual empirical Bayes estimator. (Note that both estimators are minimax, so $\rho(\pi^N,\delta)$ is zero in either case.) The range of dominance of $\delta^{2(p-2),A_0}$ over the positive part estimator will decrease as A becomes small or p large, but for moderate p and A typically encountered in practice, there may be real advantages to using $\delta^{2(p-2),A_0}$.

Actually, $\delta^{2(p-2),A_0}$ and the positive part estimator are at two extremes, the first using an entirely subjective estimate of τ^2 and the second an entirely empirical estimate. The optimal estimator is probably a Bayesian compromise between the two. The main value of this discussion is thus the indication that, for small or moderate p, subjective prior information should not be ignored even in standard empirical Bayes settings. (Of course, it may be appealing to some to give the <u>appearance</u> of objectivity by using the positive part estimator, which does not formally incorporate subjective prior information. If appearance, instead of accuracy, is important, so be it.)

<u>Comment 4.</u> When not in an empirical Bayes setting, it is probably unwise to simultaneously estimate a large number of coordinates. If faced with a large p, it would probably be best to divide the coordinates into similar

small groups, and use the relevant robust Bayesian estimators on each group. Indeed from Table 3 it seems that there is little reason to combine more than 5 or 6 coordinates (in the symmetric case). The danger of combining more is that the true prior probably has fatter tails than π^N, and $RSR(\pi, \delta^{C,A})$ can be shown to <u>increase</u> with p (for p beyond some moderate value) if π has fat enough tails. Material relevant to this issue can be found in Efron and Morris (1973b), Stein (1974), Dey (1980), and Dey and Berger (1980).

<u>Comment 5.</u> The Stein effect can be used in conjunction with Bayes estimators other than δ^N. The idea is the same: if the Bayes estimator is δ^π, select a Stein type (often minimax) estimator δ^S, and use δ^π or δ^S according as to whether the observation, x, does or does not support the prior beliefs. Unfortunately, for nonsymmetric situations it is not clear how to choose δ^S.

<u>Comment 6.</u> The new estimators in (4.5) and (5.1) are not admissible, since they are not analytic. (Brown (1971) establishes that analyticity is needed.) Nevertheless, as with the positive part James-Stein estimator, it is unlikely that admissible improvements are significantly better.

<u>Comment 7.</u> This paper has, for the most part, been addressed to Bayesians, attempting to demonstrate the power of the Stein effect in achieving Bayesian robustness. Non-Bayesians, however, should take note of the extent to which it is possible to satisfy (perhaps repressed) Bayesian urges <u>at no cost</u>, and of the great gains that can be achieved in doing so.

APPENDIX

Proof of Theorem 1.

By making the linear transformation $Y = (X-\mu)/\sigma$ (also transform θ and the loss), it is easy to see that we may assume that $\mu=0$, $X \sim \mathcal{N}_p(\theta,I)$, and π^N is $\mathcal{N}(0,AI)$. Observe that

$$r(\pi^N,\delta^{c+}) = E^{m(x)}E^{\pi(\theta|x)}|\theta-\delta_0^{c+}(x)|^2 . \tag{A1}$$

where $\pi(\theta|x)$ is $\mathcal{N}_p(\delta^N(x), \frac{A}{(1+A)} I)$ (the posterior distribution of θ given x; see (4.5) for a definition of $\delta^N(x)$), and $m(x)$ is $\mathcal{N}_p(0,(1+A)I)$ (the marginal or unconditional distribution of X). Adding and subtracting the posterior mean $\delta^N(x)$ in (A1), and then expanding the quadratic, gives (since the expectation of the cross product term is zero)

$$\begin{aligned} r(\pi^N,\delta^{c+}) &= E^{m(x)}E^{\pi(\theta|x)}\{|\theta-\delta^N(x)|^2+|\delta^N(x)-\delta^{c+}(x)|^2\} \\ &= \frac{PA}{(1+A)} + E^{m(x)}|\delta^N(x)-\delta^{c+}(x)|^2 . \end{aligned} \tag{A2}$$

Now

$$\begin{aligned} &E^{m(x)}|\delta^N(x)-\delta^{c+}(x)|^2 \qquad\qquad (A3)\\ &= \frac{A^2}{(1+A)^2} \int_{|x|^2\le c} |x|^2 m(x)dx + \int_{|x|^2\ge c} [\frac{1}{(1+A)} - \frac{c}{|x|^2}]^2 |x|^2 m(x)dx \\ &= \frac{A^2}{(1+A)^2} p(1+A) + \int_{|x|^2>c} \{[\frac{1}{(1+A)} - \frac{c}{|x|^2}]^2 - \frac{A^2}{(1+A)^2}\}|x|^2 m(x)dx \\ &= \frac{PA^2}{(1+A)} + \int_{c/2(1+A)}^{\infty} \frac{2(1+A)}{\Gamma(p/2)}\{[\frac{1}{(1+A)} - \frac{c}{2(1+A)y}]^2 - \frac{A^2}{(1+A)^2}\} y^{p/2}e^{-y}dy . \end{aligned}$$

Expanding the quadratic expression in the integral above and using the identities (valid for $p\neq 2$)

$$\int_b^\infty y^{(p-2)/2}e^{-y}dy = -\frac{2}{p}b^{p/2}e^{-b} + \frac{2}{p}\int_b^\infty y^{p/2}e^{-y}dy ,$$

$$\int_b^\infty y^{(p-4)/2}e^{-y}dy = -\frac{2}{(p-2)}b^{(p-2)/2}e^{-b} - \frac{4}{p(p-2)}b^{p/2}e^{-b} + \frac{4}{p(p-2)}\int_b^\infty y^{p/2}e^{-y}dy , \quad (A4)$$

a little algebra shows that

$$E^{m(x)}|\delta^N(x)-\delta^{c+}(x)|^2$$

$$= \frac{PA^2}{(1+A)} + \frac{2c^{(p+2)/2}e^{-c/2(1+A)}}{\Gamma(p/2)[2(1+A)]^{p/2}(1+A)}\left\{\frac{2}{p} - \frac{(1+A)}{(p-2)} - \frac{c}{p(p-2)}\right\}$$

$$+ \left\{1-A^2-\frac{2c}{p} + \frac{c^2}{p(p-2)}\right\}\frac{2}{\Gamma(p/2)(1+A)}\int_{c/2(1+A)}^\infty y^{p/2}e^{-y}dy . \quad (A5)$$

Noting that $r(\pi^N,\delta^0) = p$, $r(\pi^N,\delta^N) = PA/(1+A)$, and

$$[\Gamma(1+p/2)]^{-1}\int_{c/2(1+A)}^\infty y^{p/2}e^{-y}dy = 1 - \psi_{(p+2)}\left(\frac{c}{1+A}\right) , \quad (A6)$$

it follows from (A2) and (A5) that

$$RSR(\pi^N,\delta^{c+}) = \frac{r(\pi^N,\delta^{c+})-r(\pi^N,\delta^N)}{r(\pi^N,\delta^0)-r(\pi^N,\delta^N)}$$

is as in (4.3). ||

Proof of Theorem 2.

The proof is exactly analagous to that of Theorem 1. Equation (A3) is instead

$$E^{m(x)}|\delta^N(x)-\delta^{c,A_0}(x)|^2 = \frac{p(A-A_0)^2}{(1+A)(1+A_0)^2}$$

$$+ \int_{c(1+A_0)/2(1+A)}^\infty \frac{2(1+A)}{\Gamma(p/2)}\left\{\left[\frac{1}{(1+A)} - \frac{c}{2(1+A)y}\right]^2 - \frac{(A-A_0)^2}{(1+A_0)^2(1+A)^2}\right\} y^{p/2}e^{-y}dy , \quad (A7)$$

while the analog of (A5) is

$$E^{m(x)}|\delta^N(x)-\delta^{c,A_0}(x)|^2 = \frac{p(A-A_0)^2}{(1+A)(1+A_0)^2}$$

$$+ \frac{c^{(p+2)/2}e^{-c(1+A_0)/2(1+A)}}{2^{(p-2)/2}\Gamma(p/2)(1+A)} \left(\frac{1+A_0}{1+A}\right)^{p/2}\left\{\frac{2}{p} - \frac{(1+A)}{(p-2)(1+A_0)} - \frac{c}{p(p-2)}\right\}$$

$$+ \left\{1-\frac{2c}{p}+\frac{c^2}{p(p-2)}-\frac{(A-A_0)^2}{(1+A_0)^2}\right\}\frac{2}{\Gamma(p/2)(1+A)}\int_{c(1+A_0)/2(1+A)}^{\infty} y^{p/2}e^{-y}dy \quad (A8)$$

Recalling that $\lambda = (1+A_0)/(1+A)$, using (A6), and observing that $(A-A_0)^2/(1+A_0)^2 = [1-\lambda^{-1}]^2$, the conclusion follows. ||

Proof of Theorem 4.

The proof is analogous to that of Theorem 1. We will do only the case $p \geq 3$, and will again assume (without loss of generality) that $\mu=0$. Note that $\pi(\theta|x)$ is $\mathcal{N}_p(\delta^N(x), (\Sigma^{-1}+A^{-1})^{-1})$ and $m(x)$ is $\mathcal{N}_p(0,\Sigma+A)$. Equation (A3) should be replaced by

$$\Lambda = E^{m(x)}|\delta^N(x)-\delta^{c,A}(x)|^2 = \operatorname{tr} Q(\Sigma^{-1}+A^{-1})^{-1}$$

$$+ (2\pi)^{-p/2}|\Sigma+A|^{-\frac{1}{2}} \int_{||x||^2>c} \left(1-\frac{c}{||x||^2}\right)^2[x^t(\Sigma+A)^{-1}\Sigma Q\Sigma(\Sigma+A)^{-1}x]e^{-\frac{1}{2}x^t(\Sigma+A)^{-1}x}dx. \quad (A9)$$

Now make the transformation $y = \mathcal{O}(\Sigma+A)^{-\frac{1}{2}}x$, where $\mathcal{O}$ is an orthogonal matrix such that

$$\mathcal{O}(\Sigma+A)^{-\frac{1}{2}}\Sigma Q\Sigma(\Sigma+A)^{-\frac{1}{2}}\mathcal{O}^t = D,$$

D being a diagonal matrix with diagonal elements $\{d_1,\ldots,d_p\}$. Then (A9) becomes,

$$\Lambda = \operatorname{tr} Q(\Sigma^{-1}+A^{-1})^{-1} + (2\pi)^{-p/2}\int_{|y|^2>c}\left(1-\frac{c}{|y|^2}\right)^2\left(\sum_{i=1}^{p} d_i y_i^2\right)e^{-|y|^2/2}dy .$$

By symmetry,

$$\int_{|y|^2>c} (1-\frac{c}{|y|^2})^2 y_i^2 e^{-|y|^2/2} dy = \int_{|y|^2>c} (1-\frac{c}{|y|^2})^2 (\frac{1}{p}\sum_{i=1}^{p} y_i^2) e^{-|y|^2/2} dy ,$$

so that

$$\Lambda = \mathrm{tr}\, Q\,(\ddagger^{-1}+A^{-1})^{-1}+(2\pi)^{-p/2}\,\frac{(\mathrm{tr}\, D)}{p} \int_{|y|^2>c} (1-\frac{c}{|y|^2})^2 |y|^2 e^{-|y|^2/2} dy$$

$$= \mathrm{tr}\, Q\,(\ddagger^{-1}+A^{-1})^{-1}+ \frac{(\mathrm{tr}\, D)}{p} \int_{c/2}^{\infty} \frac{2}{\Gamma(p/2)} (1-\frac{c}{2z})^2 z^{p/2} e^{-z} dz .$$

The last integral above is exactly the integral in (A7) with $A = A_0$, so the result of that calculation yields

$$\Lambda = \mathrm{tr}\, Q\,(\ddagger^{-1}+A^{-1})^{-1}+(\mathrm{tr}\, D)\, g(p,c) ,$$

where $g(p,c)$ is the function in (4.8).

To complete the argument, observe that $r(\pi^N,\delta^N) = \mathrm{tr}\, Q\,(\ddagger^{-1}+A^{-1})^{-1}$, $r(\pi^N,\delta^0) = \mathrm{tr}\, Q\,\ddagger$, and $\mathrm{tr}\, D = \mathrm{tr}\,\ddagger Q \ddagger(\ddagger+A)^{-1}$. It follows that

$$RSR(\pi^N,\delta^{c,A}) = \frac{\mathrm{tr}\,\ddagger Q\ddagger(\ddagger+A)^{-1}}{[\mathrm{tr}\, Q\ddagger - \mathrm{tr}\, Q\,(\ddagger^{-1}+A^{-1})^{-1}]}\, g(p,c) .$$

Use of the matrix identity

$$(\ddagger^{-1}+A^{-1})^{-1} = \ddagger - \ddagger(\ddagger+A)^{-1}\ddagger$$

establishes the result. ||

REFERENCES

[1] Baranchik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. Ann. Math. Statist. 41, 642-645.

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

[3] Berger, J. (1980a). Statistical Decision Theory: Foundations, Concepts, and Methods. Springer-Verlag, New York.

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

[5] Berger, J. (1980c). Selecting a minimax estimator of a multivariate normal mean. Technical Report #80-24, Purdue University.

[6] Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. Ann. Math. Statist. 42, 855-904.

[7] Copas, J. B. (1969). Compound decisions and empirical Bayes (with discussion). J. R. Statist. Soc. B 31, 397-425.

[8] Dey, D. K. (1980). On the choice of coordinates in simultaneous estimation of normal means. Ph.D. Thesis, Purdue University.

[9] Dey, D. and Berger, J. (1980). Combining coordinates in simultaneous estimation of normal means. Technical Report 32, Department of Statistics, Purdue University.

[10] Efron, B. and Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators - Part 1: the Bayes case. J. Amer. Statist. Assoc. 66, 807-815.

[11] Efron, B. and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators - Part 2: the empirical Bayes case. J. Amer. Statist. Assoc. 67, 130-139.

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

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

[14] Hill, B. M. (1974). On coherence, inadmissibility, and inference about many parameters in the theory of least squares. In Studies in Bayesian Econometrics and Statistics (S. Fienberg and A. Zellner, eds.) North Holland, Amsterdam.

[15] Hodges, J. L., Jr. and Lehmann, E. L. (1952). The use of previous experience in reaching statistical decisions. Ann. Math. Statist. 23, 392-407.

[16] James, W. and Stein, C. (1961). Estimation with quadratic loss. Proc. Fourth Berkeley Symposium Math. Stat. Prob. 1, 361-379. University of California Press.

[17] Leonard, T. (1976). Some alternative approaches to multiparameter estimation. Biometrika 63, 69-76.

[18] Rubin, H. (1977). Robust Bayesian estimators. S. S. Gupta and D. S. Moore (eds.) Statistical Decision Theory and Related Topics II, Academic Press.

[19] Shapiro, S. H. (1972). A compromise between the Bayes and minimax approaches to estimation. Technical Report No. 31. Department of Statistics, Stanford University.

[20] Shapiro, S. H. (1975). Estimation of location and scale parameters - a compromise. Communications in Statistics 4(12), 1093-1108.

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