On Truncation of Shrinkage Estimators In Simultaneous Estimation of Normal Means*

by

Dipak K. Dey
Texas Tech University

and

James O. Berger
Purdue University

Technical Report #82-31

October 1982

Purdue University
West Lafayette, IN 47907

*Research supported by the National Science Foundation, Grants MCS 78-02300 and MCS 81-01670A1.

In estimating a multivariate normal mean $\theta = (\theta_1,\ldots,\theta_k)^t$ under sum of squares error loss, it is well known that Stein estimators improve upon the usual estimator (in terms of expected loss) if $k \geq 3$. The improvement obtained is significant, however, only if the θ_i are fairly close to the point towards which the Stein estimator shrinks. When extreme θ_i are likely (such as when the θ_i are thought to arise from a possibly

improvement over the usual estimator. Stein (1981) proposed a limited translation Stein estimator to correct this deficiency. This estimator is analyzed herein for a number of fat-tailed prior distributions. An adaptive version of the estimator is also discussed.

Key Words and Phrases: James-Stein estimator, Stein's truncation, Fat-tailed prior, Adaptive estimator.

1. INTRODUCTION

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

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

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

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

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

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

has uniformly smaller risk than σ^2. Estimators having uniformly smaller risk than δ^0 for the general situation above have been found by many authors. See Berger (1982) for references.

A key feature of any Stein type estimator is that it has risk significantly better than that of δ^0 only in a relatively small region (or subspace) of the parameter space (see Berger (1980, 1982) for discussion). For Stein estimation to result in significant improvement, therefore, one must carefully select an estimator designed to do well in the region in which θ is thought likely to lie. This is essentially done by finding a Stein type estimator which shrinks toward the desired region. If any of the θ_i happen to fall substantially outside this region, the usual Stein estimators will collapse back to X and offer little imporvement over δ^0. The estimator in (1.1), for example, is designed to do well for θ near zero, and if any of the θ_i are far from zero then X^tX will be large and δ will be approximately X.

The above problem was noted by Stein (1981), who considered a modification of the estimator in (1.1) to partially alleviate the difficulties. He proposed the estimator defined coordinatewise by

$$\delta_i^{(\ell)}(X) = \left(1 - \frac{(\ell-2)\sigma^2\min\{1,\ Z_{(\ell)}/|X_i|\}}{\sum_{j=1}^{k} X_j^2 \wedge Z_{(\ell)}^2}\right)X_i, \tag{1.2}$$

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

$$Z_i = |X_i|,\ i=1,\ldots,k \tag{1.3}$$

and

$$Z_{(1)} < Z_{(2)} < \cdots < Z_{(\ell)} < \cdots < Z_{(k)} \tag{1.4}$$

are the order statistics of $Z_1,\ldots,Z_k$. Stein (1981) proved that this estimator is minimax if $\ell \geq 3$. The estimator provides a reasonable solution to the extreme θ_i problem, as is indicated by the observation that

$$\sum_{j=1}^{k} X_j^2 \wedge Z_{(\ell)}^2$$

is fairly small even if $(k-\ell)$ of the θ_i are very extreme.

In section 2, versions of the estimaotr (1.2) will be evaluated for the symmetric situation. Since the symmetric situation mainly occurs when the θ_i are felt to arise from a common prior distribution (the empirical Bayes situation discussed, for example, in Efron and Morris (1972)), this evaluation will be in terms of Bayes risk $r(\pi,\delta) = E^{\pi}[R(\theta,\delta)]$ for a variety of prior distributions π. The goal is to determine sensible choices of the truncation point ℓ. It will also be indicated that the resulting estimator has quite small maximum component risk, a desirable feature as discussed in Efron and Morris (1972).

In section 3, the estimator in (1.2) is considered when ℓ is chosen adaptively by the data, an appealing possibility which obviates the necessity to consider the prior distribution of the θ_i.

In this paper we will mainly be concerned with results for large dimension k. Extreme θ_i are clearly more of a danger when k is large. Also, when k is small the "loss" of dimensions in going from k to ℓ can be quite harmful, so truncation is less appealing.

2. OPTIMAL CHOICE OF THE TRUNCATION POINT

The estimator (1.1) is the empirical Bayes estimator if $X \sim N_k(\theta,\sigma^2 I_k)$ and $\theta \sim N_k(0,\tau^2 I_k)$, where τ^2 is unknown (see Efron and Morris (1972)). In general it is a reasonable estimator if the θ_i are thought to be independent realizations from a common symmetric prior distribution π having median zero. If the common prior is symmetric about its median $\mu \neq 0$, the estimator

$$\delta(X) = X - \frac{(k-2)\,\sigma^2 (X-(\mu,\ldots,\mu)^t)}{\sum_{i=1}^{k} (X_i-\mu)^2} \tag{2.1}$$

would be appropriate. If μ is unknown the (the realistic case), replacing μ in (2.1) by $\bar{X} = k^{-1} \sum_{i=1}^{k} X_i$ (or some more robust estimate if π could have fat tails) would suffice. We will be considering mainly large k situations, and hence can assume that μ is known, and without loss of generality is zero.

Since the estimator (1.2) is minimax (for $\ell \geq 3$ and sum of squares error loss), the choice of ℓ should be based on the overall expected gain from use of the estimator. The most reasonable measure of this average gain is the improvement in Bayes risk of the estimator over that of δ^0. Using Stein's unbiased estimator of risk this can easily be shown (as in Berger and Dey (1980)) to be

$$\Delta_k^{(\ell)} = r(\pi,\delta^0) - r(\pi,\delta^{(\ell)}) = \sigma^4 E^m\left[\frac{(\ell-2)^2}{\sum_{j=1}^{k} X_j^2 \wedge Z_{(\ell)}^2}\right], \tag{2.2}$$

where E^m stands for expectation under the marginal density of X. Note that the X_i are marginally independent (since the θ_i are) with common marginal density

$$m(x_i) = \int (2\pi)^{-1/2}\sigma^{-1} \exp\{-(x_i-\theta_i)^2/2\sigma^2\}d\pi(\theta_i). \tag{2.3}$$

Unfortunately, $\Delta_k^{(\ell)}$ is rarely obtainable in closed form. Since the large k cases are of primary interest, however, reasonable approximations can be obtained by choosing

$$\ell = [yk], \tag{2.4}$$

where $0 < y \leq 1$ and $[v]$ denotes the nearest integer to v, and letting $k \to \infty$. The optimal fraction $(1-y)$ of observations to be truncated can then be determined in this limiting case, and should prove reasonable for more moderate k.

For use in the following, let

$$g(z) = m(z) + m(-z) \tag{2.5}$$

denote the marginal density of the $Z_i = |X_i|$, and observe from (2.3) that

$$g(z) \leq 2(2\pi)^{-1/2}\sigma^{-1} \equiv B. \tag{2.6}$$

Also, let $\alpha(y)$ denote the yth percentile of g, that is

$$\int_0^{\alpha(y)} g(z)dz = y. \tag{2.7}$$

Lemma 2.1. For $Z_{(\ell)}$ defined by (1.4) and ℓ as in (2.4),

$$Z_{(\ell)} \to \alpha(y) \text{ almost surely} \tag{2.8}$$

as $k \to \infty$.

Proof. See Rao (1973). ||

Theorem 2.1. For ℓ as in (2.4)

$$\lim_{k\to\infty} k^{-1}E^m\left[\frac{(\ell-2)^2}{\sum_{j=1}^k X_j^2 \wedge Z_{(\ell)}^2}\right] = y^2/\mu_2, \tag{2.9}$$

where

$$\begin{aligned}\mu_2 &= E^m[X_j^2 \wedge \alpha^2(y)] \\ &= 2\int_0^{\alpha(y)} x^2 m(x)dx + \alpha^2(y)(1-y).\end{aligned} \tag{2.10}$$

(Note that $\mu_2 < \infty$ if $0 < y < 1$.)

Proof. Given in the appendix. ||

Define,

$$r(y) = \frac{y^2}{2\int_0^{\alpha(y)} x^2 m(x)dx + \alpha^2(y)(1-y)} \quad . \tag{2.11}$$

Thus $r(y)$ is the asymptotic ($k\to\infty$) Bayes risk improvement (normalized by $\sigma^4 k^{-1}$) of the truncated James-Stein estimator over δ^0. This representation for $r(y)$ was essentially given for a normal prior π in Stein (1981). The following lemma will give the behavior of $r(y)$ near 0 and 1.

Lemma 2.2. If $r(y)$ is defined as in (2.11) then

(i) $\lim_{y\to 0} r(y) = 4[m(0)]^2$

and

$$\text{(ii)} \lim_{y\to 1} r(y) = \begin{cases} 1/v & \text{if } v < \infty \\ 0 & \text{if } v = \infty, \end{cases} \tag{2.12}$$

where v is the marginal variance of X_i.

Proof. Let $G(z)$ be the distribution function of $Z_i = |X_i|$. Then

$$G(z) = P(-z < X_i < z) = 2\int_0^z m(x)dx. \tag{2.13}$$

Define,

$$G^{-1}(u) = \inf\{z\colon G(z) \geq u\}. \tag{2.13}$$

Then clearly the $y\underline{th}$ fractile of G is given by

$$\alpha(y) = G^{-1}(y). \tag{2.15}$$

Hence,

$$\lim_{y\to 0} r(y) = \lim_{\alpha(y)\to 0} \frac{[G(\alpha(y))]^2}{2\int_0^{\alpha(y)} x^2 m(x)dx + \alpha^2(y)[1-G(\alpha(y))]}$$

$$= \lim_{\alpha(y)\to 0} \frac{4G(\alpha(y))m(\alpha(y))}{2\alpha(y)[1-G(\alpha(y))]} \text{ (by L'Hospital's rule)}$$

$$= \lim_{\alpha(y)\to 0} \frac{8[m(\alpha(y))]^2 + 4G(\alpha(y))m'(\alpha(y))}{2[1-G(\alpha(y))] - 4\alpha(y)m(\alpha(y))} \text{ (by L'Hospital's rule)}$$

$$= 4[m(0)]^2,$$

which completes the proof of part (i) of the lemma.

To prove part (ii), notice that as $y \to 1$, $\alpha(y) \to \infty$. Now clearly,

$$v = 2\int_0^\infty x^2 m(x)dx > 2\int_0^{\alpha(y)} x^2 m(x)dx + [\alpha(y)]^2[1-G(\alpha(y))]$$

$$= 2\int_0^{\alpha(y)} x^2 m(x)dx + [\alpha(y)]^2(1-y).$$

Therefore if $v < \infty$, then

$$\lim_{\alpha(y)\to\infty} [\alpha(y)]^2(1-y) = 0. \tag{2.16}$$

Using (2.16) in (2.11), part (ii) of the lemma follows. ||

Generally, $r(y)$ will be a concave function with a maximum occuring between 0 and 1. To give a feeling for the behavior of $r(y)$, we evalute the function when m is a t-distribution with p-degrees of freedom; that is, when the X_i have marginal density

$$m(x) = C_{p,\sigma}\left(1 + \frac{x^2}{p\sigma^2}\right)^{-(p+1)/2}, \quad p \geq 1, \tag{2.17}$$

where

$$C_{p,\sigma} = \frac{\Gamma((p+1)/2)}{\sqrt{p\pi}\,\Gamma(p/2)\sigma}.$$

Observe that

$$E(X_i) = 0 \text{ and } Var(X_i) = \frac{p}{p-2}\,\sigma^2,\ (p\geq 2),\ i=1,\ldots,k. \tag{2.18}$$

For the cases p = 1,2,3,4; r(y) is given as follows. (The proofs are given in the Appendix.)

<u>Case 1.</u> p = 1 (canchy marginal). Here

$$r(y) = y^2/[\sigma^2\{(2/\pi)\tan(\pi y/2) - y + (1-y)\tan^2(\pi y/2)\}]. \tag{2.19}$$

Also,

$$\lim_{y\to 0} r(y) = 4/\sigma^2\pi^2 \text{ and } \lim_{y\to 1} r(y) = 0.$$

<u>Case 2.</u> p = 2 (t-marginal with 2-degrees of freedom). Here

$$r(y) = y^2/[2\sigma^2\{\log(1+y)/(1-y^2)^{1/2} - y/(1+y)\}]. \tag{2.20}$$

Also

$$\lim_{y\to 0} r(y) = 1/2\sigma^2 \text{ and } \lim_{y\to 1} r(y) = 0.$$

Case 3. $p = 3$ (t-marginal with 3-degrees of freedom). Here

$$r(y) = y^2/[(6\sigma^2/\pi)\{\text{Arc tan } b(y) - b(y)/(1+b^2(y))\} + 3\sigma^2 b^2(y)(1-y)], \tag{2.21}$$

where $b(y) = \alpha(y)/\sigma\sqrt{3}$. Also

$$\sigma^2 \lim_{y\to 0} r(y) = 0.54 \text{ and } \sigma^2 \lim_{y\to 1} r(y) = 0.33.$$

Case 4. $p = 3$ (t-marginal with 4-degrees of freedom). Here

$$r(y) = y^2/[2\sigma^2\{b(y)/\{b^2(y) + 4\}^{1/2}\}^3 + \sigma^2 b^2(y)(1-y)], \tag{2.22}$$

where $\ell(y) = \alpha(y)/\sigma$. Also

$$\lim_{y\to 0} r(y) = 0.56 \text{ and } \lim_{y\to 1} r(y) = 0.5.$$

Table 1 presents values of $r(y)$ for each of these cases, and for the situation of a $N(0,1)$ marginal, which would arise from the usual empirical Bayes normal prior. (The marginal variance $(\sigma^2 + \tau^2)$ in this case would simply be a scale factor for $r(y)$.) As intuition would have suggested, the fatter the tail of the marginal (or the prior), the smaller y should be chosen. It is, of course, unlikely that explicit knowledge of the tail will be available, but Table 1 suggests that choosing y to be about .7 or .6 would be reasonable. The choice $y = .6$, for instance, leads to a Bayes risk about 13% worse than optimal at the extreme normal case, and about 20% worse than optimal at the extreme Cauchy case, a reasonable compromise between the two extremes. Of course, these results are asymptotic results as $k \to \infty$, and must be modified for smaller k. A choice of ℓ such as

$$\ell = 3 + [.7(k-3)] \tag{2.23}$$

seems reasonable for general use.

Table 1

Values of $r(y)$ For Cauchy, t_2, t_3, t_4 and normal marginals

y	Cauchy	t_2	t_3	t_4	Normal
0	.40	.50	.54	.56	.63
.1	.43	.53	.57	.59	.70
.2	.44	.55	.61	.63	.74
.3	.43	.57	.63	.65	.76
.4	.42	.58	.64	.68	.79
.5	.39	.57	.65	.69	.83
.6	.35	.56	.66	.70	.87
.7	.29	.53	.65	.71	.91
.8	.21	.49	.62	.69	.94
.9	.11	.40	.56	.67	.97
1	0	0	.33	.50	1

Remark 1. It should be mentioned that, if a specific fat-tailed prior is suspected, one might want to use an empirical Bayes estimator suitable for that prior. For example, the natural empirical Bayes estimator for a t-prior differs substantially from Stein-type estimators. The Stein type estimators have the advantage of being guaranteed to be minimax, however.

Remark 2. The truncated estimator given in (1.2) also has the advantage of sharply limiting the component risks of estimating each θ_i. While the usual James-Stein estimator has good overall risk $R(\theta,\delta)$, the risk of estimating individual θ_i, i.e.

$$R_i(\theta,\delta_i) = E_\theta(\theta_i - \delta_i(X))^2,$$

can be huge. Indeed, Efron and Morris (1972) show that this component risk can be as large as k/4, and suggest limited translation Stein-type estimators which have much smaller maximum component risk while maintaining good overall risk.

Finding the maximum component risk of $\delta^{(\ell)}$ seems quite difficult in general. An indication of its component behavior can be obtained, however, by looking at $R_1(\theta^*,\delta_1^{(\ell)})$ for $\theta^* = (|\theta|,0,\ldots,0)^t$ and large $|\theta|$. Although for $\delta^{(\ell)}$ this is not necessarily the least favorable configuration of θ in terms of maximum component risk (as it is for the James-Stein and Efron-Morris estimators), it should indicate the magnitude of the problem. Very large $|\theta|$ seem likely to be worst for $\delta^{(\ell)}$ (also in contrast to the James-Stein and Efron-Morris estimators).

Lemma 2.3. If $\ell = [yk]$, $0 < y < 1$, then

$$\lim_{k\to\infty} \lim_{|\theta|\to\infty} R_1(\theta^*, \delta^{(\ell)}) = 1 + y^2\alpha^2(y)/\mu_2^2 , \tag{2.24}$$

where $\phi(\alpha(y)) - \phi(-\alpha(y)) = y$, ϕ being the cumulative distribution function of the standard normal distribution, and

$$\mu_2 = E[X_i^2 \wedge \alpha^2(y)] = -(2/\pi)^{1/2}\alpha(y)\exp\{-\alpha^2(y)/2\} + y + \alpha^2(y)(1-y),$$

the expectation being taken assuming X_i is $N(0,1)$.

Proof. Clearly as $|\theta|\to\infty$ and when $\theta = \theta^*$,

$$\delta_1^{(\ell)}(X) \to X_1 - \frac{(\ell-2)Z_{(\ell)}}{Z_{(\ell)}^2 + \sum_{j=2}^{k} X_j^2 \wedge Z_{(\ell)}^2} ,$$

the last term being independent of X_1 with probability going to one, since X_1 will be truncated with probability approaching one. It can thus easily be shown that

$$\lim_{|\theta|\to\infty} R_1(\theta^*,\delta_1^{(\ell)}) = E_{\theta^*}(X_1 - \frac{(\ell-2)Z_{(\ell)}}{Z_{(\ell)}^2 + \sum_{j=2}^{k} X_j^2 \wedge Z_{(\ell)}^2} - \theta_1)^2$$

$$= 1 + E_{\theta^{**}}[\frac{(\ell-2)^2 Z_{(\ell)}^2}{\{Z_{(\ell)}^2 + \sum_{j=2}^{k} X_j^2 \wedge Z_{(\ell)}^2\}^2}],$$

where θ^{**} is the (k-1) dimensional zero vector. An argument similar to that of Theorem 2.1 then completes the proof of (2.24).

The calculation of μ_2 is straightforward. ||

A short table of the limiting component risks as a function of y is given below.

Table 2

Component Risks

y	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$1+y^2\alpha^2(y)/\mu_2^2$	1	1.6850	1.8577	1.9498	2.0980	2.2744	2.5010	2.8134	3.2773	4.1712

Not only are these values substantially better than the maximum component risk of k/4 for the usual James-Stein estimator, but they are only slightly worse than the maximum component risk of the Efron-Morris limited translation rules (for the choice of y leading to equivalent Bayes risk performance with respect to a normal prior). Of course the values in Table 2 are not necessarily the maximum component risk of $\delta^{(\ell)}$, but the indication is that $\delta^{(\ell)}$ has very satisfactory maximum component risk.

3. AN ADAPTIVE CHOICE OF THE TRUNCATION POINT

A very appealing possibility is to let the data select ℓ in the estimator (1.2). Since we are trying to maximize (2.2), the obvious method of selection is to choose that $\ell \geq 3$ (say ℓ^*) which maximizes

$$(\ell-2)^2 / \sum_{j=1}^{k} (X_j^2 \wedge Z_{(\ell)}^2). \tag{3.1}$$

In actual use of Stein-type estimators, the positive part versions should always be employed, so the suggested adaptive estimator is given componentwise by

$$\delta_i^*(X) = \left(1 - \frac{(\ell^*-2)\sigma^2 \min\{1, Z_{(\ell)}^*/|X_i|\}}{\sum_{j=1}^{k} X_j^2 \wedge Z_{(\ell)}^2}\right)^+ X_i, \tag{3.2}$$

where $a^+ = \max(a,0)$.

Theoretical analysis of this estimator is immensely difficult, due to the complicated dependence of ℓ^* on X. We did, however, perform a numerical study of $R(\theta,\delta^*)$ for various θ. Table 3 presents the risk (for simplicity σ^2 is taken to be equal to one) along a coordinate axis. The table only goes up to $|\theta| = 6$, since the risk is essentially constant beyond this point. (The first coordinate is always being truncated.) Observe that the risk of the usual James-Stein estimator would go to k as $|\theta| \to \infty$, so δ^* is performing very well indeed. Of course, this is the most favorable case for a truncated estimator, so in Table 4 the intuitively least favorable case, in which θ lies on the diagonal

Table 3

Values of $R(\theta,\delta^*)$ for $\theta = (|\theta|,0,\ldots,0)^t$

$\lvert\theta\rvert$ \ k	5	10	15	20	25	30	35	40	45	50
0	0.95	1.18	1.38	1.61	1.92	2.23	2.49	2.71	2.99	3.27
1	1.70	1.98	2.15	2.40	2.65	2.81	3.02	3.24	3.47	3.83
2	2.09	2.31	2.62	3.05	3.50	3.95	4.41	4.82	5.31	5.75
3	3.57	3.71	3.94	4.15	4.42	4.63	4.88	5.09	5.49	5.89
4	3.60	3.80	3.98	4.36	5.14	5.48	5.78	6.17	6.56	6.98
5	3.62	3.85	3.98	4.46	5.15	5.50	5.78	6.23	6.56	7.02
6	3.62	3.86	3.99	4.46	5.15	5.51	5.78	6.23	6.56	7.02

$(|\theta|k^{-1/2})(1,\ldots,1)^t$, is considered. (The normalization by k is done so that the resulting risks can be given as functions of $|\theta|$.) Table 4 provides strong evidence that δ^* is indeed minimax (even for small k) and shows that δ^* compares favorably with other Stein-type estimators even when all of the θ_i are similar.

These risks were calcualted by simulation, with between 1000 and 2000 random vectors being used for each value of $|\theta|$. The standard errors of the values found for the risk were found to be about .05.

Table 4

Values of $R(\theta,\delta^*)$ for $\theta = \frac{|\theta|}{\sqrt{k}}(1,1,\ldots,1)^t$

$\lvert\theta\rvert$ \ k	5	10	15	20	25	30	35	40	45	50
0	0.95	1.18	1.38	1.61	1.92	2.23	2.49	2.71	2.99	3.27
1	2.35	2.70	2.95	3.35	3.70	3.91	4.15	4.45	4.70	4.95
2	3.21	3.62	3.92	4.31	4.65	4.97	5.37	5.71	5.28	6.76
3	4.07	4.89	6.24	7.13	8.04	9.01	9.94	10.82	11.75	12.66
4	4.39	5.59	6.97	8.66	10.13	11.65	13.11	14.43	15.78	16.92
5	4.57	6.47	8.46	10.51	12.62	14.57	16.67	18.62	20.42	22.35
6	4.68	7.10	9.39	11.57	13.67	15.60	17.77	19.99	21.94	24.96
7	4.75	7.78	9.91	12.37	15.74	17.89	21.04	24.85	27.83	29.54
8	4.80	7.99	11.41	14.00	17.35	21.10	24.62	27.05	30.09	32.05
9	4.83	9.10	12.84	16.60	20.34	23.38	26.45	29.74	32.05	34.29
10	4.86	9.23	13.16	17.03	20.91	23.96	27.24	30.33	33.19	36.50
11	4.88	9.34	13.37	17.42	21.37	24.73	28.18	31.07	34.09	37.15
12	4.90	9.43	13.58	17.77	21.85	25.35	29.03	32.18	35.27	38.56
13	4.91	9.51	13.76	18.04	22.25	25.93	29.76	33.13	36.36	39.28
14	4.92	9.57	13.91	18.25	22.59	26.40	30.37	33.92	37.29	40.41
15	4.93	9.62	14.03	18.45	22.86	26.81	30.87	34.55	38.09	41.40
16	4.93	9.66	14.13	18.62	23.10	27.15	31.32	35.12	38.79	42.25
17	4.94	9.69	14.22	18.76	23.30	27.43	31.70	35.60	39.39	43.14
18	4.95	9.72	14.29	18.88	23.47	27.68	32.02	36.01	39.91	43.62
19	4.95	9.74	14.35	18.98	23.61	27.90	32.29	36.38	40.37	44.18
20	4.96	9.76	14.41	19.07	23.74	28.09	32.54	36.70	40.77	44.67

APPENDIX

Proof of Theorem 2.1.

Define

$$U_{j,k} = X_j^2 \wedge Z_{(\ell)}^2$$

and

$$V_j = X_j^2 \wedge \alpha^2(y).$$

Observe that

$$|U_{j,k} - V_j| \le |Z_{(\ell)}^2 - \alpha^2(y)|, \tag{A1}$$

and, using (2.8),

$$|Z_{(\ell)}^2 - \alpha^2(y)| \to 0 \text{ almost surely as } k \to \infty. \tag{A2}$$

Now, by the strong law of large numbers,

$$\frac{1}{k} \sum_{j=1}^{k} V_j \to \mu_2 \text{ almost surely as } k \to \infty. \tag{A3}$$

Combining (A1) through (A3), it follows that

$$\frac{1}{k}\sum_{j=1}^{k} U_{j,k} \to \mu_2 \text{ almost surely as } k \to \infty. \tag{A4}$$

Next choose $0 < \lambda < \alpha^2(y)$ such that $e\lambda B/y < 1$ where B is given in (2.6). Then using the fact

$$\sum_{j=1}^{k} X_j^2 \wedge Z_{(\ell)}^2 \geq (k-\ell)Z_{(\ell)}^2$$

it is clear that, on the set $\{z_{(\ell)}^2 > \lambda\}$,

$$\frac{k}{\sum_{j=1}^{k} u_{j,k}} \leq \frac{k}{(k-\ell)z_{(\ell)}^2} \leq \frac{k}{\lambda(k-\ell)} \leq c/\lambda \tag{A5}$$

where c is a constant $\geq k/(k-\ell)$ for all k. Using (A3) through (A5) and the dominated convergence theorem,

$$\int_{z_{(\ell)}^2>\lambda} \frac{k}{\sum_{j=1}^{k} u_{j,k}} \prod_{j=1}^{k} g(z_j)dz_j \to \mu_2^{-1} P(Z_{(\ell)}^2 > \lambda). \tag{A6}$$

Now, on the set $\{z_{(\ell)}^2 < \lambda\}$,

$$\int_{z_{(\ell)}^2<\lambda} \frac{k}{\sum_{j=1}^{k} u_{j,k}} \prod_{j=1}^{k} g(z_j)dz_j \leq \int_{z_{(\ell)}^2<\lambda} \frac{c}{z_{(\ell)}^2} \prod_{j=1}^{k} g(z_j)dz_j. \tag{A7}$$

Furthermore,

$$\int_{z^2_{(\ell)}<\lambda} \frac{1}{z^2_{(\ell)}} \prod_{j=1}^{k} g(z_j)dz_j$$

$$= \frac{k!}{(\ell-1)!(k-\ell)!} \int_0^\lambda z^{-2}[G(z)]^{\ell-1}[1-G(z)]^{k-\ell}g(z)dz$$

$$\leq \frac{k!}{(\ell-1)!(k-\ell)!} \frac{(\lambda B)^{\ell-2}}{\ell-2} B^2$$

$$= \frac{k!}{(k-\ell)!k^\ell} \frac{k^\ell}{\ell!} \frac{\ell}{\ell-2} (\lambda B)^{\ell-2}B^2. \tag{A8}$$

Now using Stirling's approximation for $\ell!$ and the fact $e\lambda B/y < 1$, (A8) goes to zero as $k \to \infty$. Combining (A6) through (A8) and letting $\lambda \to 0$, shows that

$$E^m\left[\frac{k}{\sum_{j=1}^{k} U_{j,k}} \right] \to \frac{1}{\mu^2} \text{ as } k \to \infty. \tag{A9}$$

Thus equation (2.9) is immediate. The representation for μ_2 follows from the definition of $\alpha(y)$. ||

Proof of (2.19)

Substituting $p = 1$ in (2.17) it follows from (2.13) that

$$G(z) = (2/\pi) \text{ arc tan } (z/\sigma)$$

and

$$\int_0^{\alpha(y)} x^2 m(x)dx = (1/\pi\sigma) \int_0^{\alpha(y)} x^2(1 + x^2/\sigma^2)^{-1}dx.$$

Using (2.15) it follows that

$$y = G(\alpha(y)) = (2/\pi) \text{ arc tan } \alpha(y)/\sigma.$$

Thus

$$Z(y) = \sigma \tan(\pi y/2). \tag{A16}$$

Now to evaluate r(y) as defined in (2.11), clearly

$$2 \int_0^{\alpha(y)} x^2 m(x)dx = (2/\pi) \int_0^{\alpha(y)/\sigma} \sigma^2 x^2/(1+x^2)dx$$

$$= (2\sigma^2/\pi)\{\alpha(y)/\sigma - \text{arc tan}(\alpha(y)/\sigma)\}$$

$$= \sigma^2\{(2/\pi)\tan(\pi y/2) - y\}. \tag{A17}$$

Substituting (A16) and (A17) in (2.11) it follows that

$$r(y) = y^2/\sigma^2\{(2/\pi)\tan(\pi y/2) - y + (1-y)\tan^2(\pi y/2)\}.$$

Finally using Lemma 2.2, it is clear that

$$\lim_{y\to 0} r(y) = 4/\sigma^2\pi^2 \quad \text{and} \quad \lim_{y\to 1} r(y) = 0.$$

Proof of (2.20).

Substituting p = 2 in (2.17) it follows that

$$G(z) = \int_0^{\arctan(z/\sqrt{2}\,\sigma)} \cos\theta d\theta = \sin \arctan(z/\sigma\sqrt{2})$$

$$= z/(z^2 + 2\sigma^2)^{1/2},$$

and

$$\int_0^{\alpha(y)} X^2 m(x)dx = \sigma^2 \int_0^{\arctan\alpha(y)/\sigma\sqrt{2}} \tan^2\theta\cos\theta d\theta$$

$$= \sigma^2 \int_0^{\arctan\alpha(y)/\sigma\sqrt{2}} (\sec\theta - \cos\theta)d\theta, \tag{A18}$$

where $\alpha(y) = G^{-1}(y)$. Thus

$$y = G(\alpha(y)) = \alpha(y)/(\alpha^2(y) + 2\sigma^2)^{1/2},$$

and hence

$$\alpha(y) = y\sigma\sqrt{2}/(1-y^2)^{1/2}. \tag{A19}$$

Clearly, it follows from (A18) that

$$\int_0^{\alpha(y)} x^2 m(x)dx = \sigma^2[\ell n\{(1+y)/(1-y^2)^{1/2}\} - y]. \tag{A20}$$

Substituting (A19) and (A20) in (2.11), it follows that

$$r(y) = y^2/2\sigma^2[\ell n\{(1+y)/(1-y^2)^{1/2}\} - y + 2(1-y)\sigma^2 y^2/(1-y^2)]$$

$$= y^2/2\sigma^2[\ell n\{(1+y)/(1-y^2)^{1/2}\} - y/(1+y)].$$

Finally, using Lemma 2.2,

$$\lim_{y \to 0} r(y) = 1/2\sigma^2 \quad \text{and} \quad \lim_{y \to 1} r(y) = 0.$$

<u>Proof of (2.21).</u>

Substituting p = 3, in (2.17), it is clear that

$$G(z) = (2/\pi\sigma) \int_0^{\arctan(z/\sigma\sqrt{3})} \cos^2\theta d\theta$$

$$= (2/\pi)[\arctan(z/\sigma\sqrt{3}) + \sigma z\sqrt{3}/(z^2 + 3\sigma^2)]$$

and

$$\int_0^{\alpha(y)} x^2 m(x)dx = (6/\pi) \int_0^{\arctan\alpha(y)/\sigma\sqrt{3}} \sin^2\theta d\theta, \tag{A21}$$

where $\alpha(y) = G^{-1}(y)$. Therefore

$$y = G(\alpha(y)) = (2/\pi)[\text{arc tan}(\alpha(y)/\sigma\sqrt{3}) + \sigma\alpha(y)\sqrt{3}/(\alpha^2(y) + 3\sigma^2)]$$

$$= (2/\pi)[\text{arc tan } (\alpha(y)/\sigma)/\sqrt{3} + \{\alpha(y)\sqrt{3}/\sigma\}/\{(\alpha(y)/\sigma)^2 + 3\}]$$

$$\equiv H(\alpha(y)/\sigma) \text{ (definition).}$$

Then

$$\alpha(y) = \sigma H^{-1}(y). \tag{A22}$$

It follows from (A9) that

$$\int_0^{\alpha(y)} x^2 m(x)dx = (3\sigma^2/\pi)[\text{arc tan}(\alpha(y)/\sigma\sqrt{3}) - \alpha(y)\sigma\sqrt{3}/(\alpha^2(y) + 3\sigma^2)].$$

Clearly from (A22), $\alpha(y)/\sigma$ is independent of σ. Therefore,

$$\int_0^{\alpha(y)} x^2 m(x)dx = (3\sigma^2/\pi)[\text{arc tan}b(y) - b(y)/(1 + b^2(y))], \tag{A23}$$

where $b(y)$ is defined in the statement of Case 3 and will be obtained from

$$y = (2/\pi)[\text{arc tan}b(y) + b(y)/(1 + b^2(y))].$$

The proof is complete using (A23) in (2.11). Finally, using Lemma 2.2, it follows that

$$\sigma^2 \lim_{y\to 0} r(y) = 16/3\pi^2 = 0.54 \quad \text{and} \quad \sigma^2 \lim_{y\to 1} r(y) = 1/v = 0.33.$$

Proof of (2.22).

Substituting p = 4 in (2.17), it follows that

$$G(z) = (3\sigma/2) \int_0^{\arctan(z/2\sigma)} \cos^3\theta d\theta$$

$$= (3/2) \int_0^{\arctan(z/2\sigma)} (\cos 3\theta + 3\cos\theta)/4 d\theta$$

$$= [(3\sin\theta - \sin^3\theta/2]^{\arctan(z/2\sigma)}$$

$$= z(z^2 + 6\sigma^2)/(z^2 + 4\sigma^2)^{3/2},$$

and

$$\int_0^{\alpha(y)} x^2 m(x)dx = 3\sigma^2 \int_0^{\tan^{-1}\alpha(y)/2\sigma} (\cos\theta - \cos^2\theta)d\theta, \tag{A24}$$

where $\alpha(y) = G^{-1}(y)$. Therefore

$$y = G(\alpha(y)) = \alpha(y)\{\alpha^2(y) + 6\sigma^2\}/\{\alpha^2(y) + 4\sigma^2\}^{3/2}$$

$$= [(\alpha(y)/\sigma)\{(\alpha(y)/\sigma)^2 + 6\}]/[(\alpha(y)/\sigma)^2 + 4]^{3/2}$$

$$\equiv H(\alpha(y)/\sigma) \text{ (definition).}$$

Thus,

$$\alpha(y) = \sigma H^{-1}(y). \tag{A25}$$

It follows from (A24) that

$$\int_0^{\alpha(y)} x^2 m(x)dx = \sigma^2[3\sin \text{arc} \tan(\alpha(y)/2\sigma) - 3\int_0^{\text{arc} \tan\alpha(y)/2\sigma} \cos^3\theta d\theta]$$

$$= \sigma^2\sin^3 \text{arc} \tan(\alpha(y)/2\sigma) = \sigma^2\{\alpha(y)/(\alpha^2(y) + 4\sigma^2)^{1/2}\}^3. \tag{A26}$$

Clearly from (A25), $\alpha(y)/\sigma$ is independent of σ. Then from (A26)

$$\int_0^{\alpha(y)} x^2 m(x)dx = \sigma^2 \{b(y)/(b^2(y) + 4)^{1/2}\}^3, \tag{A27}$$

where b(y) is defined in the statement of case 4 and will be obtained from

$$y = b(y)\{b^2(y) + 6\}/\{b^2(y) + 4\}^{3/2}.$$

Now (2.22) follows using (A27) in (2.11). Finally, using Lemma 2.2, it follows that

$$\lim_{y\to 0} r(y) = 9/16 = 0.56 \quad \text{and} \quad \lim_{y\to 1} r(y) = 1/v = 0.5.$$

BIBLIOGRAPHY

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

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

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

Chung, K.L. (1968). A Course in Probability Theory. Harcourt, Brace & World, Inc.

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

Berger, J. and Dey, D.K. (1980). Combining coordinates in simultaneous estimation of normal means. To appear in J. Statist. Planning and Inference.

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

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

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

Hudson, M. (1974). Empirical Bayes Estimation. Technical Report #58, Stanford University.

James, W. and Stein, C. (1960). Estimation with quadratic loss. Proc. Fourth Berkeley Symposium Math. Stat. Prob. 1, 316-379. University of California Press.

Rao, C.R. (1973). Linear Statistical Inference and Its Applications. Second Edition, John Wiley and Sons.

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

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

On Truncation of Shrinkage Estimators in Simultaneous Estimation of Normal Means

DIPAK K. DEY and JAMES O. BERGER*

In estimating a multivariate normal mean $\theta = (\theta_1, \ldots, \theta_k)'$ under sum of squares error loss, it is well known that Stein estimators improve upon the usual estimator (in terms of expected loss) if $k \geq 3$. The improvement obtained is significant, however, only if the θ_i are fairly close to the point towards which the Stein estimator shrinks. When extreme θ_i are likely (such as when the θ_i are thought to arise from a possibly heavy-tailed prior distribution), the standard Stein estimators may offer little improvement over the usual estimator. Stein (1981) proposed a limited translation Stein estimator to correct this deficiency. This estimator is analyzed herein for a number of heavy-tailed prior distributions. An adaptive version of the estimator is also discussed.

KEY WORDS: James-Stein estimator; Stein's truncation; Heavy-tailed prior; Adaptive estimator.

1. INTRODUCTION

Let $X = (X_1, \ldots, X_k)'$ have a k-variate normal distribution with mean vector $\theta = (\theta_1, \ldots, \theta_k)'$ and covariance matrix $\sigma^2 I$. For convenience, σ^2 will be assumed to be known but, as usual in symmetric Stein estimation, results remain essentially unchanged if σ^2 is replaced by a suitable estimate, say $S^2/(m + 2)$, where S^2/σ^2 has a chi-squared distribution with m degrees of freedom (df). It is desired to estimate θ, using an estimator $\delta(X) = (\delta_1(X), \ldots, \delta_k(X))'$, under sum of squares error loss $L(\theta, \delta) = \sum_{i=1}^{k} (\theta_i - \delta_i)^2$. An estimator will be evaluated by its frequentist risk or expected loss, to be denoted by $R(\theta, \delta) = E_\theta[L(\theta, \delta(X))]$. James and Stein (1961) showed that the usual estimator $\delta^0(X) = X$ is inadmissible when $k \geq 3$ and that the estimator

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

has uniformly smaller risk than δ^0. For recent references concerning Stein estimation see Stein (1981), Berger (1982), and Morris (1983).

A key feature of any Stein-type estimator is that it has risk significantly better than that of δ^0 only in a relatively small region (or subspace) of the parameter space (see Berger 1980,1982, for discussion). For Stein estimation to result in significant improvement, therefore, one must carefully select an estimator designed to do well in the region in which θ is thought likely to lie. This is essentially done by finding a Stein-type estimator that shrinks toward the desired region. If any of the θ_i happen to fall substantially outside this region, the usual Stein estimators will collapse back to X and offer little improvement over δ^0. The estimator in (1.1), for example, is designed to do well for θ near zero, and if any of the θ_i are far from zero, then $X'X$ will be large and δ will be approximately X.

The above problem was noted by Stein (1981), who considered a modification of the estimator in (1.1) to partially alleviate the difficulties. He proposed the estimator defined coordinate-wise by

$$\delta_i^{(l)}(X) = \left(1 - \frac{(l-2)\sigma^2 \min\{1, Z_{(l)}/|X_i|\}}{\sum_{j=1}^{k} X_j^2 \wedge Z_{(l)}^2}\right)X_i, \tag{1.2}$$

where l is a large fraction of k, $a \wedge b$ denotes the minimum of a and b, $Z_i = |X_i|$, $i = 1, \ldots, k$, and $Z_{(1)} < Z_{(2)} < \ldots < Z_{(l)} < \ldots < Z_{(k)}$ are the order statistics of $Z_1, \ldots, Z_k$. Stein (1981) proved that this estimator is minimax if $l \geq 3$. The estimator provides a reasonable solution to the extreme θ_i problem, as is indicated by the observation that $\sum_{j=1}^{k} X_j^2 \wedge Z_{(l)}^2$ is fairly small even if $(k - l)$ of the θ_i are very extreme.

In Section 2, we evaluate versions of the estimator (1.2) for the symmetric situation. Since the symmetric situation mainly occurs when the θ_i are felt to arise from a common prior distribution (the empirical Bayes situation discussed, for example, in Efron and Morris 1972), this evaluation will be in terms of Bayes risk $r(\pi, \delta) = E^\pi[R(\theta, \delta)]$ for a variety of prior distributions π. The goal is to determine sensible choices of the truncation point l. We also indicate that the resulting estimator has quite small maximum component risk, a desirable feature as discussed in Efron and Morris (1972).

In Section 3, the estimator in (1.2) is considered when l is chosen adaptively by the data, an appealing possibility that obviates the necessity to consider the prior distribution of the θ_i.

In this article our main concern is with results for large dimension k. Extreme θ_i are clearly more of a danger when k is large. Also, when k is small the "loss" of di-

* Dipak K. Dey is Assistant Professor, Department of Mathematics, Texas Tech University, Lubbock, TX 79409. James O. Berger is Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907. This research was supported by the National Science Foundation under Grants MCS 78-02300 and MCS 81-01670A1. The authors are grateful to the referees and the Associate Editor for valuable comments on the earlier draft.

Reprinted from: © Journal of the American Statistical Association
December 1983, Volume 78, Number 384
Theory and Methods Section

mensions in going from k to l can be quite harmful, so that truncation is less appealing.

2. OPTIMAL CHOICE OF THE TRUNCATION POINT

The estimator (1.1) is the empirical Bayes estimator if $X \sim N_k(\theta, \sigma^2 I_k)$ and $\theta \sim N_k(0, \tau^2 I_k)$, where τ^2 is unknown (see Efron and Morris 1972). In general it is a reasonable estimator if the θ_i are thought to be independent realizations from a common symmetric prior distribution π having median zero. If the common prior is symmetric about its median $\mu \neq 0$, the estimator

$$\delta(X) = X - \frac{(k-2)\sigma^2(X - (\mu, \ldots, \mu)')}{\sum_{i=1}^{k} (X_i - \mu)^2} \quad (2.1)$$

would be appropriate. If μ is unknown (the realistic case), replacing μ in (2.1) by $\bar{X} = k^{-1} \sum_{i=1}^{k} X_i$ (or some more robust estimate if π could have heavy tails) and $(k-2)$ by $(k-3)$ would suffice. We will be considering mainly large k situations, and hence can assume that μ is known, and, without loss of generality, is zero.

Since the estimator (1.2) is minimax (for $l \geq 3$ and sum of squares error loss), the choice of l should be based on the overall expected gain from use of the estimator. The most reasonable measure of this average gain is the improvement in Bayes risk of the estimator over that of δ^0. Using Stein's unbiased estimator of risk, this improvement in Bayes risk can easily be shown (as in Berger and Dey 1983) to be

$$\Delta_k^{(l)} = r(\pi, \delta^0) - r(\pi, \delta^{(l)})$$

$$= \sigma^4 E^m \left[\frac{(l-2)^2}{\sum_{j=1}^{k} X_j^2 \wedge Z_{(l)}^2} \right], \quad (2.2)$$

where E^m stands for expectation under the marginal density of X. Note that the X_i are marginally independent (since the θ_i are) with common marginal density

$$m(x_i) = \int (2\pi)^{-1/2}\sigma^{-1} \exp\{-(x_i - \theta_i)^2/2\sigma^2\} d\pi(\theta_i). \quad (2.3)$$

Unfortunately, $\Delta_k^{(l)}$ is rarely obtainable in closed form. Since the large k cases are of primary interest, however, reasonable approximations can be obtained by choosing $l = [yk]$, where $0 < y \leq 1$ and $[v]$ denotes the nearest integer to v, and letting $k \to \infty$. The optimal fraction $(1 - y)$ of observations to be truncated can then be determined in this limiting case and should prove reasonable for more moderate k.

For use in the following, let $\alpha(y)$ denote the yth percentile of the marginal density of $|X_i|$, that is, $\int_{-\alpha(y)}^{\alpha(y)} m(x)dx = y$.

Theorem 2.1. For $l = [yk]$ and $0 < y \leq 1$,

$$\lim_{k\to\infty} k^{-1} E^m \left[(l-2)^2 \Big/ \sum_{j=1}^{k} X_j^2 \wedge Z_{(l)}^2 \right] = y^2/\mu_2, \quad (2.4)$$

where

$$\mu_2 = E^m[X_j^2 \wedge \alpha^2(y)]$$

$$= 2\int_0^{\alpha(y)} x^2 m(x)dx + \alpha^2(y)(1-y). \quad (2.5)$$

(Note that $\mu_2 < \infty$ if $0 < y < 1$.)

Proof. Given in the Appendix.

Define,

$$r(y) = \frac{y^2}{2\int_0^{\alpha(y)} x^2 m(x)dx + \alpha^2(y)(1-y)}. \quad (2.6)$$

Thus $r(y)$ is the asymptotic $(k \to \infty)$ Bayes risk improvement (normalized by $\sigma^4 K^{-1}$) of the truncated James-Stein estimator over δ^0. This representation for $r(y)$ was essentially given for a normal prior π in Stein (1981). Lemma 2.1 indicates the behavior of $r(y)$ near 0 and 1. The lemma can be established by application of L'Hospital's rule (see Dey 1980 for details).

Lemma 2.1. If $r(y)$ is defined as in (2.6) then

(i) $\lim_{y\to 0} r(y) = 4[m(0)]^2$

and

(ii) $\lim_{y\to 1} r(y) = 1/v \quad$ if $v < \infty$

$= 0 \quad$ if $v = \infty$,

where v is the marginal variance of X_i.

Generally, $r(y)$ will be a concave function with a maximum occurring between 0 and 1. To give a feeling for the behavior of $r(y)$, we evaluate the function when m is a scaled t distribution with p df, that is, when the X_i have marginal density

$$m(x) = C_{p,\rho}\left(1 + \frac{x^2}{p\rho^2}\right)^{-(p+1)/2}, \ p \geq 1, \quad (2.7)$$

where $C_{p,\rho} = \Gamma((p+1)/2)/[\sqrt{p\pi}\,\Gamma(p/2)\rho]$.

Densities of this form are convenient to work with and provide a suitably wide range of heavy-tailed distributions. They are, however, not necessarily true marginals, in that there is probably no prior π that when used in (2.3) results in (2.7). A heavy-tailed prior will have essentially the same marginal tail, however; hence, to obtain a rough idea of the effect of heavy tails, consideration of the m in (2.7) should suffice. If a specific heavy-tailed prior were of interest, $r(y)$ could be calculated numerically and minimized over y.

For the densities in (2.7) it is easy to see that $r(y) = \rho^2 r_1(y)$, where $r_1(y)$ refers to $r(y)$ for $\rho^2 = 1$. Thus the scale factor ρ is irrelevant for determination of the optimal y. (This reinforces the idea that it is mainly the heaviness of the tail of the prior, and hence m, that is of concern.) For the cases $p = 1, 2, 3, 4$, and ∞ (which corresponds to a normal prior and hence normal m), $r_1(y)$ is as follows. (The derivations of these can be found in Dey 1980.)

(i) $p = 1$ (Cauchy marginal). Here

$$r_1(y) = y^2/[(2/\pi)\tan(\pi y/2) - y + (1 - y)\tan^2(\pi y/2)].$$

(ii) $p = 2$ (t marginal with 2 df). Here

$$r_1(y) = (.5)y^2/[\log\{(1 + y)/(1 - y^2)^{1/2}\} - y/(1 + y)].$$

(iii) $p = 3$ (t marginal with 3 df). Here

$$r_1(y) = y^2/[(6/\pi)\{\text{Arctan } b(y) - b(y)/(1 + b^2(y))\} + 3b^2(y)(1 - y)],$$

where $b(y) = \alpha(y)/\sqrt{3}$.

(iv) $p = 4$ (t marginal with 4 df). Here

$$r_1(y) = y^2/[2\{\alpha(y)/(\alpha^2(y) + 4)^{1/2}\}^3 + \alpha^2(y)(1 - y)].$$

(v) $p = \infty$ (normal marginal). Here

$$r_1(y) = y^2/[-(2/\pi)^{1/2}\alpha(y)\exp\{-\alpha^2(y)/2\} + y + \alpha^2(y)(1 - y)].$$

These five functions are graphed in Figure 1. As intuition would suggest, the heavier the tail of the marginal (or the prior), the smaller y should be chosen. It is, of course, unlikely that explicit knowledge of the tail will be available, but Figure 1 suggests that choosing y to be about .7 or .6 would be reasonable. The choice $y = .6$, for instance, leads to a Bayes risk about 13% worse than optimal at the extreme normal case, and about 20% worse than optimal at the extreme Cauchy case, which is a reasonable compromise between the two extremes. Of course, these results are asymptotic results as $k \to \infty$ and must be modified for smaller k. A choice of l such as

$$l = 3 + [.7(k - 3)] \tag{2.8}$$

seems reasonable for general use.

Remark 1. It should be mentioned that, if a *specific* heavy-tailed functional form for the prior is suspected,

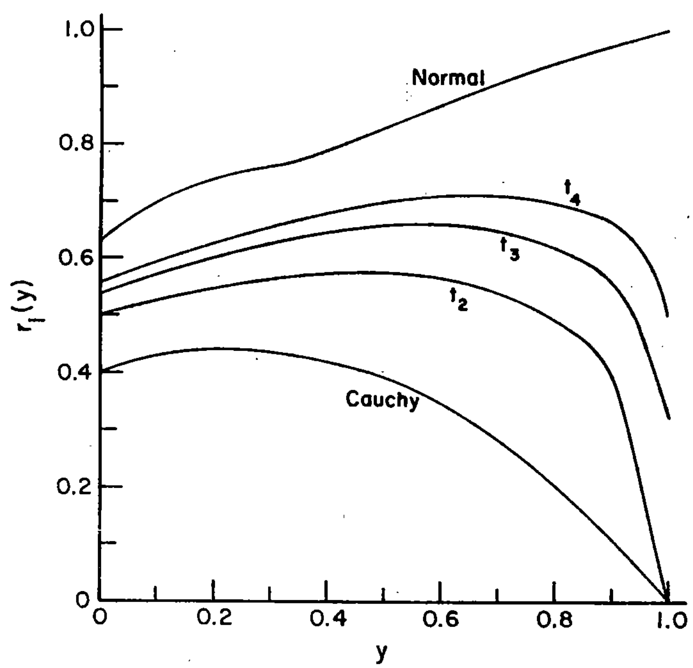

Figure 1. $r_1(y)$ for Various Marginals.

one might want to use an empirical Bayes estimator suitable for that prior instead of (1.2). For example, the natural empirical Bayes estimator for the class of t priors (with a given df) differs substantially from Stein-type estimators. The Stein-type estimators have the advantage of being guaranteed to be minimax, however, which may be an important consideration for non-Bayesians.

Remark 2. The truncated estimator given in (1.2) also has the advantage of sharply limiting the component risks of estimating each θ_i. While the usual James-Stein estimator has good overall risk $R(\theta, \delta)$, the risk of estimating individual θ_i, that is, $R_i(\theta, \delta_i) = E_\theta(\theta_i - \delta_i(X))^2$, can be huge. Indeed, Efron and Morris (1972) show that this component risk can be as large as $k/4$ and suggest limited translation Stein-type estimators that have much smaller maximum component risk while maintaining good overall risk.

Finding the maximum component risk of $\delta^{(l)}$ seems quite difficult in general. An indication of its component behavior can be obtained, however, by looking at $R_1(\theta^*, \delta_1^{(l)})$ for $\theta^* = (|\theta|, 0, \ldots, 0)'$ and large $|\theta|$. Although for $\delta^{(l)}$ this is not necessarily the least favorable configuration of θ in terms of maximum component risk (as it is for the James-Stein and Efron-Morris estimators), it should indicate the magnitude of the problem. For this configuration, very large $|\theta|$ seem likely to be worst for $\delta_1^{(l)}$ (also in contrast to the James-Stein and Efron-Morris estimators). This is because, as $|\theta|$ increases, so does the probability that $\delta_1^{(l)}$ is far from X_1 (the cause of a large component risk).

Lemma 2.2. If $l = [yk]$, $0 < y < 1$, then

$$\bar{R}_1 \equiv \lim_{k\to\infty} \lim_{|\theta|\to\infty} R_1(\theta^*, \delta_1^{(l)}) = 1 + y^2\alpha^2(y)/\mu_2^2, \tag{2.9}$$

where $\phi(\alpha(y)) - \phi(-\alpha(y)) = y$, ϕ being the cumulative distribution function of the standard normal distribution, and

$$\mu_2 = E[X_i^2 \wedge \alpha^2(y)] = -(2/\pi)^{1/2}\alpha(y)\exp\{-\alpha^2(y)/2\} + y + \alpha^2(y)(1 - y),$$

the expectation being taken assuming X_i is $N(0, 1)$.

Proof. Clearly as $|\theta| \to \infty$ and when $\theta = \theta^*$,

$$\delta_1^{(l)}(X) \to X_1 - \frac{(l - 2)Z_{(l)}}{Z_{(l)}^2 + \sum_{j=2}^{k} X_j^2 \wedge Z_{(l)}^2},$$

the last term being independent of X_1 with probability going to one, since X_1 will be truncated with probability approaching one. It can thus easily be shown that

$$\lim_{|\theta|\to\infty} R_1(\theta^*, \delta_1^{(l)}) = E_{0^*}\left(X_1 - \frac{(l - 2)Z_{(l)}}{Z_{(l)}^2 + \sum_{j=2}^{k} X_j^2 \wedge Z_{(l)}^2} - \theta_1\right)^2$$

$$= 1 + E_{\theta^{**}}\left[\frac{(l-1)^2 Z_{(l)}^2}{\left\{Z_{(l)}^2 + \sum_{j=2}^{k} X_j^2 \wedge Z_{(l)}^2\right\}^2}\right],$$

where θ^{**} is the $(k-1)$-dimensional zero vector. An argument similar to that of Theorem 2.1 then completes the proof of (2.9). The calculation of μ_2 is straightforward.

Table 1 gives the limiting component risks as a function of y. Not only are these values substantially better than the maximum component risk of $k/4$ for the usual James-Stein estimator, but they are only slightly worse than the maximum component risk of the Efron-Morris limited translation rules (for the choice of y leading to equivalent Bayes risk performance with respect to a normal prior). Of course, the values in Table 1 are not necessarily the maximum component risk of $\delta^{(l)}$, but the indication is that $\delta^{(l)}$ has very satisfactory maximum component risk.

3. AN ADAPTIVE CHOICE OF THE TRUNCATION POINT

A very appealing possibility is to let the data select l in the estimator (1.2). Since we are trying to maximize (2.2), the obvious method of selection is to choose that $l \geq 3$ (say l^*), which maximizes

$$(l-2)^2 \Big/ \sum_{j=1}^{k} (X_j^2 \wedge Z_{(l)}^2). \tag{3.1}$$

In actual use of Stein-type estimators, the positive part versions should always be employed, so the suggested adaptive estimator is given component-wise by

$$\delta_i^*(X) = \left(1 - \frac{(l^*-2)\sigma^2 \min\{1, Z_{(l^*)}/|X_i|\}}{\sum_{j=1}^{k} X_j^2 \wedge Z_{(l^*)}^2}\right)^+ X_i, \tag{3.2}$$

where $a^+ = \max(a, 0)$.

Theoretical analysis of this estimator is immensely difficult, due to the complicated dependence of l^* on X. We did, however, perform a numerical study of $R(\theta, \delta^*)$ for various θ. Table 2 presents the risk (for simplicity σ^2 is taken to be equal to one) along a coordinate axis. The table only goes up to $|\theta| = 6$, since the risk is essentially constant beyond this point. (The first coordinate is always being truncated.) Observe that the risk of the usual James-Stein estimator would go to k as $|\theta| \to \infty$, so δ^* is performing very well indeed. Of course, this is the most favorable case for a truncated estimator, thus in Table 3 the intuitively least favorable case, in which θ lies on the diagonal $(|\theta| k^{-1/2})(1, \ldots, 1)'$, is considered. (The normalization by k is done so that the resulting risks can be given as functions of $|\theta|$.) Table 3 provides strong evidence that δ^* is indeed minimax (even for small k) and shows that δ^* compares favorably with other Stein-type estimators even when all of the θ_i are similar.

These risks were calculated by simulation, with between 1,000 and 2,000 random vectors being used for each value of $|\theta|$. The standard errors of the values found for the risk were found to be about .05.

Table 1. Component Risks ($\bar{R}_1$)

y	$\bar{R}_1$	y	$\bar{R}_1$
0	1	.5	2.274
.1	1.685	.6	2.501
.2	1.858	.7	2.813
.3	1.950	.8	3.277
.4	2.098	.9	4.171

Table 2. Values of $R(\theta, \delta^*)$ for $\theta = (|\theta|, 0, \ldots, 0)^t$.

	k					
$\lvert\theta\rvert$	5	10	20	30	40	50
0	.95	1.18	1.61	2.23	2.71	3.27
1	1.70	1.98	2.40	2.81	3.24	3.83
2	2.09	2.31	3.05	3.95	4.82	5.75
3	3.57	3.71	4.15	4.63	5.09	5.89
4	3.60	3.80	4.36	5.48	6.17	6.98
5	3.62	3.85	4.46	5.50	6.23	7.02
6	3.62	3.86	4.46	5.51	6.23	7.02

APPENDIX: PROOF OF THEOREM 2.1

Define $U_{j,k} = X_j^2 \wedge Z_{(l)}^2$ and $V_j = X_j^2 \wedge \alpha^2(y)$. Observe that

$$|U_{j,k} - V_j| \leq |Z_{(l)}^2 - \alpha^2(y)|,$$

and that $|Z_{(l)}^2 - \alpha^2(y)| \to 0$ almost surely as $k \to \infty$ (cf. Rao 1973). Also, by the strong law of large numbers

$$\frac{1}{k}\sum_{j=1}^{k} U_j \to \mu_2 \quad \text{almost surely as } k \to \infty. \tag{A.1}$$

From these facts it follows that

$$\frac{1}{k}\sum_{j=1}^{k} U_{j,k} \to \mu_2 \quad \text{almost surely as } k \to \infty. \tag{A.2}$$

Next, choose $0 < \lambda < \alpha^2(y)$ such that $e\lambda B/y < 1$, where

Table 3. Values of $R(\theta, \delta^*)$ for $\theta = k^{-1/2}|\theta|(1, \ldots, 1)^t$.

	k					
$\lvert\theta\rvert$	5	10	20	30	40	50
0	.95	1.18	1.61	2.23	2.71	3.27
1	2.35	2.70	3.35	3.91	4.45	4.95
2	3.21	3.62	4.31	4.97	5.71	6.76
3	4.07	4.89	7.13	9.01	10.82	12.66
4	4.39	5.59	8.66	11.65	14.43	16.92
5	4.57	6.47	10.51	14.57	18.62	22.35
6	4.68	7.10	11.57	15.60	19.99	24.96
8	4.80	7.99	14.00	21.10	27.05	32.05
10	4.86	9.23	17.03	23.96	30.33	36.50
12	4.90	9.43	17.77	25.35	32.18	38.56
14	4.92	9.57	18.25	26.40	33.92	40.41
16	4.93	9.66	18.62	27.15	35.12	42.25
18	4.95	9.72	18.88	27.68	36.01	43.62
20	4.96	9.76	19.07	28.09	36.70	44.67

$B = 2(2\pi)^{-1/2}\sigma^{-1}$. Then, since $\sum_{j=1}^{k} X_j^2 \wedge Z_{(l)}^2 \geq (k - l)Z_{(l)}^2$, it is clear that, when $Z_{(l)}^2 > \lambda$,

$$k \Big/ \sum_{j=1}^{k} U_{j,k} \leq \frac{k}{(k-l)Z_{(l)}^2} \leq \frac{k}{\lambda(k-l)} \leq \frac{c}{\lambda}, \quad \text{(A.3)}$$

where c is a constant $\geq k/(k - l)$ for all k. Using (A.1) through (A.3) and the dominated convergence theorem, and defining $g(z) = m(z) + m(-z)$, establishes that

$$\int_{z_{(l)}^2>\lambda} \left[k \Big/ \sum_{j=1}^{k} u_{j,k}\right] \prod_{j=1}^{k} [g(z_j)dz_j] \to \mu_2^{-1} P(Z_{(l)}^2 > \lambda), \quad \text{(A.4)}$$

$$\int_{z_{(l)}^2<\lambda} \left[k \Big/ \sum_{j=1}^{k} u_{j,k}\right] \prod_{j=1}^{k} [g(z_j)dz_j] \leq \int_{z_{(l)}^2<\lambda} [c/z_{(l)}^2] \prod_{j=1}^{k} [g(z_j)dz_j]. \quad \text{(A.5)}$$

Furthermore, letting $G(z) = \int_0^z g(t)dt$ and observing from (2.3) that $g(z) \leq B$, it follows that

$$\int_{z_{(l)}^2<\lambda} [c/z_{(l)}^2] \prod_{j=1}^{k} [g(z_j)dz_j]$$

$$= \frac{ck!}{(l-1)!(k-l)!} \int_0^\lambda z^{-2}[G(z)]^{l-1}[1 - G(z)]^{k-l} g(z)dz$$

$$\leq \frac{ck!}{(k-l)!k^l} \cdot \frac{k^l}{l!} \cdot \frac{l}{(l-2)} \cdot (\lambda B)^{l-2} B^2. \quad \text{(A.6)}$$

Using Stirling's approximation for $l!$ and the fact that $e\lambda B/y < 1$, it is easy to show that the expression in (A.6) goes to zero as $k \to \infty$. Combining this with (A.4) and (A.5) and letting $\lambda \to 0$ proves that

$$E^m\left[k \Big/ \sum_{j=1}^{k} U_{j,k}\right] \to 1/\mu_2 \quad \text{as} \quad k \to \infty.$$

Equation (2.4) follows immediately. The representation for μ_2 follows from the definition of $\alpha(y)$.

[Received November 1982. Revised April 1983.]

REFERENCES

BERGER, J. (1980), "A Robust Generalized Bayes Estimator and Confidence Region for a Multivariate Normal Mean," *Annals of Statistics*, 8, 716–761.

——— (1982), "Selecting a Minimax Estimator of a Multivariate Normal Mean," *Annals of Statistics*, 10, 81–92.

BERGER, J., and DEY, D.K. (1983), "Combining Coordinates in Simultaneous Estimation of Normal Means," *Journal of Statistical Planning and Inference*, to appear.

DEY, D.K. (1980), "On the Choice of Coordinates in Simultaneous Estimation of Normal Means," Ph.D. thesis, Purdue University, Statistics Dept.

EFRON, B., and MORRIS, C. (1972), "Limiting the Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case," *Journal of the American Statistical Association*, 67, 130–139.

——— (1973), "Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach," *Journal of the American Statistical Association*, 68, 117–130.

JAMES, W., and STEIN, C. (1961), "Estimation With Quadratic Loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, 316–379.

MORRIS, C. (1983), "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78, 47–54.

RAO, C.R. (1973), *Linear Statistical Inference and Its Applications*, New York: John Wiley.

STEIN, C. (1981), "Estimation of the Parameters of a Multivariate Normal Distribution: I. Estimation of the Means," *Annals of Statistics*, 9, 1135–1151.