© Journal of the American Statistical Association
March 1973, Volume 68, Number 341
Theory and Methods Section

# Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach

BRADLEY EFRON and CARL MORRIS*

Stein's estimator for $k$ normal means is known to dominate the MLE if $k \geq 3$. In this article we ask if Stein's estimator is any good in its own right. Our answer is yes: the positive part version of Stein's estimator is one member of a class of "good" rules that have Bayesian properties and also dominate the MLE. Other members of this class are also useful in various situations. Our approach is by means of empirical Bayes ideas. In the later sections we discuss rules for more complicated estimation problems, and conclude with results from empirical linear Bayes rules in non-normal cases.

## 1. A BAYESIAN DERIVATION OF THE JAMES–STEIN ESTIMATOR

Our purpose in this article is to show that the James-Stein estimator for $k \geq 3$ normal means does more than just demonstrate the inadequacy of the maximum likelihood estimator. It, or more precisely, its "positive part" version is a reasonable estimator in its own right. We will use Bayesian ideas to show that it is a member of a class of "good" rules, all of which dominate the MLE, and all of which may be useful in different estimation situations. We begin with a Bayesian derivation of the James-Stein estimator which illustrates our general approach.

Suppose we have $k$ parameters $\theta_1, \theta_2, \theta_3, \cdots, \theta_k$, $k \geq 3$ and for each one we observe an independent normal variate

$$x_i|\theta_i \sim \mathfrak{n}(\theta_i, 1). \tag{1.1}$$

(Actually, each $x_i$ might be the mean of $n$ independent observations $Y_{ij} \sim \mathfrak{n}(\theta_i, \sigma^2)$. Then $x_i \sim \mathfrak{n}(\theta_i, \sigma^2/n)$, and a change of scale transforms $\sigma^2/n$ to the more convenient value 1. This assumes $\sigma^2$ known, but we shall see that this does not greatly affect the results which follow. We wish to estimate the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)$ using squared error loss

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = \frac{1}{k}\|\boldsymbol{\delta} - \boldsymbol{\theta}\|^2 = \frac{1}{k}\sum_{i=1}^{k}(\delta_i - \theta_i)^2 \tag{1.2}$$

to assess the performance of an estimation rule

$$\boldsymbol{\delta}(\mathbf{x}) = (\delta_1(\mathbf{x}), \delta_2(\mathbf{x}), \cdots, \delta_k(\mathbf{x})),$$

where $\delta_i(\mathbf{x})$ is the estimate of $\theta_i$. We allow our estimate of $\theta_i$ to depend on the entire vector of observations $\mathbf{x} = (x_1, x_2, \cdots, x_k)$, even though only $x_i$ seems to be relevant to $\theta_i$.

The maximum likelihood estimator $\delta_i^0(\mathbf{x}) = x_i$ has risk function $R(\boldsymbol{\theta}, \boldsymbol{\delta}^0) \equiv E_\theta L(\boldsymbol{\theta}, \boldsymbol{\delta}^0(\mathbf{x})) = 1$ for every value of $\boldsymbol{\theta}$. James and Stein [9] showed that the estimator

$$\delta_i^1(x) = \left(1 - \frac{k-2}{S}\right)x_i \quad (S = \|\mathbf{x}\|^2) \tag{1.3}$$

has $R(\boldsymbol{\theta}, \boldsymbol{\delta}^1) < 1$ for all values of $\boldsymbol{\theta}$. We will now present an elementary proof of this fact which begins by temporarily making the Bayesian assumption that the $\theta_i$ are themselves independently normally distributed with mean 0 and variance $A$,

$$\theta_i \sim \mathfrak{n}(0, A) \tag{1.4}$$

Recall that under the distributional assumptions (1.1) and (1.4) the Bayes estimator of $\theta_i$ is

$$\delta_i^* = (1 - B)x_i \tag{1.5}$$

where

$$B = 1/(A + 1). \tag{1.6}$$

(Equivalently $A = (1/B) - 1$. $B$ decreases from 1 to 0 as $A$ increases from 0 to $\infty$.) This follows from the conditional distribution of $\theta_i$ given $x_i$,

$$\theta_i|x_i \sim \mathfrak{n}((1 - B)x_i, 1 - B). \tag{1.7}$$

Formula (1.7) also shows that

$$R(B, \boldsymbol{\delta}^*) = (1 - B) \tag{1.8}$$

where $R(B, \boldsymbol{\delta}^*)$ is somewhat abused notation for the Bayes risk of $\boldsymbol{\delta}^*$, that is $E_B R(\boldsymbol{\theta}, \boldsymbol{\delta}^*)$, "$E_B$" indicating expectation under the distribution (1.4) with $A = (1/B) - 1$. This should be compared with the Bayes risk of $\boldsymbol{\delta}^0$, $R(B, \boldsymbol{\delta}^0) = 1$. The "savings" obtained by using $\boldsymbol{\delta}^*$ instead of $\boldsymbol{\delta}^0$ are $R(B, \boldsymbol{\delta}^0) - R(B, \boldsymbol{\delta}^*) = B$. If $A$ is large then $B$ is small, but as $A$ approaches zero $B$ approaches 1 and the savings become considerable.

If the statistician does not know $B$, or equivalently $A$, he cannot use the Bayes rule $\delta_i^*$. However he can attempt to estimate $B$ from the data. Under (1.4) $S = \sum_{i=1}^{k} x_i^2$ is a

* Bradley Efron is professor, Department of Statistics, Stanford University, Stanford, Calif. 94305. Carl Morris is statistician, Department of Economics, Rand Corporation, Santa Monica, Calif. 90406.

sufficient statistic for $B$, with marginal distribution $S \sim (1/B)\chi^2_k$. Suppose the statistician decides to use some estimator $\hat{B}(S)$ for $B$ in (1.5), that is he uses the estimation rule

$$\delta_i = (1 - \hat{B}(S))x_i. \tag{1.9}$$

It is obvious that a price must be paid for estimating $B$ rather than knowing its true value. The following lemma expresses this risk in terms of the "relative savings loss,"

$$\begin{aligned}\mathrm{RSL}(B, \boldsymbol{\delta}) &\equiv \frac{R(B, \boldsymbol{\delta}) - R(B, \boldsymbol{\delta}^*)}{R(B, \boldsymbol{\delta}^0) - R(B, \boldsymbol{\delta}^*)} \\ &= [R(B, \boldsymbol{\delta}) - (1 - B)]/B\end{aligned} \tag{1.10}$$

the proportion of the savings lost by using $\boldsymbol{\delta}$ instead of $\boldsymbol{\delta}^*$.

*Lemma 1:* If $\theta_i \sim \mathfrak{n}(0, A)$, $x_i|\theta_i \sim \mathfrak{n}(\theta_i, 1)$ independently for $i = 1, 2, \cdots, k$, and $\delta_i = (1 - \hat{B}(S))x_i$, $S = \|\mathbf{x}\|^2$, then

$$\mathrm{RSL}(B, \boldsymbol{\delta}) = \tilde{E}_B\left[\frac{\hat{B}(S) - B}{B}\right]^2. \tag{1.11}$$

Here $\tilde{E}_B$ indicates expectation with respect to

$$S \sim (1/B)\chi^2_{k+2}.$$

*Proof:* $E_B[L(\boldsymbol{\theta}, \boldsymbol{\delta})|\mathbf{x}] = E_B[\|(1 - \hat{B}(S))\mathbf{x} - \boldsymbol{\theta}\|^2/k|\mathbf{x}] = (1 - B) + (\hat{B} - B)^2S/k$

by (1.7). Therefore,

$$R(B, \boldsymbol{\delta}) - R(B, \boldsymbol{\delta}^*) = E_B(\hat{B} - B)^2S/k,$$

so that

$$\mathrm{RSL}(B, \boldsymbol{\delta}) = E_B\left[\frac{\hat{B} - B}{B}\right]^2\frac{B\cdot S}{k}.$$

However for any function $g(S)$ we have

$$E_{B}g(S)\frac{B\cdot S}{k} = \tilde{E}_Bg(S).$$

$$(E_{B}g(S)\frac{B\cdot S}{k} = \int_0^\infty g(S)\frac{B\cdot S}{k}\left(\frac{B\cdot S}{2}\right)^{k/2-1}\frac{e^{-BS/2}BdS}{\Gamma\left(\frac{k}{2}\right)2}$$

$$= \int_0^\infty g(S)\left(\frac{B\cdot S}{2}\right)^{(k+2)/2-1}\frac{e^{-BS/2}}{\Gamma\left(\frac{k+2}{2}\right)}\frac{B}{2}dS = \tilde{E}_Bg(S)).$$

*Theorem 1*: The James-Stein rule

$$\delta_i^1 = \left(1 - \frac{k-2}{S}\right)x_i$$

has

$$\mathrm{RSL}(B, \boldsymbol{\delta}^1) = \frac{2}{k} \tag{1.12}$$

for every value of $B$. In terms of the Bayes risk function,

$$R(B, \boldsymbol{\delta}^1) = 1 - \frac{k-2}{k}B. \tag{1.13}$$

*Proof:* Let $\hat{B}(S) = (k-2)/S$. Then by Lemma 1,

$$\begin{aligned}\mathrm{RSL}(B, \boldsymbol{\delta}^1) &= \tilde{E}_B\left[\frac{\hat{B}}{B} - 1\right]^2 \\ &= \tilde{E}_B\left[\left(\frac{(k-2)}{BS}\right)^2 - 2\left(\frac{k-2}{BS}\right) + 1\right] \\ &= E\left[\frac{(k-2)^2}{(\chi^2_{k+2})^2} - \frac{2(k-2)}{\chi^2_{k+2}} + 1\right] \\ &= \frac{(k-2)^2}{k(k-2)} - \frac{2(k-2)}{k} + 1 = \frac{2}{k}.\end{aligned}$$

Equation (1.13) follows from the definition (1.10).

*Corollary 1:* The risk of the James-Stein rule as a function of $\boldsymbol{\theta}$ is

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}^1) = 1 - \frac{(k-2)}{k}E_\theta\frac{k-2}{S}. \tag{1.14}$$

*Proof:* It is easy to see that $R(\boldsymbol{\theta}, \boldsymbol{\delta}^1)$ is a function only of $\|\boldsymbol{\theta}\|^2$, say $f(\|\boldsymbol{\theta}\|^2)$, and likewise the right side of (1.14) is a function of $\|\boldsymbol{\theta}\|^2$, say $g(\|\boldsymbol{\theta}\|^2)$. By definition

$$E_Bf(\|\boldsymbol{\theta}\|^2) = R(B, \boldsymbol{\delta}^1) = 1 - \frac{k-2}{k}B.$$

Also

$$\begin{aligned}E_Bg(\|\boldsymbol{\theta}\|^2) &= 1 - \frac{(k-2)^2}{k}E_BE_\theta\frac{1}{S} = 1 - \frac{(k-2)^2}{k}E_B\frac{1}{S} \\ &= 1 - \frac{(k-2)^2}{k}E\frac{B}{\chi^2_k} = 1 - \frac{k-2}{k}B.\end{aligned}$$

Therefore,

$$E_Bf(\|\boldsymbol{\theta}\|^2) = E_Bg(\|\boldsymbol{\theta}\|^2) \tag{1.15}$$

for every value of $B$. This proves that $f$ and $g$ must be the same function of $\|\boldsymbol{\theta}\|^2$ since the distributions of $\|\boldsymbol{\theta}\|^2$, $\|\boldsymbol{\theta}\|^2 \sim A\chi^2_k$, are complete as a function of $A$ and therefore also as a function of $B$.

Corollary 1 shows that $R(\boldsymbol{\theta}, \boldsymbol{\delta}^1) < 1$ for all $\boldsymbol{\theta}$, with $1 - R(\boldsymbol{\theta}, \boldsymbol{\delta}^1) = ((k-2)^2/k)E_\theta(1/S)$. If we define

$$\lambda = \|\boldsymbol{\theta}\|^2/2 \tag{1.16}$$

then given $\boldsymbol{\theta}$, $S \sim \chi^2_k(2\lambda)$, a noncentral chi-square distribution with $k$ degrees of freedom and noncentrality parameter $\sum_{i=1}^k \theta_i^2 = 2\lambda$. Define

$$r(\lambda, \boldsymbol{\delta}^1) = R(\boldsymbol{\theta}, \boldsymbol{\delta}^1), \quad \|\boldsymbol{\theta}\|^2/2 = \lambda. \tag{1.17}$$

*Corollary 2:* $r(\lambda, \boldsymbol{\delta}^1)$ is an increasing concave function of $\lambda$, increasing from $r(0, \boldsymbol{\delta}^1) = 2/k$ to $r(\infty, \boldsymbol{\delta}^1) = 1$. It has the power series expansion

$$r(\lambda, \boldsymbol{\delta}^1) = 1 - \frac{k-2}{k}\sum_{j=0}^\infty\frac{\Gamma(k/2)}{\Gamma(k/2 + j)}(-\lambda)^j \tag{1.18}$$

which for $k$ even is

$$\begin{aligned}r(\lambda, \boldsymbol{\delta}^1) = 1 - \frac{k-2}{k}&\left[\frac{(k/2-1)!}{(-\lambda)^{k/2-1}}\right] \\ &\cdot\left[e^{-\lambda} - \sum_{j=0}^{k/2-2}\frac{(-\lambda)^j}{j!}\right].\end{aligned} \tag{1.19}$$

It also has the integral representation

$$r(\lambda, \delta^1) = 1 - \frac{(k-2)^2}{2k} \int_0^1 e^{-\lambda t}(1-t)^{k/2-2} dt. \quad (1.20)$$

*Proof:* Under the Bayesian assumption $\theta_i \sim \pi(0, A)$, we have $\lambda \sim AG_{k/2}$, where $G_{k/2}$ indicates a gamma distribution with parameter $k/2$, that is $\lambda$ has density function

$$p_A(\lambda) = \left(\frac{\lambda}{A}\right)^{k/2-1} \frac{e^{-\lambda/A}}{A\Gamma(k/2)}. \quad (1.21)$$

By definition

$$R(B, \delta^1) = E_B r(\lambda, \delta^1) = \int_0^\infty r(\lambda, \delta^1) p_A(\lambda) d\lambda,$$

which can be rewritten as

$$R(B, \delta^1) = \int_0^\infty \frac{r(A\lambda, \delta^1)\lambda^{k/2-1}e^{-\lambda}}{\Gamma(k/2)} d\lambda. \quad (1.22)$$

Differentiating both sides of (1.22) $j$ times with respect to $A$ gives

$$\left.\frac{d^j R(B, \delta^1)}{dA^j}\right|_{A=0} = \frac{\Gamma(k/2+j)}{\Gamma(k/2)} \left.\frac{d^j r(\lambda, \delta^1)}{d\lambda^j}\right|_{\lambda=0}. \quad (1.23)$$

However (1.13) expressed in terms of $A$ is

$$R(B, \delta^1) = 1 - \frac{k-2}{k}\frac{1}{A+1}, \quad (1.24)$$

so that the derivative on the left side of (1.23) is $-((k-2)/k)(-1)^j j!$ (1.18) is therefore the power series expansion of $r(\lambda, \delta^1)$ about $\lambda = 0$. This is a special form of the confluent hypergeometric function, and the integral form (1.20) is given in [1, p. 505, formula 13.31]. Differentiating (1.20) with respect to $\lambda$ reveals the increasing concave nature of $r(\lambda, \delta^1)$. Letting $u = 1 - t$, expanding $e^{\lambda u}$ as a power series, and integrating the resulting form of (1.20) term by term gives

$$r(\lambda, \delta^1) = 1 - \left(\frac{k-2}{k}\right)^2 \sum_{j=0}^\infty \frac{1}{k-2+2j}\frac{e^{-\lambda}\lambda^j}{j!} \quad (1.25)$$

an expression given in [15].

Much of Corollary 2 is already known, cf. [10, 15]. The results are presented here both for convenient reference and to illustrate how information about $R(\theta, \delta^1)$ can be obtained from knowledge of $R(B, \delta^1)$. Multivariate versions of the results in this section, where $\theta_i$ and $x_i$ are themselves vectors, appear in [5].

## 2. AN EMPIRICAL BAYES APPROACH TO SIMULTANEOUS ESTIMATION

We have used the Bayesian assumption $\theta_i \sim \pi(0, A)$ and its consequence, Lemma 1, as a convenient mathematical tool to investigate the risk function $R(\boldsymbol{\theta}, \delta^1)$ of Stein's rule. We will be investigating more complicated estimation rules $\boldsymbol{\delta}(\mathbf{x})$, and this convenience will become almost a necessity. It turns out that it is usually much easier to work with $R(B, \delta)$ than $R(\boldsymbol{\theta}, \boldsymbol{\delta})$. In principle all information about $R(\boldsymbol{\theta}, \boldsymbol{\delta})$ is contained in $R(B, \delta)$, at least if $R(\boldsymbol{\theta}, \boldsymbol{\delta})$ is a function only of $\|\boldsymbol{\theta}\|^2$: the relationship $R(B, \delta) = E_B R(\boldsymbol{\theta}, \boldsymbol{\delta})$ expresses $R(B, \delta)$ as the "gamma transform" of $R(\boldsymbol{\theta}, \boldsymbol{\delta})$, as in (1.22), and completeness guarantees that different $R(\boldsymbol{\theta}, \boldsymbol{\delta})$ functions transform into different $R(B, \delta)$ functions. In practice most of our calculations will stop with $R(B, \delta)$.

There is another way to look at Lemma 1. One can take the assumption $\theta_i \sim \pi(0, A)$ seriously, but stop short of full Bayesianhood by assuming that $A$ is unknown and must be estimated from the data. That is, one can be an empirical Bayesian. Lindley in the discussion of [14] makes a cogent argument for this point of view.

The virtue of Lemma 1 is that it reduced this empirical Bayes problem to more familiar form: on the basis of $S \sim (1/B)\chi^2_{k+2}$ one wishes to estimate $B$ with loss function $L(B, \hat{B}) = ((\hat{B} - B)/B)^2$. (This is essentially the problem of the estimation of a variance component encountered in Model II ANOVA.) The question, "Is the James-Stein estimator any good in its own right?" becomes "Is $\mathrm{RSL}(B, \delta^1) = 2/k$ for all $B$ a good risk function in the problem above?" A partial answer is given by the following theorem:

*Theorem 2:* For estimating $B$, $0 < B \leq 1$, from $S \sim (1/B)\chi^2_{k+2}$ with loss function

$$L(B, \hat{B}) = ((\hat{B} - B)/B)^2,$$

the minimax value of the risk is $2/k$.

*Proof:* Let $g_a(B)$ be the prior distribution on $B$ having density $B^{-a} \cdot [1-a]$, on $(0, 1]$, $a < 1$. Then the Bayes rule is

$$\hat{B}_a(S) = \frac{\int_0^1 B^{k/2-a} e^{-BS/2} dB}{\int_0^1 B^{k/2-a-1} e^{-BS/2} dB} = \frac{k-2a}{S}\frac{I_{k/2-a+1}(S/2)}{I_{k/2-a}(S/2)} \quad (2.1)$$

where

$$I_\iota(t) = \int_0^t s^{\iota-1}e^{-s}/\Gamma(\ell) ds,$$

the incomplete gamma function (cf. Section 4). Define $\tau(S) = \hat{B}_a(S)/\hat{B}^1(S)$, where $\hat{B}^1(S) = (k-2)/S$, so that

$$\tau(S) = \frac{k-2a}{k-2}\frac{I_{k/2-a+1}(S/2)}{I_{k/2-a}(S/2)}. \quad (2.2)$$

Since the ratio of the gamma densities

$$i_{k/2-a+1}(s)/i_{k/2-a}(s) = s/(k/2-a),$$

an increasing function of $s$, it is easily known that $\tau(S)$ increases monotonically from 0 to $(k-2a)/(k-2)$ as $S$ increases from 0 to $\infty$.

We have

$$E_a\left[\frac{\hat{B}_a(S) - B}{B}\right]^2 = E_a\left[\tau^2(S)\left(\frac{\hat{B}^1(S)}{B}\right)^2 - 2\tau(S)\frac{\hat{B}^1(S)}{B} + 1\right]. \quad (2.3)$$

As $a$ increases to 1 then $g_a(B)$ converges in distribution to 0, and therefore $S \sim (1/B)\chi^2_{k+2}$ converges in distribution to $\infty$. By the remark following (2.2) we therefore have $\tau(S)$ converging in distribution to 1, and by (2.3), $\lim_{a\to 1} E_a[(\hat{B}_a(S) - B)/B]^2 = 2/k$. We have found a sequence of proper Bayes rules with Bayes risk approaching $2/k$, which shows that $2/k$ is the minimax value for this situation.

The fact that the James-Stein rule is minimax does not mean that we cannot find a rule $\boldsymbol{\delta}$ with $R(B, \boldsymbol{\delta}) < 2/k$ for all $B$ in $(0, 1]$. We can take advantage of the knowledge that $B$ is in $(0, 1]$ to improve the estimator $\hat{B}^1(S) = (k - 2)/S$, which is greater than 1 for $S < k - 2$. The estimator $\hat{B}^{1+}(S) = \min\{1, \hat{B}^1(S)\}$ uniformly improves on the James-Stein rule (cf. Section 5 and Corollary 3).

So far we have discussed the problem of estimating $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)$ solely in terms of the loss function $(1/k)\|\boldsymbol{\delta} - \boldsymbol{\theta}\|^2$. However in many situations we will be interested in calculating the risk for a single component $\theta_i$ rather than the whole vector $\boldsymbol{\theta}$. We will now give a componentwise form of Lemma 1, after first broadening our assumptions and definitions to allow for distributional differences in the $k$ coordinates.

Assume that given $\theta_i$, $x_i$ is normal with mean $\theta_i$, variance $D_i$,

$$x_i|\theta_i \sim \mathfrak{n}(\theta_i, D_i) \text{ independently } i = 1, 2, \cdots, k \quad (2.4)$$

and that $\theta_i$ has prior distribution

$$\theta_i \sim \mathfrak{n}(0, A_i) \text{ independently } i = 1, 2, \cdots, k. \quad (2.5)$$

Also, define

$$B_i = D_i/(A_i + D_i) \quad (2.6)$$

and $\mathbf{B} = (B_1, B_2, \cdots, B_k)$. Let the loss incurred for estimating $\theta_i$ by $\delta_i$, be

$$L_i(\theta_i, \delta_i) = (\delta_i - \theta_i)^2 \quad (2.7)$$

and for any estimation rule $\delta_i(\mathbf{x})$ let

$$R_i(\boldsymbol{\theta}, \delta_i) \equiv E_{\boldsymbol{\theta}} L_i(\theta_i, \delta_i), \quad R_i(\mathbf{B}, \delta_i) = E_{\mathbf{B}} R_i(\boldsymbol{\theta}, \delta_i), \quad (2.8)$$

$E_{\boldsymbol{\theta}}$ indicating expectation with respect to the distribution (2.4) of $\mathbf{x}$ given $\boldsymbol{\theta}$, $E_{\mathbf{B}}$ indicating expectation with respect to the distribution (2.5) of $\boldsymbol{\theta}$,

$$A_i = D_i\left(\frac{1}{B_i} - 1\right), \quad i = 1, 2, \cdots, k.$$

Given $x_i$, $\theta_i$ is normally distributed,

$$\theta_i|x_i \sim \eta((1 - B_i)x_i, D_i(1 - B_i)) \quad (2.9)$$

which shows that the Bayes rule for the $i$th component is

$$\delta_i^*(x_i) = (1 - B_i)x_i, \quad (2.10)$$

with risk function $R_i(\mathbf{B}, \delta_i^*) = D_i(1 - B_i)$. This should be compared with the risk function for the maximum likelihood rule $\delta_i^0(x_i) = x_i$, $R_i(\mathbf{B}, \delta_i^0) = D_i$. Maintaining the terminology of Section 1, we define the $i$th coordinate "relative savings loss" to be

$$\begin{aligned}\mathrm{RSL}_i(\mathbf{B}, \delta_i) &\equiv \frac{R_i(\mathbf{B}, \delta_i) - R_i(\mathbf{B}, \delta_i^*)}{R_i(\mathbf{B}, \delta_i^0) - R_i(B, \delta_i^*)} \\ &= [R_i(\mathbf{B}, \delta_i) - (1 - B_i)]/B_iD_i.\end{aligned} \quad (2.11)$$

It is obvious that a sufficient statistic for the unknown vector $\mathbf{B}$ is $\mathbf{S} = (S_1, S_2, \cdots, S_k)$, where $S_i \equiv x_i^2 \sim (D_i/B_i)\chi^2_1$. Let $\delta_i(\mathbf{x})$ be of the form

$$\delta_i(\mathbf{x}) = (1 - \hat{B}_i(\mathbf{S}))x_i. \quad (2.12)$$

*Lemma 2:* The rule $\delta_i(x)$ has

$$\mathrm{RSL}_i(\mathbf{B}, \delta_i) = \tilde{E}_{\mathbf{B}}^{(i)}\left[\frac{\hat{B}_i(\mathbf{S}) - B_i}{B_i}\right]^2 \quad (2.13)$$

where $\tilde{E}_{\mathbf{B}}^{(i)}$ indicates expectation with respect to the distribution of $\mathbf{S}$

$$S_i \sim \frac{D_i}{B_i}\chi^2_3, \quad S_j \sim \frac{D_j}{B_j}\chi^2_1 \quad j \neq i, \quad (2.14)$$

all $k$ components of $\mathbf{S}$ being independent.

The proof of Lemma 2 is almost identical with that of Lemma 1 and will not be presented here. Notice that if $A_i \equiv A$, $D_i \equiv D$, for $i = 1, 2, \cdots, k$, then $B_i \equiv B = D/(A + D)$, and if $\hat{B}_i(\mathbf{S}) = \hat{B}(S)$, where $S = \sum_{i=1}^k x_i^2$, then Lemma 2 reduces to Lemma 1. That is, $\mathrm{RSL}_i(\mathbf{B}, \delta_i) = \mathrm{RSL}(B, \boldsymbol{\delta})$ as given by (1.11).

A simple but interesting result of Lemma 2 is that in the case where all the $A_i$ are equal and all the $D_i$ are equal, the observation on the $i$th coordinate is worth three times that of the other coordinates in estimating $\theta_i$. More precisely, the best rule of the form

$$\delta_i = (1 - c/\textstyle\sum_{j\neq i} S_j)x_i$$

has RSL $= 2/(k - 3)$ instead of $2/k$.

Suppose we have an estimation rule of the form (2.12) and that there is positive probability that $\hat{B}_i(\mathbf{S}) > 1$. Corollary 3 shows that we can always improve the RSL of such a rule by replacing $\hat{B}_i(\mathbf{S})$ with

$$\hat{B}_i^+(\mathbf{S}) = \min\{1, \hat{B}_i(\mathbf{S})\}. \quad (2.15)$$

*Corollary 3:* Let $\delta_i^+(\mathbf{x}) = (1 - \hat{B}_i^+(\mathbf{S}))x_i$. Then,

$$\begin{aligned}\mathrm{RSL}_i(\mathbf{B}, \delta_i) - \mathrm{RSL}_i(\mathbf{B}, \delta_i^+) = \frac{1}{B_i^2}\tilde{E}^{(i)}\{[\{\hat{B}_i(\mathbf{S}) - \hat{B}_i^+(\mathbf{S})\} \\ + (1 - B_i)]^2 - [1 - B_i]^2\}.\end{aligned} \quad (2.16)$$

The function of $\mathbf{S}$ in the large brackets is nonnegative, and strictly greater than 0 if $\hat{B}_i(\mathbf{S}) > 1$, so that $\mathrm{RSL}_i(\mathbf{B}, \delta_i) \geq \mathrm{RSL}_i(\mathbf{B}, \delta_i^+)$ for all $\mathbf{B}$ with strict inequality if $\mathrm{prob}_B\{\hat{B}_i(\mathbf{S}) > 1\} > 0$ for any value of $B$. The proof is obtained immediately from Lemma 2 by writing

$$\mathrm{RSL}_i(\mathbf{B}, \delta_i) = \tilde{E}_{\mathbf{B}}^{(i)}\left[\frac{\{\hat{B}_i(\mathbf{S}) - \hat{B}_i^+(\mathbf{S})\} + \{\hat{B}_i^+(\mathbf{S}) - B_i\}}{B_i}\right]^2.$$

## 3. A CLASS OF ESTIMATORS WHICH DOMINATE THE MLE

We assume once again that

$$x_i|\theta_i \sim \mathfrak{n}(\theta_i, 1), \text{ independently, } i = 1, 2, \cdots, k \quad (3.1)$$

and for any Bayesian calculation that

$$\theta_i \sim \mathfrak{n}(0, A), \text{ independently, } i = 1, 2, \cdots, k. \quad (3.2)$$

Using Lemma 1 it is easy to show that the rule $\boldsymbol{\delta}^t$ defined by

$$\delta_i^t = \left(1 - \frac{(k-2)t}{S}\right)x_i \quad (3.3)$$

has $\mathrm{RSL}(B, \delta^t)$ constant and equal to

$$\mathrm{RSL}(B, \boldsymbol{\delta}^t) \equiv \mathrm{RSL}^t \equiv \frac{2}{k} + \frac{k-2}{k}(t-1)^2. \quad (3.4)$$

Thus $R(B, \boldsymbol{\delta}^t) = 1 - (1 - \mathrm{RSL}^t)B$. The proof of Corollary 1 can be used again to give

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}^t) = 1 - (1 - \mathrm{RSL}^t)E_\theta\frac{k-2}{S}. \quad (3.5)$$

We see that for any value of $t$ in $(0, 2)$ the rule $\boldsymbol{\delta}^t$ dominates the maximum likelihood estimator, $R(\boldsymbol{\theta}, \boldsymbol{\delta}^t) < 1$. The uniformly best value of $t$ is the James-Stein choice $t = 1$. However, the situation is not so simple for the plus-rules $\delta_i^{t+} = (1 - \min\{1, (k-2)t/S\})x_i$. We show in Section 5 that the risk functions $R(\boldsymbol{\theta}, \boldsymbol{\delta}^{t+})$ are noncomparable, i.e., do not dominate one another, for $t \in [1, 2]$. (The rules with $t < 1$ are dominated.) In the next two sections we shall develop a wide class of rules, including the $\boldsymbol{\delta}^{t+}$ rules, which are mutually non-comparable and all of which dominate the MLE.

Baranchik [2] developed a useful sufficient condition for an estimator of the form

$$\boldsymbol{\delta}(\mathbf{x}) = \left(1 - \frac{k-2}{S}\tau(S)\right)\mathbf{x} \quad (3.6)$$

to dominate the MLE:

*Baranchik's Theorem:* For $0 < t \leq 2$, define $\mathfrak{S}_t$ as the class of rules (3.6) with $\tau(S)$ satisfying

(i) $\tau(S)$ in nondecreasing in $S$, $\tau(S) \geq 0$

and

(ii) $\lim_{S\to\infty}\tau(S) = t$.

Then all the rules in $\mathfrak{S}_t$ have $R(\boldsymbol{\theta}, \boldsymbol{\delta}) < 1$ for all $\boldsymbol{\theta}$. (Actually Baranchik required $t \leq 1$, but Strawderman [16] pointed out that the proof goes through for $t \leq 2$. In the two extreme cases $\tau(S) \equiv 0$ or $\tau(S) \equiv 2$, we actually have $R(\boldsymbol{\theta}, \boldsymbol{\delta}) = 1$ for all $\boldsymbol{\theta}$.)

$R(\boldsymbol{\theta}, \boldsymbol{\delta}) < 1$ for all $\boldsymbol{\theta}$ implies that $R(B, \boldsymbol{\delta}) < 1$ for all $B$, which is equivalent to $\mathrm{RSL}(B, \boldsymbol{\delta}) < 1$ for all $B$. We can prove this weaker result very easily from our previous theory. Using

$$\mathrm{RSL}(B, \delta) = E_B\left[\frac{\hat{B} - B}{B}\right]^2\frac{B\cdot S}{k},$$

an intermediate step in the proof of Lemma 1, gives

$$\mathrm{RSL}(B, \boldsymbol{\delta}) = 1 - 2\frac{k-2}{k}E_B\tau(S) \cdot\left\{1 - \frac{k-2}{2}\frac{\tau(S)}{BS}\right\} \quad (3.7)$$

$$\leq 1 - 2\frac{k-2}{k}E_B\tau(S)\left\{1 - \frac{k-2}{BS}\right\} \quad (3.8)$$

by Condition ii and $\tau(S) \geq 0$. The term in brackets has expectation 0, since $BS \sim \chi_k^2$, and is an increasing function of $S$. $\tau(S)$ is nondecreasing in $S$ by Condition i. Therefore $\tau(S)$ and $\{1 - (k-2)/S\}$ have nonnegative correlation and the expectation in (3.8) is positive.

Given Condition i, one can see that the restriction $t \leq 2$ is necessary as well as sufficient: if $\lim_{S\to\infty}\tau(S) > 2$ then as $B \to 0$ the right-hand side of (3.7) will become greater than 1. The condition $\tau(S)$ be nondecreasing in $S$ is not necessary, at least not for the result $R(B, \boldsymbol{\delta}) < 1$ for all $B$, but no convenient substitute has been found.

Suppose we have a rule of the form (3.6) satisfying (i), and $\lim_{S\to\infty}\tau(S) \equiv \bar{t}$ ($\bar{t}$ possibly $>2$). Let $\tilde{S}$ be distributed as $\chi_{k+2}^2$. Then by Lemma 1,

$$\lim_{B\to 0}\mathrm{RSL}(B, \boldsymbol{\delta}) = \lim_{B\to 0}E\left[\frac{(k-2)\tau(\tilde{S}/B)}{\tilde{S}} - 1\right]^2 = \mathrm{RSL}^{\bar{t}}. \quad (3.9)$$

Since $R(B, \boldsymbol{\delta}) = 1 - (1 - \mathrm{RSL}\ (B, \boldsymbol{\delta}))B$, we have

$$1 - R(B, \boldsymbol{\delta}) \approx (1 - \mathrm{RSL}^{\bar{t}})B \text{ as } B \to 0. \quad (3.10)$$

## 4. TRUNCATED BAYES RULES

The Baranchik class of rules is too large since it contains some rules, such as $\boldsymbol{\delta}^t$, $t \neq 1$, which are dominated by others in the class. One way to guarantee that the rules we use are admissible is to make them Bayes rules.

Let $h(B)$ be a prior distribution on $B$ putting all of its probability on the interval $(0, 1]$. For convenience we will treat $h(B)$ as a probability density function, but the calculations which follow remain true, with the obvious changes, for general distribution functions. It is not necessary for $h$ to be a proper distribution on $(0, 1]$, but we do require the weaker condition that

$$E_h B < \infty. \quad (4.1)$$

To summarize the situation, first $B \sim h$ is selected, then $\theta_i \sim \mathfrak{n}(0, A)$ independently $i = 1, 2, \cdots, k$, $A = (1/B) - 1$, and finally $x_i \sim \mathfrak{n}(\theta_i, 1)$ independently $i = 1, 2, \cdots, k$. The Bayes rule under our loss function $(1/k)\|\boldsymbol{\delta} - \boldsymbol{\theta}\|^2$ is calculated to be

$$\begin{aligned}\delta_h^*(\mathbf{x}) &= E_h(\boldsymbol{\theta}|\mathbf{x}) = E_h[E(\boldsymbol{\theta}|\mathbf{x}, B)] \\ &= \int_0^1 [1 - B]\mathbf{x}h_S(B)dB \\ &= [1 - B_h^*(S)]\mathbf{x}\end{aligned} \tag{4.2}$$

where $E_h$ indicates expectation with respect to the situation described above, $h_S(B)$ is the conditional density of $B$ given $S$, and $B_h^*(S)$ is the conditional expectation of $B$ given $S$.

We can state the definition of the Bayes rule in several equivalent forms:

1. $\delta_h^*$ is that $\delta$ which minimizes $\int_0^1 R(B,\delta)h(B)dB$,
2. $\delta_h^*$ is that $\delta$ which minimizes

$$\int_0^1 \{[R(B,\delta) - (1 - B)]/B\}g(B)dB,$$

where

$$g(B) \equiv Bh(B) \tag{4.3}$$

3. $\delta_h^*$ is that $\delta = (1 - \hat{B}(S))\mathbf{x}$ for which $\hat{B}(S)$ minimizes

$$\tilde{E}_\theta\left(\frac{\hat{B} - B}{B}\right)^2 \equiv \int_0^1 \tilde{E}_B\left[\frac{\hat{B}(S) - B}{B}\right]^2 g(B)dB.$$

Form (3) is derived from (2) by Lemma 1, and of course the $\hat{B}(S)$ which accomplishes the minimization in (3) will be $B_h^*(S)$. An explicit formula[1] is

$$\begin{aligned}B_h^*(S) &= \frac{\int_0^1 g(B)B^{k/2}e^{-BS/2}dB}{\int_0^1 g(B)B^{k/2-1}e^{-BS/2}dB} \\ &= \frac{\int_0^1 h(B)B^{k/2+1}e^{-BS/2}dB}{\int_0^1 h(B)B^{k/2}e^{-BS/2}dB}.\end{aligned} \tag{4.4}$$

If we define

$$\tilde{h}(B) \equiv B^{k/2}h(B) \tag{4.5}$$

and

$$e^{-\psi(S)} \equiv \int_0^1 \tilde{h}(B)e^{-BS/2}dB,$$

then

$$B_h^*(S) = 2\frac{d\psi(S)}{dS} \tag{4.6}$$

which allows us to use known Laplace transforms to calculate $B_h^*(S)$ in some cases.

If we take the improper prior $h(B) = 1/B^2$, $g(B) = 1/B$, defined for all $B > 0$ and not just for $0 < B \leq 1$, then we get as a formal Bayes rule

$$B_h^*(S) = \frac{\int_0^\infty B^{k/2-1}e^{-BS/2}dB}{\int_0^\infty B^{k/2-2}e^{-BS/2}dB} = \frac{k - 2}{S} \tag{4.7}$$

which is the James-Stein estimator.

Now of course most Bayes rules will not dominate the MLE, a property which we are interested in preserving. Strawderman [16] has shown that the prior $h(B) = B^{-a}[1 - a]$ for suitably chosen $a$ and $k \geq 5$ results in

$$\tau_h^*(S) \equiv B_h^*(S) \Big/ \left(\frac{k - 2}{S}\right) \tag{4.8}$$

being in $\mathfrak{B}_t$ for $t \leq 2$, and therefore does dominate the MLE. (See Section 6.) This calculation is rather special, but in the next two sections we shall show that it is easy to derive Bayes rules with $B_h^*(S)$ satisfying Condition i of Baranchik's theorem. We now show that the class of rules obtained by modifying such Bayes rules so that they are in $\mathfrak{B}_t$ for some $t \leq 2$ has certain desirable properties.

*Lemma 3:* Suppose that $h(B)$ is such that $\tau_h^*(S)$ defined by (4.8) is nonnegative and satisfies Condition i of Baranchik's theorem. Then the rule in $\mathfrak{B}_t$ which minimizes the Bayes risk versus $h$, $E_hR(B, \boldsymbol{\delta})$, is given by $\boldsymbol{\delta}_h^t(\mathbf{x}) = (1 - ((k - 2)/S)\tau_h^t(S))\mathbf{x}$ where

$$\tau_h^t(S) = \min\{t, \tau_h^*(S)\}. \tag{4.9}$$

*Proof:* The calculation is done conditionally on $S$. Let $g_S(B) = Bh_S(B)$, so $g_S(B) \propto g(B)B^{k/2+1}e^{-BS/2}$ where $g(B) = Bh(B)$ as before. For any rule $\boldsymbol{\delta} = (1 - \hat{B}(S))\mathbf{x}$,

$$\begin{aligned}&\int_0^1 \left[\frac{\hat{B}(S) - B}{B}\right]^2 g_S(B)dB \\ &\quad = \int_0^1 \left[\frac{\hat{B}(S) - B_h^*(S)}{B}\right]^2 g_S(B)dB \\ &\qquad + \int_0^1 \left[\frac{B_h^*(S) - B}{B}\right] g_S(B)dB\end{aligned} \tag{4.10}$$

the cross-term vanishing because of (4.4). In terms of $\tau(S) = \hat{B}(S)/((k - 2)/S)$ this is written as

$$\begin{aligned}&\int_0^1 \left[\frac{\hat{B}(S) - B}{B}\right]^2 g_S(B)dB \\ &\quad = \int_0^1 [\tau(S) - \tau_h^*(S)]^2 \left[\frac{k - 2}{BS}\right]^2 g_S(B)dB \\ &\qquad + \int_0^1 \left[\frac{B_h^*(S) - B}{B}\right]^2 g_S(B)dB.\end{aligned} \tag{4.11}$$

It is obvious that $\tau_h^t(S)$ minimizes the right side of (4.11) for every $S$, among $\tau$ functions giving rules in $\mathfrak{B}_t$. Inte-

[1] If $E_h 1 = \infty$ then the last expression in (4.4) is only a formal Bayes rule. However the middle expression and condition (4.1) show that $B_h^*(S)$ is also proper Bayes. Thus $\delta_h^*$ will be admissible in terms of the RSL function, and this in turn implies that $\delta_h^*$ will be admissible in terms of the original risk function $R(\theta, \delta)$. This last result depends on two facts; that rules of the form (1.9) are a complete class in terms of $R(B, \delta)$; and that if $R(\theta, \delta') \leq R(\theta, \delta)$ for all $\theta$ with strict inequality for some $\theta$, then $R(B, \delta') \leq R(B, \delta)$ for all $B$ with strict inequality for some $B$.

grating over the marginal distribution of $S$,

$$\int_0^1 g(B)B^{k/2+1}S^{k/2}e^{-BS/2}dB/(2^{k/2+1}\Gamma(k/2+1)),$$

and applying Lemma 1 shows that $\boldsymbol{\delta}_h^t(\mathbf{x})$ minimizes

$$\int_0^1 \mathrm{RSL}(B, \boldsymbol{\delta})g(B)dB$$

in $\mathcal{C}_t$. As before this is equivalent to minimizing

$$\int_0^1 R(B, \boldsymbol{\delta})h(B)dB$$

in $\mathcal{C}_t$. (This last assertion follows only if

$$\int_0^1 h(B)dB < \infty.$$

If not, but if (4.1) holds, then we have still shown that $\boldsymbol{\delta}_h^t$ minimizes

$$\int_0^1 \mathrm{RSL}(B, \boldsymbol{\delta})g(B)dB$$

in $\mathcal{C}_t$.)

We will call $\boldsymbol{\delta}_h^t(\mathbf{x})$ a "Truncated Bayes Rule." We also define

$$B_h^t(S) \equiv \frac{k-2}{S}\tau_h^t(S). \tag{4.12}$$

To avoid trivialities we will always assume

$$t \leq \bar{t} \equiv \lim_{S\to\infty} \tau_h^*(S). \tag{4.13}$$

*Lemma 4:* $E_hR(B, \boldsymbol{\delta}_h^t)$ is a strictly decreasing convex function of $t$.

*Proof:* $\{\tau_h^t(S) - \tau_h^*(S)\}^2 = \max^2\{0, \tau_h^*(S) - t\}$ is a decreasing convex function of $t$, and the result follows from (4.11).

*Theorem 3:* The risk functions of two distinct truncated Bayes rules, $\boldsymbol{\delta}_{h_1}^{t_1}$ and $\boldsymbol{\delta}_{h_2}^{t_2}$, $t_1, t_2 \in [1, 2]$, are noncomparable. That is, there exists $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ such that $R(\boldsymbol{\theta}_1, \boldsymbol{\delta}_{h_1}^{t_1}) < R(\boldsymbol{\theta}_1, \boldsymbol{\delta}_{h_2}^{t_2})$ and $R(\boldsymbol{\theta}_2, \boldsymbol{\delta}_{h_2}^{t_2}) < R(\boldsymbol{\theta}_2, \boldsymbol{\delta}_{h_1}^{t_2})$.

*Proof:* The theorem is equivalent to the seemingly weaker statement that $R(B, \boldsymbol{\delta}_{h_1}^{t_1})$ and $R(B, \boldsymbol{\delta}_{h_2}^{t_2})$ are noncomparable as functions of $B$. This in turn is equivalent to the statement that $\mathrm{RSL}(B, \boldsymbol{\delta}_{h_1}^{t_1})$ and $\mathrm{RSL}(B, \boldsymbol{\delta}_{h_2}^{t_1})$ are noncomparable as functions of $B$. (Note that the converse statement is *not* true: $R(B, \boldsymbol{\delta}_1) < R(B, \boldsymbol{\delta}_2)$ for all $B \in (0, 1]$ does not imply $R(\boldsymbol{\theta}, \boldsymbol{\delta}_1) < R(\boldsymbol{\theta}, \boldsymbol{\delta}_2)$ for all $\boldsymbol{\theta}$.)

*If* $t_1 = t_2$: $E_{h_1}R(B, \boldsymbol{\delta}_{h_1}^{t_1}) < E_{h_1}R(B, \boldsymbol{\delta}_{h_2}^{t_1})$ by the definition of $\boldsymbol{\delta}_{h_1}^{t_1}$, so there exists $B_1$ such that $R(B_1, \boldsymbol{\delta}_{h_1}^{t_1}) < R(B_1, \boldsymbol{\delta}_{h_2}^{t_1})$, and likewise there exists $B_2$ such that $R(B_2, \boldsymbol{\delta}_{h_2}^{t_1}) < R(B_2, \boldsymbol{\delta}_{h_1}^{t_1})$.

*If* $t_1 \neq t_2$: Assume, without loss of generality, that $t_1 < t_2$.

$$\lim_{B\to 0} \mathrm{RSL}(B, \boldsymbol{\delta}_{h_1}^{t_1}) = \mathrm{RSL}^{t_1} < \mathrm{RSL}^{t_2} = \lim_{B\to 0} \mathrm{RSL}(B, \boldsymbol{\delta}_{h_2}^{t_2}).$$

Therefore, there exists $B_1$ such that

$$\mathrm{RSL}(B_1, \boldsymbol{\delta}_{h_1}^{t_1}) < \mathrm{RSL}(B_1, \boldsymbol{\delta}_{h_2}^{t_2}) \Rightarrow R(B_1, \boldsymbol{\delta}_{h_1}^{t_1}) < R(B_1, \boldsymbol{\delta}_{h_2}^{t_2}).$$

Conversely, $E_{h_2}R(B, \boldsymbol{\delta}_{h_1}^{t_1}) > E_{h_2}R(B, \boldsymbol{\delta}_{h_1}^{t_2})$ by Lemma 4, and $E_{h_2}R(B, \boldsymbol{\delta}_{h_1}^{t_2}) > E_{h_2}R(B, \boldsymbol{\delta}_{h_2}^{t_2})$, proving the result as in the case $t_1 = t_2$.

*Corollary 4:* If

$$\boldsymbol{\delta}_h^t(\mathbf{x}) = (1 - ((k-2)/S)\tau_h^t(S))\mathbf{x}$$

is a truncated Bayes rule, then so is

$$\boldsymbol{\delta}_{h,b}^t(\mathbf{x}) = (1 - ((k-2)/S)\tau_h^t(bS))\mathbf{x}$$

for any constant $b$, $0 < b \leq 1$. Moreover, $R(B, \boldsymbol{\delta}_{h,b}^t) = R(B/b, \boldsymbol{\delta}_h^t)$ for $B \in (0, b]$. (Equivalently, $\mathrm{RSL}(B, \boldsymbol{\delta}_{h,b}^t) = \mathrm{RSL}(B/b, \boldsymbol{\delta}_h^t)$ for $B \in (0, b]$.)

*Proof:* Let $h_b(B) = (1/b)h(B/b)$ for $B \in (0, b]$, and equal 0 for $B \in (b, 1]$. It is then a straightforward calculation from (4.3) that $\boldsymbol{\delta}_{h,b}^t = \boldsymbol{\delta}_{hb}^t$. In other words, $\boldsymbol{\delta}_{h,b}^t$ is the truncated Bayes rule versus the prior on $B$ obtained by scaling the original distribution by a factor of $b$. The last statement of the corollary follows immediately from Lemma 1 and the fact that definition (4.12) implies

$$B_{h,b}^t(S) \equiv B_{hb}^t(S) = bB_h^t(bS). \tag{4.14}$$

## 5. JAMES–STEIN POSITIVE PART ESTIMATOR AS A TRUNCATED BAYES RULE

Suppose that $h(B)$, the prior distribution on $B$, puts all its probability mass on $B = 1$. Then, obviously,

$$B_h^*(S) = 1 \quad \text{for all } B \tag{5.1}$$

giving the Bayes rule

$$\boldsymbol{\delta}_h^*(S) = \left(1 - \frac{k-2}{S}\tau_h^*(S)\right)\mathbf{x} \tag{5.2}$$

with

$$\tau_h^*(S) = S/(k-2). \tag{5.3}$$

Applying Lemma 3, the rule in $\mathcal{C}_t$ which minimizes $E_hR(B, \boldsymbol{\delta})$ has $\tau_h^t(S) = \min\{t, S/(k-2)\}$, i.e., it is the rule

$$\boldsymbol{\delta}_h^t(S) = \left(1 - \min\left(1, \frac{(k-2)t}{S}\right)\right)\mathbf{x}. \tag{5.4}$$

This is the positive part version of the estimator

$$(1 - (k-2)t/S)\mathbf{x},$$

which for $t = 1$ is the James-Stein estimator.

We see that the rule defined by

$$\hat{B}^{t+}(S) = \min\{1, (k-2)t/S\}, \quad 1 \leq t \leq 2 \tag{5.5}$$

is a truncated Bayes rule and in the class $\mathcal{C}_t$. By applying Corollary 4, most conveniently in the form (4.14), we get that

$$\hat{B}_b^{t+}(S) = \min\left\{b, \frac{(k-2)t}{S}\right\}, \quad \begin{matrix} 1 \leq t \leq 2 \\ 0 < b \leq 1 \end{matrix} \tag{5.6}$$

is also truncated Bayes and in the class $\mathcal{C}_t$. (Actually

$\hat{B}_b^{t+}$ is truncated Bayes against the prior on $B$ which puts all of its mass on $B = b$.) All the rules

$$\boldsymbol{\delta}_b^{t+} \equiv (1 - \hat{B}_b^{t+}(S))\mathbf{x}$$

are mutually noncomparable by Theorem 3.[2] In the case $b = 1$ we will drop the $b$ subscript and write $\boldsymbol{\delta}^{t+}$ for $\boldsymbol{\delta}_1^{t+}$.

We can write down an explicit formula for $\mathrm{RSL}(B, \boldsymbol{\delta}_b^{t+})$:

*Theorem 4:* Define

$$c = (k - 2)t. \tag{5.7}$$

Then for the case $b = 1$,

$$\mathrm{RSL}(B, \boldsymbol{\delta}^{t+}) = \mathrm{RSL}^t - \left\{[\mathrm{RSL}^t - A^2]I_{k/2}\left(\frac{cB}{2}\right) + \left[\frac{c}{k-2} - 1 + A\right]\frac{c}{k} i_{k/2}\left(\frac{cB}{2}\right)\right\} \tag{5.8}$$

with $\mathrm{RSL}^t$ defined by (3.4), $I_{k/2}(s)$ the incomplete Gamma function defined after (2.1), and

$$i_{k/2}(s) = (d/ds)I_{k/2}(s).$$

For values of $b < 1$,

$$\mathrm{RSL}(B, \boldsymbol{\delta}_b^{t+}) = \mathrm{RSL}(B/b, \boldsymbol{\delta}^{t+}), \tag{5.9}$$

where $\mathrm{RSL}(B/b, \boldsymbol{\delta}^{t+})$ is calculated by (5.8) even if $B/b > 1$.

The proof of (5.8) is tedious but straightforward from Lemma 1. It will not be presented here. Equation (5.9) also follows from Lemma 1, and the fact that

$$\hat{B}_b^{t+}(S) = b\hat{B}^{t+}(bS),$$

as in Corollary 4.

Since $\mathrm{RSL}^t = \mathrm{RSL}(B, \boldsymbol{\delta}^t)$, the term in brackets in (5.8) is equal to $\mathrm{RSL}(B, \boldsymbol{\delta}^{t+}) - \mathrm{RSL}(B, \boldsymbol{\delta}^t)$. As expected, this quantity is greatest at $B = 1$:

*Corollary 5:* As a function of $B$, $\mathrm{RSL}(B, \boldsymbol{\delta}_b^{t+})$ as given by (5.8) and (5.9) achieves its minimum value over $B \in (0, \infty)$ at $B = b$,

$$\mathrm{RSL}(b, \boldsymbol{\delta}_b^{t+}) = \mathrm{RSL}^t - \left\{\mathrm{RSL}^t I_{k/2}(c/2) + \left(\frac{c}{k-2} - 1\right)\frac{c}{k} i_{k/2}(c/2)\right\}. \tag{5.10}$$

*Proof:* By (5.9) we can just consider the case $b = 1$. By the definition of a truncated Bayes rule,

$$\mathrm{RSL}(1, \boldsymbol{\delta}^{t+}) = \min_{\delta \in \mathfrak{B}_t} \mathrm{RSL}(1, \boldsymbol{\delta}). \tag{5.11}$$

(For any rule, $\mathrm{RSL}(1, \boldsymbol{\delta}) = R(1, \boldsymbol{\delta})$, and $\boldsymbol{\delta}^{t+}$ minimizes $R(1, \boldsymbol{\delta})$ for $\boldsymbol{\delta} \in \mathfrak{B}_t$.) For any value of $B > 1$,

$$\mathrm{RSL}(1, \boldsymbol{\delta}^{t+}) < \mathrm{RSL}(1, \boldsymbol{\delta}_{1/B}^{t+}) = \mathrm{RSL}(B, \boldsymbol{\delta}^{t+}) \tag{5.12}$$

while for $B < 1$,

$$\mathrm{RSL}(1, \boldsymbol{\delta}^{t+}) = \mathrm{RSL}(B, \boldsymbol{\delta}_B^{t+}) < \mathrm{RSL}(B, \boldsymbol{\delta}_1), \tag{5.13}$$

the last inequality following since we can generalize (5.12) to

$$\mathrm{RSL}(b, \boldsymbol{\delta}_b^{t+}) = \min_{\delta \in \mathfrak{B}_t} \mathrm{RSL}(b, \boldsymbol{\delta}). \tag{5.14}$$

The situation as concerns the extended class of Stein rules (5.6), can be stated very simply: for each value of $t$, $1 \leq t \leq 2$, the function defined by (5.8) for $B \in (0, \infty)$ has its lowest point at $B = 1$. Scale changes on the argument of this function give $\mathrm{RSL}(B, \boldsymbol{\delta}_b^{t+})$ for all values of $b$ via (5.9). If the statistician thinks that $B = .5$ is a particularly likely value, e.g., he might use $\boldsymbol{\delta}_{.5}^{t+}$, with $t$ chosen to give a satisfactorily small value of $\lim_{B \to 0} R(B, \boldsymbol{\delta}_{.5}^{t+}) = \mathrm{RSL}^t$. If he works with the rules $\boldsymbol{\delta}^{t+}$ the statistician is favoring the value $B = 1$, i.e., the "null hypothesis" $A = 0$ which implies $\theta_i = 0$, $i = 1, 2, \cdots, k$.

There is another interpretation to the constant $t$ in the case $b = 1$:

*Corollary 6:* $\mathrm{prob}_B\{\hat{B}^{t+} \geq B\} = \mathrm{prob}\{\chi_k^2 \leq c\}$, where $c = t \cdot (k - 2)$, as before.

*Proof:*

$$\mathrm{prob}_B\{\hat{B}^{t+} \geq B\} = \mathrm{prob}_B\left\{\frac{t \cdot (k-2)}{S} \geq B\right\} = \mathrm{prob}_B\{c \geq \chi_k^2\}$$

where

$$S \sim \frac{1}{B}\chi_k^2.$$

That is, the probability that $\hat{B}^{t+}(S)$ overestimates the true value of $B$ is equal to $\mathrm{prob}\{\chi_k^2 \leq c\}$ for all values of $B$. A reasonable choice of $c$ would seem to be the median of the $\chi_k^2$ distribution, which is nearly equal $k - 2 + 1.34$ for $k \geq 4$. (The actual value, for $k = 3$ is $2.37 > 2(k - 2)$ and hence not allowable.) The choice $k - 2$ is too conservative if one wishes to favor small values of $B$ since the probability of underestimating the true $B$ is greater than 50 percent. This is particularly true for small values of $k$. Figure A shows the RSL and $R$ functions for various choices of $c = t(k - 2)$ in the case $k = 4$, and one can see that there is a good case for using $c$ at least as great as 3.

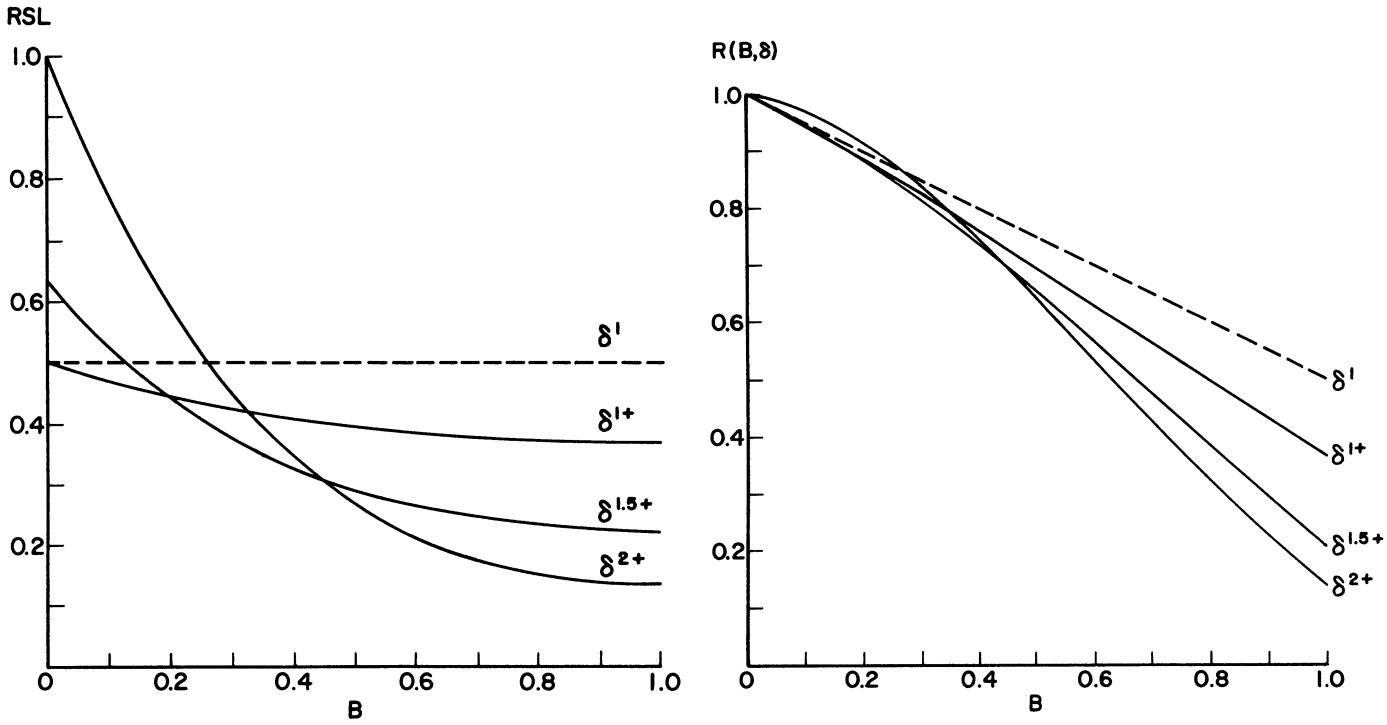
## 6. SOME OTHER TRUNCATED BAYES RULES

If we choose $h(B) \propto B^{-(a+1)}$ on $(0, 1]$, then by definition (4.3), $g(B) \propto B^{-a}$. In the proof of Theorem 2 we have already showed that the resulting Bayes rule satisfies Condition i of Baranchik's theorem, and that $\lim_{S \to \infty} \tau_h^*(S) = (k - 2a)/(k - 2)$. We need $a < 1$ in order to satisfy condition (4.1). If, in addition, we have

$$a \geq -\left(\frac{k}{2} - 2\right) \tag{6.1}$$

[2] It can be shown that if $t_1 \in (0, 1)$ and $t_2 \in [1, 2 - t_1]$ then $\mathrm{RSL}(B, \boldsymbol{\delta}_b^{t_1+}) > \mathrm{RSL}(B, \boldsymbol{\delta}_b^{t_2+})$ for every value of $B$. This fact is derived easily from Corollary 3 and (3.4) for the case $b = 1$, and extends to the general case by (5.9). We see there is no point in choosing $t < 1$.

A. RSL(B, $\delta^{t+}$) and R(B, $\delta^{t+}$), $k = 4$, $t = 1, 1.5, 2$ ($c = 2,3,4$)




then $(k - 2a)/(k - 2) \leq 2$ and Condition ii of Baranchik's theorem is also fulfilled. For $t \geq (k - 2a)/(k - 2)$ the Bayes rule is in $\mathfrak{S}_t$, without need for truncation. Such a rule is admissible as well as dominating the MLE. (Compare with Strawderman [16].)

An extended class of such rules is obtained by taking $h(B) \propto B^{-(a+1)} \exp(-cB/2)$ for $B \in (0, d)$, $h(B) = 0$ for $B \notin (0, d)$, with $a$ chosen as above, $c > 0$, and $0 < d \leq 1$. In this case,

$$\tau_h^*(S) = [(k - 2a)/(k - 2)] \cdot [S/(S + c)] \cdot [I_{k/2-a+1}(d\{S + c\}/2)/I_{k/2-a}(d\{S + c\}/2)].$$

In Section 5 we derived the James-Stein positive part estimator as a truncated Bayes rule versus a one point prior. It seems natural to investigate two point priors, three point priors, etc. To this end, let $Y$ be a binomial random variable with parameters $n$ and $p$, $Y \sim Bi(n, p)$, $0 \leq p < 1$, and suppose that $\tilde{h}(B)$ defined by (4.5) is an $n + 1$ point discrete distribution with probability masses proportional to those of a linear function of $Y$, any

$$\tilde{h}(B) \sim aY + c, \qquad P_{\tilde{h}}(B = ay + c) = \binom{n}{y} p^y (1 - p)^{n-y} \tag{6.2}$$

with $a > 0$, $c > 0$, $an + c \leq 1$. The inequalities guarantee that $\tilde{h}(B)$ is a distribution on $(0, 1]$.

Using (4.6) it is easy to show that

$$B_h^*(S) = c + anr/(r + e^{aS/2}) \tag{6.3}$$

where

$$r \equiv p/(1 - p). \tag{6.4}$$

Moreover $\tau_h^*(S)$ will satisfy Condition i of Baranchik's theorem if and only if

$$c/na \geq L_r \tag{6.5}$$

where

$$L_r = \sup_{y \leq 0} \frac{(y - 1)e^y/r - 1}{(e^y/r + 1)^2}. \tag{6.6}$$

$L_r$ increases monotonically from 0 to $\infty$ as $r$ increases from 0 to $\infty$. Assuming that (6.5) is satisfied, the resulting truncated Bayes rule is

$$\delta_h^t(S) = \left(1 - \min\left\{c + \frac{anr}{r + e^{aS/2}}, \frac{(k - 2)t}{S}\right\}\right)\mathbf{x}. \tag{6.7}$$

A short tabulation of $L_r$ as a function of $r$ is given next where $L_r$ is defined by (6.6) as a function of $r = p/(1 - p)$:

| $r$ | 0 | .1 | .3 | .5 | 1.0 | 1.5 | 2.0 | 3.0 | 4.0 | 5.0 | 6.0 | 8.0 | 16.0 | $\infty$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $L_r$ | 0 | .013 | .036 | .057 | .100 | .134 | .164 | .211 | .249 | .281 | .309 | .355 | .476 | $\infty$ |

Another class of estimators which have been mentioned often, [3, 13], employ $\hat{B}(S)$ of the form

$$\hat{B}(S) = c/(c + S) \tag{6.8}$$

$0 < c \leq 2(k - 2)$. Then $\hat{B}(S) \in B_t$, $t = c/(k - 2)$. We can easily show that these estimators are not truncated Bayes for any choice of $h(B)$ on $(0, 1]$. As a matter of fact, let $\tilde{h}(B)$ as defined by (4.5) be proportional to a scaled Gamma density on $(0, \infty)$,

$$\tilde{h}(B) \propto B^{c/2-1}e^{-cB/2} \quad B \in (0, \infty). \tag{6.9}$$

Then it is easily derived from (4.5) that

$$B_h^*(S) = c/(c + S),$$

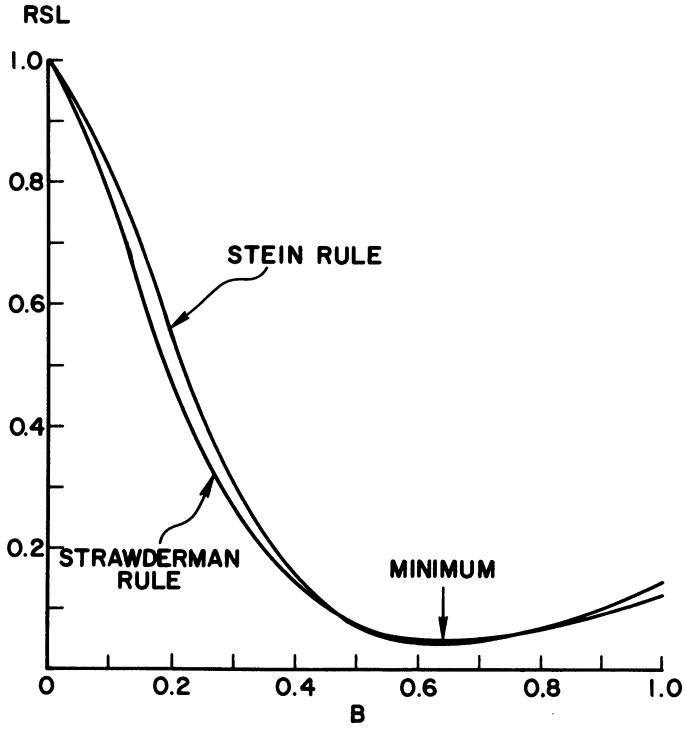
and by the uniqueness theorem for Laplace transforms this choice of $\tilde{h}$ is the only one yielding (6.8) as the Bayes rule. A look at the proof of Theorem 3 shows that none of these rules can dominate a truncated Bayes rule. The converse may not be true.

Figure B graphs $\mathrm{RSL}(B, \boldsymbol{\delta}_h^*)$ for the Bayes rule versus the uniform prior $h(B) = 1$ on $(0, 1]$. The dimension is $k = 6$. By (6.1) such a rule dominates the MLE. Also graphed is $\mathrm{RSL}(B, \boldsymbol{\delta}_{.64}^{2+})$. This is the Stein rule which has its minimum at the same place as $\boldsymbol{\delta}_h^*$ and has the same limiting value of the RSL as $B \to 0$, namely 1. It can be seen that the two curves are very similar, $\boldsymbol{\delta}_{.64}^{2+}$ having a lower minimum, as it must, and higher values near $B = 0$ and $B = 1$. This same phenomenon was observed in other numerical comparisons of Stein rules with Bayes rules of the type discussed in the first paragraph of this section, and also with the binomial rules (6.3). One can obtain lower values of the RSL over a considerable proportion of the $B$ range at the expense of a higher minimum RSL. However the lowering effect was never large enough to be considered useful. The authors feel that the extended class of Stein rules described in Section 5, besides being simple and flexible, cannot be substantially improved on in terms of their risk functions.

## 7. RULES FOR MORE COMPLICATED SITUATIONS

In a practical application of the James-Stein rule (or any of the competitors so far mentioned) all of the potential savings will disappear if any of the $|\theta_i|$ are very large. In that case $S$ will be large with high probability and $\boldsymbol{\delta}^1 = (1 - ((k-2)/S)\mathbf{x}$ will be close to the MLE $\boldsymbol{\delta}^0 = \mathbf{x}$. $\boldsymbol{\delta}^1$ shrinks $\mathbf{x}$ toward the origin, and reaping the savings depends on the origin being well chosen for the problem at hand.

*B. RSL(B, $\boldsymbol{\delta}_h^*$), h UNIFORM ON (0, 1], COMPARED WITH RSL(B, $\boldsymbol{\delta}^{2+}$); DIMENSION $k = 6$*




Of course we can choose any origin we want by subtracting arbitrary constants from the data. Or we can let the data choose the origin for us by shrinking toward a central value for the $k$ observed values $x_i$, a natural choice being $\bar{x} \equiv (1/k)\sum_{i=1}^k x_i$:

$$\delta_i^{1\prime}(x) = \bar{x} + \left(1 - \frac{k-3}{S'}\right)(x_i - \bar{x}) \quad \text{where} \quad S' = \sum_{i=1}^k (x_i - \bar{x})^2. \tag{7.1}$$

We can see that $\boldsymbol{\delta}^{1\prime}$ estimates one linear combination of the $\theta_i$, namely $\bar{\theta}$, by its maximum likelihood estimator $\bar{x}$, and applies $\boldsymbol{\delta}^1$ to the residual vector $(x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_k - \bar{x})$. The dimension of the space in which $\boldsymbol{\delta}^1$ is applied is reduced from $k$ to $k - 1$, which is reflected in the use of the constant $k - 3$ instead of $k - 2$. (We now require $k \geq 4$.) The RSL is easily calculated:

$$R(B, \boldsymbol{\delta}^{1\prime}) = \frac{1}{k}\left[1 + (k-1)\left(1 - \frac{(k-3)}{(k-1)}B\right)\right] = 1 - \frac{k-3}{k}B,$$

from (1.13) with $k$ reduced to $k - 1$, so that

$$\mathrm{RSL}(B, \boldsymbol{\delta}^{1\prime}) = 3/k. \tag{7.2}$$

The price we pay for letting the data choose the origin is an increase of the RSL from $2/k$ to $3/k$. This price is often very small compared to the reduction in the length of the vector $\theta$, in the Bayesian formulation the reduction in the effective size of $A$, which increases the possible Bayes savings $1/(A + 1)$.

From any rule $\boldsymbol{\delta} = (1 - \hat{B}(S))\mathbf{x}$ we can obtain

$$\delta_i' = \bar{x} + (1 - \hat{B}(S'))(x_i - \bar{x}) \tag{7.3}$$

with

$$\mathrm{RSL}(B, \boldsymbol{\delta}') = \frac{1}{k} + \frac{k-1}{k}\mathrm{RSL}(B, \boldsymbol{\delta})_{(k-1)} \tag{7.4}$$

where $\mathrm{RSL}(B, \boldsymbol{\delta})_{(k-1)}$ is calculated from Lemma 1, with $k$ replaced by $k - 1$.

By using $\boldsymbol{\delta}'$ instead of $\boldsymbol{\delta}$ we are expressing a suspicion that one component of the $\theta$ vector, $\bar{\theta}$, will be far from 0. A more difficult situation is that where we want to protect our savings against the possibility that one coordinate is large, but we do not know *a priori* which coordinate it will be. Our assumptions might be those of Lemma 1, but with the added reservation that one of the $\theta$, might be $\mathfrak{n}(0, A_i)$ instead of $\mathfrak{n}(0, A)$, with $A_i \gg A$. We would be trying to estimate $B$ as in Lemma 1, but with the possibility of outlier among the observations.

The authors have found the following class of rules

useful for a variety of special situations:

$$\delta_i(\mathbf{x}) = \left(1 - \frac{k-2}{S}\rho_i(\mathbf{V})\right)x_i \tag{7.5}$$

where

$$\mathbf{V} = (V_1, V_2, \cdots, V_k) \equiv \frac{1}{S}(S_1, S_2, \cdots, S_k). \tag{7.6}$$

$S_i = x_i^2$, $S = \sum_{i=1}^k x_i^2$, as before. The function $\rho_i(\mathbf{V})$ is used to express the relevance of the other observations $x_j$, $j \neq i$, to the estimation of $\theta_i$, and is called the "relevance function" in [6]. The $i$th coordinate relative savings loss, $\mathrm{RSL}_i$, given by (2.10) can be calculated under the assumptions of Lemma 1.

*Lemma 5:* Assuming $x_i|\theta_i \sim \mathfrak{n}(\theta_i, 1)$, $\theta_i \sim \mathfrak{n}(0, A)$, independently, $i = 1, 2, \cdots, k$,

$$\mathrm{RSL}_i(B, \delta_i) = \frac{2}{k} + \frac{k-2}{k}\tilde{E}^{(i)}[\rho_i(\mathbf{V}) - 1]^2 \tag{7.7}$$

$\tilde{E}^{(i)}$ indicating expectation with respect to the Dirichlet distribution on $\mathbf{V}$,

$$V_i^{\frac{1}{2}}\textstyle\prod_{j\neq i} V_j^{-\frac{1}{2}}/[\Gamma((k+2)/2)/(\Gamma(\frac{3}{2})^{k-1}\Gamma(\frac{1}{2}))].$$

(Lemma 5 is a special case of Lemma 1, except that we have let $\hat{B}$ depend on $\mathbf{V}$ as well as $S$. This does not affect the proof given for Lemma 1.) If $\rho_i$ is the same function $\rho$ for $i = 1, 2, \cdots, k$, then $\mathrm{RSL}(B, \boldsymbol{\delta}) \equiv \mathrm{RSL}^\rho$ say, given by (7.7), and does not depend on $i$. The proof of Corollary 1 no longer applies since $R(\boldsymbol{\theta}, \boldsymbol{\delta})$ is not just a function of $\|\boldsymbol{\theta}\|^2$. However, if we define $r(\|\boldsymbol{\theta}\|^2, \boldsymbol{\delta})$ as the average of $R(\boldsymbol{\theta}, \boldsymbol{\delta})$ over the sphere of radius $\|\boldsymbol{\theta}\|$, averaged with respect to the uniform distribution on the sphere, the proof gives

$$r(\boldsymbol{\theta}, \boldsymbol{\delta}) = 1 - (1 - \mathrm{RSL}^\rho)E_\theta\frac{k-2}{S}. \tag{7.8}$$

The uniformly best choice of $\rho_i(\mathbf{V})$ seems to be $\rho_i(\mathbf{V}) \equiv 1$, but this is only true if the assumptions of Lemma 5 hold. In [6], e.g., the authors took $\rho_i$ to be a function that went to 0 as $V_i$ approached 1. This choice tended to exempt an $x_i$ that was observed to be very large in magnitude compared to the other $x_j$ from shrinkage toward the origin. For a modest increase in $\mathrm{RSL}(B, \boldsymbol{\delta})$ this sharply reduced the maximum risk possible for an individual component $\theta_i$.

The authors have not attached the "outlier" problem mentioned earlier. To do so one must find an estimator $\hat{B}(\mathbf{S})$, $\mathbf{S} = (S_1, S_2, \cdots, S_k)$, that is more robust against outliers than $(k-2)/S$. If the proposed estimator is scale invariant, $\hat{B}(c\mathbf{S}) = (1/c)\hat{B}(\mathbf{S})$, then the rule will be of the form $\boldsymbol{\delta} = (1 - ((k-2)/S)\rho(\mathbf{V}))\mathbf{x}$ and Lemma 5 will apply. (Of course once a satisfactory choice of $\hat{B}(\mathbf{S})$ is found it can then be improved on by $\hat{B}^+(\mathbf{S}) = \min\{1, \hat{B}(\mathbf{S})\}$.) The simplicity of Lemma 5 makes it makes it useful in searching for a satisfactory $\hat{B}(\mathbf{S})$.

If we change the first assumption of Lemma 5 to

$$x_i|\theta_i \sim \mathfrak{n}(\theta_i, D), \text{ independently, } i = 1, 2, \cdots, k \tag{7.9}$$

then the lemma remains true if the rule $\boldsymbol{\delta}$ is redefined to be

$$\delta_i(\mathbf{x}) = \left(1 - \frac{D(k-2)}{S}\rho_i(\mathbf{V})\right)x_i. \tag{7.10}$$

If $D$ is unknown, but we observe a statistic

$$W \sim D\chi_n^2, \tag{7.11}$$

$W$ independent of $\mathbf{S}$, then the rule

$$\hat{\delta}_i(\mathbf{x}) = \left(1 - \frac{\hat{D}(k-2)}{S}\rho_i(\mathbf{V})\right)x_i \tag{7.12}$$

has

$$\mathrm{RSL}_i(B, \hat{\delta}_i) = \frac{2}{n+2} + \frac{n}{n+2}\mathrm{RSL}_i(B, \delta_i). \tag{7.13}$$

Here

$$\hat{D} = W/(n+2) \tag{7.14}$$

is the best scale invariant estimator of $D$ versus the loss function $(\hat{D} - D)^2/D^2$. Formula (7.12) says that comparisons among different rules $\delta$ do not change if $D$ is estimated from the data. It is easy to prove (7.12) in the same way as Lemma 1, and once again the details will not be given here.

## 8. THE CASE $x_i \sim \mathfrak{n}(\theta_i, D_i)$

In this section we discuss the situation

$$x_i|\theta_i \sim \mathfrak{n}(\theta_i, D_i), \text{ independently, } i = 1, 2, \cdots, k \tag{8.1}$$

where the $D_i$ are known, but are different from one another. (For example, if the original data was

$$Y_{ij} \sim \mathfrak{n}(\theta_i, \sigma^2),$$

$j = 1, 2, \cdots, n_i$, then $x_i = \sum_{j=1}^{n_i} Y_{ij}/n_i \sim \mathfrak{n}(\theta_i, \sigma^2/n_i)$ and $D_i = \sigma^2/n_i$.) We continue to assume

$$\theta_i \sim \mathfrak{n}(0, A), \text{ independently, } i = 1, 2, \cdots, k \tag{8.2}$$

which is a Bayesian statement of belief that the $\theta_i$ are of comparable magnitude. We will investigate the individual component relative savings loss of a generalized form of the James-Stein rule, $\mathrm{RSL}_i$, as defined in (2.4)-(2.11).

A very simple way to generalize the James-Stein rule for this situation is to define $\tilde{x}_i = D_i^{-\frac{1}{2}}x_i$, $\tilde{\theta}_i = D_i^{-\frac{1}{2}}\theta_i$, so that $\tilde{x}_i \sim \mathfrak{n}(\tilde{\theta}_i, 1)$, apply (1.3) to the transformed data, and then transform back to the original coordinates. The resulting rule estimates $\theta_i$ by

$$\delta_i^1 = \left(1 - \frac{k-2}{\sum_{j=1}^k S_j/D_j}\right)x_i, \quad S_j = x_j^2. \tag{8.3}$$

This is unappealing since each $x_i$ is shrunk toward the origin by the same factor. By (2.10) we know that the shrinkage factor is $(1 - B_i)$,

$$B_i = D_i/(A + D_i) \tag{8.4}$$

so that the larger $D_i$ is the more shrinkage there should be.

Another approach to the problem, which avoids the objection of the previous paragraph, is to estimate $A$

from the data, say by $\hat{A}$, substitute $\hat{A}$ into (8.4) to get $\hat{B}_i = D_i/(\hat{A} + D_i)$, and estimate $\theta_i$ by $(1 - \hat{B}_i)x_i$. Actually we will modify the details of this procedure so that it coincides exactly with the James-Stein estimator in the case where all the $D_i$ are equal.

Suppose that

$$S_j \sim (A + D_j)\chi^2_{d_j}, \text{ independently, } j = 1, 2, \cdots, k. \quad (8.5)$$

Then,

$$E_j = (S_j - d_jD_j)/d_j \quad (8.6)$$

is an unbiased estimator of $A$ with variance $1/I_j(A)$, $I_j(A)$ being the Fisher information for $A$ in $S_j$

$$I_j(A) = d_j/2(A + D_j)^2. \quad (8.7)$$

The maximum likelihood estimator of $A$ based on $\mathbf{S} = (S_1, S_2, \cdots, S_k)$ is directly calculated to satisfy the equation

$$\hat{A} = (\textstyle\sum_{j=1}^k E_j \cdot I_j(\hat{A}))/\sum_{j=1}^k I_j(\hat{A}). \quad (8.8)$$

We see that $\hat{A}$ is the minimum variance linear combination of the unbiased estimators $E_j$, except that the optimum weights $I_j(A)$ are themselves being estimated. The form (8.8) is particularly convenient for computation.

Define

$$d_i^* = 2(A + D_i)^2 \cdot \textstyle\sum_j \dfrac{d_j}{2(A + D_j)^2}. \quad (8.9)$$

The interpretation of $d_i^*$ is that (if it were an integer) $d_i^*$ independent identically distributed observations of $S_i \sim (A + D_i)\chi^2_{d_i}$ would have the same Fisher information for $A$ as does (8.5).

By Lemma 2 the estimation of $B_i$ is in terms of $\mathbf{S} = (S_1, S_2, \cdots, S_k)$ with

$$d_i = 3, \quad d_j = 1, \quad j \neq i. \quad (8.10)$$

We propose estimating $A$ by $\hat{A}_i$ as given in (8.8), (the subscript $i$ being necessary since $\mathbf{d} = (d_1, d_2, \cdots, d_k)$ depends on $i$), defining $\hat{d}_i^*$ by (8.9) with $\hat{A}$ substituted for $A$, and finally defining

$$\hat{B}_i = \frac{\hat{d}_i^* - 4}{\hat{d}_i^*} \frac{D_i}{\hat{A}_i + D_i}. \quad (8.11)$$

The estimate of $\theta_i$ is then given by $\delta_i = (1 - \hat{B}_i)x_i$.

In the case where all the $D_i$ equal $D$, say, we have $\hat{A} = S/(k + 2)$ and $d_i^* = k + 2$. Substituting into (8.11) gives $\hat{B}_i = (k - 2)D/S$. This is the James-Stein estimator (1.3), with the variance of $x_i$ given $\theta_i$ changed from 1 to $D$.

It is impossible to calculate $\mathrm{RSL}_i$ exactly for $\delta_i$, but from the definition of $d_i^*$ we expect

$$2/(d_i^* - 2) \quad (8.12)$$

to be a reasonable approximation. Actually, we would not use $\hat{B}_i$ as defined in (8.11) but rather $\hat{B}_i^+$ which chooses the nearest point in $[0, 1]$ to $\hat{B}_i$. (Notice, in this case there is the possibility of $\hat{B}_i < 0$ as well as $\hat{B}_i > 1$.)

Suppose that in a given experimental situation described by (8.1) and (8.2) one has the data $\mathbf{x}$ at hand, and has estimated $B_i$ by $\hat{B}_i$, not necessarily by the method described above. It will be convenient to define

$$C_i = 1 - B_i, \quad \hat{C}_i = 1 - \hat{B}_i \quad (8.13)$$

so that

$$\theta_i | x_i \sim \mathfrak{n}(C_ix_i, D_iC_i), \quad \text{independently, } i = 1, 2, \cdots, k. \quad (8.14)$$

The following lemmas compare the rule $\delta_i = \hat{C}_ix_i$ with the MLE $\delta_i^0 = x_i$ *conditionally on* $\mathbf{x}$.

*Lemma 6:* For the rule $\delta_i = \hat{C}_ix_i$, define $R_i(B_i, \delta_i|\mathbf{x}) \equiv E_{B_i}[L_i(\theta_i, \delta_i)|\mathbf{x}]$. Then,

$$\mathrm{RSL}_i(B_i, \delta_i|\mathbf{x}) \equiv \frac{R_i(B_i, \delta_i|\mathbf{x}) - R_i(B_i, \delta_i^*|\mathbf{x})}{R_i(B_i, \delta_i^0|x) - R_i(B_i, \delta_i^*|\mathbf{x})} = \left[\frac{\hat{B}_i}{B_i} - 1\right]^2. \quad (8.15)$$

*Proof:*

$$E_{B_i}[L_i(\theta_i, \delta_i)|\mathbf{x}] = E_{B_i}\left[\frac{(\theta_i - \hat{C}_ix_i)^2}{D_i}|\mathbf{x}\right] = D_iC_i + (\hat{C}_i - C_i)^2x_i^2$$

by (8.14). For $\delta_i^0 = x_i$ the same calculation gives

$$E_{B_i}[L_i(\theta_i, \delta_i^0)|\mathbf{x}] = D_iC_i + (1 - C_i)^2x_i^2$$

while for

$$\delta_i^* = C_ix_i, \quad E_{B_i}[L_i(\theta_i, \delta_i^*)|\mathbf{x}] = D_iC_i.$$

Lemma 6 follows from the definition of $\mathrm{RSL}_i(B_i, \delta_i|\mathbf{x})$.

Notice that $\mathrm{RSL}_i(B_i, \delta_i|\mathbf{x}) \leq 1$ if $\hat{B}_i \in [0, 2B_i]$. If we underestimate $B_i$ we will always beat the MLE, while an overestimate has to be too large by a factor of two before the MLE is preferred. The authors have considered a practical example, [4], where it is possible to assign a rough confidence interval for $B_i$ as well as the point estimate $\hat{B}_i$. In that particular case it turned out to be very unlikely that $\hat{B}_i > 2B_i$.

Formula (8.15) depends on $\hat{A}$ and $D_i$ only through $\hat{B}_i$, so it also applies if $D_i$ has to be estimated from additional data rather than being known to the statistician.

Let $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_k) = (\hat{C}_1x_1, \hat{C}_2x_2, \cdots, \hat{C}_kx_k)$ be the vector of estimates. Lemma 7 expresses the probability that $\boldsymbol{\delta}$ will be closer to the true $\boldsymbol{\theta}$ than $\boldsymbol{\delta}^0 = (x_1, x_2, \cdots, x_k)$, again conditional on the observed data $\mathbf{x}$. In the case where the $\hat{C}_i$ are different, as in (8.11), $\boldsymbol{\delta}$ will change not only the magnitude but the relative ordering of the estimated values. Lemma 8 expresses the probability, conditional on $\mathbf{x}$, that the ordering given by $\boldsymbol{\delta}$ is superior to that given by $\boldsymbol{\delta}^0$.

*Lemma 7:*

$$\text{prob}_B\{\|\boldsymbol{\delta} - \boldsymbol{\theta}\| < \|\boldsymbol{\delta}^0 - \boldsymbol{\theta}\| \,|\mathbf{x}\} = \Phi\left[\frac{1}{\sqrt{A}} \frac{\sum_{i=1}^k \hat{B}_i(B_i - \hat{B}_i/2)x_i^2}{\sqrt{\sum_{i=1}^k B_i \hat{B}_i^2 x_i^2}}\right]$$

where

$$\Phi(t) = (1/\sqrt{2\pi}) \int_{-\infty}^{t} e^{-\frac{1}{2}s^2} ds.$$

*Lemma 8:* Let

$$\rho(\boldsymbol{\delta}, \boldsymbol{\theta}) = \frac{\boldsymbol{\delta} \cdot \boldsymbol{\theta}}{\|\boldsymbol{\delta}\| \, \|\boldsymbol{\theta}\|}.$$

Then

$$\text{prob}_B\{\rho(\boldsymbol{\delta}, \boldsymbol{\theta}) > \rho(\boldsymbol{\delta}^0, \boldsymbol{\theta}) \,|\mathbf{x}\} = \Phi\left[\frac{1}{\sqrt{A}} \frac{\sum_{i=1}^k (\hat{C}_i - \bar{C}) C_i x_i^2}{\sqrt{\sum_{i=1}^k (\hat{C}_i - \bar{C})^2 (1 - C_i) x_i^2}}\right]$$

where

$$\bar{C} = \sqrt{\sum_{i=1}^k \hat{C}_i^2 x_i^2 / \sum_{i=1}^k x_i^2}.$$

The proofs of Lemmas 7 and 8 are easy multivariate normal calculations, from (8.14), and will not be given here. For an application of these ideas, the reader is referred to [4].

## 9. NON-NORMAL THEORY: EMPIRICAL LINEAR BAYES RULES IN GENERAL

The normal theory we have developed has a "wide-sense" extension to non-normal cases. We start with with very general assumptions:

$$\theta_i \sim (M_i, A_i)$$

and

$$i = 1, 2, \cdots, k \tag{9.1}$$

$$x_i|\theta_i \sim (\theta_i, D_i(\theta_i)),$$

the notation $z \sim (\mu, \sigma^2)$ indicating $z$ has mean $\mu$ and variance $\sigma^2$, with no other assumptions about the distribution of $z$. We also assume

$$ED_i(\theta_i) \equiv D_i \tag{9.2}$$

exists, the expectation being taken over the distribution of $\theta_i$ that is (partially) specified in (9.1). As before we let $B_i = D_i/(A_i + D_i)$, and use the vector notation $\mathbf{B} = (B_1, B_2, \cdots, B_k)$, $\mathbf{M} = (M_1, M_2, \cdots, M_k)$, etc. We also let $f_{\mathbf{M},\mathbf{B}}(\mathbf{x})$ indicate the marginal probability density function of $\mathbf{x} = (x_1, x_2, \cdots, x_k)$, and

$$f^{(i)}_{\mathbf{M},\mathbf{B}}(\mathbf{x}) = (B_i/D_i)\,(x_i - M_i)^2 f_{\mathbf{M},\mathbf{B}}(\mathbf{x}). \tag{9.3}$$

Since $x_i \sim (M_i, D_i/B_i)$ under (9.1), (9.2),

$$E_{\mathbf{M},\mathbf{B}}(x_i - M_i)^2 \equiv \int_{E^k} (x_i - M_i)^2 f_{\mathbf{M},\mathbf{B}}(\mathbf{x}) d\mathbf{x} = A_i + D_i = D_i/B_i,$$

so that $\tilde{f}^{(i)}_{\mathbf{M},\mathbf{B}}$ is a bona fide probability density. (The assumption that $\mathbf{x}$ has a density is for convenience only, and the results below hold in general.)

Suppose now that we wish to estimate $\theta_i$ with squared error loss function $L_i(\theta_i, \delta_i) = (\theta_i - \delta_i)^2$, and we restrict attention to rules which are linear functions of $x_i$. It is easy to see, [7, 8], that the best linear rule, i.e., the "linear Bayes rule," is

$$\delta_i^* = M_i + C_i(x_i - M_i), \tag{9.4}$$

where $C_i = 1 - B_i = A_i/(A_i + D_i)$, and that this rule has risk

$$R_i((M_i, A_i), \delta_i^*) = D_i C_i \tag{9.5}$$

exactly as in the normal case.

If we do not know $M_i$ and $C_i$, we can try to estimate them from the data $\mathbf{x}$, and use the "empirical linear Bayes rule"

$$\hat{\delta}_i = \hat{M}_i(\mathbf{x}) + \hat{C}_i(\mathbf{x}) \cdot [x_i - \hat{M}_i(\mathbf{x})]. \tag{9.6}$$

Lemma 9 compares the performance of $\hat{\delta}_i$ with that of $\delta_i^0 = x_i$, and will be shown to be the wide-sense analogue of Lemma 2.

*Lemma 9:*

$$\frac{E_{\mathbf{M},\mathbf{B}}(\hat{\delta}_i - \delta_i^*)^2}{E_{\mathbf{M},\mathbf{B}}(\delta_i^0 - \delta_i^*)^2} = \tilde{E}^{(i)}_{\mathbf{M},\mathbf{B}}\left[\frac{\hat{B}_i}{B_i} \frac{\hat{M}_i - x_i}{M_i - x_i} - 1\right]^2 \tag{9.7}$$

with $\hat{B}_i(\mathbf{x}) = 1 - \hat{C}_i(x)$. (9.7) can also be written as

$$\frac{B_i}{D_i} E_{\mathbf{M},\mathbf{B}}\left[(\hat{M}_i - M_i) + \left(\frac{\hat{B}_i}{B_i} - 1\right)(\hat{M}_i - x_i)\right]^2. \tag{9.8}$$

*Proof:* Algebra gives

$$(\hat{\delta}_i - \delta_i^*)^2 = \frac{B_i}{D_i}(x_i - M_i)^2 \left[\frac{\hat{B}_i}{B_i} \frac{\hat{M}_i - x_i}{M_i - x_i} - 1\right]^2 D_i B_i. \tag{9.9}$$

We can write $\delta_i^0 = M_i^0 + C_i^0(x_i - M_i^0)$ with $M_i^0 = M_i$ and $C_i^0 = 1$, so that $B_i^0 = 0$ and (9.9) gives

$$(\delta_i^0 - \delta_i^*)^2 = (B_i/D_i)\,(x_i - M_i)^2 D_i B_i.$$

This proves (9.7). (9.8) follows from (9.3) and (9.7).

Now define the relative savings loss as

$$\text{RSL}_i((M_i, B_i), \hat{\delta}_i) = \frac{R_i((M_i, B_i), \hat{\delta}_i) - R_i((M_i, B_i), \delta_i^*)}{R_i((M_i, B_i), \delta_i^0) - R_i((M_i, B_i), \delta_i^*)}. \tag{9.10}$$

*Theorem 5:* Suppose that the pairs $(\theta_i, x_i)$ are mutually independent of each other $i = 1, 2, \cdots, k$. Under either

of the following two conditions,

1. $\delta_i^* = E_{M_i,B_i}(\theta_i|x_i)$, i.e., that $\delta_i^*$ is Bayes as well as linear Bayes, or
2. $\hat{\delta}(\mathbf{x})$ is linear in $x_i$ (but possibly nonlinear in $x_j$, $j \neq i$), we have

$$\mathrm{RSL}_i((M_i, B_i), \hat{\delta}_i) = \frac{E_{\mathbf{M},\mathbf{B}}(\hat{\delta}_i - \delta_i^*)^2}{E_{\mathbf{M},\mathbf{B}}(\delta_i^0 - \delta_i^*)^2} = \bar{E}^{(i)}_{\mathbf{M},\mathbf{B}}\left[\frac{\hat{B}_i}{B_i}\frac{\hat{M}_i - x_i}{M_i - x_i} - 1\right]^2. \quad (9.11)$$

*Proof*:

$$E_{\mathbf{M},\mathbf{B}}(\hat{\delta}_i - \theta_i)^2 = E_{\mathbf{M},\mathbf{B}}(\hat{\delta}_i - \delta_i^*)^2 + E_{\mathbf{M},\mathbf{B}}(\delta_i^* - \theta_i)^2 + 2E_{\mathbf{M},\mathbf{B}}(\hat{\delta}_i - \delta_i^*)(\delta_i^* - \theta_i).$$

It is easily shown that the last term has expectation 0 under (1) or (2), proving the result.

*Note 1:* Condition 1 holds if the prior distribution of $\theta_i$ is conjugate to the distribution of $x_i$ given $\theta_i$, [12].

*Note 2:* Condition 2 holds if $\hat{M}_i$ and $\hat{C}_i$ are not functions of $x_i$. This has been the case in many proposed empirical Bayes estimation schemes, [7, 11], though it is definitely less efficient in the normal case as remarked after Lemma 2.

*Note 3:* If $M_i$ is known to the statistician he can take $\hat{M}_i = M_i$. In that case the right-hand side of (9.7) reduces to $\bar{E}^{(i)}_{\mathbf{M},\mathbf{B}}[\hat{B}_i/B_i - 1]^2$, just as in Lemma 2.

*Note 4:* If $B_i$ is known to the statistician he can take $\hat{B}_i = B_i$, in which case (9.8) reduces to

$$(B_i/D_i)E_{\mathbf{M},\mathbf{B}}[\hat{M}_i - M_i]^2.$$

In this case the empirical Bayes problem again reduces to a standard estimation problem.

*Note 5:* In the special case where all the $M_i = M$ and all the $B_i = B$, say, $\hat{M}$ and $\hat{B}$ are symmetric functions of the $x_i$ and the pairs $(\theta_i, x_i)$ are i.i.d., $i = 1, 2, \cdots, k$, formula (9.8) can be reduced to

$$\frac{B}{D}E_{M,B}\left[(\hat{M} - M)^2 + 2(\hat{M} - \bar{x})(\hat{M} - M)\left(\frac{\hat{B}}{B} - 1\right) + \frac{\hat{S}}{k}\left(\frac{\hat{B}}{B} - 1\right)^2\right] \quad (9.12)$$

where $\hat{S} = \sum_{i=1}^k (x_i - \hat{M})^2$. In particular, if $\hat{M} = \bar{x}$ this becomes

$$\frac{1}{k} + \frac{k-1}{k}E'_{M,B}\left(\frac{\hat{B}}{B} - 1\right)^2 \quad (9.13)$$

where $E'_{M,B}$ indicates expectation with respect to the density $(B/D)(S'/(k-1))f_{M,B}(\mathbf{x})$, $S' = \sum_{i=1}^k (x_i - \bar{x})^2$. This is a generalization of (7.4).

*Note 6:* Under the assumptions leading to (9.12), let $k \to \infty$ and for each $k$ let $\hat{M}$ and $\hat{B}$ be the maximum likelihood estimates of $M$ and $B$ based on $x$. Under sufficiently stringent regularity conditions we then have

$$\lim_{k\to\infty} k \cdot \mathrm{RSL}((M, B), \hat{\delta}) = \frac{B}{D}I^{MM} + \frac{1}{B^2}I^{BB}, \quad (9.14)$$

where $I^{MM}$ and $I^{BB}$ are the diagonal elements of the inverse of the Fisher information matrix for $(M, B)$ based on the marginal distribution of one $x_i$. In the normal case $\theta_i \sim \mathfrak{n}(M, A)$, $x_i|\theta_i \sim \mathfrak{n}(\theta_i, D)$, we have $x_i \sim \mathfrak{n}(M, D/B)$, $I^{MM} = D/B$, $I^{BB} = 2B^2$, and $(B/D)I^{MM} + (1/B^2)I^{BB} = 3$. From (7.2), this value is achieved by the James-Stein rule.

[*Received July 1972. Revised September 1972.*]

## REFERENCES


[1] Atramowitz, M. and Stegun, I., *Handbook of Mathematical Functions*, National Bureau of Standards, Applied Mathematics Series, 55, 1964.

[2] Baranchik, A., "Multiple Regression and Estimation of the Mean of the Multivariate Normal Distribution," Technical Report No. 51, Department of Statistics, Stanford University, 1964.

[3] Brown, L., "On the Admissibility of Invariant Estimators of One or More Location Parameters," *Annals of Mathematical Statistics*, 37 (October 1966), 1087–136.

[4] Efron, B. and Morris, C., "Data Analysis Using Stein's Estimator and Its Generalizations" (in preparation).

[5] ——— and Morris, C., "Empirical Bayes on Vector Observations: An Extension of Stein's Method," *Biometrika*, 59 (August 1972).

[6] ——— and Morris, C., "Limiting the Risk of Bayes and Empirical Bayes Estimators–Part II: The Empirical Bayes Case," *Journal of the American Statistical Association*, 67 (March 1972), 130–9.

[7] Griffin, B. and Krutchkoff, R., "Optimal Linear Estimators: An Empirical Bayes Version with Application to the Binomial Distribution," *Biometrika*, 58 (April 1971), 195–203.

[8] Hartigan, J., "Linear Bayes Methods," *Journal of the Royal Statistical Society*, Ser. B, 31 (December 1969), 446–54.

[9] James, W. and Stein, C., "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium*, University of California Press, 1 (1961), 361–79.

[10] Kantor, M., "Estimating the Mean of a Multivariate Normal Distribution with Applications to Time Series and Empirical Bayes Estimation," Unpublished Ph.D. thesis, Columbia University, 1967.

[11] Maritz, J., "Empirical Bayes Estimation for the Poisson Distribution," *Biometrika*, 54 (August 1967) 367–74.

[12] Raiffa, H., and Schlaifer, R., *Applied Statistical Decision Theory*, Cambridge, Massachusetts, Harvard University Press, 1961.

[13] Stein, C., "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," *Proccedings of the Third Berkeley Symposium*, University of California Press, 1 (1955) 197–206.

[14] ———, "Confidence Sets for the Mean of a Multivariate Normal Distribution," *Journal of the Royal Statistical Society*, Series B, 24 (1962) 265–96.

[15] ———, "An Approach to the Recovery of Inter-block Information in Balanced Incomplete Block Designs," *Festschrift for J. Neyman*, in F.N. David, ed., New York: John Wiley & Sons, Inc., 1966, 351–66.

[16] Strawderman, W., "Proper Bayes Minimax Estimators of the Multivariate Normal Mean," *Annals of Mathematical Statistics*, 42 (December 1971) 385–8.