The Annals of Statistics
1980, Vol. 8, No. 3, 572–585

EXAMPLES OF BERGER'S PHENOMENON IN THE ESTIMATION OF INDEPENDENT NORMAL MEANS[1]

BY L. BROWN

Rutgers University and Cornell University

Two examples are presented. In each, p independent normal random variables having unit variance are observed. It is desired to estimate the unknown means, θ_i, and the loss is of the form $L(\theta, a) = (\Sigma_{i=1}^p v(\theta_i))^{-1}\Sigma_{i=1}^p v(\theta_i)(\theta_i - a_i)^2$. The usual estimator, $\delta_0(x) = x$, is minimax with constant risk.

In the first example $v(t) = e^{rt}$. It is shown that when $r \neq 0$, δ_0 is inadmissible if and only if $p > 2$ whereas when $r = 0$ it is known to be inadmissible if and only if $p > 3$.

In the second example $v(t) = (1 + t^2)^{r/2}$. It is shown that δ_0 is inadmissible if $p > (2 - r)/(1 - r)$ and admissible if $p < (2 - r)/(1 - r)$. (In particular δ_0 is admissible for all p when $r > 1$ and only for $p = 1$ when $r < 0$.) In the first example the first order qualitative description of the better estimator when δ_0 is inadmissible depends on r, while in the second example it does not.

An example which is closely related to the first example, and which has more significance in applications, has been described by J. Berger.

1. Introduction. J. Berger (1980) has discovered a surprising phenomenon connected with the simultaneous estimation of two or more coordinate parameters. This phenomenon is more fully described below. Berger's illustration of this phenomenon involved the simultaneous estimation of gamma scale parameters. We show in Example 1 that the same unexpected behavior can occur in the context of simultaneous estimation of normal means. It will thus be seen that the phenomenon is caused by the form of the loss function, and does not rely for its existence on the type of coordinate distribution (e.g. gamma distribution).

The calculations for our normal example are simpler and more generally familiar than those for Berger's gamma example. Because of this our example makes it easier to understand Berger's phenomenon. It also makes it possible to discuss some peripheral aspects of this phenomenon which are not addressed in Berger's paper. Especially, Example 2 exhibits a form of this phenomenon which was not previously observed in the context of estimating scale parameters, but may have been observed in the simultaneous estimation of Poisson parameters.

On the other hand, the loss functions which appeared in Berger (1980) were fairly natural for his gamma problem whereas the loss functions we use below are, of necessity, artificial from a practical point of view. The results below are offered as mathematically simple illustrations and explanations of a phenomenon whose statistical importance is found in other contexts.

Received October 1978; revised March 1979.

[1]Research supported in part by NSF Research Grant Number MCS 75-23343-A01.

AMS 1970 *subject classification.* Primary 62C15, 62F10; secondary 62C10.

Key words and phrases. Estimation, admissibility, estimating several normal means.

Berger's phenomenon. Let X_i, $i = 1, \cdots, p$ be independent random variables having distributions F_{θ_i}, $i = 1, \cdots, p$, $\theta_i \in \Theta^* \subset R$. For the moment consider the ith coordinate problem of estimating θ_i, on the basis of X_i, with loss function $L^*(\theta_i, a_i)$, $L^* \geqslant 0$. Let $\delta_i^* : X_i \to \mathcal{Q}_i$ denote an admissible constant-risk estimator (nonrandomized) for this problem. (Thus δ_i^* is minimax.)

Now, consider the problem of estimating $\theta = (\theta_1, \cdots, \theta_p)$, on the basis of $X = (X_1, \cdots, X_p)$ with loss function

$$L(\theta, a) = (\Sigma_{i=1}^p v(\theta_i))^{-1} \Sigma_{i=1}^p v(\theta_i) L^*(\theta_i, a_i). \tag{1}$$

The function v is a given positive weight function. (The initial normalization factor $(\Sigma v(\theta_i))^{-1}$ has no bearing on admissibility considerations.) A reasonable procedure for such a problem is the coordinatewise procedure $\delta_0 = (\delta_1^*, \cdots, \delta_p^*)$. This procedure will have constant risk (because of the normalization in (1)). In most situations, including all those discussed below, it will be minimax. But—is it admissible?

It is now well known that the answer to this question may be no. This is Stein's phenomenon. Stein's original example (Stein (1956), James and Stein (1960)) concerned the normal means problem in which F_θ is the normal distribution with mean θ and variance one, $\Theta^* = R$, and the loss is squared error—that is, $L^*(\theta_i, a_i) = (\theta_i - a_i)^2$, and $v \equiv 1$. The procedure δ_0 is given by $\delta_0(x) = x$. This procedure is inadmissible if (and only if) $p \geqslant 3$. If $p \geqslant 3$ δ_0 is dominated by another procedure, δ', say. A first order description of the behavior of δ' is that δ' pulls the estimate of δ_0 in towards a specific point (e.g., the origin).

In Berger's example (Berger (1980)) $\{F_\theta\}$ is a gamma scale family with scale parameter $1/\theta$ and

$$L^*(\theta_i, a_i) = (1 - a_i\theta_i)^2. \tag{2}$$

Berger found that (i) *the critical dimension for inadmissibility of δ_0 may be either $p = 2$ or $p = 3$ depending on the choice of v, and* (ii) *the first order description of the δ' which improves on δ_0 also depends on the choice of v.* More specifically, Berger found that when $v(t) = t^m$ the critical dimension $p = 2$ when $m \neq 0$ and $p = 3$ when $m = 0$. Considering only first order terms we may describe δ' as an estimator obtained from δ_0 by pulling towards $\mathbf{0}$ when $m > 0$, towards a specific point, say $\mathbf{1} = (1, \cdots, 1)$, when $m = 0$, and towards $\boldsymbol{\infty} = (\infty, \cdots, \infty)$ when $m < 0$.

Example 1, below, exhibits both of these characteristics of Berger's phenomenon in the context of estimating several normal means. Part of the discussion which follows the example explains the intimate relationship between Berger's example and our own.

Example 2 exhibits the first general characteristic above in an exaggerated form. The critical dimension for inadmissibility of δ_0 may be any value of $p \geqslant 2$, depending on the choice of v. It is even possible to choose v so that δ_0 is admissible for any value of p. Example 2 does not exhibit the second general characteristic found in both Berger's example and Example 1.

2. Berger's phenomenon in the normal means problem. The following example shows that in the normal means problem the critical dimension for inadmissibility may be $p = 2$ instead of the value $p = 3$ which is familiar from Stein's example. The following example differs from Stein's familiar one only in the choice of the weight function, v. Furthermore the competing estimators have different first order descriptions. It is thus an instance of Berger's phenomenon.

EXAMPLE 1. Consider the normal means problem with $L^*(\theta_i, a_i) = (\theta_i - a_i)^2$ and L defined by (1) and with $v(t) = e^{rt}$. If $r \neq 0$ then δ_0 is inadmissible if and only if $p \geqslant 2$. A better estimator is $\delta' = (\delta_1', \cdots, \delta_p')$ with

$$\delta_i'(x) = x_i + (\operatorname{sgn}(r))c\frac{e^{-rx_i}}{\Sigma_{j=1}^p e^{-rx_j}} \tag{3}$$

where $0 < c < 2|r|(p-1)/e^{3r^2}$.

PROOF. Write $\gamma(x) = \delta'(x) - x$. Then Stein's unbiased estimator of the risk yields

$$\begin{aligned} \Delta(\theta) &= (\Sigma e^{r\theta_j})[R(\theta, \delta_0) - R(\theta, \delta')] \\ &= E_\theta(-\Sigma e^{r\theta_i}(2\gamma_i^{(i)}(x) + \gamma_i^2(x))) \end{aligned} \tag{4}$$

where $\gamma_i^{(j)} = \partial/(\partial x_j)\gamma_i$. See Stein (1973).

Now, let $g : R^p \to R^1$ and let e^i denote the ith coordinate vector in R^p. Suppose $E_\theta(g)$ exists. Then

$$\begin{aligned} \int e^{r\theta_i}g(x)e^{-\|x-\theta\|^2/2}\,dx &= \int e^{rx_i}g(x)e^{r(\theta_i - x_i) - \|x-\theta\|^2/2}\,dx \\ &= \int e^{rx_i}g(x)e^{r^2/2}e^{-\|x+re^i-\theta\|^2/2}\,dx \\ &= e^{r^2/2}\int e^{r(y_i - r)}g(y - re^i)e^{-\|y-\theta\|^2/2}\,dy \\ &= e^{-r^2/2}\int e^{ry_i}g(y - re^i)e^{-\|y-\theta\|^2/2}\,dy. \end{aligned} \tag{5}$$

Hence, (4) may be rewritten as

$$\Delta(\theta) = e^{-r^2/2}E_\theta(-\Sigma_{i=1}^p e^{rx_i}(2\gamma_i^{(i)}(x - re^i) + \gamma_i^2(x - re_i)) \tag{6}$$

A direct calculation yields

$$\gamma_i^{(i)}(x) = -\frac{c|r|e^{-rx_i}}{\Sigma e^{-rx_j}}\left(\frac{\Sigma_{j\neq i}e^{-rx_j}}{\Sigma e^{-rx_j}}\right)$$

so that (for $c \geqslant 0$)

$$\begin{aligned} -2e^{rx_i}\gamma_i^{(i)}(x - re^i) &= 2c|r|\frac{e^{r^2}\Sigma_{j\neq i}e^{-rx_j}}{\left(e^{-rx_i + r^2} + \Sigma_{j\neq i}e^{-rx_j}\right)^2} \\ &\geqslant \frac{2c|r|\Sigma_{j\neq i}e^{-rx_j}}{e^{r^2}(\Sigma e^{-rx_j})^2} \end{aligned} \tag{7}$$

while

$$-e^{rx_i}\gamma_i^2(x - re^i) = -c^2\frac{e^{-rx_i+2r^2}}{\left(e^{-rx_i+r^2} + \Sigma_{j\neq i}e^{-rx_j}\right)^2} \tag{8}$$

$$\geqslant -c^2e^{2r^2}\frac{e^{-rx_i}}{(\Sigma e^{-rx_j})^2}.$$

It follows that

$$\Delta(\theta) \geqslant E_\theta\left(\frac{e^{-r^2/2}}{\Sigma e^{-rx_j}}\left(\frac{2c|r|(p-1)}{e^{r^2}} - c^2e^{2r^2}\right)\right). \tag{9}$$

When $0 < c < 2|r|(p-1)/e^{3r^2}$ then the expectand in (7) is always positive and hence $\Delta(\theta) > 0$ for all θ. This completes the proof that δ' is better than δ_0. $\square$

Note that the inequalities (7) and (8) are somewhat crude. Thus it may be that an estimator of the form of δ' (see (3)) dominates δ_0 even when c is somewhat larger than $2|r|(p-1)/e^{3r^2}$. It is also true therefore that the inequality in (9) will be rather crude for many values of $\theta \in R^p$. However, note that the expression on the right of (9) does at least give the correct order of magnitude for $\Delta(\theta)$ when $c < 2|r|(p-1)/e^{3r^2}$ since for any $c \geqslant 0$

$$\Delta(\theta) \leqslant E_\theta\left(\frac{e^{-r^2/2}}{\Sigma e^{-rx_j}}(2c|r|(p-1) - c^2)\right).$$

(This can be seen by writing reverse inequalities analogous to (7), (8).)

The behavior of δ'. Berger's phenomenon, as we previously described it, also involves the dependence on the choice of v of the qualitative behavior of δ'. The procedure δ' of the above example exhibits this type of dependence. When $r > 0$ then $\gamma_i(x) = \delta_i'(x) - \delta_{0i}(x) > 0$ for all x and all $i = 1, \cdots, p$. Thus the first order description of δ' is that it pulls δ_0 toward ∞. Similarly when $r < 0$ the first order description of δ' is that it pulls δ_0 toward $-\infty$. And, again, when $r = 0$ δ_0 is inadmissible only when $p \geqslant 3$, and then the dominating δ' pulls δ_0 toward a given point, e.g., $\mathbf{0}$.

This first order description does not explain why δ' dominates δ_0. To see this it is necessary to examine more closely the behavior of δ'. Consider the case $p = 2$ and $r > 0$. Transform coordinates to $z_1 = (x_1 + x_2)/2^{\frac{1}{2}}$, $z_2 = (x_1 - x_2)/2^{\frac{1}{2}}$ and make the same transformation in the decision space. In these new coordinates γ becomes $\gamma(z) = c(2^{-\frac{1}{2}}, -2^{-\frac{1}{2}}\tanh(rz_2/2^{\frac{1}{2}}))$. Note that in this system the first coordinate of γ is a constant while the second is a strictly decreasing odd function of z_2 only, and the magnitude of the second coordinate is smaller than of the first. In particular, $\gamma(z)$ depends only on z_2 (i.e., the distance of x from the line $x_1 = x_2$).

It is now possible to better understand from some calculations similar to those in the proof of Example 1 why δ' dominates δ_0. (The calculations are left for the reader.) Qualitatively the situation is as follows: when $\theta_1 = \theta_2$ the loss, L, is just a multiple of the usual squared error loss. Considering the behavior of γ in the z-system we see a constant bias $(2^{-\frac{1}{2}}c)$ in the z_1 direction and a pulling inward of γ in the z_2 direction toward the true second coordinate value of zero. Breaking the risk into a z_1 and a z_2 coordinate shows a deterioration of risk in the z_1 direction when using δ' and an improvement in the z_2 direction which outweighs this deterioration. Qualitatively the same idea holds when $|\theta_1 - \theta_2|$ is (very) small, although, of course, it is not strictly possible to break the risk into z_1 and z_2 coordinates since the loss is no longer a multiple of squared error.

When $|\theta_1 - \theta_2|$ is large δ' improves on δ_0 for a different reason. To be specific, consider the case where θ_1 is large and θ_2 is not. Then the x_2 coordinate of γ (in the original coordinate system) is approximately a constant (c) and this results in a considerable deterioration of quadratic risk in the x_2 coordinate direction. However in computing the contribution to L this deterioration is weighted by $e^{r\theta_2}/(e^{r\theta_1} + e^{r\theta_2})$, which is extremely small. The coordinate risk in the x_1 direction gives most of the contribution to the overall risk under L. This coordinate risk is more complex to analyze. Briefly, there is a small first order positive bias in the x_1 coordinate of δ' (i.e., γ_1 is small and positive near $x = \theta$) but there is also a second order bias toward the true value of θ_1. This second order bias is also small, but it is enough so that in the computation of risk it lends enough improvement to outweigh the deterioration due to the small first order bias in the x_1-direction and it also outweighs the deterioration of risk in the x_2-direction *because the weight factor* $e^{r\theta_2}/(e^{r\theta_1} + e^{r\theta_2})$ *is very small.*

Relation to Berger's gamma example. Example 1 is intimately related to Berger's example involving simultaneous estimation of gamma scale parameters (Berger (1980)). If x_i is gamma with scale parameter $1/\theta_i$ then $y_i = \ln x_i$ has distribution from a location family "centered" on $\xi_i = -\ln \theta_i$. Berger's loss functions are

$$L_B(\theta, a) = \Sigma \theta_i^m (1 - \alpha_i \theta_i)^2$$

for various values of m.

In the transformed problem these become

$$L(\xi, b) = \Sigma e^{-m\xi_i}(1 - e^{b_i - \xi_i})^2 \tag{10}$$

where $b_i = \ln a_i$. The weight functions here are $v(\xi) = e^{-m\xi}$ which are the same as those in our Example 1 with $r = -m$. The other part of the loss (10) (namely $(1 - e^{b_i - \xi_i})^2$) is, of course, not the same as the quadratic term in our Example 1, (nor is y_i a normal variable). But there is considerable evidence that terms of the form $(1 - e^{b_i - \xi_i})^2$ in the loss function of a location problem have the same admissibility characteristics as simple quadratic terms. See, for example, Farrell (1964), Brown (1966), and especially Brown (1979).

Not only are the settings of Berger's example and ours closely related in this way, but the results bear the corresponding relation, as should be expected. In particular, when $m \neq 0$ (corresponding to $r \neq 0$) the usual coordinatewise estimator is inadmissible if and only if $p \geqslant 2$ while the critical dimension is $p = 3$ when $m = 0$ (corresponding to $r = 0$). The first order description of the improving δ' in the original coordinate system is that it pulls δ_0 toward $\boldsymbol{\infty}$, $\mathbf{1}$, or $\mathbf{0}$ depending on whether $m < 0$, $= 0$, or > 0 (corresponding to pulls toward $\boldsymbol{\infty}$, $\mathbf{0}$, $-\boldsymbol{\infty}$ when $r > 0$, $= 0$, < 0).

Relation to the Poisson example. A form of Berger's phenomenon has also been observed when the X_i are independent Poisson (λ_i) variables and

$$L(\lambda, a) = \Sigma_{i=1}^{p} \lambda_i^{-r}(a_i - \lambda_i)^2.$$

When $r = \frac{1}{2}$ the critical dimension for inadmissibility is $p = 2$. (See Clevenson and Zidek (1977) and Tsui and Press (1978).) When $r = 0$ the critical dimension appears to be $p = 3$. (Inadmissibility for $p \geqslant 3$ is proved in Peng (1975). See also Hudson (1978) and Tsui (1978). Admissibility for $p = 2$ is *conjectured* in Brown (1979, Section 2.3) and proved in Peng (1975).) The estimators δ' which have been found to improve on $\delta_0(x) = x$ in the two cases are functionally different, but this difference cannot be described as a difference in their first order behavior. Brown (1979, Section 2.3) contains a detailed heuristic discussion of this problem. The type of heuristics used there are related to the heuristics used above for the discussion of Berger's example, but the situation appears to be more complex to explain. (Perhaps it is only that we do not understand it as well.) We will not attempt to paraphrase that discussion here. However, it has led us to feel that this Poisson example is more closely related to our Example 2, to follow, than to Example 1.

3. A second phenomenon. As previously noted, the Poisson example exhibits the first characteristic of Berger's phenomenon (the dividing dimension for inadmissibility depends on v) but not the second (a change in the first order description of the improved estimators). Example 2, below, shows that this same behavior can be exhibited in a normal means problem. By heuristic extension it should, therefore, be observable in various other contexts as well.

The most novel feature of Example 2, moreover, is that the dividing dimension for inadmissibility may be any value of p (including $p = \infty$) depending on the norming exponent in the definition of v.

EXAMPLE 2. Consider the normal means problem with $L^*(\theta_i, a_i) = (\theta_i - a_i)^2$ and L defined by (1) with $v(t) = (1 + t^2)^{r/2}$. Then, the usual estimator, $\delta_0(x) = x$, is admissible if $r < 1$ and $1 \leqslant p < (2 - r)/(1 - r)$. If $r \geqslant 1$ δ_0 is admissible for any p. Conversely, when $r < 1$ and $p > (2 - r)/(1 - r)$ then δ_0 is inadmissible.

(Some aspects of the above example should be noted. The loss functions used above were chosen for illustrative purposes only. Although loss functions of this nature may not be realistic in practical applications they do at least appear more reasonable than those used in Example 1.

An estimator, δ', is described in (29), below, which dominates δ_0 when $0 \leqslant r < 1$ and $p > (2 - r)/(1 - r)$. However no attempt was made to find an estimator yielding the greatest improvement in any sense. Indeed, unless the loss functions above are found to be of practical interest such an endeavor would probably be pointless, although it might be of some interest to investigate the variety of possible functional forms for δ'.

The statement of the example leaves open the question of admissibility when $p = (2 - r)/(1 - r)$. Heuristics, as well as the known result for the case $r = 0$ ($p = 2$), suggest that δ_0 is admissible when $p = (2 - r)/(1 - r)$. Since this question is of secondary importance and would require more complex calculations than those in the proof below, it has not been investigated.)

PROOF. Consider the sequence of priors, H_σ, $\sigma = 1, 2, \cdots$, having densities

$$h_\sigma(\theta) = \frac{c_\sigma(\Sigma v(\theta_i))\exp(-\|\theta\|^2/2\sigma^2)}{\Pi v(\theta_i)}. \tag{11}$$

Familiar calculations show that the Bayes procedure for the prior G_σ is $\delta_\sigma(x) = \sigma^2 x/(1 + \sigma^2)$. Also,

$$R(\theta, \delta_0) - R(\theta, \delta_\sigma) = (\Sigma v(\theta_i))^{-1}\Sigma v(\theta_i)\left(\frac{2\sigma^2 + 1 - \theta_i^2}{(1 + \sigma^2)^2}\right). \tag{12}$$

A sufficient condition for admissibility is that

$$I_\sigma = \int c_\sigma^{-1}(R(\theta, \delta_0) - R(\theta, \delta_\sigma))h_\sigma(\theta)\, d\theta \to 0 \quad \text{as } \sigma \to \infty. \tag{13}$$

See Blyth (1951), Stein (1959), Farrell (1964). Then,

$$\begin{aligned} I_\sigma &= p\left(\int \frac{2\sigma^2 + 1 - \theta_i^2}{(1 + \sigma^2)^2} e^{-\theta_i^2/2\sigma^2}\, d\theta_i\right)\left(\int \frac{1}{v(\theta_j)} e^{-\theta_j^2/2\sigma^2}\, d\theta_j\right)^{p-1} \\ &= p\frac{\sigma}{1 + \sigma^2}\left(\int \frac{1}{(1 + \theta^2)^{r/2}} e^{-\theta^2/2\sigma^2}\, d\theta\right)^{p-1} \qquad (14) \\ &= p0(\sigma^{-1})0\big((1 + \sigma^2)^{(p-1)(1-r)^+/2}\big) \to 0 \end{aligned}$$

when $p < (2 - r)/(1 - r)$ or when $r \geqslant 1$.

This proves the admissibility assertion.

Motivation for converse. An argument which motivates the converse result is as follows: Stein's unbiased estimator of the risk yields

$$\begin{aligned} \Delta(\theta) &= \big(\Sigma(1 + \theta_i^2)^{r/2}\big)(R(\theta, \delta_0) - R(\theta, \delta')) \qquad (15) \\ &= \Sigma(1 + \theta_i^2)^{r/2}E_\theta(-2\gamma_i^{(i)}(x) - \gamma_i^2(x)) \end{aligned}$$

where $\gamma(x) = \delta'(x) - x$ and $\gamma_i^{(j)} = (\partial/\partial x_j)\gamma_i$. Now, if γ_i and $\gamma_i^{(i)}$ are very smooth functions then one would expect that $E_\theta(\gamma_i^2(x)) \cong \gamma_i^2(\theta)$ and $E_\theta(-2\gamma_i^{(i)}(x)) \cong$

$-2\gamma_i^{(i)}(\theta)$. Hence one would expect that

$$\Delta(\theta) \cong \Sigma(1+\theta_i^2)^{r/2}(-2\gamma_i^{(i)}(\theta) - \gamma_i^2(\theta)) = \Delta_{\text{est}}(\theta) \qquad \text{(say)}. \tag{16}$$

Berger has examined the general differential inequality of the form $\Delta_{\text{est}} \geqslant 0$. See Berger (1980, Theorem 1). He finds that defining γ by

$$\gamma_i(x) = -\frac{cg(x_i)}{b + \Sigma|g(x_i)|^{(2-r)/(1-r)}} \tag{17}$$

where

$$g(t) = \textstyle\int_0^t (1+v^2)^{-r/2}\, dv \tag{18}$$

yields $\Delta_{\text{est}}(\theta) \geqslant 0$ for all θ when $c > 0$ is sufficiently small. The factor $b \geqslant 0$ is inserted in (17) to (hopefully!) yield the desired smoothness in γ when b is sufficiently large. A crucial hypothesis of Berger's (1980, Theorem 1) is that

$$K(b) = \sup \frac{(1+\theta_i^2)^{r/2} g^2(\theta_i)}{b + \Sigma|g(\theta_i)|^{(2-r)/(1-r)}} < \infty.$$

This hypothesis is satisfied here since g defined by (18) satisfies $|g(t)| \sim (1-r)^{-1} t^{1-r}$ as $|t| \to \infty$.

Unfortunately the choice of γ in (17) does not seem to us to be smooth enough to guarantee the validity of (16) (except when $r = 0$) no matter how $b \geqslant 0$ is chosen. (The difficulty with (16) occurs when one, or more, coordinates of θ are near 0.) Hence we use a somewhat different argument.

PROOF OF CONVERSE. Consider only the case $0 \leqslant r < 1$. Let

$$\gamma_i(x) = \frac{-c(\operatorname{sgn} x_i)|x_i|^{1-r}}{D(x)} \tag{19}$$

where

$$D(x) = \Sigma|x_i|^{2-r}.$$

(This is of the form (17) with $b = 0$ but with $g(t) = \int_0^t (v^2)^{-r/2}\, dv$ replacing (18).) Let $\Delta(\theta)$ be defined by (15). We will first show

$$\liminf_{\|\theta\| \to \infty} D(\theta)\Delta(\theta) > 0 \tag{20}$$

where $c > 0$ is sufficiently small. We will then invoke the necessary condition for admissibility of Brown (1980) in order to show that δ_0 is inadmissible when $p > (2-r)/(1-r)$.

Now,

$$\gamma_i^{(i)}(x) = -c(1-r)\frac{|x_i|^{-r}}{D(x)} + c(2-r)\frac{|x_i|^{2-r}}{D(x)}\frac{|x_i|^{-r}}{D(x)}. \tag{21}$$

Let $Z_1, \cdots, Z_p$ be independent $N(0, 1)$ variables. Then

$$(1+\theta_i^2)^{r/2} D(\theta) E_\theta\left(\frac{|x_i|^{-r}}{D(x)}\right) = E\left(\frac{|\theta_i + z_i|^{-r} D(\theta)}{(1+\theta_i^2)^{-r/2} D(\theta + z)}\right).$$

Note that $D(\theta)/D(\theta + z) \to 1$ as $\|\theta\| \to \infty$ for each fixed z. Furthermore

$$\frac{|\theta_i + z_i|^{-r} D(\theta)}{\left(1 + \theta_i^2\right)^{-r/2} D(\theta + z)} = 0(1 + \Sigma |z_i|^{-r})$$

uniformly in θ. Note that $E(1 + \Sigma |z_i|^{-r}) < \infty$. Also note that $(\theta_i + z_i)^2/(1 + \theta_i^2) \to 1$ if $|\theta_i| \to l = \infty$ and $\to (l + z_i)^2/(1 + l^2)$ if $\theta_i \to l < \infty$. Hence

$$\left(1 + \theta_i^2\right)^{r/2} D(\theta) E_\theta\left(\frac{|x_i|^{-r}}{D(x)}\right) \to E\left[\left(\frac{(l + z_i)^2}{(1 + l^2)}\right)^{-r/2}\right]$$

as $\|\theta\| \to \infty$, $\theta_i \to l$, by the dominated convergence theorem. Another way of expressing the above is

$$(22) \qquad \left(1 + \theta_i^2\right)^{r/2} D(\theta) E_\theta\left(\frac{|x_i|^{-r}}{D(x)}\right) \sim E_\theta\left(\frac{|x_i|^{-r}}{\left(1 + \theta_i^2\right)^{-r/2}}\right) \quad \text{as } \|\theta\| \to \infty.$$

Now, $|x_i|^{-r} = (x_i^2)^{-r/2}$ is convex in x_i^2, and $E_\theta(x_i^2) = 1 + \theta_i^2$. Hence

$$(23) \qquad E_\theta(|x_i|^{-r}) \geqslant \left(1 + \theta_i^2\right)^{-r/2}$$

by Jensen's inequality. Combining (22) and (23) yields

$$(24) \qquad \liminf_{\|\theta\| \to \infty} \left(1 + \theta_i^2\right)^{r/2} D(\theta) E_\theta\left(\frac{|x_i|^{-r}}{D(x)}\right) \geqslant 1.$$

In a similar fashion it is possible to conclude that if $\|\theta\| \to \infty$ so that $|\theta_i|^{2-r}/D(\theta) \to l$ then

$$(25) \qquad D(\theta)\left(1 + \theta_i^2\right)^{r/2} E_\theta\left(\frac{|x_i|^{2-r}}{D(x)} \frac{|x_i|^{-r}}{D(x)}\right) \to \frac{|\theta_i|^{2-r}}{D(\theta)} = l.$$

(The cases $l = 0$ and $l > 0$ can most conveniently be handled separately.) Also,

$$(26) \quad D(\theta)\left(1 + \theta_i^2\right)^{r/2} E_\theta\left(\gamma_i^2(x)\right) = D(\theta)\left(1 + \theta_i^2\right)^{r/2} E_\theta\left(\frac{c^2 |x_i|^{2-r} |x_i|^{-r}}{D^2(x)}\right) \sim \frac{|\theta_i|^{2-r}}{D(\theta)}$$

when $\|\theta\| \to \infty$.

Combining (15), (24), (25), (26) yields

$$(27) \qquad \liminf_{\|\theta\| \to \infty} D(\theta) \Delta(\theta) \geqslant c[2p(1 - r) - 2(2 - r) - c] > 0$$

for $0 < c < 2(p(1 - r) - (2 - r))$. (Here we obviously need the condition $p > (2 - r)/(1 - r)$.) This verifies (20).

Now, let $G(d\theta)$ be any nonnegative locally finite measure (generalized prior) on R^p. The corresponding generalized Bayes procedure relative to the loss function L

is given by $\delta_G = x + \gamma$ where

$$\gamma_i(x) = \frac{\hat{g}_i^{(i)}(x)}{\hat{g}_i(x)}$$

with

$$\hat{g}_i(x) = \int \left(\Sigma\left(1 + \theta_j^2\right)^{r/2}\right)^{-1}\left(1 + \theta_i^2\right)^{r/2} e^{-\|\theta - x\|^2/2} G(d\theta)$$

and $\hat{g}_i^{(i)} = (\partial/\partial x_i)\hat{g}_i$. If δ_0 is generalized Bayes then $(\partial/\partial x_i) \ln \hat{g}_i \equiv 0$ so that $\hat{g}_i$ is independent of x_i. A standard completeness argument shows that the conditional distribution of $(\Sigma(1 + \theta_j^2)^{r/2})^{-1}(1 + \theta_i^2)^{r/2}G(d\theta)$ given $\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_p$ must be proportional to Lebesgue measure; $i = 1, \cdots, p$. It follows that G must have density $g(\theta) = k(\Sigma(1 + \theta_j^2)^{\frac{1}{2}})(\Pi(1 + \theta_j^2)^{-r/2})$ relative to Lebesgue measure. To apply the criterion for inadmissibility in Brown (1980) let $h(\theta) = 1 + D(\theta)\Sigma(1 + \theta_j^2)^{r/2}$ so that $\liminf_{\|\theta\| \to \infty} h(\theta)(R(\theta, \delta_0) - R(\theta, \delta)) > 0$ by (20). Note that $D(\theta)\Pi(1 + \theta_j^2)^{r/2} = 0(\|\theta\|^{2-r+pr})$ when $\|\theta\| > 1$. Thus, for G as above,

$$(28)\quad \int h^{-1}(\theta)G(d\theta)$$
$$= k\int\left(1 + D(\theta)\Sigma(1 + \theta^2)^{r/2}\right)^{-1}\Sigma\left(1 + \theta_j^2\right)^{r/2}\Pi\left(1 + \theta_j^2\right)^{-r/2} d\theta$$
$$\geqslant k'\int_{\|\theta\| \geqslant 1} \frac{1}{D(\theta)\Pi\left(1 + \theta_j^2\right)^{r/2}}\, d\theta$$
$$\geqslant k''\int_{\|\theta\| \geqslant 1} \|\theta\|^{2-r+pr}\, d\theta$$
$$= \infty$$

when $p \geqslant (2 - r)/(1 - r)$. This implies that δ_0 is inadmissible by Brown (1980). This concludes the proof for the case $0 \leqslant r < 1$. (This proof is also valid for $r \leqslant -2$.)

NOTE. Considerably more calculations, somewhat similar to those involved in (24)–(26), suffice to show that if $\delta' = x + \gamma$ where

$$(29)\qquad \gamma_i(x) = -\frac{c(\operatorname{sgn} x_i)|x_i|^{1-r}}{b + \Sigma|x_i|^{2-r}},$$

if $1 > r \geqslant 0$, $p > (2 - r)/(1 - r)$, $0 < c < 2[p(1 - r) - (2 - r)]$, and b is sufficiently large then δ' dominates δ_0. In fact

$$(30)\quad \left(b + \Sigma|\theta_i|^{2-r}\right)\Sigma\left(1 + \theta_i^2\right)^{r/2}\left(R(\theta, \delta_0) - R(\theta, \delta')\right)$$
$$\geqslant c\left[2p(1 - r) - 2(2 - r) - c\right] \qquad \text{for all } \theta.$$

When $-2 \leqslant r < 0$ it is not true that $(x_i^2)^{-r/2}$ is convex in x_i^2; and thus the argument leading to (23) need not be valid. The preceding arguments therefore need some modification. We sketch a suitable argument below which is valid for all $r < 0$.

Let $r < r' < 0$ and define γ by (19) but with the value r' in place of r. Computations like the preceding yield

$$(31)\qquad D(\theta)\Delta(\theta) \sim \Bigg[2c(1-r')\Sigma(1+\theta_i^2)^{(r-r')/2}E\left[\frac{|x_i|^{-r'}}{(1+\theta_i^2)^{-r'/2}}\right] - 2c(2-r')\Sigma(1+\theta_i^2)^{(r-r')/2}\frac{|\theta_i|^{2-r'}}{D(\theta)} - c^2\Sigma(1+\theta_i^2)^{(r-r')/2}\frac{|\theta_i|^{2-r'}}{D(\theta)}\Bigg] \quad \text{as } \|\theta\| \to \infty$$

so long as c is chosen to make the right side of (31) always positive for large θ. Now,

$$E\left[\frac{|x_i|^{-r'}}{(1+\theta_i^2)^{-r'/2}}\right] > \varepsilon > 0 \quad \text{and} \quad \to 1 \text{ as } |\theta_i| \to \infty.$$

Calculations then show that

$$\liminf_{\|\theta\|\to\infty}\left[\Sigma(1+\theta_i^2)^{(r-r')/2}E\left[\frac{|x_i|^{-r'}}{(1+\theta_i^2)^{-r'/2}}\right]\right]\left(\Sigma(1+\theta_i^2)^{(r-r')/2}\frac{|\theta_i|^{2-r'}}{D(\theta)}\right)^{-1} \geqslant p$$

since $r - r' < 0$. Thus

$$(32)\qquad \liminf_{\|\theta\|\to\infty}\left(\Sigma(1+\theta_i^2)^{(r-r')/2}\frac{|\theta_i|^{2-r'}}{D(\theta)}\right)^{-1} D(\theta)\Delta(\theta) \geqslant \left[2cp(1-r') - 2c(2-r') - c^2\right] > 0$$

for $\theta < c < p(1-r') - (2-r')$. (Thus, incidentally, validating (31) for these values of c.) The inadmissibility criterion of Brown (1980) can be used as before, except that here, of course, the obvious appropriate definition of h is

$$h(\theta) = 1 + \left(\Sigma(1+\theta_j^2)^{r'/2}\right)\left(\Sigma(1+\theta_i^2)^{(r-r')/2}\frac{|\theta_i|^{2-r'}}{D(\theta)}\right)^{-1} D(\theta).$$

With this definition of h the proof may be completed using Brown (1980) following a calculation similar to (28). □

Qualitative explanation. There is a qualitative explanation for the type of behavior seen in Example 2. Consider the coordinatewise risk in the problem with $v \equiv 1$ (i.e., $r = 0$). Suppose $\theta = (\|\theta\|, 0, \cdots, 0)$ and $\|\theta\|$ is large. Consider the usual James-Stein estimator, $\delta(x) = (1 - (c/\|x\|^2))x$. It is well known that the coordinatewise risk of this estimator is worse than that of δ_0 in the first coordinate direction, and better in the other directions. (See, e.g., James and Stein (1960) or Berger (1977, page 29).) Formally, $\Delta_1(\theta) = E_\theta(L^*(\theta_1, \delta_{01}) - L^*(\theta_1, \delta_1')) < 0$ and $\Delta_j(\theta) = E_\theta(L^*(\theta_j, \delta_{0j}) - L^*(\theta_j, \delta_j)) > 0$ for $j \neq 1$. Furthermore, when $c > 0$ is small $|\Delta_1(\theta)| \approx |\Delta_j(\theta)|$—more precisely

$$\Delta_1(\theta) = \frac{-2 - c^2}{\|\theta\|^2} + o\left(\frac{1}{\|\theta\|^2}\right)$$

$$\Delta_2(\theta) = \frac{2}{\|\theta\|^2} + o\left(\frac{1}{\|\theta\|^2}\right).$$

With the above in mind consider now the problem of this section in dimension $p = 2$ and with $r < 0$. Then $v(\|\theta\|)/v(0) \to 0$ as $\|\theta\| \to \infty$. When $\theta = (\|\theta\|, 0)$ with $\|\theta\|$ large the weight function thus gives greatest weight to the coordinate direction in which the James-Stein estimator performs well. Asymptotically the James-Stein estimator is, therefore, better than the usual estimator since

$$\Delta(\theta) = (v(0) + v(\|\theta\|))^{-1}(v(\|\theta\|)\Delta_1(\theta) + v(0)\Delta_2(0))$$

$$\to \Delta_2(\theta) = \frac{2}{\|\theta\|^2} + o\left(\frac{1}{\|\theta\|^2}\right)$$

as $\|\theta\| \to \infty$. This suggests that the usual estimator is dominated by the James-Stein estimator when $r < 0$ and $p = 2$. Actually, this is not quite so, since for $\theta = (\|\theta\|/2^{\frac{1}{2}}, \|\theta\|/2^{\frac{1}{2}})$ calculations like the above show that

$$\Delta(\theta) = -\frac{c^2}{\|\theta\|^2} + o\left(\frac{1}{\|\theta\|^2}\right).$$

In order to dominate the usual estimator one must, therefore, modify the James-Stein estimator in order to spread some of the advantage at points near $(\|\theta\|, 0)$ around to points near $(\|\theta\|/2^{\frac{1}{2}}, \|\theta\|/2^{\frac{1}{2}})$. The preceding proof formally shows this can be done and suggests a functional form for the modification of the James-Stein estimator (see above (31)).

The preceding explanation provides less explicit information about the case $r > 0$. But it does at least indicate that it should be increasingly harder to dominate the usual estimator as r increases, since then the weight function puts greater weight on those coordinate directions for which the James-Stein estimator performs badly.

4. Concluding remarks. The above results add understanding to apparent peculiarities discovered in the gamma scale problem by Berger (1980) and in the Poisson problem by Clevenson and Zidek (1977) and by Peng (1975). It appears that these results are not at all peculiar, but rather are part of a general pattern to be anticipated in a wide variety of simultaneous estimation problems.

These results indicate that the choice of the coordinate loss functions is more critical for practical applications than had previously been supposed. Alteration of the weight functions (v) may drastically affect the admissibility of the obvious minimax estimator (δ_0). (See Brown (1975) for a discussion of another respect in which the determination of the loss function is important.)

The above results *suggest* that when the coordinate losses are weighted so that $E(L^*(\theta_i, \delta_i^*)) = E(L^*(\theta_j, \delta_j^*))\ \forall i, j$ then one should expect that the critical dimension for inadmissibility of δ_0 is 3 (as in Stein's phenomenon). (By heuristic extension and the evidence of Brown (1975) this should be the case whenever $0 < \inf v(\theta_i)R^*(\theta_i, \delta_i^*)/v(\theta_j)R^*(\theta_j, \delta_j^*)\ \forall i, j$.) However, when the coordinate losses are not comparably weighted then other types of admissibility phenomena may occur.

The explanations at the end of Sections 2 and 3 help explain the reasons for the special admissibility characteristics exposed in Examples 1 and 2, respectively. Presumably, one should expect similar characteristics from problems possessing qualitatively similar behaviour of the ratios $v(\theta_i)L^*(\theta_i, \delta_i^*)/v(\theta_j)L^*(\theta_j, \delta_j^*)$.

There are undoubtedly qualifications and exceptions to the above heuristic guidelines. Perhaps there are also some completely different general types of admissibility behaviour. A more precise (but more complicated) heuristic tool for discovering such behaviour is described in Brown (1979).

REFERENCES

[1] BERGER, J. (1977). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. Technical report, Statist. Depart., Purdue Univ.

[2] BERGER, J. (1980). Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters. *Ann. Statist.* **8** 545–571.

[3] BLYTH, C. R. (1951). On minimax statistical decision procedures and their admissibility. *Ann. Math. Statist.* **22** 22–42.

[4] BROWN, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Statist.* **37** 1087–1136.

[5] BROWN, L. D. (1975). Estimation with incompletely specified loss functions (the case of several location parameters) *J. Amer. Statist. Assoc.* **70** 417–427.

[6] BROWN, L. D. (1979) A heuristic method for determining admissibility of estimators—with applications. *Ann. Statist.* **7** 960–994.

[7] BROWN, LAWRENCE D. (1980). A necessary condition for admissibility. *Ann. Statist.* **8** 540–544.

[8] CLEVENSON, M. and ZIDEK, J. (1977). Simultaneous estimation of the mean of independent Poisson laws. *J. Amer. Statist. Assoc.* **70** 698–705.

[9] FARRELL, R. (1964). Estimators of a location parameter in the absolutely continuous case. *Ann. Math. Statist.* **35** 949–998.

[10] HUDSON, H. (1978). A natural identity for exponential families with applications in multiparameter estimation. *Ann. Statist.* **6** 473–484.

[11] JAMES, W. and STEIN, C. (1960). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probability* **1** 361–379.

[12] PENG, J. C. M. (1975). Simultaneous estimation of the parameters of independent Poisson distributions. Technical Report No. 78 (MPS 74-21416), Stanford Univ.

[13] STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probability* **1** 197–206.

[14] STEIN, C. (1959). The admissibility of Pitman's estimator for a single location parameter. *Ann. Math. Statist.* **30** 970–979.

[15] STEIN, C. (1973). Estimation of the mean of a multivariate distribution. *Proc. Prague Symp. Asymptotic Statist.* 345–381.

[16] TSUI, K. W. and Press, S. J. (1978). Simultaneous estimation of several Poisson parameters under K-normalized squared error loss. Unpublished manuscript.

DEPARTMENT OF MATHEMATICS
CORNELL UNIVERSITY
WHITE HALL
ITHACA, NEW YORK 14853