Estimation of the Mean of a Multivariate Normal Distribution

Charles M. Stein

Annals of Statistics, Volume 9, Issue 6 (Nov., 1981), 1135-1151.

Your use of the JSTOR database indicates your acceptance of JSTOR's Terms and Conditions of Use. A copy of JSTOR's Terms and Conditions of Use is available at http://www.jstor.org/about/terms.html, by contacting JSTOR at jstor-info@umich.edu, or by calling JSTOR at (888)388-3574, (734)998-9101 or (FAX) (734)998-9113. No part of a JSTOR transmission may be copied, downloaded, stored, further transmitted, transferred, distributed, altered, or otherwise used, in any form or by any means, except: (1) one stored electronic and one paper copy of any article solely for your personal, non-commercial use, or (2) with prior written permission of JSTOR and the publisher of the article or other text.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Annals of Statistics is published by Institute of Mathematical Statistics. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ims.html.

Annals of Statistics
©1981 Institute of Mathematical Statistics

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2000 JSTOR

http://www.jstor.org/
Wed Mar 15 21:01:45 2000

The Annals of Statistics
1981, Vol. 9, No. 6, 1135–1151

ESTIMATION OF THE MEAN OF A MULTIVARIATE NORMAL DISTRIBUTION

BY CHARLES M. STEIN

Stanford University

Estimation of the means of independent normal random variables is considered, using sum of squared errors as loss. An unbiased estimate of risk is obtained for an arbitrary estimate, and certain special classes of estimates are then discussed. The results are applied to smoothing by use of moving averages and to trimmed analogs of the James-Stein estimate. A suggestion is made for calculating approximate confidence sets for the mean vector centered at an arbitrary estimate.

1. Introduction. The central problem studied in this paper is that of estimating the mean of a multivariate normal distribution with the squared length of the error vector as loss when the covariance matrix is known to be the identity matrix. First an unbiased estimate is obtained for the risk of a nearly arbitrary estimate of the mean. This is specialized to a class of estimates that includes all formal Bayes estimates. Among these a large class of minimax estimates is found when the dimension is at least three. A small number of special problems are considered.

The present work arose from a question raised by Malcolm Hudson in connection with his dissertation (Hudson, 1974, 1978). It was also inspired in part by the work of Efron and Morris (1971, 1972a, b, 1973a, b) who modified the estimate of James and Stein (1961) in several different ways to obtain estimates that are more appropriate in practical situations.

Sections 2 and 3 constitute a slightly expanded version of Section 2 of an earlier paper, Stein (1973). The basic formulas for unbiased estimation of the risk are obtained in Section 2. The Bayes and formal Bayes estimates are computed in Section 3 and studied in the light of the results of Section 2. In Sections 4–6, two special problems, related to papers of Efron and Morris (1971, 1972a, 1973a), are considered. First, if X is normally distributed with unknown mean ξ and the identity as covariance matrix, estimates of the form $\hat{\xi} = X - \lambda(X)AX$ are studied, where A is a given symmetric matrix. Under certain conditions, a choice of the real-valued function λ is found in Section 4 that yields a minimax estimate which is optimum in the sense that the risk of the estimate cannot be improved at any point by multiplication of λ by a constant factor. An application is then considered in Section 5. In Section 6 a modification of the James-Stein estimate is studied which limits the amount by which any coordinate $\hat{\xi}_i$ of the estimate $\hat{\xi}$ can differ from the corresponding X_i. Section 7 considers the modification needed when the common variance is unknown but an independent estimate of the variance is available in an important special case. In Section 8, an unbiased estimate is obtained for the expected squared difference between the squared length of the error vector and the unbiased estimate of its expectation. This suggests rough confidence sets for the mean.

Since this paper was written (in 1974), there has been a great deal of research concerning the estimation of a multivariate mean. No attempt will be made to discuss these subsequent contributions extensively, although papers having a direct bearing on issues raised here will be mentioned. The references in Berger (1980) may be helpful.

Received January 1981.

AMS 1980 *subject classifications.* Primary 62F15; secondary 62F10, 62F25.

Key words and phrases. Minimax estimate, Bayes estimate, multivariate normal mean, moving average, James-Stein estimate, confidence region, trimmed mean, simultaneous estimation.

Throughout the discussion we take the zero vector as the focal point for both X and ξ, this being the canonical form of the general problem. In practice X will often be the residual from a fitted model; see, e.g., Efron and Morris (1975).

2. Basic formulas. A simple identity concerning expectations of functions of a normal random variable is proved in Lemma 1 and extended to functions of several independent normal random variables in Lemma 2. This result is used in Theorem 1 to obtain an unbiased estimate of the risk (expected squared length of the error vector) of a nearly arbitrary estimate $X + g(X)$ of the mean of a multivariate normal distribution with the identity as covariance matrix. This is specialized in Theorem 2 to a class of estimates that contains all formal Bayes estimates, yielding a large class of minimax estimates. Although versions of the lemmas have subsequently appeared in Hudson (1978), they are included here for completeness and the generality obtained.

LEMMA 1. *Let Y be a $N(0, 1)$ real random variable and let $g: \mathscr{R} \to \mathscr{R}$ be an indefinite integral of the Lebesgue measurable function g', essentially the derivative of g. Suppose also that $E\,|\,g'(Y)\,| < \infty$. Then*

$$E\{g'(Y)\} = E\{Yg(Y)\}. \tag{2.1}$$

PROOF. Write $\phi(y)$ for the standard normal density, with derivative $\phi'(y) = -y\phi(y)$.

$$\begin{aligned} E\{g'(Y)\} &= \int_{-\infty}^{\infty} g'(y)\phi(y)\,dy \\ &= \int_0^{\infty} g'(y)\left\{\int_y^{\infty} z\phi(z)\,dz\right\} dy - \int_{-\infty}^{0} g'(y)\left\{\int_{-\infty}^{y} z\phi(z)\,dz\right\} dy \\ &= \int_0^{\infty} z\phi(z)\left\{\int_0^{z} g'(y)\,dy\right\} dz - \int_{-\infty}^{0} z\phi(z)\left\{\int_z^{0} g'(y)\,dy\right\} dz \qquad (2.2) \\ &= \left(\int_0^{\infty} + \int_{-\infty}^{0}\right)[z\phi(z)\{g(z) - g(0)\}]\,dz \\ &= \int_{-\infty}^{\infty} zg(z)\phi(z)\,dz = E\{Yg(Y)\}. \end{aligned}$$

The third equality in (2.2) uses Fubini's Theorem. This proof is essentially an application of integration by parts, but this slightly disguised form seems to make a proof of the result in the desired generality a bit easier.

In order to express this result in terms of an arbitrary normal random variable, define a new random variable X related to the random variable Y of Lemma 1 by

$$X = \sigma Y + \xi,$$

where ξ is real and σ positive so that X is $N(\xi, \sigma^2)$. If we also define a new function $h: \mathscr{R} \to \mathscr{R}$ by

$$h(x) = g\left(\frac{x - \xi}{\sigma}\right),$$

then (2.1) yields

$$Eh'(X) = \frac{1}{\sigma} Eg'\left(\frac{X-\xi}{\sigma}\right) = \frac{1}{\sigma} Eg'(Y) = \frac{1}{\sigma} E\{Yg(Y)\}$$

(2.3)

$$= E\left\{\frac{X-\xi}{\sigma^2} g\left(\frac{X-\xi}{\sigma}\right)\right\} = E\left\{\frac{X-\xi}{\sigma^2} h(X)\right\}.$$

Next let us indicate the notation and regularity conditions needed for the extension of Lemma 1 to the multidimensional case. For $x, y \in \mathcal{R}^p$ we define

$$x \cdot y = \sum_{i=1}^p x_y y_i, \quad \|x\|^2 = x \cdot x = \sum_{i=1}^p x_i^2. \tag{2.4}$$

DEFINITION 1. A function $h: \mathcal{R}^p \to \mathcal{R}$ will be called almost differentiable if there exists a function $\nabla h: \mathcal{R}^p \to \mathcal{R}^p$ such that, for all $z \in \mathcal{R}^p$,

$$h(x+z) - h(x) = \int_0^1 z \cdot \nabla h(x+tz)\, dt$$

for almost all $x \in \mathcal{R}^p$. A function $g: \mathcal{R}^p \to \mathcal{R}^p$ is almost differentiable if all its coordinate functions are. Essentially ∇ is the vector differential operator of first partial derivatives with ith coordinate

$$\nabla_i = \frac{\partial}{\partial x_i}. \tag{2.5}$$

Let us now extend Lemma 1 to functions of a normal random vector with the identity as covariance matrix. Throughout the remainder of this paper, X will denote a p-dimensional random coordinate vector with mean ξ and the identity as covariance matrix, with some change of point of view in Section 3. In order to indicate the dependence of expectations on ξ, I write E_ξ rather than E.

LEMMA 2. *If $h: \mathcal{R}^p \to \mathcal{R}$ is an almost differentiable function with $E_\xi \| \nabla h(X) \| < \infty$, then*

$$E_\xi \nabla h(X) = E_\xi\{(X-\xi)h(X)\}. \tag{2.6}$$

PROOF. For $i \in \{1 \cdots p\}$, let $\mathcal{B}_i$ be the σ-algebra generated by X_i alone, and let $\mathcal{B}_{-i}$ be the σ-algebra generated by all the X_j for $j \neq i$. Let X_{-i} be the random $(p-1)$-dimensional coordinate vector with index set $\{1, \ldots, p\} \cap \{i\}^c$ having the jth coordinate X_j for $j \neq i$. Somewhat imprecisely, to express the fact that X determines and is determined by X_i and X_{-i}, I write

$$X = (X_i, X_{-i}).$$

Then, using the independence of $\mathcal{B}_i$ and $\mathcal{B}_{-i}$, and also Lemma 1, we find that, for almost all ω in the set Ω of the underlying probability space,

$$\begin{aligned}[E\{(X_i - \xi_i)h(X) \mid \mathcal{B}_{-i}\}](\omega) &= [E\{(X_i - \xi_i)h(X_i, X_{-i}(\omega)) \mid \mathcal{B}_{-i}\}](\omega) \\ &= [E\{(\nabla h)_i(X_i, X_{-i}(\omega)) \mid \mathcal{B}_{-i}\}](\omega) \\ &= [E\{(\nabla h)_i(X) \mid \mathcal{B}_{-i}\}](\omega).\end{aligned}$$

Thus

$$E\{(\nabla h)_i(X) \mid \mathcal{B}_{-i}\} = E\{(X_i - \xi_i)h(X) \mid \mathcal{B}_{-i}\},$$

and, taking the expectation of both sides, we find that

$$E_\xi\{(\nabla h)_i(X)\} = E_\xi\{(X_i - \xi_i)h(X)\},$$

which yields (2.6).

We shall need the definition and some elementary properties of harmonic and superharmonic functions.

DEFINITION 2. A lower semicontinuous function $f: \mathscr{R}^p \to \mathscr{R} \cup \{+\infty\}$ is superharmonic at a point $x^0 \in \mathscr{R}^p$ if, for every $r > 0$, the average of f over the sphere

$$S_r(x^0) = \{x: \| x - x^0 \|^2 = r^2\}$$

of radius r centered at x^0 is not greater than $f(x^0)$. The function f is superharmonic in $\mathscr{R}^p$ if it is superharmonic at each $x^0 \in \mathscr{R}^p$.

LEMMA 3. *If $f: \mathscr{R}^p \to \mathscr{R}$ is twice continuous differentiable, then f is superharmonic in $\mathscr{R}^p$ if and only if, for all $x \in \mathscr{R}^p$,*

$$\nabla^2 f(x) \leq 0$$

where ∇^2 is the Laplacian $\nabla^2 = \Sigma \nabla_i^2$ with ∇_i as in (2.5). *The proof of Lemma* 3 *is given in Theorem* 4.8 *of Helms* (1969, *page* 63).

DEFINITION 3. The twice continuously differentiable function $f: \mathscr{R}^p \to \mathscr{R}$ is harmonic at $x^0 \in \mathscr{R}^p$ if

$$\nabla^2 f(x^0) = 0. \tag{2.7}$$

It is harmonic in $\mathscr{R}^p$ if it is harmonic at each $x^0 \in \mathscr{R}^p$.

Lemma 2 will be used to obtain an unbiased estimate of the risk of a nearly arbitrary estimate $X + g(X)$ of the mean of a multivariate normal distribution with the identity as covariance matrix. In accordance with (2.4) and (2.5), if $g: \mathscr{R}^p \to \mathscr{R}^p$ is almost differentiable, I shall write

$$\nabla \cdot g = \sum \nabla_i g_i.$$

THEOREM 1. Consider the estimate $X + g(X)$ for ξ such that $g: \mathscr{R}^p \to \mathscr{R}^p$ is an almost differentiable function for which

$$E_\xi \Sigma | \nabla_i g_i(X) | < \infty.$$

Then, for each $i \in \{1, \ldots, p\}$,

$$E_\xi \{X_i + g_i(X) - \xi_i\}^2 = 1 + E_\xi \{g_i^2(X) + 2 \nabla_i g_i(X)\}, \tag{2.8}$$

and consequently

$$E_\xi \| X + g(X) - \xi \|^2 = p + E_\xi \{\| g(X) \|^2 + 2 \nabla \cdot g(X)\}. \tag{2.9}$$

PROOF. From formula (2.6) with $h = g_i$, it follows that

$$\begin{aligned} E_\xi \{X_i + g_i(X) - \xi_i\}^2 &= E_\xi \{(X_i - \xi_i)^2 + 2(X_i - \xi_i) g_i(X) + g_i^2(X)\} \\ &= 1 + 2E_\xi \nabla_i g_i(X) + E_\xi g_i^2(X), \end{aligned}$$

which is (2.8). Summing over i we obtain (2.9). We observe that the latter formula asserts that $p + \| g(X) \|^2 + 2 \nabla \cdot g(X)$ is an unbiased estimate of the risk of the nearly arbitrary estimate $X + g(X)$ for ξ.

When the dimension $p \geq 3$, we shall obtain a large collection of minimax estimates of ξ by specializing Theorem 1 to a class of estimates which, as we shall see in Section 3, contains all formal Bayes estimates.

THEOREM 2. *Let $f: \mathscr{R}^p \to \mathscr{R}^+ \cap \{0\}^c$ be an almost differentiable function for which $\nabla f: \mathscr{R}^p \to \mathscr{R}^p$ can be taken to be almost differentiable, and suppose also that*

$$E_\xi\left\{\frac{1}{f(X)}\Sigma|\nabla_i^2 f(X)|\right\}<\infty, \tag{2.10}$$

and

$$E_\xi\|\nabla \log f(X)\|^2<\infty. \tag{2.11}$$

Then

$$\begin{aligned}E_\xi\|X+\nabla\log f(X)-\xi\|^2 &= p+E_\xi\left\{2\frac{\nabla^2 f(X)}{f(X)}-\frac{\|\nabla f(X)\|^2}{f^2(X)}\right\}\\ &= p+4E_\xi\left\{\frac{\nabla^2\sqrt{f(X)}}{\sqrt{f(X)}}\right\}.\end{aligned} \tag{2.12}$$

PROOF. Let $g:\mathscr{R}^p\to\mathscr{R}^p$ be defined by

$$g=\nabla\log f=\frac{\nabla f}{f}.$$

Then

$$\nabla\cdot g=\nabla\cdot\nabla\log f=\frac{\nabla^2 f}{f}-\frac{\|\nabla f\|^2}{f^2},$$

and thus it follows from equation (2.9) that

$$\begin{aligned}E_\xi\|X+\nabla\log f(X)-\xi\|^2 &= p+E_\xi\left\{\frac{\|\nabla f(X)\|^2}{f^2(X)}+2\left(\frac{\nabla^2 f(X)}{f(X)}-\frac{\|\nabla f(X)\|^2}{f^2(X)}\right)\right\}\\ &= p+E_\xi\left\{2\frac{\nabla^2 f(X)}{f(X)}-\frac{\|\nabla f(X)\|^2}{f^2(X)}\right\},\end{aligned}$$

which is the first form of (2.12). Also

$$\nabla^2\sqrt{f}=\nabla\cdot\nabla\sqrt{f}=\nabla\cdot\frac{\nabla f}{2\sqrt{f}}=\frac{1}{2\sqrt{f}}\nabla^2 f-\frac{1}{4f^{3/2}}\|\nabla f\|^2.$$

The final expression of (2.12) follows.

COROLLARY 1. *If $f:\mathscr{R}^p\to\mathscr{R}^+\cap\{0\}^c$ is twice continuously differentiable and its square root is superharmonic and* (2.10) *and* (2.11) *are satisfied, then $X+\nabla\log f(X)$ is a minimax estimate of ξ, that is, for all ξ,*

$$\begin{aligned}E_\xi\|X+\nabla\log f(X)-\xi\|^2 &= p+4E_\xi\left\{\frac{\nabla^2\sqrt{f(X)}}{\sqrt{f(X)}}\right\}\\ &\le p=\inf_g\sup_\xi E_\xi\|X+g(X)-\xi\|^2.\end{aligned} \tag{2.13}$$

PROOF. The first equality in (2.13) follows from (2.12), the inequality follows from the defining property (2.7) of superharmonic functions, and the final equality is well known.

Efron and Morris (1976), and Berger (1976) for the nonsymmetric situation, give conditions under which estimates of forms other than $X+\nabla\log f(X)$ are minimax. All admissible estimators must be of this form, however, as was shown by Brown (1971).

3. Formal Bayes estimates. Some easy known results about Bayes estimates, and also formal Bayes estimates, are now recalled, including the fact that they are all of the form considered in Theorem 2. The unbiased estimate of the risk of a formal Bayes estimate given in Theorem 2 is compared with the formal posterior risk, and it is found

that if the formal prior density is superharmonic, the formal posterior risk is always larger. Finally we make some remarks on L. Brown's deep admissibility results for the present problem.

The notation used in this section will differ slightly from that of the rest of the paper. Let ξ be a random p-dimensional coordinate vector distributed according to the prior probability measure Π. Let X be a random vector in $\mathscr{R}^p$, conditionally normally distributed given ξ with conditional mean ξ, and the identity as conditional covariance matrix. Then the unconditional density of X with respect to Lebesgue measure in $\mathscr{R}^p$ is given by

$$f(x) = \frac{1}{(2\pi)^{p/2}} \int e^{-1/2\|x-\xi\|^2}\, d\Pi(\xi). \tag{3.1}$$

I shall use E_ξ to denote conditional expectation given ξ and E^X to denote conditional expectation given X. The formulas of Section 2 involving E_ξ remain valid, although their interpretation is different. The Bayes estimate $\phi_\Pi(X)$ of ξ, which is defined by the condition that $\phi = \phi_\Pi$ minimizes

$$E\|\xi - \phi(X)\|^2 = EE^X\|\xi - \phi(X)\|^2 = E\left\{\frac{\int \|\xi - \phi(X)\|^2 e^{-1/2\|X-\xi\|^2}\, d\Pi(\xi)}{\int e^{-1/2\|X-\xi\|^2}\, d\Pi(\xi)}\right\}, \tag{3.2}$$

is given by

$$\begin{aligned}\phi_\Pi(X) = E^X\xi &= X + E^X(\xi - X) = X + \frac{\int (\xi - X) e^{-1/2\|X-\xi\|^2}\, d\Pi(\xi)}{\int e^{-1/2\|X-\xi\|^2}\, d\Pi(\xi)} \\ &= X + \nabla \log f(X),\end{aligned} \tag{3.3}$$

where f is given by (3.1). In equation (3.2), E denotes unconditional expectation. More generally if Π is a possibly infinite measure for which f defined by (3.1) is everywhere finite, we define the formal Bayes estimate $\phi_\Pi(X)$ by (3.3). Formal posterior expectation E^X is defined by the formula that yields posterior expectation in the case where Π is a probability measure:

$$E^X g(X, \xi) = \frac{\int g(X, \xi) e^{-1/2\|X-\xi\|^2}\, d\Pi(\xi)}{\int e^{-1/2\|X-\xi\|^2}\, d\Pi(\xi)}.$$

Next let us compare the unbiased estimate of the risk of the formal Bayes estimate $\phi_\Pi(X)$ of ξ given by Theorem 2 with the formal posterior risk $E^X\|\xi - \phi_\Pi(X)\|^2$. From Theorem 2, the unbiased estimate of the risk is given by

$$\rho(X) = p + 2\frac{\nabla^2 f(X)}{f(X)} - \frac{\|\nabla f(X)\|^2}{f^2(X)}. \tag{3.4}$$

For the formal posterior risk we have

$$\begin{aligned}E^X\|\xi - \phi_\Pi(X)\|^2 &= E^X\|\xi - X - \nabla \log f(X)\|^2 \\ &= E^X\{\|X - \xi\|^2 - \|\nabla \log f(X)\|^2\} \\ &= p + \frac{\nabla^2 f(X)}{f(X)} - \|\nabla \log f(X)\|^2.\end{aligned} \tag{3.5}$$

The second equality in (3.5) uses essentially the theorem of Pythagoras in the appropriate Hilbert space. The squared distance from ξ to X is the sum of the squared distance from ξ to the closest X-measurable random variable $X + \nabla \log f(X)$ and the squared distance $\|\nabla \log f(X)\|^2$ from $X + \nabla \log f(X)$ to X. Here squared distance is to be interpreted as formal posterior expectation of squared geometric distance. The final equality in (3.5) uses the fact that

$$\frac{\nabla^2 f(X)}{f(X)} = \frac{\nabla^2 \int e^{-1/2\|X-\xi\|^2}\, d\Pi(\xi)}{\int e^{-1/2\|X-\xi\|^2}\, d\Pi(\xi)} = \frac{\int (\|X-\xi\|^2 - p) e^{-1/2\|X-\xi\|^2}\, d\Pi(\xi)}{\int e^{-1/2\|X-\xi\|^2}\, d\Pi(\xi)}$$
$$= E^X(\|X-\xi\|^2 - p).$$

Comparing (3.4) and (3.5) we see that

$$E^X\|\xi - \phi_\Pi(X)\|^2 = \rho(X) - \frac{\nabla^2 f(X)}{f(X)}.$$

This shows that if f is superharmonic then the formal posterior risk $E^X\|\xi - \phi_\Pi(X)\|^2$ is an overestimate of the risk of the estimate $\phi_\Pi(X)$ given by (3.3) in the sense that

$$E^X\|\xi - \phi_\Pi(X)\|^2 \geq \rho(X), \tag{3.6}$$

and thus, for all ξ,

$$E_\xi E^X\|\xi - \phi_\Pi(X)\|^2 \geq E_\xi \rho(X) = E_\xi\|\xi - \phi_\Pi(X)\|^2. \tag{3.7}$$

Of course this inequality cannot hold in a non-trivial way if Π is a probability measure. For possible implications of this see Morris (1977) and Berger (1980).

Let us also observe that, if the formal prior measure Π has a superharmonic density π with respect to Lebesgue measure, then f defined by (3.1) is also superharmonic and thus $\phi_\Pi(X)$ is a minimax estimate of ξ and (3.6) and (3.7) hold. To see this, write

$$f(x) = \frac{1}{(2\pi)^{p/2}} \int e^{-1/2\|x-\xi\|^2} \pi(\xi)\, d\xi = \frac{1}{(2\pi)^{p/2}} \int e^{-1/2\|y\|^2} \pi(x-y)\, dy.$$

If π is superharmonic so is the mapping $x \mapsto \pi(x-y)$, and thus also f, which is a convex combination of these functions.

It may be of some theoretical interest to observe that, with the aid of the results of Brown (1971), it is not difficult to obtain a fairly large class of admissible minimax estimates of ξ. For, it seems to follow from his main theorem (ibid, page 884) that a formal Bayes estimate with respect to a prior density π of the form

$$\pi(\xi) = \int \|\xi - \eta\|^{-(p-2)}\, d\rho(\eta),$$

with ρ a finite measure, is admissible; and since π is superharmonic and thus also the corresponding f given by (3.1) with $d\Pi(\xi) = \pi(\xi)\, d\xi$, it follows from formula (2.12) that the formal Bayes estimate $X + \nabla \log f(X)$ is also minimax.

4. Choice of a scalar factor. We shall see that, in a fairly convincing sense, there is a best choice of the magnitude of the correction to be made on the naive estimate X of ξ if we have decided in advance on the direction of the correction, related linearly to X. Detailed application to the use of three-term moving averages will be discussed in Section 5.

Let us look at estimates of the form

$$\hat{\xi} = X - \lambda(X)AX, \tag{4.1}$$

where A is a preassigned symmetric matrix, and $\lambda: \mathscr{R}^p \to \mathscr{R}^+$ is to be chosen appropriately. It is convenient now to think of the $x \in \mathscr{R}^p$ as column vectors and to write, for example, x^T for the row vector transpose of x. We observe that, if

$$2A < (\operatorname{tr} A)I, \tag{4.2}$$

in the sense that the largest characteristic root of A is less than 1/2 tr A, then the risk of the estimate $\hat{\xi}$ defined by (4.1) with

$$\lambda(x) = \frac{1}{x^T B x}, \tag{4.3}$$

where

$$B = \{(\operatorname{tr} A)I - 2A\}^{-1}A^2, \tag{4.4}$$

is given by

$$\begin{aligned} E_\xi \| X - \frac{1}{X^TBX} AX - \xi \|^2 &= p + E_\xi\left\{\frac{X^TA^2X}{(X^TBX)^2} - 2\nabla \cdot \frac{AX}{X^TBX}\right\} \\ &= p + E_\xi\left\{\frac{X^TA^2X}{(X^TBX)^2} - 2\frac{\operatorname{tr} A}{X^TBX} + 4\frac{X^TABX}{(X^TBX)^2}\right\} \\ &= p - E_\xi\left\{\frac{X^TA^2X}{(X^TBX)^2}\right\}. \end{aligned} \tag{4.5}$$

In the final equality we have used the particular choice (4.4) of B. Condition (4.2) is needed for B given by (4.4) to be positive definite. If B is not positive definite, the expectations do not exist and the formal computations are incorrect. Formula (4.5) shows that the estimate $\hat{\xi}$ defined by (4.1)–(4.4) is minimax.

It may be of some interest to observe that this estimate has a mild optimum property. Any estimate that changes the choice of λ in (4.3) by a constant factor cannot be better at any parameter point. For, by a simple modification of (4.5) we see that, for any real constant β,

$$\begin{aligned} E_\xi \left\| X - \frac{\beta}{X^TBX} AX - \xi \right\|^2 &= p + E_\xi\left[\frac{\beta^2X^TA^2X - 2\beta X^T\{(\operatorname{tr} A)I - 2A\}BX}{(X^TBX)^2}\right] \\ &= p + (\beta^2 - 2\beta)E_\xi\left\{\frac{X^TA^2X}{(X^TBX)^2}\right\} \end{aligned}$$

For all ξ, this is minimized by $\beta = 1$. We observe that the special case $A = I$ is the nontruncated estimate considered by James and Stein (1961).

5. Application to symmetric moving averages. Let us apply the general results of Section 4 to the question of the appropriate choice of the weight in a three-term symmetric moving average, first in the cyclic case. Let $X_1, \cdots, X_p$ be independently normally distributed with means $\xi_1, \cdots, \xi_p$ and variance 1, and suppose we plan to estimate the ξ_i by

$$\hat{\xi}_i = X_i - \lambda(X)\{X_i - \tfrac{1}{2}(X_{i-1} + X_{i+1})\},$$

where it is understood that $X_0 = X_p$ and $X_{p+1} = X_1$, and similarly for the ξ's. This is the special case of

$$A_{ij} = \begin{cases} -\tfrac{1}{2} & \text{if } j - i \equiv \pm 1 \pmod p \\ 1 & \text{if } j - i \equiv 0 \pmod p \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

The characteristic roots and vectors of A, the solutions α_j and y_j of

$$Ay_j = \alpha_j y_j,$$

with α_j real and $y_j \in \mathscr{R}^p$ are given, with j varying over the integers such that

$$-\left[\frac{p}{2}\right] \leq j < \left[\frac{p}{2}\right] \tag{5.2}$$

by

$$\alpha_j = 1 - \cos\left(2\pi \frac{j}{p}\right), \tag{5.3}$$

and for $i \in \{1 \cdots p\}$

$$y_{ij} = \begin{cases} \dfrac{1}{\sqrt{p}} & \text{if } j = 0 \\ \dfrac{(-1)^i}{\sqrt{p}} & \text{if } j = -\left[\dfrac{p}{2}\right] \\ \sqrt{\dfrac{2}{p}} \cos(2\pi ij/p) & \text{if } -\left[\dfrac{p}{2}\right] < j < 0 \\ \sqrt{\dfrac{2}{p}} \sin(2\pi ij/p) & \text{if } 0 < j < \left[\dfrac{p}{2}\right], \end{cases} \tag{5.4}$$

this being the ith coordinate of y_j. No difficulty is caused by the different ranges of i and j in (5.4); see, for example, Anderson (1971, pages 278–284). The matrix A can be expressed as

$$A = y\alpha y^T,$$

where α is the diagonal matrix with jth diagonal element α_j (for j satisfying (5.2)) and y is the orthogonal matrix with ijth element y_{ij}. From the definition (5.1) of A we have

$$\operatorname{tr} A = p,$$

and from (5.3) we see that the largest characteristic root of A is less than or equal to 2, and equal to 2 when p is even. Thus condition (4.2) is satisfied if and only if $p \geq 5$. In this case the matrix B, given by (4.4), to be used in (4.1) and (4.3) is

$$B = \{(\operatorname{tr} A)I - 2A^{-1}\}A^2 = y(pI - 2\alpha)^{-1}\alpha^2 y^T. \tag{5.5}$$

It is unreasonable to use a three-term moving average with weights more extreme than (⅓, ⅓, ⅓). Thus it seems appropriate to modify our estimate to

$$\hat{\xi} = X - \lambda_1(X)AX,$$

where

$$\lambda_1(X) = \min\left(\frac{1}{X^TBX}, \frac{2}{3}\right).$$

Of course A is given by (5.1) and B by (5.5). The unbiased estimate of the improvement in the risk is changed from

$$\Delta(X) = \frac{X^TA^2X}{(X^TBX)^2},$$

given by (4.5), to

$$\Delta_1(X) = \begin{cases} \Delta(X) & \text{if } X^TBX > \frac{3}{2} \\ \frac{4p}{3} - \frac{4}{9}\sum\left\{X_i - \frac{1}{2}(X_{i-1} + X_{i+1})\right\}^2 & \text{if } X^TBX \leq \frac{3}{2}. \end{cases} \tag{5.6}$$

The second expression in (5.6) is obtained by applying Theorem 1 to the form g takes on $\{x: x^TBx \leq 3/2\}$, namely

$$g_i(x) = -\tfrac{2}{3}\{x_i - \tfrac{1}{2}(x_{i-1} + x_{i+1})\}.$$

Next let us look at the case of a (nearly) symmetric three-term moving average in the case where the indices are ordered rather than cyclic. We consider estimates of the form (4.1) with λ given by (4.3) and (4.4) and the $p \times p$ matrix A given by

$$A_{ij} = \begin{cases} \frac{1}{2} & \text{if } i = j = 1 \text{ or } i = j = p \\ 1 & \text{if } i = j \neq 1, p \\ -\frac{1}{2} & \text{if } |i - j| = 1 \\ 0 & \text{if } |i - j| \neq 0, 1, \end{cases} \tag{5.7}$$

that is

$$\hat{\xi} = \begin{cases} \{1 - \lambda(X)\}X_i + \frac{1}{2}\lambda(X)(X_{i-1} + X_{i+1}) & \text{if } i \neq 1, p \\ \{1 - \frac{1}{2}\lambda(X)\}X_1 + \frac{1}{2}\lambda(X)X_2 & \text{if } i = 1 \\ \{1 - \frac{1}{2}\lambda(X)\}X_p + \frac{1}{2}\lambda(X)X_{p-1} & \text{if } i = p. \end{cases}$$

The characteristic roots α_j and vectors y_j of A are given, for $j \in \{1, \cdots, p\}$, by

$$\alpha_j = 1 - \cos\left\{\frac{\pi(j-1)}{p}\right\} \tag{5.8}$$

and, for $i \in \{1, \cdots, p\}$,

$$y_{ij} = \begin{cases} \frac{1}{\sqrt{p}} & \text{if } j = 1 \\ \frac{2}{\sqrt{p}} \cos \frac{\pi(2i-1)(j-1)}{2p} & \text{if } j \neq 1; \end{cases} \tag{5.9}$$

see Anderson (1971, pages 284–290). The matrix A can be expressed as

$$A = y\alpha y^T$$

where α is the diagonal matrix with jth diagonal element α_j given by (5.8) and y is the orthogonal matrix with i, j element given by (5.9) for $i, j \in \{1, \cdots, p\}$. By (5.7),

$$\operatorname{tr} A = p - 1,$$

and, by (5.8), the largest characteristic root of A is less than 2. Thus, according to condition (4.2), the estimate given by (4.1), (4.3) and (4.4) is applicable for $p \geq 5$. Again the appropriate choice of B is given by (5.5), but with A, α, and y given by (5.7) to (5.9). Again it seems appropriate to replace λ in (4.1) by λ_1, given by

$$\lambda_1(x) = \min\left(\frac{1}{x^TBx}, \frac{2}{3}\right).$$

The unbiased estimate of the improvement in the risk is changed from $\Delta(X)$ given by (4.5) to $\Delta_1(X)$ given by

$$\Delta_1(X) = \begin{cases} \Delta(X) & \text{if } X^TBX \geq \frac{3}{2} \\ \frac{4(p-1)}{3} - \frac{4}{9}\sum_{i=2}^{p-1}\Sigma\left\{X_i - \frac{1}{2}(X_{i-1} + X_{i+1})\right\}^2 & \\ \quad - \frac{1}{9}(X_1 - X_2)^2 - \frac{1}{9}(X_p - X_{p-1})^2 & \text{if } X^TBX < \frac{3}{2}. \end{cases}$$

The second expression is obtained by applying Theorem 1 to the form g takes on $\{x: x'Bx \leq \frac{3}{2}\}$, namely

$$g_i(x) = \begin{cases} -\frac{1}{3}(x_1 - x_2) & \text{if } i = 1 \\ -\frac{2}{3}\{x_i - \frac{1}{2}(x_{i-1} + x_{i+1})\} & \text{if } 2 \leq i \leq p-1 \\ -\frac{1}{3}(x_p - x_{p-1}) & \text{if } i = p. \end{cases}$$

6. Estimates in which the modification of individual coordinates is sharply limited. This section treats a modification of an idea of Efron and Morris (1971, 1972a). Roughly speaking, their idea was to modify the James-Stein estimate

$$\hat{\xi}_0 = \left(1 - \frac{p-2}{\| X \|^2}\right)_+ X \tag{6.1}$$

by requiring that no coordinate be changed by more than a preassigned quantity c. This leads to an improvement on the James-Stein estimate when the empirical distribution of the $|\xi_i|$ is long-tailed, and at worst only a relatively unimportant deterioration if the prior distribution of ξ has spherical symmetry. We consider a modification of the Efron and Morris procedure, based on order statistics, that may permit a somewhat larger improvement over the James-Stein estimate when the empirical distribution of the $|\xi_i|$ is long-tailed.

Let us look at the simplest case of the estimate based on order statistics. Let

$$Z_i = |X_i|,$$

and let

$$Z_{(1)} < \ldots < Z_{(p)}$$

be the rearrangement of $Z_1, \ldots, Z_p$ in increasing order, and let k be a positive integer, a large fraction of p, the appropriate choice of which will be discussed later. Let

$$\hat{\xi} = X + g(X), \tag{6.2}$$

where $g: \mathscr{R}^p \to \mathscr{R}^p$ is defined by

$$g_i(X) = \begin{cases} -\dfrac{a}{\Sigma(X_j^2 \wedge Z_{(k)}^2)} X_i & \text{if } |X_i| \leq Z_{(k)} \\ -\dfrac{a}{\Sigma(X_j^2 \wedge Z_{(k)}^2)} Z_{(k)} \operatorname{sgn} X_i & \text{if } |X_i| > Z_{(k)} \end{cases} \tag{6.3}$$

with a a constant to be determined and $a \wedge b = \min(a, b)$.

$$\begin{aligned} E_\xi \| \hat{\xi} - \xi \|^2 &= p + E_\xi \left[\frac{a^2}{\Sigma(X_j^2 \wedge Z_{(k)}^2)} - 2a \sum_{i=1}^{k} \frac{1}{\Sigma(X_j^2 \wedge Z_{(k)}^2)} + 4a \sum_{i=1}^{k-1} \frac{X_i^2}{\{\Sigma(X_j^2 \wedge Z_{(k)}^2)\}^2} \right. \\ &\qquad \left. + 4a(p-k+1) \frac{Z_{(k)}^2}{\{\Sigma(X_j^2 \wedge Z_{(k)}^2)\}^2} \right] \\ &= p + \{a^2 - 2(k-2)a\} E_\xi \left\{ \frac{1}{\Sigma(X_j^2 \wedge Z_{(k)}^2)} \right\}. \end{aligned} \tag{6.4}$$

The optimum choice of a is

$$a = k - 2, \tag{6.5}$$

and, for this choice, the risk given by (6.4) becomes

$$E_\xi \| \hat{\xi} - \xi \|^2 = p - (k-2)^2 E_\xi \left\{ \frac{1}{\Sigma(X_j^2 \wedge Z_{(k)}^2)} \right\}.$$

As a guide to the choice of k, let us compute, for large p, the relative efficiency of the estimate $\hat{\xi}^{(k)}$ given by (6.2), (6.3) and (6.5) compared to the James-Stein estimate $\hat{\xi}_0$ given by (6.1) in the case most favorable to the James-Stein estimate, that where the ξ_i are themselves independently normally distributed with variance τ^2. We shall see that if $\tau > 0$, the asymptotic relative efficiency e_y when $k/p \to y$ is given by

$$e_y = \frac{y^2}{y + (1-y)q^2 - 2q\phi(q)}, \qquad \text{where } q = \Phi^{-1}\left\{\frac{1}{2}(1+y)\right\}. \tag{6.6}$$

Some numerical values are given in Table 1.

TABLE 1
Efficiency e_y of trimmed estimate relative to James-Stein estimate for large p, $k \approx yp$

y	.5	.6	.7	.8	.9
e_y	.827	.873	.909	.943	.974

To derive (6.6), we observe that the estimated improvement in the risk for the estimate $\hat{\xi}^{(k)}$, given by (6.2), (6.3) and (6.5) is

$$\Delta^{(k)}(X) = \frac{(k-2)^2}{\Sigma(X_j^2 \wedge Z_{(k)}^2)},$$

and the estimated improvement in the risk for the James-Stein estimate is

$$\Delta(X) = \frac{(p-2)^2}{\Sigma X_j^2}.$$

The truncation in (6.1) is ignored because with $\tau^2 > 0$ fixed and $p \to \infty$, the probability that truncation will occur approaches 0. For large p, $\Delta(X)$ and $\Delta^{(k)}(X)$ are approximately constant with high probability:

$$\Delta(X) \approx \frac{p^2}{E\Sigma X_j^2} = \frac{p}{1+\tau^2}$$

and

$$\Delta^{(k)}(X) \approx \frac{k^2}{pE(X_j^2 \wedge Z_{(k)}^2)} \approx \frac{k^2}{2p(1+\tau^2)\left\{\int_0^q x^2\phi(x)\,dx + q^2\int_q^\infty \phi(x)\,dx\right\}}, \tag{6.7}$$

where $q = \Phi^{-1}\{\frac{1}{2}(1+y)\}$ is an approximation to $Z_{(k)}/\sqrt{1+\tau^2}$.
The first integral in the denominator of (6.7) can be evaluated by integration by parts, so that we obtain

$$e_y \approx \frac{\Delta^{(k)}(X)}{\Delta(X)} \approx \frac{y^2}{2\left\{\int_0^q \phi(x)\,dx - q\phi(q) + q^2\int_q^\infty \phi(x)\,dx\right\}}$$

which is (6.6).

The numerical efficiencies given in Table 1 suggest that in the case most favorable to the choice $k = p$, the loss due to taking k even as small as $0.7\,p$ is small enough so that it will ordinarily be more than compensated for by the possibility that the empirical c.d.f. of the ξ_i is long-tailed. For small p, the possible loss in efficiency is likely to be somewhat larger for a given value of $y = k/p$. Of course, we must have $k \geq 3$ for the formulas to be meaningful.

7. The case of unknown variance. The formulation of the problem up to this point may be unrealistic in that the variance has been assumed known and taken to be 1. Here a partial treatment is given of the more common case where the variance is unknown but can be estimated by an independent multiple of a χ^2 random variable. We consider only the case where, when the variance is 1, $\hat{\xi} - X$ is chosen to be homogeneous of degree -1 in X. Then it is not difficult to decide on an appropriate proportionality factor when the variance is unknown in a way completely analogous to that of James and Stein (1961). The problem has been treated more thoroughly by Efron and Morris (1973a). Since there seem to be no complications in the special cases of Sections 4 and 5, they are discussed only briefly.

The problem considered here differs from the basic formulation of Section 2 in that X is now a random p-dimensional coordinate vector, normally distributed with unknown mean ξ and covariance matrix $\sigma^2 I$, where σ^2 is unknown but we also observe a real random variable S, distributed independently of X as $\sigma^2\chi^2_n$. If, in the case where σ^2 is known to be 1, we would use the estimate

$$\hat{\xi}_0 = X + g(X)$$

where $g: \mathscr{R}^p \to \mathscr{R}^p$ is homogeneous of degree -1, that is

$$g(\lambda x) = \frac{1}{\lambda} g(x)$$

for all real $\lambda \neq 0$, we consider, for the present problem, the modified estimate

$$\hat{\xi} = X + cSg(X),$$

where c is a constant to be determined. Let

$$Y = \frac{X}{\sigma}, \qquad \eta = \frac{\xi}{\sigma}, \qquad S^* = \frac{S}{\sigma^2}.$$

Then, using the independence of S^* and Y, from Theorem 1 we obtain

$$\begin{aligned} E_{\xi,\sigma} \| X + cSg(X) - \xi \|^2 &= \sigma^2 E \| Y + cS^* g(Y) - \eta \|^2 \\ &= \sigma^2 E[p + c^2 S^{*2} \| g(Y)\|^2 + 2cS^* \nabla^* \cdot g(Y)] \\ &= \sigma^2 E[p + n(n+2)c^2 \| g(Y)\|^2 + 2nc\nabla^* \cdot g(Y)] \end{aligned} \tag{7.1}$$

where ∇^* is the vector of first partial derivatives with respect to Y. If we choose $c = 1/(n+2)$, (7.1) now becomes

$$E_{\xi,\sigma} \| X + \frac{S}{n+2} g(X) - \xi \|^2 = \sigma^2 E\left[p + \frac{n}{n+2} (\| g(Y)\|^2 + 2\nabla^* \cdot g(Y)) \right].$$

If g has been chosen so as to make $\| g(Y)\|^2 + 2\nabla^* \cdot g(Y)$ everywhere negative and, roughly speaking, as negative as possible, by the methods of the preceding sections, this should be a satisfactory estimate. Observe that we lose only the proportion $2/(n+2)$ of the reduction in risk that we would have achieved if we had known σ^2.

For some purposes it may be useful to have an unbiased estimate of the expected squared length of the error vector. Using formula (2.3) we have

$$\begin{aligned} E_{\xi,\sigma^2} \| X + \frac{S}{n+2} g(X) - \xi \|^2 &= p\sigma^2 + E_{\xi,\sigma^2}\left\{ \frac{S^2}{(n+2)^2} \| g(X)\|^2 + 2\frac{S}{n+2} (X - \xi)' g(X) \right\} \\ &= E_{\xi,\sigma^2}\left\{ p\frac{S}{n} + \frac{S^2}{(n+2)^2} \| g(X)\|^2 + 2\sigma^2 \frac{S}{n+2} \nabla \cdot g(X) \right\} \\ &= E_{\xi,\sigma^2}\left[p\frac{S}{n} + \frac{S^2}{(n+2)^2} \{\| g(X)\|^2 + 2\nabla \cdot g(X)\} \right] \end{aligned}$$

where ∇ is the vector of first partial derivatives with respect to X.

8. An identity suggesting approximate confidence sets for the mean. By repeated application of formula (2.3), it is possible to obtain, in the one-dimensional case, an unbiased estimate of the expected value of a power of $X - \xi$ times a function of X. This is done for small powers of $X - \xi$ and the result is applied, in the p-dimensional case, to obtain, for a nearly arbitrary estimate of ξ, an unbiased estimate of the variance of the difference between the squared length of the error vector and the unbiased estimate of the risk. This suggests a method of obtaining, approximately, spherical confidence sets for the mean centered as a nearly arbitrary estimate. No attempt is made here to study the validity of this approximation.

LEMMA 4. *If X is a $N(\xi, 1)$ random variable then*

$$E_\xi\{(X - \xi)g(X)\} = E_\xi g'(X),$$

(8.1)

$$E_\xi\{(X - \xi)^2 g(X)\} = E_\xi\{g(X) + g''(X)\}$$

and

$$E_\xi\{(X - \xi)^4 g(X)\} = E_\xi\{3g(X) + 6g''(X) + g^{(iv)}(X)\},$$

where, in each case, all the derivatives involved are assumed to exist in the sense that an indefinite integral of each is the next preceding one, and to have finite expectations. The first through fourth derivatives of g are denoted by g', g'', g''', $g^{(iv)}$.

PROOF. Clearly it suffices to consider the special case $\xi = 0$. Formula (8.1) is the same as equation (2.1). The remaining formulas follow by repeated application of (8.1). Somewhat imprecisely we write $(f(x))'$ as well as $f'(x)$ for the derivative of f at x. Then, formula (8.1) can be rewritten

$$E\{Xg(X)\} = E\{g(X)\}'. \tag{8.2}$$

By repeated application of (8.2) we obtain

$$E\{X^2 g(X)\} = E[X\{Xg(X)\}] = E\{Xg(X)\}' = E\{g(X) + Xg'(X)\} = E\{g(X) + g''(X)\},$$

and similarly

$$E\{X^3 g(X)\} = E\{3g'(X) + g'''(X)\}$$

and

$$E\{X^4 g(X)\} = E\{3g(X) + 6g''(X) + g^{(iv)}(X)\}.$$

The following corollary will not be used here, but may be of some interest.

COROLLARY 2. *If X is a $N(\xi, 1)$ random variable, then, with the derivatives interpreted as in Lemma* 1,

$$\xi E_\xi g(X) = E_\xi\{Xg(X) - g'(X)\}$$

$$\xi^2 E_\xi g(X) = E_\xi\{(X^2 - 1)g(X) - 2Xg'(X) + g''(X)\} \tag{8.3}$$

$$\xi^3 E_\xi g(X) = E_\xi\{(X^2 - 3)Xg(X) - 3(X^2 - 1)g'(X) + 3X^2 g''(X) - g'''(X)\}$$

and

$$\xi^4 E_\xi g(X) = E_\xi\{(X^4 - 6X^2 + 3)g(X) - 4(X^2 - 3)Xg'(X) + 6(X^2 - 1)g''(X) - 4Xg'''(X) + g^{(iv)}(X)\}.$$

It is not difficult to derive these formulas from Theorem 1 by computations similar to those used in the proof of that theorem. However, it may be simpler, or at least more

systematic, to proceed in the following way. Let D denote the operation of differentiation:

$$(Dg)x = g'(x)$$

and T the operation of multiplying by x:

$$(Tg)x = xg(x)$$

Then equation (8.1) can be rewritten

$$\xi E_\xi g(X) = E_\xi\{(D - T)g\}(X). \tag{8.4}$$

By induction, for k a positive integer

$$\xi^k E_\xi g(X) = E_\xi\{(D - T)^k g\}(X). \tag{8.5}$$

Equation (8.3) is simply another form of (8.1) or (8.4), and the equations following (8.3) are obtained by expanding (8.5) for $k = 2$, 3, and 4, using

$$D^j T = TD^j + jD^{j-1}$$

and

$$DT^j = T^j D + jT^{j-1}$$

which follow by induction from the special case $j = 1$ of either.

Next we look at an expression for the mean square of the difference between the squared norm of the error vector and the unbiased estimate of its expectation. The regularity conditions assumed here may be stronger than needed.

For an application of the above ideas in finding minimax estimators of ξ with quartic loss, see Berger (1978).

THEOREM 3. *Let X be a random p-dimensional coordinate vector, normally distributed with mean ξ and the identity as covariance matrix. Let $g:\mathscr{R}^p \to \mathscr{R}^p$ be a twice continuously differentiable function such that*

$$E_\xi\{\| g(X)\|^2 + \textstyle\sum_{i,j} g_{ij}^2(X) + \sum_{i,j} g_{ij}^2(X)\} < \infty,$$

where $g_{ij} = \nabla_j g_i$ and $g_{iij} = \nabla_i \nabla_j g_i$. Then

$$E_\xi[\| X + g(X) - \xi \|^2 - \{p + \| g(X)\|^2 + 2\nabla^T g(X)\}]^2 = 2p + 4E_\xi[\| g(X)\|^2 + 2\nabla^T g(X) + \operatorname{tr}\{\nabla g^T(X)\}^2], \tag{8.6}$$

where g^T denotes the vector-valued function whose value is the transpose of the value of the function g.

PROOF. Expanding the left hand side of (8.6), we obtain

$$\begin{aligned} &E_\xi[\| X + g(X) - \xi \|^2 - \{p + \| g(X)\|^2 + 2\nabla^T g(X)\}]^2 \\ &\quad = E_\xi[(\| X - \xi \|^2 - p) + 2\{(X - \xi)^T g(X) - \nabla^T g(X)\}]^2 \\ &\quad = E_\xi[(\| X - \xi \|^2 - p)^2 + 4\{(X - \xi)^T g(X) - \nabla^T g(X)\}^2 \\ &\qquad + 4(\| X - \xi \|^2 - p)\{(X - \xi)^T g(X) - \nabla^T g(X)\}] \qquad (8.7) \\ &\quad = 2p + 4E_\xi[\{(X - \xi)^T g(X)\}^2 + \{\nabla^T g(X)\}^2 - 2\{(X - \xi)^T g(X)\}\nabla^T g(X) \\ &\qquad + \| X - \xi \|^2 (X - \xi)^T g(X) - \| X - \xi \|^2 \nabla^T g(X)]. \end{aligned}$$

We can express the expectation of the first term in brackets on the right hand side as the expectation of a function of X alone in the following way:

$$
\begin{aligned}
E_\xi[(X-\xi)^T g(X)]^2 &= E_\xi \textstyle\sum_i \sum_j (X_i-\xi_i)(X_j-\xi_j) g_i(X) g_j(X) \\
&= E_\xi \textstyle\sum_i \sum_j \dfrac{\partial}{\partial X_i}\{(X_j-\xi_j) g_i(X) g_j(X)\} \\
&= E_\xi \textstyle\sum_i \sum_j [\delta_{ij} g_i(X) g_j(X) + (X_j-\xi_j)\{g_{ii}(X) g_j(X) + g_i(X) g_{ji}(X)\}] \\
&= E_\xi \textstyle\sum_i \sum_j \{\delta_{ij} g_i(X) g_j(X) + g_{iij}(X) g_j(X) + g_{ii}(X) g_{jj}(X) \\
&\quad + g_{ij}(X) g_{ji}(X) + g_i(X) g_{jij}(X)\} \\
&= E_\xi[\| g(X)\|^2 + \{\nabla^T g(X)\}^2 + \operatorname{tr}\{\nabla g^T(X)\}^2 + 2\textstyle\sum_i \sum_j g_i(X) g_{jij}(X)].
\end{aligned} \tag{8.8}
$$

The second term in brackets on the right hand side of (8.7) is already in the desired form. For the third term we have

$$
\begin{aligned}
E_\xi[\{(X-\xi)^T g(X)\}\nabla^T g(X)] &= E_\xi(\nabla^T[\{\nabla^T g(X)\} g(X)]) \\
&= E_\xi[\{\nabla^T g(X)\}^2 + \textstyle\sum_i \sum_j g_{jj\cdot i}(X) g_i(X)].
\end{aligned} \tag{8.9}
$$

For the fourth term in brackets on the right hand side of (8.7) we have

$$
\begin{aligned}
E_\xi\{\| X-\xi\|^2 (X-\xi)^T g(X)\} &= E_\xi\{\textstyle\sum_i \sum_j (X_i-\xi_i)^2 (X_j-\xi_j) g_j(X)\} \\
&= E_\xi[\textstyle\sum_i \sum_j \dfrac{\partial}{\partial X_j}\{(X_i-\xi_i)^2 g_j(X)\}] \\
&= E_\xi[\textstyle\sum_i \sum_j \{2\delta_{ij}(X_i-\xi_i) g_i(X) + (X_i-\xi_i)^2 g_{jj}(X)\}] \\
&= E_\xi\{2\nabla^T g(X) + \| X-\xi\|^2 \nabla^T g(X)\}.
\end{aligned}
$$

Thus for the combined fourth and fifth terms we have

$$
\begin{aligned}
&E_\xi\{\| X-\xi\|^2 (X-\xi)^T g(X) - \| X-\xi\|^2 \nabla^T g(X)\} \\
&\quad = E_\xi\{2\nabla^T g(X) + \| X-\xi\|^2 \nabla^T g(X) - \| X-\xi\|^2 \nabla^T g(X)\} = 2E_\xi\{\nabla^T g(X)\}.
\end{aligned} \tag{8.10}
$$

Combining (8.7) through (8.10) we obtain (8.6).

It seems plausible that under appropriate conditions, with p large, the random variable in brackets on the left hand side of (8.6) is approximately normally distributed with mean 0, and that the random variable in braces on the right hand side is approximately constant. This suggests as confidence sets for ξ with approximate probability $1-\alpha$ of covering ξ

$$
\begin{aligned}
S_X = \{\xi : \| \xi - (X + g(X))\|^2 &< p + \| g(X)\|^2 + 2\nabla^T g(X) \\
&\quad + c_\alpha \sqrt{2p + 4\{\| g(X)\|^2 + 2\nabla^T g(X) + \operatorname{tr}[\nabla g^T(X)]^2\}}\}
\end{aligned} \tag{8.11}
$$

where

$$\Phi(c_\alpha) \doteq 1-\alpha.$$

Actually it may be better to choose c in such a way that

$$P\{\chi_p^2 < p + c_\alpha \sqrt{2p}\} = 1-\alpha.$$

With this choice of c_α and reasonable choice of g, the probability that $\xi \in S_X$ approaches $1-\alpha$ as $\xi \to \infty$.

Different approaches to obtaining improved confidence sets for ξ are described by Morris (1977), Faith (1978) and Berger (1980).

REFERENCES

ANDERSON, T. W. (1971). *The Statistical Analysis of Time Series.* Wiley, New York.

BERGER, J. (1979). Minimax estimation of a multivariate normal mean under polynomial loss. *J. Multivariate Anal.* **8** 173–180.

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

BROWN, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903.

EFRON B. and MORRIS, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators, Part I: The Bayes case. *J. Amer. Statist. Assoc.* **66** 807–815.

EFRON, B. and MORRIS, C. (1972a). Limiting the risk of Bayes and empirical Bayes estimators, Part II: The empirical Bayes case. *J. Amer. Statist. Assoc.* **67** 130–139.

EFRON, B. and MORRIS, C. (1972b). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika* **59** 335–347.

EFRON, B. and MORRIS, C. (1973a). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130.

EFRON, B. and MORRIS, C. (1973b). Combining possibly related estimation problems (with discussion). *J. Roy. Statist. Soc. Ser. B* **35** 379–421.

EFRON, B. and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311–319.

EFRON, B. and MORRIS, C. (1976). Families of minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.* **4** 11–21.

FAITH, R. E. (1978). Minimax Bayes estimators of a multivariate normal mean. *J. Multivariate Anal.* **8** 372–379.

FERGUSON, T. (1967). *Mathematical Statistics, A Decision-Theoretic Approach.* Academic, New York.

HELMS, L. (1969). *Introduction to Potential Theory.* Wiley, New York.

HUDSON, H. M. (1974). *Empirical Bayes Estimation.* Ph.D. thesis, Department of Statistics, Stanford University, Stanford, California.

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

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361–380. Univ. California Press.

MORRIS, C. (1977). Interval estimation for empirical Bayes generalizations of Stein's estimator. *Proc. 22nd. Conf. on Design of Experiments in Army Res. Dev. and Testing*, ARO Report 77-2.

STEIN, C. (1973). Estimation of the mean of a multivariate normal distribution. *Proc. Prague Symp. on Asymptotic Statistics* 345–381.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305