ESTIMATION OF THE PARAMETERS OF A
MULTIVARIATE NORMAL DISTRIBUTION

I. ESTIMATION OF THE MEANS

BY

CHARLES M. STEIN

TECHNICAL REPORT NO. 63
NOVEMBER 6, 1974

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT GP-30711X-2

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ESTIMATION OF THE PARAMETERS OF A
MULTIVARIATE NORMAL DISTRIBUTION

I. ESTIMATION OF THE MEANS

by

Charles M. Stein

TECHNICAL REPORT NO. 63
NOVEMBER 6, 1974

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT GP-30711X-2

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ESTIMATION OF THE PARAMETERS OF A
MULTIVARIATE NORMAL DISTRIBUTION

I. ESTIMATION OF THE MEANS

by

Charles M. Stein

1. Introduction

The aim of this series of papers is to study the problem of estimating
certain functions of the parameters of a multivariate normal distribution
with the aid of formulas for unbiased estimates of the risk. The problem
studied in this first paper is that of estimating the mean 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
difficult and vague problem of how to choose among the estimates considered
here will be discussed in later papers in the series.

The present work arose from a question raised by Malcolm Hudson, a
Stanford student, in connection with his dissertation (1974). It was
also inspired, in part, by the work of Efron and Morris 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 (1974). 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 and 5 two special problems,
related to papers of Efron and Morris (1971, 1972a, 1973a) are considered.
First, if \( X \) is normally distributed with unknown mean \( \xi \) 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 \( \lambda \) is found that yields
a minimax estimate that is optimum in the sense that the risk of the
estimate cannot be improved at any point by multiplication of \( \lambda \) by a
constant factor. In Section 5 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 \( X_i \). Other special
problems will be treated in the third paper of this series. In Section
6, the modification needed when the common variance is unknown but an
independent estimate of the variance is available is considered in an
important special case. In Section 7, 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.

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 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.

**Lemma 1:** Let \( Y \) be a normally distributed real random variable with mean \( 0 \) and variance \( 1 \), and let \( g: \mathbb{R} \to \mathbb{R} \) be an indefinite integral of the Lebesgue measurable function \( g' \), that is, for all \( a \) and \( b \) with \( a < b \),

\[
(1) \quad g(b) - g(a) = \int_a^b g'(x) \, dx.
\]

Suppose also that

\[
(2) \quad E|g'(Y)| < \infty.
\]

Then

\[
(3) \quad Eg'(Y) = E[g(Y)].
\]

**Proof:**

\[
(4) \quad Eg'(Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g'(y) e^{-\frac{1}{2}y^2} \, dy
\]

\[
= \frac{1}{\sqrt{2\pi}} \left\{ \int_0^{\infty} g'(y) \left( \int_y^{\infty} ze^{-\frac{1}{2}z^2} \, dz \right) \, dy \right. \\
- \left. \int_{-\infty}^0 g'(y) \left( \int_{-\infty}^{y} ze^{-\frac{1}{2}z^2} \, dz \right) \, dy \right\}
\]
\[
\frac{1}{\sqrt{2\pi}} \left\{ \int_0^\infty ze^{-\frac{1}{2}z^2} \left( \int_0^z g'(y) dy \right) dz \right.
- \int_{-\infty}^0 ze^{-\frac{1}{2}z^2} \left( \int_z^0 g'(y) dy \right) dz \right\}
\]
\[
= \frac{1}{\sqrt{2\pi}} \left\{ \int_0^\infty ze^{-\frac{1}{2}z^2} [g(z) - g(0)] dz
+ \int_{-\infty}^0 ze^{-\frac{1}{2}z^2} [g(z) - g(0)] dz \right\}
\]
\[
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty z g(z) e^{-\frac{1}{2}z^2} dz = Eyg(y). 
\]

The third equality in (4) 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, let us define a new random variable \( X \) related to the random variable \( Y \) of Lemma 1 by

(5) \[ X = \sigma Y + \xi, \]

where \( \xi \) is real and \( \sigma \) positive so that \( X \) is normally distributed with mean \( \xi \) and variance \( \sigma^2 \). If we also define a new function \( h : \mathbb{R} \to \mathbb{R} \) by

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

the derivative \( h' \) of \( h \), in the same sense as before, is given by

(7) \[ h'(x) = \frac{1}{\sigma} g'\left(\frac{x-\xi}{\sigma}\right), \]

4
and consequently formula (3) yields

\[
\begin{align*}
\mathbb{E}h'(x) &= \frac{1}{\sigma} \mathbb{E}g'(\frac{x - \mu}{\sigma}) = \frac{1}{\sigma} \mathbb{E}g'(y)
= \mathbb{E} \frac{X - \mu}{\sigma} \mathbb{E}g(X)
= \mathbb{E} \frac{X - \mu}{\sigma} \mathbb{E}g(\frac{X - \mu}{\sigma})
= \mathbb{E} \frac{X - \mu}{\sigma} \mathbb{E}h(x).
\end{align*}
\]

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

\[
(x \cdot y) = \sum_{i=1}^{p} x_i y_i
\]

and

\[
\|x\|^2 = x \cdot x = \sum_{i=1}^{p} x_i^2.
\]

**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. Roughly speaking \( \nabla \) is the vector differential operator of first partial derivatives with \( i \)th coordinate

\[
\nabla_i = \frac{\partial}{\partial x_i}.
\]

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 \( \mu \) 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 \( \mu \) we write \( \mathbb{E}^\mu \) rather than \( \mathbb{E} \).
Lemma 2: If \( h: \mathcal{R}^p \to \mathcal{R} \) is an almost differentiable function such that

\[
E^\xi ||\nabla h(x)|| < \infty ,
\]

then

\[
E^\xi h(x) = E^\xi (x-\xi)h(x) .
\]

Proof: For \( i \in \{1 \ldots p\} \), let \( \mathcal{B}_i \) be the \( \sigma \)-algebra generated by \( X_i \) alone, and let \( \mathcal{B}_{-i} \) be the \( \sigma \)-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 \( j \)-th coordinate \( X_j \) for \( j \neq i \). Somewhat imprecisely, to express the fact that \( X \) determines and is determined by \( X_i \) and \( X_{-i} \) we 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 \( \omega \) in the set \( \Omega \) of the underlying probability space,

\[
[E^{-i}(X_i-\xi_i) h(X)](\omega) = E^{-i}(X_{-i} h(X)) h(X_i, X_{-i}(\omega))](\omega)
\]

\[
= [E^{-i}(\nabla h)_i (X_i, X_{-i}(\omega)))](\omega) = [E^{-i}(\nabla h)_i (X)](\omega) .
\]

Thus

\[
E^{-i}(\nabla h)_i (X) = E^{-i}(X_i-\xi_i) h(X) ,
\]

and, taking the expectation of both sides, we find
(18) \[ E^S(\mathcal{V}h)_1(x) = E^S(x_1-x_1)h(x), \]

which yields (14).

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

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

(19) \[ S_r(x^0) = \{ x: ||x-x^0||^2 = r^2 \} \]
of radius \( r \) centered at \( x^0 \) is not greater than \( f(x^0) \). The function \( f \) is superharmonic in \( \mathcal{R}^p \) if it is superharmonic at each \( x^0 \in \mathcal{R}^p \).

**Lemma 3:** If \( f: \mathcal{R}^p \to \mathcal{R} \) is twice continuously differentiable, then \( f \) is superharmonic in \( \mathcal{R}^p \) if and only if, for all \( x \in \mathcal{R}^p \),

(20) \[ \nabla^2 f(x) \leq 0 \]

where \( \nabla^2 \) is the Laplacian

(21) \[ \nabla^2 = \sum_{i=1}^{p} \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_i} = \sum_{i=1}^{p} \frac{\partial^2 f}{\partial x_i^2}. \]

The proof of Lemma 3 is given in Theorem 4.8 of Helms (1969), p. 63.

**Definition 3:** The twice continuously differentiable function \( f: \mathcal{R}^p \to \mathcal{R} \) is harmonic at \( x^0 \in \mathcal{R}^p \) if

(22) \[ \nabla^2 f(x^0) = 0. \]

It is harmonic in \( \mathcal{R}^p \) if it is harmonic at each \( x^0 \in \mathcal{R}^p \).

We shall apply Lemma 2 to obtain an unbiased estimate of the risk of a nearly arbitrary estimate of the mean of a multivariate normal
distribution with the identity as covariance matrix. In accordance with (9) and (12), if \( g: \mathcal{F}^p \to \mathcal{F}^p \) is almost differentiable, we shall write

\[
\nabla \cdot g = \Sigma \nabla_i g_i.
\]

**Theorem 1:** Let \( g: \mathcal{F}^p \to \mathcal{F}^p \) be an almost differentiable function for which

\[
E \xi_{\Sigma} |\nabla_i g_i(x)| < \infty.
\]

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

\[
E_{\Sigma} [X_i + g_i(x) - \xi_i]^2 = 1 + E_{\Sigma} [g_i^2(x)] + 2 \nabla_i g_i(x),
\]

and consequently

\[
E_{\Sigma} |x + g(x) - \xi|^2 = p + E_{\Sigma} [|g(x)|^2 + 2 \nabla \cdot g(x)].
\]

**Proof:** From formula (14) of Lemma 2 with \( h = g_1 \), it follows that

\[
E_{\Sigma} [X_1 + g_1(x) - \xi_1]^2 = E_{\Sigma} [(x_1 - \xi_1)^2 + 2(x_1 - \xi_1 \cdot g_1(x) + g_1^2(x)]
\]

\[= 1 + 2E_{\Sigma} \nabla_1 g_1(x) + E_{\Sigma} g_1^2(x),\]

which is (25). Summing over \( i \) we obtain formula (26). 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 \( \xi \).

When the dimension \( p \geq 3 \), we shall obtain a large collection of minimax estimates of \( \xi \) 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 $f: \mathbb{R}^p \to \mathbb{R} \cap \{0\}^c$ be an almost differentiable function for which $\nabla f: \mathbb{R}^p \to \mathbb{R}^p$ can be taken to be almost differentiable, and suppose also that

\begin{equation}
E^5 \frac{1}{f(x)} \sum_i \nabla_i^2 f(x) < \infty,
\end{equation}

and

\begin{equation}
E^5 \left| \nabla \log f(x) \right|^2 < \infty.
\end{equation}

Then

\begin{equation}
E^5 \left| x + \nabla \log f(x) - \xi \right|^2 = p + E^5 \left[ 2 \frac{\nabla^2 f(x)}{f(x)} - \frac{||\nabla f(x)||^2}{f^2(x)} \right] = p + 4 E^5 \frac{\nabla^2 f(x)}{f(x)}.
\end{equation}

Proof: Let $g: \mathbb{R}^p \to \mathbb{R}^p$ be defined by

\begin{equation}
g = \nabla \log f = \frac{\nabla f}{f}.
\end{equation}

Then

\begin{equation}
\nabla \cdot g = \nabla \cdot \nabla \log f = \frac{\nabla^2 f}{f} - \frac{||\nabla f||^2}{f^2},
\end{equation}

and thus it follows from equation (26) of Theorem 1 that

\begin{equation}
E^5 \left| x + \nabla \log f(x) - \xi \right|^2 = E^5 \left| x + g(x) - \xi \right|^2
\end{equation}

\begin{equation}
= p + E^5 \left[ ||g(x)||^2 + 2 \nabla \cdot g(x) \right]
\end{equation}

\begin{equation}
= p + E^5 \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]
\end{equation}

\begin{equation}
= p + E^5 \left[ 2 \frac{\nabla^2 f(x)}{f(x)} - \frac{||\nabla f(x)||^2}{f^2(x)} \right],
\end{equation}

9
which is the first form of (30). Also

\[
(34) \quad \nabla^2 \sqrt{f} = \nabla \cdot \nabla \sqrt{f} = \nabla \cdot \frac{\nabla \sqrt{f}}{2 \sqrt{f}} = \frac{1}{2} \nabla^2 \sqrt{f} - \frac{1}{4f^{3/2}} \| \nabla f \|^2.
\]

The final expression of (30) follows.

**Corollary 1:** If \( f: \mathcal{X}^{p} \to \mathcal{X}^{p} \cap \{0\}^c \) is twice continuously differentiable and its square root is superharmonic and (28) and (29) are satisfied, then \( X + \nabla \log f(X) \) is a minimax estimate of \( \xi \), that is, for all \( \xi \),

\[
(35) \quad E^\xi \| X + \nabla \log f(X) - \xi \|^2 = p + 4E^\xi \frac{\nabla^2 \sqrt{f(X)}}{\sqrt{f(X)}}
\]

\[
\leq p = \inf_{g} \sup_{\xi} E^\xi \| X + g(X) - \xi \|^2.
\]

**Proof:** The first equality in (35) follows from (30), the inequality follows from the defining property (22) of superharmonic functions, and the final equality is well known.

3. **Formal Bayes Estimates**

Some easy known results about Bayes estimates, and also formal Bayes estimates, are recalled, including the fact that they are all of the form considered in Theorem 2.2. The unbiased estimate of the risk of a formal Bayes estimate given in Theorem 2.2 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 \( \xi \) be a random \( p \)-dimensional coordinate vector distributed according to the probability measure \( \Pi \), called the
prior probability measure. Let \( X \) be a random vector in \( \mathcal{R}^P \), conditionally normally distributed given \( \xi \) with conditional mean \( \xi \), and the identity as conditional covariance matrix. Then the unconditional density \( f \) of \( X \) with respect to Lebesgue measure in \( \mathcal{R}^P \) is given by

\[
f(x) = \frac{1}{(2\pi)^{P/2}} \int e^{-\frac{1}{2}\|x-\xi\|^2} \, d\Pi(\xi) .
\]

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

\[
E \|\xi - \phi(X)\|^2 = E^X \|\xi - \phi(X)\|^2
\]

\[
= E \int \|\xi - \phi(X)\|^2 e^{-\frac{1}{2}\|x-\xi\|^2} \, d\Pi(\xi)
\]

\[
\int e^{-\frac{1}{2}\|x-\xi\|^2} \, d\Pi(\xi)
\]

is given by

\[
\phi_\Pi(X) = E^X \xi = X + E^X (\xi - X) = X + \int (\xi - X) e^{-\frac{1}{2}\|x-\xi\|^2} \, d(\xi)
\]

\[
\int e^{-\frac{1}{2}\|x-\xi\|^2} \, d\Pi(\xi)
\]

\[
= X + \nabla \log f(X) ,
\]

where \( f \) is given by (1). In equation (2), \( E \) denotes unconditional expectation. More generally if \( \Pi \) is a possibly infinite measure for which \( f \) defined by (1) is everywhere finite, we define the formal Bayes estimate \( \phi_\Pi(X) \) by (3). Formal posterior expectation \( E^X \) is defined by the formula that yields posterior expectation in the case where \( \Pi \) is a probability measure.
(4) \[ E^X g(X, \xi) = \frac{\int g(X, \xi) e^{-\frac{1}{2}||X-\xi||^2} \, d\pi(\xi)}{\int e^{-\frac{1}{2}||X-\xi||^2} \, d\pi(\xi)} \, . \]

Next let us compare the unbiased estimate of the risk of the formal Bayes estimate \( \phi_\Pi(X) \) of \( \xi \) given by Theorem 2.2 with the formal posterior risk \( E^X ||\xi - \phi_\Pi(X)||^2 \). From Theorem 2.2, the unbiased estimate of the risk is given by

(5) \[ \rho(X) = p + 2 \frac{\nabla^2 f(X)}{f(X)} - \frac{||\nabla f(X)||^2}{f^2(X)} \, . \]

For the formal posterior risk we have

(6) \[ E^X ||\xi - \phi_\Pi(X)||^2 = E^X ||\xi - X - \nabla \log f(X)||^2 \]
\[ = E^X \{||X - \xi||^2 - ||\nabla \log f(X)||^2 \} \]
\[ = p + \frac{\nabla^2 f(X)}{f(X)} - ||\nabla \log f(X)||^2 \, . \]

The second equality in (6) uses essentially the theorem of Pythagoras in the appropriate Hilbert space. The squared distance from \( \xi \) to \( X \) is the sum of the squared distance from \( \xi \) 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 (6) uses the fact that

(7) \[ \frac{\nabla^2 f(X)}{f(X)} = \frac{\nabla^2 \int e^{-\frac{1}{2}||X-\xi||^2} \, d\pi(\xi)}{\int e^{-\frac{1}{2}||X-\xi||^2} \, d\pi(\xi)} = \frac{\int \left[ \frac{||X-\xi||^2 - p}{e^{-\frac{1}{2}||X-\xi||^2}} \right] e^{-\frac{1}{2}||X-\xi||^2} \, d\pi(\xi)}{\int e^{-\frac{1}{2}||X-\xi||^2} \, d\pi(\xi)} \]
\[ = E^X [||X-\xi||^2 - p] \, . \]
Comparing (5) and (6) we see that

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

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) in the sense that

$$E^X | | \xi - \phi_\Pi (x) | |^2 \geq \rho (x) ,$$

and thus, for all $\xi$,

$$E^\xi E^X | | \xi - \phi_\Pi (x) | |^2 \geq E^\xi \rho (x) = E^\xi | | \xi - \phi_\Pi (x) | |^2 .$$

Of course this inequality cannot hold in a non-trivial way if $\Pi$ is a probability measure.

Let us also observe that, if the formal prior measure $\Pi$ has a superharmonic density $\pi$ with respect to Lebesgue measure, then $f$ defined by (1) is also superharmonic and thus $\phi_\Pi (x)$ is a minimax estimate of $\xi$ and (9) and (10) hold. For

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

If $\pi$ is superharmonic so is $x \mapsto \pi(x-y)$, 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 $\xi$. For, it follows from his Theorem that a formal Bayes estimate with respect to a prior density $\pi$ of the form
\[ \pi(\xi) = \int \frac{1}{||\xi - \eta||^{(p-2)}} \, d\rho(\eta) \]

with \( \rho \) a finite measure is admissible, and since \( \pi \) is superharmonic and thus also the corresponding \( f \) given by (1) with \( d\Pi(\xi) = \pi(\xi) d\xi \), it follows from formula (2.30) 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 \( \xi \) if we have decided in advance on the direction of the correction, related linearly to \( X \). In this spirit we shall look at the use of three-term moving averages, in both the cyclical and the ordered case, and also a case of observations assigned different weights.

Let us look at estimates of the form

\[ \hat{\xi} = X - \lambda(X)AX \]

where \( A \) is a preassigned symmetric matrix, and \( \lambda: \mathcal{R}^p \to \mathcal{R} \) is to be chosen appropriately. It is convenient, in this section, to use matrix notation, to think of the \( x \in \mathcal{R}^p \) as column vectors and to write, for example, \( x' \) for the row vector that is the transpose of the column vector \( x \). We observe that, if

\[ 2A < (\text{tr } A) I, \]

in the sense that the largest characteristic root of \( A \) is less than \( \frac{1}{2} \text{tr } A \), then the risk of the estimate \( \hat{\xi} \) defined by (1) with
\( \lambda(x) = \frac{1}{x'Bx} \),

where

\( B = [(tr\ A)I - 2A]^{-1} A^2 \),

is given by

\[
\mathbb{E}^\xi \| x - \frac{1}{x'Bx} Ax - \xi \|^2 = p + \mathbb{E}^\xi \left[ \frac{X^'A^2X}{(x'Bx)^2} - 2\nu \cdot \frac{AX}{x'Bx} \right] \\
= p + \mathbb{E}^\xi \left[ \frac{X^'A^2X}{(x'Bx)^2} - 2\frac{tr\ A}{x'Bx} + 4\frac{X'ABX}{(x'Bx)^2} \right] = p - \mathbb{E}^\xi \frac{X^'A^2X}{(x'Bx)^2}.
\]

In the final equality we have used the particular choice (4) of \( B \). Condition (2) is needed for \( B \) given by (4) to be positive definite. If \( B \) is not positive definite, the expectations do not exist and the formal computations are incorrect. Formula (5) shows that, under the conditions indicated, the estimate \( \hat{\xi} \) defined by (1), (3), and (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 \( \lambda \) in (3) by a constant factor cannot be better at any parameter point. For, by a simple modification of (5) we see that, for any real constant \( \beta \),

\[
\mathbb{E}^\xi \| x - \frac{\beta}{x'Bx} Ax - \xi \|^2 = p + \mathbb{E}^\xi \frac{\beta^2X^'A^2X - 2\beta X'[(tr\ A)I - 2A]BX}{(x'Bx)^2} \\
= p + (\beta^2 - \beta) \mathbb{E}^\xi \frac{X^'A^2X}{(x'Bx)^2}.
\]

For all \( \xi \), this is minimized by \( \beta = 1 \). We observe that the special case \( A = I \) is the non-truncated estimate considered by James and Stein (1961).
Let us apply this to the question of the appropriate choice of the weight in a three-term symmetric moving average, first in the cyclic case. Let $X_1, \ldots, X_p$ be independently normally distributed with means $\xi_1, \ldots, \xi_p$ and variance 1, and suppose we plan to estimate the $\xi_i$ by

$$\hat{\xi}_i = X_i - \lambda(X)\left[X_i - \frac{1}{2}(X_{i-1} + X_{i+1})\right],$$

where it is understood that $X_0 = X_p$ and $X_{p+1} = X_1$, and similarly for the $\xi^l$s. This is the special case of (1) with $A$ given by

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

The characteristic roots and vectors of $A$, the solutions $\alpha_j$ and $y_j$ of

$$Ay_j = \alpha_j y_j,$$

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

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

by

$$\alpha_j = 1 - \cos 2\pi \frac{j}{p},$$

and for $i \in \{1 \ldots p\}$
\[ y_{ij} = \begin{cases} \frac{1}{\sqrt{p}} & \text{if } j = 0 \\ \frac{(-1)^i}{\sqrt{p}} & \text{if } j = \left[-\frac{p}{2}\right] \\ \sqrt{\frac{1}{p}} \cos 2\pi \frac{i}{p} & \text{if } -\left[\frac{p}{2}\right] < j < 0 \\ \sqrt{\frac{1}{p}} \sin 2\pi \frac{i}{p} & \text{if } 0 < j < \left[\frac{p}{2}\right] \end{cases} \]

this being the \( i^{\text{th}} \) coordinate of \( y_{j} \). No difficulty is caused by the different ranges of \( i \) and \( j \) in (12). See, for example, Anderson (1971), pp. 278-284. The matrix \( A \) can be expressed as

(13)
\[ A = \alpha y y', \]

where \( \alpha \) is the diagonal matrix with \( j^{\text{th}} \) diagonal element \( \alpha_j \) (for \( j \) satisfying (10)) and \( y \) is the (orthogonal) matrix with \( ij \) element \( y_{ij} \). From the definition (8) of \( A \) we have

(14)
\[ \text{tr } A = p, \]

and from (11) 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 (2) is satisfied if and only if \( p \geq 5 \). In this case the matrix \( B \), given by (4), to be used in (1) and (3) is

(15)
\[ B = [(\text{tr } A)I - 2A^{-1}]A^2 = y(pI - 2\alpha)^{-1} \alpha^2 y'. \]

For reasons that will be indicated later, it is unreasonable to use a three-term moving average with weights more extreme than \( (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \). Thus it seems appropriate to modify our estimate to

17
\[ \hat{\xi} = x - \lambda_1(x)AX, \]

where

\[ \lambda_1(x) = \frac{1}{x'BX} \wedge \frac{2}{3}. \]

Of course \( A \) is given by (8) and \( B \) by (15). The unbiased estimate of the improvement in the risk is changed from

\[ \Delta(x) = \frac{x'A^2x}{(x'BX)^2}, \]

given by (5), to

\[ \Delta_1(x) = \begin{cases} 
\Delta(x) & \text{if } x'BX > \frac{3}{2} \\
\frac{4p}{3} - \frac{4}{9} \sum_{i=1}^{p} \left( x_i - \frac{x_{i-1} + x_{i+1}}{2} \right)^2 & \text{if } x'BX \leq \frac{3}{2}.
\end{cases} \]

The second expression in (19) is obtained by applying Theorem 2.1 to the form \( g \) takes on \( \{x: x'Ex \leq \frac{3}{2}\} \):

\[ g_1(x) = -\frac{2}{3} \left[ x_i - \frac{x_{i-1} + x_{i+1}}{2} \right]. \]

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 (1) with \( \lambda \) given by (3) and (4) and the \( p \times p \) matrix \( A \) given by

\[ A_{ij} = \begin{cases} 
\frac{1}{2} & \text{if } i=j=1 \text{ or } i=j=p \\
1 & \text{if } i=j \neq 1, p \\
-\frac{1}{2} & \text{if } |i-j| = 1 \\
0 & \text{if } |i-j| \neq 0,1,
\end{cases} \]
that is

\[
\hat{\xi}_i = \begin{cases} 
(1 - \lambda(x)\hat{x}_1 + \frac{3}{2}\lambda(x)(\hat{x}_{i-1} + \hat{x}_{i+1}) & \text{if } i \neq 1, p \\
(1 - \frac{3}{2}\lambda(x))\hat{x}_1 + \frac{1}{2}\lambda(x)\hat{x}_2 & \text{if } i = 1 \\
(1 - \frac{3}{2}\lambda(x))\hat{x}_p + \frac{1}{2}\lambda(x)\hat{x}_{p-1} & \text{if } i = p.
\end{cases}
\]

The characteristic roots \( \alpha_j \) and vectors \( \gamma_j \) of \( A \) are given, for \( j \in \{1, \ldots, p\} \) by

\[
\alpha_j = 1 - \cos \frac{\pi(j-1)}{p}
\]

and, for \( i \in \{1, \ldots, p\} \),

\[
\gamma_{ij} = \begin{cases} 
\frac{1}{\sqrt{p}} & \text{if } j = 1 \\
\sqrt{\frac{2}{p}} \cos \frac{\pi(2i-1)(j-1)}{2p} & \text{if } j \neq 1.
\end{cases}
\]

See Anderson (1971), pp. 284–290. The matrix \( A \) can be expressed as

\[
A = \gamma \alpha \gamma^T
\]

where \( \alpha \) is the diagonal matrix with \( j \)th diagonal element \( \alpha_j \) given by (23) and \( \gamma \) is the orthogonal matrix with \( i, j \) element given by (24) for \( i, j \in \{1, \ldots, p\} \). By (21),

\[
\text{tr} A = p - 1,
\]

and, by (23) the largest characteristic root of \( A \) is less than 2. Thus, according to condition (2), the estimate given by (1), (3), and (4) is applicable for \( p \geq 5 \). Again the appropriate choice of \( B \) is given by (15), but with \( A, \alpha, \) and \( \gamma \) given by (21), (23), and (24).
Again it seems appropriate to replace $\lambda$ in (1) by $\lambda_1$, given by

$$
\lambda_1(x) = \frac{1}{x'Bx} \wedge \frac{2}{3} .
$$

The unbiased estimate of the improvement in the risk is changed from $\Delta(X)$ given by (5) to $\Delta_1(X)$ given by

$$
\Delta_1(X) = \begin{cases} 
\Delta(X) & \text{if } x'Bx \geq \frac{3}{2} \\
\frac{4(p-1)}{3} - \frac{4}{9} \sum_{i=2}^{p-1} \left[ x_i - \frac{x_{i-1} + x_{i+1}}{2} \right]^2 & \text{if } x'Bx < \frac{3}{2} 
\end{cases}
$$

The second expression is obtained by applying Theorem 2.1 to the form $g$ takes on \( \{ x : x'Bx \leq \frac{3}{2} \} :$

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

Let us look at the estimate given by (1), (3), and (4) when $A$ is a diagonal matrix, a case that may arise directly in practice, and which is not essentially less general than the general case, since the symmetric matrix $A$ can always be diagonalized by an orthogonal matrix. Let $A_j$ be the $j^{th}$ diagonal element of the diagonal matrix $A$. Then $B$ defined by (4) is the diagonal matrix with $j^{th}$ diagonal element given by
(30)\[ B_j = \frac{A_j^2}{\sum_{i=1}^{p} A_i - 2A_j}, \]

and \( \lambda \) defined by (30) is given by

(31)\[ \lambda(X) = \left( \frac{\sum_{j=1}^{p} A_j^2 x_j^2}{\sum_{i=1}^{p} A_i - 2A_j} \right)^{-1}. \]

The estimate (1) becomes

(32)\[ \hat{\xi}_j = x_j - \lambda(X) A_j x_j, \]

and the formula (5) giving an unbiased estimate of its risk becomes

(33)\[ \mathbb{E}[\sum_{j=1}^{p} [x_j - \lambda(X) A_j x_j - \xi_j]^2] = p - \mathbb{E}\left[ \left( \sum_{j=1}^{p} A_j^2 x_j^2 \right) \left( \sum_{i=1}^{p} \frac{A_i^2 x_i^2}{A_i - 2A_j} \right)^{-2} \right]. \]

5. 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

(1)\[ \hat{\xi} = \left( 1 - \frac{p-2}{||X||^2} \right) X \]

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 comparatively unimportant deterioration compared with the James-Stein estimate if the prior distribution of $\xi$ has spherical symmetry. We consider a modification of their 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

\begin{equation}
Z_i = |X_i|,
\end{equation}

and let

\begin{equation}
Z(1) < \cdots < Z(p)
\end{equation}

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

\begin{equation}
\hat{\xi} = X + g(X),
\end{equation}

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

\begin{equation}
g_i(x) = \begin{cases} 
- \frac{a}{\Sigma(x_j^2 \wedge Z^2(k))} X_i & \text{if } |X_i| \leq Z(k) \\
- \frac{a}{\Sigma(x_j^2 \wedge Z^2(k))} Z(k) \text{sgn } X_i & \text{if } |X_i| > Z(k)
\end{cases}
\end{equation}

where $a$ is a constant to be determined. By Theorem 2.1
\begin{align*}
\mathbb{E}^\xi \| \xi - \hat{\xi} \|^2 &= p + \mathbb{E}^\xi \left[ \frac{s^2}{\sum (x_j^2 \wedge z_{(k)}^2)} - 2a \sum_{i=1}^{k} \frac{1}{\Sigma (x_j^2 \wedge z_{(k)}^2)} ight] \\
&\quad + 4a \sum_{i=1}^{k-1} \frac{x_i^2}{(\Sigma (x_j^2 \wedge z_{(k)}^2))^2} + 4a(p-k+1) \frac{z_{(k)}^2}{(\Sigma (x_j^2 \wedge z_{(k)}^2))^2} \\
&= p + [a^2 - 2(k-2)a] \mathbb{E}^\xi \frac{1}{\Sigma (x_j^2 \wedge z_{(k)}^2)}. \\
\end{align*}

The optimum choice of \( a \) is

\begin{equation}
a = \frac{1}{k-2},
\end{equation}

and, for this choice, the risk given by (6) becomes

\begin{align*}
\mathbb{E}^\xi \| \hat{\xi} - \xi \|^2 &= p - (k-2)^2 \mathbb{E}^\xi \frac{1}{\Sigma (x_j^2 \wedge z_{(k)}^2)}. \\
\end{align*}

As a guide to the choice of \( k \), let us compute, for large \( p \), the relative efficiency of the estimate given by (4), (5), and (7) compared to the James-Stein estimate

\begin{align*}
\hat{\xi}_0 &= \left( 1 - \frac{p-2}{\| x \|^2} \right) x \\
\end{align*}

in the case most favorable to the James-Stein estimate, that where the \( \xi \) are themselves independently normally distributed with variance \( \tau^2 \). We shall see that if \( \tau > 0 \), the asymptotic relative efficiency \( e^y \) when \( \frac{k}{p} \to y \) is given by

\begin{align*}
e^y &= \frac{y^2}{y + (1-y) \left[ \Phi \left( \frac{1+y}{2} \right) \right]^2} - \frac{2 \Phi^{-1}(\frac{1+y}{2}) - \log \Phi^{-1}(\frac{1+y}{2})}{\sqrt{2\pi}} e^{-\left[ \frac{1+y}{2} \right]^2/2}
\end{align*}
Some numerical values are given in Table I.

\[
\begin{array}{cc}
y & 0.827 \\
0.5 & 0.873 \\
0.6 & 0.909 \\
0.7 & 0.943 \\
0.8 & 0.974 \\
0.9 &
\end{array}
\]

To derive (10), we observe that the estimated improvement in the risk for the estimate \( \xi^{(k)} \), given by (4), (5), and (7):

\[
\xi_i^{(k)}(x) = \begin{cases} 
\left( 1 - \frac{k-2}{\sum (x_j^2 + z_{(k)}^2)} \right) x_i & \text{if } |x_i| < z_{(k)} \\
\frac{k-2}{\sum (x_j^2 + z_{(k)}^2)} z_{(k)} \text{ sgn } x_i & \text{if } |x_i| \geq z_{(k)}
\end{cases}
\]

is

\[
\Delta^{(k)}(x) = \frac{(k-2)^2}{\sum (x_j^2 + z_{(k)}^2)},
\]

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

\[
\Delta(x) = \frac{(p-2)^2}{\sum x_j^2}.
\]

The truncation in (9) is ignored because with \( \tau^2 > 0 \) fixed and \( p \to \infty \), the probability that it will occur approaches 0. For large \( p \), \( \Delta(x) \) and \( \Delta^{(k)}(x) \) are approximately constant with high probability:

\[
\Delta(x) \approx \frac{p}{\sum x_j^2} = \frac{p}{1 + \tau^2}.
\]
and

\begin{equation}
\Delta^{(k)}(X) \approx \frac{k^2}{p B(x_j^2 \wedge Z^2_{(k)})} \approx \frac{k^2}{p(1+r^2)^2 \left[ \int_0^a x^2 e^{-\frac{1}{2}x^2} \, dx + a^2 \int_a^\infty e^{-\frac{1}{2}x^2} \, dx \right]}
\end{equation}

where \( a \) is an approximation to \( Z_{(k)}/\sqrt{1+r^2} \), given by

\begin{equation}
a = \Phi^{-1}(\frac{1+y}{2}).
\end{equation}

The first integral in the denominator of (15) can be evaluated by integration by parts:

\begin{equation}
\int_0^a x^2 e^{-\frac{1}{2}x^2} \, dx = \int_0^a x d(-e^{-\frac{1}{2}x^2}) = -ae^{-\frac{1}{2}a^2} + \int_0^a e^{-\frac{1}{2}x^2} \, dx.
\end{equation}

Thus

\begin{equation}
\frac{\Delta^{(k)}(X)}{\Delta(X)} \approx \frac{y^2}{2 \left[ \int_0^a e^{-\frac{1}{2}x^2} \, dx - ae^{-\frac{1}{2}a^2} + a^2 \int_a^\infty e^{-\frac{1}{2}x^2} \, dx \right]}
\end{equation}

which is (18).

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 \(.7 \) \( p \) is small enough so that it will ordinarily be more than compensated for by the possibility that the empirical c.d.f. of the 
\( \xi_i \)
is long-tailed. For small \( p \), the possible loss in efficiency is likely to be somewhat larger for a given value of \( y = \frac{k}{p} \). Of course, we must have \( k \geq 3 \) for the formulas to be meaningful.
6. 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 $\chi^2$.

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 $\xi$ and covariance matrix $\sigma^2 I$, where $\sigma^2$ is unknown but we also observe a real random variable $S$, distributed independently of $X$ as $\sigma^2 \chi_n^2$. If, in the case where $\sigma^2$ is known to be 1, we would use the estimate

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

where $g: \mathbb{R}^p \rightarrow \mathbb{R}^p$ 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

$$\hat{\xi} = X + \sigma S g(x)$$.
where \( c \) is a constant to be determined. Let

\[
Y = \frac{X}{\sigma}, \quad \eta = \frac{\xi}{\sigma}, \quad S^* = \frac{S}{\sigma^2}.
\]

Then, using the independence of \( S^* \) and \( Y \), from Theorem 2.1 we obtain

\[
E^{\xi, \eta}||X + cSg(X) - \xi||^2 = \sigma^2 E||Y + cS^*g(Y) - \eta||^2
\]

\[
= \sigma^2 E[p + c^2 S^*^2] ||g(Y)||^2 + 2c S^* \nabla^* \cdot g(Y)
\]

\[
= \sigma^2 E[p + n(n+2) c^2] ||g(Y)||^2 + 2nc \nabla^* \cdot g(Y)
\]

where \( \nabla^* \) is the vector of first partial derivatives with respect to \( Y \). If we choose

\[
c = \frac{1}{n+2},
\]

this becomes

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

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 \( \frac{2}{n+2} \) of the reduction in risk we would have achieved if we had known \( \sigma^2 \).

For some purposes it may be useful to have an unbiased estimate of the expected squared length of the error vector. Using formula (2.8) we have

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

where \( \nabla \) is the vector of first partial derivatives with respect to \( X \).

7. **An Identity Suggesting Approximate Confidence Sets for the Mean**

By repeated application of formula (2.8), 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 \( \xi \), 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 at a nearly arbitrary estimate. No attempt is made to study the validity of this approximation, but this problem will be considered in later papers of this series, at least in special cases.

**Lemma 1:** If \( X \) is a normally distributed real random variable with mean \( \xi \) and variance \( 1 \), then

\begin{align*}
(1) & \quad E^\xi (X-\xi) g(X) = E^\xi g'(X) \\
(2) & \quad E^\xi (X-\xi)^2 g(X) = E^\xi [g(X) + g''(X)]
\end{align*}
and

\[ E^E(x-\xi)^4 g(x) = E^E[3g(x) + 6g''(x) + g^{(iv)}(x)] , \]

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 (1) is the same as equation (2.3) of Lemma 2.1. The remaining formulas follow by repeated application of (1). Somewhat imprecisely we write \((f(x))'\) as well as \(f'(x)\) for the derivative of \( f \) at \( x \). Then, formula (1) can be rewritten as

\[ E^E[xg(x)] = E[g(x)]' . \]

By repeated application of (5) we obtain

\[ E^E[xg(x)] = E[xg(x)] = E[xg(x)]' \]

\[ = E[g(x) + xg'(x)] = E[g(x) + g''(x)] , \]

\[ E^E[xg(x)] = E[xg(x)] = E[xg(x) + (xg(x))'''] \]

\[ = E[xg(x) + g''(x) + 2g'(x)] = E[3g'(x) + g'''(x)] \]

and

\[ E^E[xg(x)] = E[xg(x)] = E[3(xg(x))' + (xg(x))''''] \]

\[ = E[3g(x) + xg'(x)] + (xg''(x) + 3g''(x))] \]

\[ = E[3g(x) + 3g''(x) + x(3g'(x) + g''(x))] \]

\[ = E[3g(x) + 6g''(x) + g^{(iv)}(x)] . \]
The following corollary will not be used here, but may be of some interest.

**Corollary 1:** If \( X \) is a normally distributed real random variable with mean \( \xi \) and variance 1, then, with the derivatives interpreted as in Lemma 1,

\[
(9) \quad \xi E^\xi g(X) = E^\xi [Xg(X) - g'(X)]
\]

\[
(10) \quad \xi^2 E^\xi g(X) = E^\xi [(X^2 - 1) g(X) - 2X g'(X) + g''(X)]
\]

\[
(11) \quad \xi^3 E^\xi g(X) = E^\xi [(X^2 - 3)X g(X) - 3(X^2 - 1) g'(X) + 3X^2 g''(X) - g'''(X)]
\]

and

\[
(12) \quad \xi^4 E^\xi g(X) = E^\xi [(X^4 - 6X^2 + 3) g(X) - 4(X^2 - 3)X g'(X)]
\]

\[
+ 6(X^2 - 1) g''(X) - 4X g'''(X) + g''''(X)]
\]

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:

\[
(13) \quad (Dg)x = g'(x)
\]

and \( T \) the operation of multiplying by \( x \):

\[
(14) \quad (Tg)x = xg(x)
\]

Then equation (1) can be rewritten

\[
(15) \quad \xi E^\xi g(X) = E^\xi [(D - T)g] (X)
\]
By induction, for \( k \) a positive integer

\[
(16) \quad \xi^k \mathbb{E}^\xi g(X) = \mathbb{E}^\xi[(D-T)^k g](X).
\]

Equation (9) is simply another form of (1) or (15). Equations (10), (11), and (12) are obtained by expanding (16) for \( k = 2, 3, \) and 4, using

\[
(17) \quad D^j T = TD^j + JD^{j-1}
\]
or

\[
(18) \quad DT^j = T^j D + JT^{j-1}
\]

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 may be stronger than those needed.

**Theorem 2:** Let \( X \) be a random p-dimensional coordinate vector, normally distributed with mean \( \xi \) and the identity as covariance matrix. Let \( g: \mathbb{R}^p \rightarrow \mathbb{R}^p \) be a twice continuously differentiable function such that

\[
(19) \quad \mathbb{E}^\xi\{||g(X)||^2 + \sum_{i,j} g_{ij}^2(X) + \sum_{i,j} g_{ij}^2(X)\} < \infty.
\]

Then

\[
(20) \quad \mathbb{E}^\xi\{||X + g(X) - \xi||^2 - (p+||g(X)||^2 + 2V' g(X))^2
\]

\[
= 2p + 4\mathbb{E}^\xi||g(X)||^2 + 2V' g(X) + \text{tr}[V g'(X)^2],
\]

where \( g' \) denotes the vector-valued function whose value is the transpose of the value of the function \( g \).
Proof: Expanding the left hand side of (20), we obtain

\[(21) \quad E^5[| |x + g(x) - \xi| |^2 - (p^+ |g(x)| |^2 + 2v' g(x))]^2 \]

\[= E^5[| |x-\xi| |^2-p + 2((x-\xi)', g(x) - v' g(x))]^2 \]

\[= E^5[| |x-\xi| |^2-p]^2 + 4[(x-\xi)', g(x) - v' g(x)]^2 \]

\[+ 4[| |x-\xi| |^2-p][(x-\xi)', g(x) - v' g(x)] \]

\[= 2p + 4E^5[((x-\xi)', g(x))^2 + (v' g(x))^2 - 2((x-\xi)', g(x)) v' g(x) \]

\[+ | |x-\xi| |^2 (x-\xi)', g(x) - | |x-\xi| |^2 v' g(x)] . \]

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:

\[(22) \quad E^5[(x-\xi)', g(x)]^2 = E^5 \sum_i \sum_j (x_i-\xi_i)(x_j-\xi_j) g_i(x) g_j(x) \]

\[= E^5 \sum_i \sum_j \frac{\partial}{\partial x_i} [(x_j-\xi_j) g_i(x) g_j(x)] \]

\[= E^5 \sum_i \sum_j \{ \delta_{ij} g_i(x) g_j(x) + (x_j-\xi_j) g_{i1}(x) g_j(x) + g_{i1}(x) g_{ij}(x) \} \]

\[= E^5 \sum_i \sum_j \{ \delta_{ij} g_i(x) g_j(x) + g_{i1}(x) g_j(x) + g_{i1}(x) g_{ij}(x) \}

\[+ g_{ij}(x) g_{i1}(x) + g_i(x) g_{i1j}(x) \}

\[= E^5[| |g(x)| |^2 + [v' g(x)]^2 + tr[v' g'(x)]^2 + 2 \sum_i \sum_j g_i(x) g_{i1j}(x)] , \]

where subscripts beyond the first denote partial derivatives. The second term in brackets on the right hand side of (21) is already in the derived form. For the third term we have
\[
(23) \quad E^\xi[(x-\xi)' g(x)] V' g(x) = E^\xi V'[V' g(x) g(x)] \\
\quad = E^\xi[V' g(x)]^2 + \sum_i \sum_j g_{ij}(x) g_{ij}(x) .
\]

For the fourth term in brackets on the right hand side of (21) we have

\[
(24) \quad E^\xi[I ||x-\xi||^2 (x-\xi)' g(x)] = E^\xi \sum_i \sum_j (x_i-\xi_i)^2 (x_j-\xi_j) g_{ij}(x) \\
\quad = E^\xi \sum_i \sum_j \frac{\partial}{\partial x_j} [(x_i-\xi_i)^2 g_{ij}(x)] \\
\quad = E^\xi \sum_i \sum_j [2 \delta_{ij} (x_i-\xi_i) g_i(x) + (x_i-\xi_i)^2 g_{ij}(x)] \\
\quad = E^\xi[2V' g(x) + ||x-\xi||^2 V' g(x)] .
\]

Thus for the combined fourth and fifth terms we have

\[
(25) \quad E^\xi[I ||x-\xi||^2 (x-\xi)' g(x) - ||x-\xi||^2 V' g(x)] \\
\quad = E^\xi[2V' g(x) + ||x-\xi||^2 V' g(x) - ||x-\xi||^2 V' g(x)] = 2E^\xi V' g(x) .
\]

Combining (21), (22), (23), and (25), we obtain

\[
(26) \quad E^\xi[I ||x + g(x) - \xi||^2 - (p+||g(x)||^2 + 2V' g(x))]^2 \\
\quad = 2p + 4E^\xi[||g(x)||^2 + 2V' g(x) + tr[V g'(x)]^2} .
\]

It seems plausible, and will be verified in a later paper, that under appropriate conditions, with \( p \) large, the random variable in brackets on the left hand side of (19) is approximately normally distributed with mean 0, and that the random variable in brackets on the right hand side is approximately constant. This suggests as confidence sets for \( \xi \) with approximate probability \( 1-\alpha \) of covering \( \xi \)
\[ S_X = \left\{ \xi : \left| |\xi - (X + g(X))|\right|^2 < p + \left| |g(X)|\right|^2 + 2V'g(X) \right. \]
\[ + c_\alpha \sqrt{2^{p+2}\left( |g(X)|^2 + 2V'g(X) + tr[Vg'(X)]^2 \right)} \left\} \]

where

\[ \phi(c_\alpha) = 1 - \alpha. \]

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

\[ P\left( \chi^2_p < p + c_\alpha \sqrt{2p} \right) = 1 - \alpha. \]

With this choice of \( c_\alpha \) and reasonable choice of \( g \), the probability that \( \xi \in S_X \) approaches \( 1 - \alpha \) as \( \xi \to \infty \).
REFERENCES


