Statistica Sinica **5**(1995), 109-127

STEIN CONFIDENCE SETS AND THE BOOTSTRAP

Rudolf Beran

University of California, Berkeley

Abstract. The Stein estimator $\hat{\xi}_S$ dominates the sample mean, under quadratic loss, in the $N(\xi, I)$ model of dimension $q \geq 3$. A Stein confidence set is a sphere of radius $\hat{d}$ centered at $\hat{\xi}_S$. The radius $\hat{d}$ is constructed to make the coverage probability converge to α as dimension q increases. This paper studies properties of Stein confidence sets for moderate to large values of q. Our main results are:

- Stein confidence sets dominate the classical confidence spheres for ξ under a geometrical risk criterion as $q \to \infty$.
- Correct bootstrap critical values for Stein confidence sets require resampling from a $N(\hat{\xi}, I)$ distribution, where $|\hat{\xi}|$ estimates $|\xi|$ well.
- Simple asymptotic or bootstrap constructions of $\hat{d}$ result in a coverage probability error of $O(q^{-1/2})$. A more sophisticated bootstrap approach reduces coverage probability error to $O(q^{-1})$. The faster rate of convergence manifests itself numerically for $q \geq 5$.

Key words and phrases: Signal, white noise, coverage probability, geometrical risk.

1. Introduction

A basic model in time-series analysis states that the observed data consists of a signal plus white noise. The goal is to estimate the signal from the data as well as possible. In Stein's (1956) formulation, we observe the random q-vector $X = (X_1, \ldots, X_q)$, which is linked to the unknown signal vector $\xi = (\xi_1, \ldots, \xi_q)$ by the model

$$X_i = \xi_i + E_i, \qquad 1 \leq i \leq q. \tag{1.1}$$

The error vector $(E_1, \ldots, E_q)$ has a standard normal distribution in R^q. The quality of an estimator $\hat{\xi} = \hat{\xi}(X)$ of the signal ξ is measured by the quadratic risk

$$R_q(\hat{\xi}, \xi) = q^{-1} E_\xi |\hat{\xi} - \xi|^2, \tag{1.2}$$

where $|\cdot|$ is Euclidean norm and E_ξ denotes expectation with respect to the distribution P_ξ of X.

This risk is the discrete-time analogue of the mean integrated squared error criterion commonly used in estimation of continuous-time signals (cf. Rice (1984), Li and Hwang (1984)). By the Cauchy-Schwarz inequality, we have

$$|\hat{\xi} - \xi|^2 = \sup\{(u'\hat{\xi} - u'\xi)^2 : u \in R^q, |u| = 1\}. \tag{1.3}$$

Thus, an estimator $\hat{\xi}$ is close to ξ in quadratic loss if and only if every normalized linear combination of $\hat{\xi}$ is close in squared error to the corresponding linear combination of ξ. This property provides a motivation for the use of risk (1.2).

Model (1.1) is completely general as regards the signal ξ, which may be any vector in R^q. More restrictive is the assumption that the scale of the white noise is known. This assumption may be relaxed if we can replicate the experiment or if we can impose conditions on the possible values of ξ. For simplicity in the asymptotics, we will retain the known scale.

Stein (1956) proved that the best unbiased estimator X is inadmissible for ξ, under quadratic risk, whenever $q \geq 3$. Methods for bettering X accept bias in return for smaller variance. One improvement is the James-Stein estimator

$$\hat{\xi}_S = [1 - (q-2)/|X|^2]X, \tag{1.4}$$

whose risk for $q \geq 3$ is

$$R_q(\hat{\xi}_S, \xi) = 1 - q^{-1}E_\xi[(q-2)^2/|X|^2] < 1. \tag{1.5}$$

The risk of $\hat{\xi}_S$ is thus strictly less than the risk of X at every ξ and achieves a minimum value of $2/q$ at $\xi = 0$ (James and Stein (1961)). A recent survey of the extensive non-asymptotic literature on shrinkage estimators is given by Brandwein and Strawderman (1990).

Stein estimation and signal extraction in time-series analysis are linked fundamentally. In a paper written from a time-series perspective, Pinsker (1980) obtained asymptotic lower bounds for the minimax mean integrated squared error incurred when estimating the mean of a continuous-time Gaussian process. Specialized to the discrete-time model (1.1), Pinsker's paper establishes two important points: First, for every finite positive c,

$$\liminf_{q\to\infty} \inf_{\hat{\xi}} \sup_{|\xi|^2 \leq qc} R_q(\hat{\xi}, \xi) \geq c/(1+c), \tag{1.6}$$

the infimum being taken over all estimators ξ. Secondly, *given* the value of c, the linear estimator $\hat{\xi}_L = cX/(1+X)$ is asymptotically minimax in that

$$\lim_{q\to\infty} \sup_{|\xi|^2 \leq qc} R_q(\hat{\xi}_L, \xi) = c/(1+c). \tag{1.7}$$

For further aspects of Pinsker's results, see Donoho and Johnstone (1994).

The simple linear estimator $\hat{\xi}_L$ is not very useful because its asymptotic minimaxity requires specifying the constant c. By contrast, the Stein estimator $\hat{\xi}_S$, which depends only on the data, is asymptotically minimax for *every* positive finite c. This follows because, if $|\xi|^2/q$ converges to finite a as q increases, then the risk of $\hat{\xi}_S$ converges to $a/(1+a)$ (see Casella and Hwang (1982)). Thus, though $\hat{\xi}_S$ is not quite admissible itself, it is very nearly minimax, when q is large, on large compact balls centered at $\xi = 0$. The unbiased estimator X lacks this minimax property because its risk equals 1 at every ξ.

Constructing good confidence sets for the signal ξ is the theme of this paper. The goal is to devise a confidence set, centered at $\hat{\xi}_S$, that has accurate coverage probability and is geometrically smaller, on average, than the classical confidence sphere centered at X. Section 2 formulates the problem technically and studies Stein confidence sets of the form

$$C(\hat{\xi}_S, \hat{d}) = \{t : |\hat{\xi}_S - t| \le \hat{d}\}. \tag{1.8}$$

The critical values $\hat{d}$ here may be obtained from asymptotic theory, as in Section 2, or from bootstrap distributions, as in Section 3. Direct asymptotic or bootstrap constructions of $\hat{d}$ result in a coverage probability error of order $O(q^{-1/2})$. A more sophisticated bootstrap construction reduces coverage probability error to $O(q^{-1})$. In the bootstrap approaches, it proves essential to resample from a $N(\hat{\xi}, I)$ distribution, where $|\hat{\xi}|$ estimates $|\xi|$ well. Resampling from $N(X, I)$ or $N(\hat{\xi}_S, I)$ does not work. For both bootstrap or asymptotic critical values, the expected maximum error of confidence sphere (1.8), viewed as a set-valued estimator of ξ, is shown to be smaller than that of the competing confidence sphere centered at X.

2. Asymptotic Stein Confidence Sets

We now seek to construct a confidence sphere for ξ,

$$C(\hat{\xi}, \hat{d}) = \{t : |\hat{\xi} - t| \le \hat{d}\}, \tag{2.1}$$

whose coverage probability $P_\xi(C \ni \xi)$ is exactly or very nearly α, whatever the true value of ξ. Here P_ξ denotes the distribution of X under model (1.1), $\hat{\xi} = \hat{\xi}(X)$ is the center of the confidence sphere and $\hat{d} = \hat{d}(X)$ is its radius. The expected geometrical error in $C(\hat{\xi}, \hat{d})$, as a set-valued estimator of ξ, will be measured by the *geometrical risk*

$$G_q(C, \xi) = q^{-1/2} E_\xi \sup_{t \in C} |t - \xi| = q^{-1/2} E_\xi |\hat{\xi} - \xi| + q^{-1/2} E_\xi \hat{d}. \tag{2.2}$$

This geometrical risk has a projection-pursuit interpretation. Let $U = \{u \in R^q: |u| = 1\}$ denote the unit sphere. By the Cauchy-Schwarz inequality, confidence set (2.1) is equivalent to the following simultaneous one-sided confidence intervals for linear combinations of ξ:

$$\begin{aligned} C(\hat{\xi}, \hat{d}) &= \{t \in R^q: \sup_{u \in U}(u't - u'\hat{\xi}) \leq \hat{d}\} \\ &= \{t \in R^q: u't \leq u'\hat{\xi} + \hat{d} \quad \forall u \in U\}. \end{aligned} \tag{2.3}$$

This is essentially Scheffé's argument for simultaneous confidence intervals.

Each one-sided confidence interval $(-\infty, u'\hat{\xi} + \hat{d}]$ overshoots the correct value $u'\xi$ by the amount $\max\{u'\hat{\xi} + \hat{d} - u'\xi, 0\}$. The maximum overshoot as u ranges over the unit sphere is

$$\begin{aligned} \sup_{u \in U} \max\{u'\hat{\xi} + \hat{d} - u'\xi, 0\} &= \sup_{u \in U} |u'(\hat{\xi} - \xi)| + \hat{d} \\ &= \sup_{t \in C} |t - \xi|. \end{aligned} \tag{2.4}$$

Thus, minimizing the geometrical risk $G_q(C, \xi)$ is the same as minimizing the expected maximum overshoot of the equivalent simultaneous one-sided confidence intervals for linear combinations of ξ.

Using expected length or expected volume to measure the geometrical size of a confidence set is a longstanding idea. Cohen and Strawderman (1973) explored decision theoretic implications of the expected length criterion. Beran and Millar (1985) analyzed geometrical risks like (2.2) for infinite dimensional parameters taking values in a normed space. By contrast, Neyman (1937) proposed minimizing $P_\xi(C \ni \xi')$ for all $\xi' \neq \xi$, subject to the coverage probability condition. Casella, Hwang and Robert (1993) treated several other criteria for set-valued estimators and gave further references.

From a technical viewpoint, the geometrical risk (2.2) extends to set-valued estimators the quadratic risk criterion that underlies Stein point estimators. In defining geometrical risk, we can replace $q^{-1/2}|t - \xi|$ by its square or any other continuous, strictly monotone function. Then, by inspection of the proof for Theorem 2.2 below, we see that the limiting risk (2.18) has immediate analogs for this entire family of geometrical risks. Thus, it is not surprising that the geometrical risk (2.2) turns out to measure a basic advantage of confidence sets based on Stein estimators.

2.1. Confidence set $C_{S,A}$

The classical confidence sphere for ξ is

$$C_C = C(X, \chi_q^{-1/2}(\alpha)), \tag{2.5}$$

where the square of $\chi_q^{-1/2}(\alpha)$ is the αth quantile of the chi-squared distribution with q degrees of freedom. Evidently, $P_\xi(C_C \ni \xi)$ is exactly α, for every ξ. By the triangular array central limit theorem, C_C is a sphere centered at X whose radius is approximately $q + (2q)^{1/2}\Phi^{-1}(\alpha)$ for large values of q (cf. Section 4). Thus,

$$\lim_{q\to\infty} \sup_{|\xi|^2 \le qc} |G_q(C_C, \xi) - 2| = 0 \tag{2.6}$$

for every positive finite c. Moreover, the approximate critical value for C_C can be found directly from the asymptotic normal distribution of the difference $q^{-1/2}\{|X - \xi|^2 - q\}$, which compares the quadratic loss of X with an unbiased estimator of its risk.

The Stein confidence sets studied in this paper are also of the form (2.1), with center $\hat{\xi} = \hat{\xi}_S$. To construct suitable critical values $\hat{d}$ for such confidence sets, we proceed by analogy with the last sentence of the previous paragraph. Consider the quantity

$$D_q(X, \xi) = q^{-1/2}\{|\hat{\xi}_S - \xi|^2 - [q - (q-2)^2/|X|^2]\}, \tag{2.7}$$

which compares the loss of $\hat{\xi}_S$ with an unbiased estimator of its risk. The next step is to approximate the distribution of $D_q(X, \xi)$ for large q. Stein (1962) introduced large q asymptotics in studying confidence sets for ξ. Theorem 2.1 below is related to ideas in the penultimate paragraph of Stein (1981). Section 3 refines the normal approximation to an Edgeworth expansion for the distribution of $D_q(X, \xi)$.

Berger (1980) explored a generalized Bayes approach to confidence sets that improve on C_C. Empirical Bayes constructions of confidence sets based on shrinkage estimators were developed and studied by Morris (1983), Casella and Hwang (1983, 1987) and others. The principal advances in this paper, described in Section 3, are sharper control of the coverage probabilities of confidence spheres centered at $\hat{\xi}_S$ and a convenient bootstrap algorithm for computing the better critical values. The remainder of this section gives pre-requisite first-order asymptotics for Stein confidence spheres.

From Theorem 3 in Stein (1981), it follows that

$$\begin{aligned} &E_\xi D_q(X, \xi) = 0 \\ &\mathrm{Var}_\xi D_q(X, \xi) = 2 + 4E_\xi[(q-2)^2/|X|^4 - (q-2)^2/(q|X|^2)]. \end{aligned} \tag{2.8}$$

By orthogonal invariance, the distribution of $D_q(X, \xi)$ depends on ξ only through $|\xi|$, and so can be written as $H_q(|\xi|^2/q)$. Let the symbol $\Rightarrow$ denote weak convergence of distributions.

Theorem 2.1. *Suppose that $\{\xi_q \in R^q\}$ is any sequence such that $|\xi_q|^2/q \to a < \infty$ as $q \to \infty$. Then*

$$H_q(|\xi_q|^2/q) \Rightarrow N(0, \sigma^2(a)), \tag{2.9}$$

where

$$\sigma^2(t) = 2 - 4t/(1+t)^2 \geq 1. \tag{2.10}$$

Moreover

$$\mathrm{Var}_{\xi_q} D_q(X, \xi_q) = \sigma^2(|\xi_q|^2/q) + O(q^{-1}). \tag{2.11}$$

The proofs of Theorem 2.1 and of all subsequent theorems are deferred to Section 4. Let

$$\hat{e} = q - (q-2)^2/|X|^2 + q^{1/2}\sigma(|\xi|^2/q)\Phi^{-1}(\alpha), \tag{2.12}$$

where Φ^{-1} is the standard normal quantile function, and let $\hat{d} = [\hat{e}]_+^{1/2}$, meaning the square root of the positive part. If we knew $|\xi|$, the confidence set $C(\hat{\xi}_S, \hat{d})$ would have asymptotic coverage probability α. Indeed, let $\{\xi_q\}$ be as in Theorem 2.1. Because $P_{\xi_q}(|\hat{\xi}_S - \xi_q| = 0) = 0$, we have

$$\begin{aligned}
P_{\xi_q}[C(\hat{\xi}_S, \hat{d}) \ni \xi_q] &= P_{\xi_q}(|\hat{\xi}_S - \xi_q|^2 \leq \hat{d}^2) \\
&= P_{\xi_q}(|\hat{\xi}_S - \xi_q|^2 \leq \hat{d}^2, \hat{e} \geq 0) + P_{\xi_q}(|\hat{\xi}_S - \xi_q|^2 = 0) \\
&= P_{\xi_q}(|\hat{\xi}_S - \xi_q|^2 \leq \hat{e}) \\
&= P_{\xi_q}[D_q(X, \xi_q) \leq \sigma(|\xi_q|^2/q)\Phi^{-1}(\alpha)].
\end{aligned} \tag{2.13}$$

By Theorem 2.1, the last probability in this display converges to α.

The uniformly minimum variance unbiased estimator of $|\xi|^2$ is $|X|^2 - q$. This and the definition of $\hat{\xi}_S$ suggest estimating $|\xi|$ by $|\hat{\xi}_{CL}|$, where

$$\hat{\xi}_{CL} = [1 - (q-2)/|X|^2]_+^{1/2} X. \tag{2.14}$$

The implied estimator of $\sigma(|\xi|^2/q)$ is then

$$\hat{\sigma}_A = \sigma(|\hat{\xi}_{CL}|^2/q), \tag{2.15}$$

the function σ being defined in (2.10). A stronger rationale for the estimator $\hat{\sigma}_A$ is the local asymptotic minimax property to be stated in Section 2.3.

The *asymptotic Stein confidence set* is defined to be $C_{S,A} = C(\hat{\xi}_S, \hat{d}_A)$, where

$$\hat{d}_A = [q - (q-2)^2/|X|^2 + q^{1/2}\hat{\sigma}_A\Phi^{-1}(\alpha)]_+^{1/2}. \tag{2.16}$$

Constructions related to (2.16) are suggested in the penultimate paragraph of Stein (1981).

Theorem 2.2. *For every finite positive* c,

$$\lim_{q\to\infty} \sup_{|\xi|^2 \leq qc} |P_\xi(C_{S,A} \ni \xi) - \alpha| = 0 \tag{2.17}$$

and

$$\lim_{q\to\infty} \sup_{|\xi|^2\le qc} |G_q(C_{S,A},\xi) - r_S(|\xi|^2/q)| = 0, \tag{2.18}$$

where

$$r_S(t) = 2(t/(1+t))^{1/2} < 2. \tag{2.19}$$

Like the classical confidence set C_C defined in (2.5), $C_{S,A}$ has correct asymptotic coverage probability α, uniformly over large compact balls about the shrinkage point $\xi = 0$. However, the geometrical risk of $C_{S,A}$ is asymptotically smaller than that of C_C, especially when ξ is near 0.

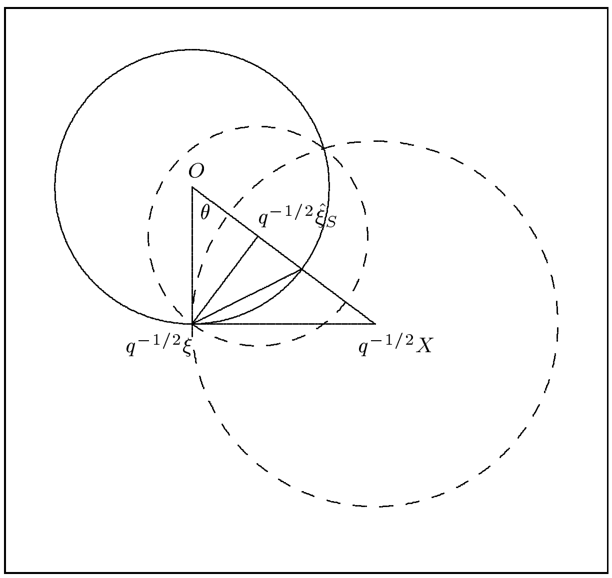


Figure 1. Asymptotic geometry of the classical confidence set C_C (larger dotted circle) and of the Stein confidence sets $C_{S,A}$, $C_{S,B}$, or $C_{S,BB}$ (smaller dotted circle).

2.2. Asymptotic geometry

The asymptotic risks of $C_{S,A}$ and of the classical confidence set C_C, given by (2.18) and (2.6) respectively, have a simple geometrical interpretation that is exhibited in Figure 1. Let $\{\xi_q \in R^q\}$ be any sequence such that $|\xi_q|^2/q$ converges to a. From model (1.1) and the weak law of large numbers, it follows that the following relations are very nearly true with high P_{ξ_q}-probability when q is large:

$$|q^{-1/2}\xi_q|^2 \approx a, \quad |q^{-1/2}X - q^{-1/2}\xi_q|^2 \approx 1, \quad |q^{-1/2}X|^2 \approx 1 + a. \tag{2.20}$$

Consequently, the triangle in Figure 1 with base vector $q^{-1/2}\xi_q$ and with hypotenuse $q^{-1/2}X_q$ is very nearly right-angled. The angle θ between $q^{-1/2}\xi_q$ and $q^{-1/2}X$ is therefore approximately determined by $\cos^2(\theta) \approx a/(1+a)$.

In seeking minimax or admissible estimators of ξ, we may consider only estimators that are equivariant under the orthogonal group. This follows from the Hunt-Stein theorem and the compactness of the orthogonal group. Any orthogonally equivariant estimator $\hat{\xi}$ has the structure $\hat{\xi}(X) = h(|X|)X$ for some real-valued measurable function h (Stein (1956)) and so lies along the vector X.

The scaled orthogonally equivariant estimator $q^{-1/2}\hat{\xi}$ that minimizes the loss $|q^{-1/2}\hat{\xi} - q^{-1/2}\xi_q|^2$ is thus the orthogonal projection of $q^{-1/2}\xi_q$ onto X. For large q, this minimizing $\hat{\xi}$ satisfies

$$q^{-1/2}\hat{\xi} \approx |q^{-1/2}\xi_q|\cos(\theta)X/|X| \approx [a/(1+a)]q^{-1/2}X \approx q^{-1/2}\hat{\xi}_S \tag{2.21}$$

with high probability for large q. This geometrical argument explains both the structure and asymptotic optimality of $\hat{\xi}_S$. By the geometry of the projection, illustrated in Figure 1, the minimized loss is asymptotically

$$q^{-1}|\hat{\xi}_S - \xi_q|^2 \approx a\sin^2(\theta) \approx a/(1+a) \tag{2.22}$$

with high probability, in agreement with the analytical evaluation by Casella and Hwang (1982).

This last calculation helps explain why the critical value of $C_{S,A}$ satisfies $q^{-1/2}\hat{d}_A \approx (a/(1+a))^{1/2}$. Indeed, if $q^{-1/2}\hat{d}_A$ were to converge in probability to a smaller or larger limit, then the asymptotic coverage probability of $C_{S,A}$ would be 0 or 1 respectively, by Figure 1. The smaller dotted circle in Figure 1 is the limiting form of $q^{-1/2}C_{S,A}$. The larger dotted circle represents the limit of $q^{-1/2}C_C$. Evidently, the asymptotic geometrical risk of $q^{-1/2}C_{S,A}$ is $r_S(a)$, in agreement with (2.18), while that of $q^{-1/2}C_C$ is 2, as in (2.6).

2.3. Asymptotic minimaxity of $\hat{\sigma}_A$

A stronger basis for the estimator $\hat{\sigma}_A$ defined in (2.15) is the following result on estimation of functions of $|\xi|^2/q$. Let u be a non-negative monotone increasing function defined on R^+, with $u(0) = 0$.

Theorem 2.3. *Suppose that $f : R^+ \to R$ is differentiable on its domain, with derivative f'. For every positive finite a,*

$$\begin{aligned} &\lim_{b\to\infty} \liminf_{q\to\infty} \inf_{\hat{f}} \sup_{||\xi|^2/q - a| \le q^{-1/2}b} E_\xi u(q^{1/2}|\hat{f} - f(|\xi|^2/q)|) \\ &\ge Eu[(2+4a)^{1/2}|f'(a)Z|] = \tau_u(a), \quad \text{say,} \end{aligned} \tag{2.23}$$

where Z has a standard normal distribution and the infimum is taken over all estimators $\hat{f}$. When u is bounded, the estimator $\hat{f}_A = f(|\hat{\xi}_{CL}|^2/q)$ attains the lower bound (2.23) *asymptotically in that*

$$\lim_{q\to\infty} \sup_{||\xi|^2/q-a|\le q^{-1/2}b} E_\xi u(q^{1/2}|\hat{f}_A - f(|\xi|^2/q)|) = \tau_u(a) \tag{2.24}$$

for every positive finite a and b.

This theorem establishes a local asymptotic minimax bound of the Hájek-LeCam type. Hasminski and Nussbaum (1984) derived a closely related bound when $f \equiv 1$. The plug-in estimator $\hat{\sigma}_A$ defined in (2.15) attains the lower bound for all estimators of $\sigma(|\xi|^2/q)$. Inequality (2.23) also implies that, for every positive finite c,

$$\liminf_{q\to\infty} \inf_{\hat{f}} \sup_{|\xi|^2\le qc} E_\xi u(q^{1/2}|\hat{f} - f(|\xi|^2/q)|) \ge \tau_u(c). \tag{2.25}$$

This minimax lower bound parallels the structure of (1.6) and is again attained asymptotically by the estimator $\hat{f}_A$ when u is bounded.

3. Bootstrap Stein Confidence Sets

Figures 2 and 3 plot the coverage probability of $C_{S,A}$ (as diamonds) versus $|\xi|^2/q$ for q between 3 and 19 and $\alpha = .90$. These coverage probabilities are estimates based on 20,000 Monte Carlo samples. Three points stand out:

- In the range $0 \le |\xi|^2/q \le 2$, the actual coverage probability varies sharply, attaining a maximum value around $|\xi|^2/q = 1$ and a minimum around $|\xi|^2/q = 2$.
- The maximum coverage probability exceeds $\alpha = .90$. For $|\xi|^2/q \ge 2$, the coverage probability stays below .90.
- The convergence of the coverage probabilities to $\alpha = .90$ as q increases is not swift.

This section offers some theoretical insight into these findings and proposes improved Stein confidence sets.

Suppose $\{\xi_q \in R^q\}$ is such that $|\xi_q|^2/q \to a$ as $q \to \infty$. By straightforward algebra (cf. the proof of Theorem 2.1),

$$\lim_{q\to\infty} q^{1/2} E_{\xi_q} D_q^3(X, \xi_q) = \mu_3(a), \tag{3.1}$$

where

$$\mu_3(t) = 8(1 - 2/(1+t))(1 - t/(1+t)^2). \tag{3.2}$$

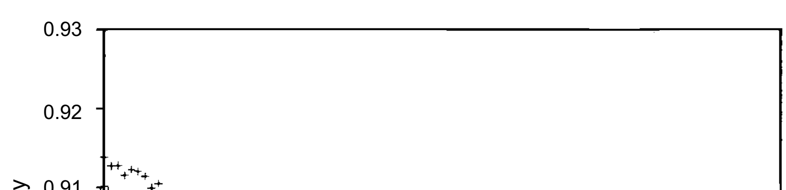
$q = 3$

NCP/dimension

$q = 5$

NCP/dimension

Figure 2. Simulated coverage probabilities of $C_{S,A}$ (diamonds), of $C_{S,B}$ (crosses), and of $C_{S,BB}$ (squares) when dimension q is 3 and 5. This Monte Carlo simulation used 20,000 normal samples, each bootstrap critical value being computed from 199 bootstrap samples. The standard error of each coverage probability plotted is thus .002.



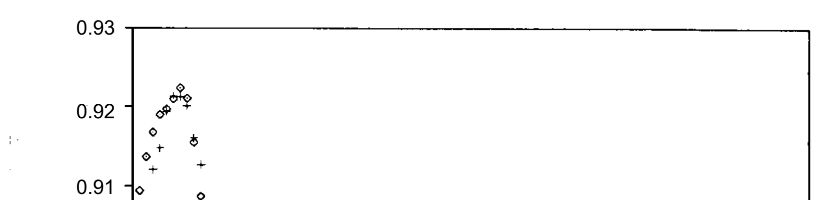
$q = 9$

NCP/dimension

$q = 19$

NCP/dimension

Figure 3. Simulated coverage probabilities of $C_{S,A}$ (diamonds), of $C_{S,B}$ (crosses), and of $C_{S,BB}$ (squares) when dimension q is 9 and 19. This Monte Carlo simulation used 20,000 normal samples, each bootstrap critical value being computed from 199 bootstrap samples. The standard error of each coverage probability plotted is thus .002.



NCP/dimension

Figure 4. The values of $\hat{\sigma}^2(|\xi|^2/q)$ (solid curve) and $\kappa_3(|\xi|^2/q)$ (dashed curve) are plotted against $|\xi|^2/q$. Both quantities affect coverage probability error of Stein confidence sets.

Let $\kappa_3(t) = \mu_3(t)/\sigma^{3/2}(t)$, for σ defined in (2.10), and let $p(x) = (1 - x^2)/6$. By the methods of Hall (1992), the cdf of $H_q(|\xi|^2/q)$ has the Edgeworth expansion

$$H_q(x, |\xi|^2/q) = \Phi[x/\sigma(|\xi|^2/q)] + q^{-1/2}\kappa_3(|\xi|^2/q)p[x/\sigma(|\xi|^2/q)] + O(q^{-1}) \quad (3.3)$$

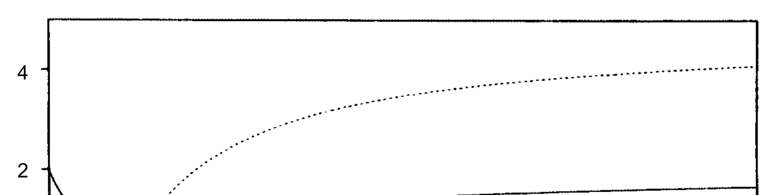
uniformly over x and the compact $\{|\xi|^2 \leq qc\}$, for every positive finite c.

The critical value $\hat{d}_A$ in (2.13) ignores the skewness term in (3.3) and estimates $\sigma(|\xi|^2/q)$ by $\hat{\sigma}_A$. Both errors introduce coverage probability errors of order $O(q^{-1/2})$. Figure 4 shows plots of $\sigma^2(|\xi|^2/q)$ and $\kappa_3(|\xi|^2/q)$ against $|\xi|^2/q$. The slope of $\sigma^2(|\xi|^2/q)$ — and hence our ability to estimate $\sigma(|\xi|^2/q)$ — changes sharply for small values of $|\xi|^2/q$. The skewness $\kappa_3(|\xi|^2/q)$ changes sign when $|\xi|^2/q = 1$ and stabilizes as $|\xi|^2/q$ increases. These factors underlie the marked fluctuations in the coverage probability of $C_{S,A}$ that occur as $|\xi|^2/q$ ranges between 0 and 2.

As a possible improvement on $C_{S,A}$, we consider, next, bootstrap critical values for Stein confidence sets. The following result clarifies how to bootstrap successfully.

Theorem 3.1. *Suppose that* $\{\xi_q \in R^q\}$ *is any sequence such that* $|\xi_q|^2/q \to a < \infty$ *as* $q \to \infty$. *Then, for* σ^2 *defined in* (2.10),

$$H_q(|\hat{\xi}_{CL}|^2/q) \Rightarrow N(0, \sigma^2(a)) \quad (3.4)$$



while

$$H_q(|X|^2/q) \Rightarrow N(0, \sigma^2(1+a)) \tag{3.5}$$

and

$$H_q(|\hat{\xi}_S|^2/q) \Rightarrow N(0, \sigma^2(a^2/(1+a))), \tag{3.6}$$

all three convergences being in P_{ξ_q}-probability.

Thus, the bootstrap estimator

$$\hat{H}_B = H_q(|\hat{\xi}_{CL}|^2/q) \tag{3.7}$$

is consistent for $H_q(|\xi|^2/q)$ while its plausible competitors in (3.5) and (3.6) are not. Let X^* be a random vector whose conditional distribution, given X, is $N(\hat{\xi}_{CL}, I)$. Then

$$\hat{H}_B = \mathcal{L}[D_q(X^*, \hat{\xi}_{CL})|X]. \tag{3.8}$$

This representation as a conditional distribution leads immediately to a simple Monte Carlo algorithm for approximating $\hat{H}_B$.

The *bootstrap Stein confidence set* is defined to be $C_{S,B} = C(\hat{\xi}_S, \hat{d}_B)$, where

$$\hat{d}_B = [q - (q-2)^2/|X|^2 + q^{1/2}\hat{H}_B^{-1}(\alpha)]_+^{1/2} \tag{3.9}$$

and $\hat{H}_B^{-1}$ denotes the quantile function of $\hat{H}_B$. It is clear from Theorem 3.1 that Theorem 2.2 holds for $C_{S,B}$ as well as $C_{S,A}$. In particular, the limiting form of $q^{-1/2}C_{S,B}$ coincides with that of $q^{-1/2}C_{S,A}$ in Figure 1.

Figures 2 and 3 plot the coverage probability of $C_{S,B}$ (as crosses) versus $|\xi|^2/q$. In the range $0 \le |\xi|^2/q \le 2$, the coverage probabilities of $C_{S,B}$ and $C_{S,A}$ are very similar. Only for $|\xi|^2/q > 2$ is $C_{S,B}$ markedly more accurate in coverage probability. The rate of convergence of coverage probability to $\alpha = .90$, as q increases, remains slow. To understand these simulation results, observe that (3.3) entails an empirical Edgeworth expansion for the cdf of $\hat{H}_B$:

$$\hat{H}_B(x) = \Phi(x/\hat{\sigma}_A) + q^{-1/2}\kappa_3(|\hat{\xi}_{CL}|^2/q)p(x/\hat{\sigma}_A) + O_p(q^{-1}). \tag{3.10}$$

Thus, $\hat{H}_B$ successfully estimates the skewness term in (3.3) with an error of $O_p(q^{-1})$, but still incurs an error of $O_p(q^{-1/2})$ in estimating the leading term on the right side of (3.3). Recall, from the discussion preceding Theorem 3.1, that the $O_p(q^{-1/2})$ error in estimating the leading term is most pronounced for small values of $|\xi|^2/q$, and that the corresponding coverage error is committed by both $C_{S,A}$ and $C_{S,B}$. On the other hand, the effect of not correcting for skewness — the additional $O(q^{-1/2})$ coverage error in $C_{S,A}$ — remains prominent for $|\xi|^2/q \ge 2$. These theoretical observations agree with the patterns found in Figures 2 and 3.

The overall coverage probability error of $C_{S,B}$, like that of $C_{S,A}$, is of order $O(q^{-1/2})$; this result forms part of Theorem 3.2. To improve the rate-of-convergence of coverage probability, it is desirable to bootstrap not $D_q(X,\xi)$ but a related quantity whose limiting distribution does not depend on the unknown parameter ξ. Lessening the dependence of the sampling distribution upon unknown parameters should enhance the accuracy with which we can estimate its quantiles. One possibility — bootstrapping $D_q(X,\xi)/\hat{\sigma}_A$ — meets the requirement just proposed but, in numerical trials, did not give accurate coverage probability for small values of q.

The author has experienced similar disappointment with bootstrapping-after -studentization in several situations where the quantity being studentized has a skewed and heavy-tailed distribution. For example, confidence sets for the mean based on nonparametric bootstrapping of the t-statistic have inaccurate coverage when the data is drawn from an exponential distribution and the sample size is modest. On the other hand, bootstrapping-after-studentization can be very successful in other contexts, such as the Behrens-Fisher problem, where the numerator has a normal distribution (cf. Beran (1988)).

Fisher's classical transformation of the sample correlation coefficient suggests an alternative strategy: bootstrapping a transformed version of $D_q(X,\xi)$ whose limiting distribution is standard normal. This approach turns out to work well in small samples, as described below . A satisfactory theoretical explanation for why transformation works better than studentization is an open problem. Current understanding of bootstrap coverage accuracy relies on Edgeworth expansions valid for q tending to infinity. These expansions need not reflect what happens for small q.

Consider the strictly monotone increasing function

$$b(t) = 2^{-1}\log[-2 + 4t + 2^{3/2}(2t^2 - 2t + 1)^{1/2}], \tag{3.11}$$

whose first derivative is

$$b'(t) = 2^{-1/2}(2t^2 - 2t + 1)^{-1/2}. \tag{3.12}$$

Let

$$E_q(X,\xi) = q^{1/2}\{b[|\hat{\xi}_S - \xi|^2/q] - b[1 - (q-2)^2/(q|X|^2)]\} \tag{3.13}$$

and let $J_q(|\xi|^2/q)$ denote the distribution of $E_q(X,\xi)$ under model (1.1). Under the conditions of Theorem 2.1, it follows from (3.13), (2.9) and (3.12) that

$$J_q(|\xi_q|^2/q) \Rightarrow N(0,1). \tag{3.14}$$

Consequently, as in Theorem 3.1, the bootstrap distribution

$$\hat{J}_B = J_q(|\hat{\xi}_{CL}|^2/q) \Rightarrow N(0,1) \tag{3.15}$$

in probability.

Further analysis (cf. Hall (1992)) shows that the cdf of $J_q(|\xi|^2/q)$ has an Edgeworth expansion analogous to (3.3) but with leading term $\Phi(x)$. Because this leading term does not depend on $|\xi|$, the cdf of $\hat{J}_B$ estimates the cdf of $J_q(|\xi|^2/q)$ with an error of $O_p(q^{-1})$. By contrast, $\hat{H}_B$ estimates $H_q(|\xi|^2/q)$ with an error of $O_p(q^{-1/2})$. For this reason, we expect that critical values based on $\hat{J}_B$ will generate Stein confidence sets with higher coverage accuracy than do critical values based on $\hat{H}_B$. The next paragraph describes these improved confidence sets and their coverage probabilities.

The *better bootstrap Stein confidence set* is defined to be $C_{S,BB} = C(\hat{\xi}_S, \hat{d}_{BB})$, where

$$\hat{d}_{BB} = [qb^{-1}\{b[1-(q-2)^2/(q|X|^2)] + q^{-1/2}\hat{J}_B^{-1}(\alpha)\}]_+^{1/2} \tag{3.16}$$

and $\hat{J}_B^{-1}$ is the quantile function of $\hat{J}_B$. In view of (3.14) and (3.15), Theorem 2.2 also holds for $C_{S,BB}$. The improvement achieved by $C_{S,BB}$ is expressed formally as follows.

Theorem 3.2. *For every finite positive c,*

$$\sup_{|\xi|^2 \le qc} |P_\xi(C_{S,BB} \ni \xi) - \alpha| = O(q^{-1}) \tag{3.17}$$

while the corresponding coverage probability errors for $C_{S,A}$ or $C_{S,B}$ are $O(q^{-1/2})$. Conclusion (2.18) of Theorem 2.2 holds for $C_{S,BB}$ as well as for $C_{S,A}$ and $C_{S,B}$.

Figures 2 and 3 show the plot of the coverage probability of $C_{S,BB}$ (as squares) versus $|\xi|^2/q$. The improvement over $C_{S,A}$ and $C_{S,B}$ is evident, especially for $|\xi|^2/q \le 2$, and suggests that the asymptotics in (3.17) take hold for $q \ge 5$. Theoretically, $C_{S,BB}$ reduces to $O(q^{-1})$ the $O(q^{-1/2})$ coverage error in both $C_{S,A}$ and $C_{S,B}$ that is caused by estimating asymptotic variance; this effect shows up most for small values of $|\xi|^2/q$, as previously discussed. At the same time, like $C_{S,B}$, the refined confidence set $C_{S,BB}$ retains the $O(q^{-1})$ coverage error due to estimating skewness.

Further refinement of the critical values for Stein confidence sets is possible by prepivoting $E_q(X,\xi)$ as follows. Let $J_q(\cdot, |\xi|^2/q)$ denote the cumulative distribution function of $E_q(X,\xi)$ under the model P_ξ. Construct a Stein confidence set for ξ by referring $F_q(X,\xi) = J_q[E_q(X,\xi), |\hat{\xi}_{CL}|^2/q]$ to the αth quantile of its bootstrap distribution. As in (3.7), this bootstrap distribution is $\mathcal{L}[F_q(X^*, \hat{\xi}_{CL})|X]$. The coverage probability of this new Stein confidence set will differ from the desired α by $O(q^{-3/2})$ as q increases. A heuristic argument for this conclusion and a double bootstrap Monte Carlo algorithm for approximating the confidence set follow the pattern in Beran (1988). It is not known whether this further asymptotic refinement in the critical value of Stein confidence sets is beneficial for small or moderate values of q.

4. Proofs

The following lemma plays an important role in our derivations. Let

$$W_q(X,\xi) = (q^{-1/2}[|X-\xi|^2 - q],\ q^{-1/2}\xi'(X-\xi)). \tag{4.1}$$

Lemma 4.1. *Suppose that $\{\xi_q \in R^q\}$ is any sequence such that $|\xi_q|^2/q \to a < \infty$ as $q \to \infty$. Then, under P_{ξ_q},*

$$W_q(X,\xi_q) \Rightarrow (2^{1/2}Z_1, a^{1/2}Z_2), \tag{4.2}$$

where Z_1, Z_2 are independent standard normal random variables.

Proof. By orthogonal invariance, the distribution of $W_q(X,\xi_q)$ depends on ξ_q only through $|\xi_q|$. Without loss of generality, take each component of ξ_q to be $q^{-1/2}|\xi_q|$. The weak convergence (4.2) now follows from model (1.1) and the Lindeberg-Feller central limit theorem.

We now prove the theorems and related remarks in Sections 2 and 3.

Proof of (2.6). Let $\{\xi_q\}$ be as in Lemma (4.1). From (2.2) and (2.5),

$$G_q(C_C,\xi_q) = q^{-1/2}E_{\xi_q}|X-\xi_q| + q^{-1/2}\chi_q^{-1}(\alpha). \tag{4.3}$$

By Lemma 4.1 and Polya's theorem,

$$q^{-1/2}\chi_q^{-1}(\alpha) = q^{-1/2}[q + (2q)^{1/2}\Phi^{-1}(\alpha) + o(q^{1/2})]^{1/2} = 1 + O(q^{-1/2}). \tag{4.4}$$

The first term on the right side of (4.3) also converges to 1, because $q^{-1/2}|X - \xi_q| \to 1$ in probability and $q^{-1}E_{\xi_q}|X-\xi_q|^2 = 1$. Assertion (2.6) follows.

Proof of Theorem 2.1. Write $p = q-2$. Then

$$q^{1/2}D_q(X,\xi_q) = |(1-p/|X|^2)X - \xi_q|^2 - (p - p^2/|X|^2) - 2. \tag{4.5}$$

Observe that

$$\begin{aligned}|(1-p/|X|^2)X - \xi_q|^2 = {} & (1-p/|X|^2)^2|X-\xi_q|^2 + p^2|\xi_q|^2/|X|^4 \\ & - 2(p/|X|^2)(1-p/|X|^2)\xi_q'(X-\xi_q)\end{aligned} \tag{4.6}$$

and

$$p - p^2/|X|^2 = p[(1-p/|X|^2)^2 + (p^2/|X|^4)(|X|^2/p - 1)] \tag{4.7}$$

and

$$|X|^2 = |X - \xi_q|^2 + |\xi_q|^2 + 2\xi_q'(X - \xi_q). \tag{4.8}$$

Substituting the last three expressions into (4.5) yields

$$q^{1/2} D_q(X, \xi_q) = (1 - 2p/|X|^2)(|X - \xi_q|^2 - p) - (2p/|X|^2)\xi_q'(X - \xi_q) - 2. \tag{4.9}$$

The weak convergence (2.9) follows from (4.9) and Lemma 4.1.

Stein's variance formula (2.8), a Taylor expansion, and Lemma 4.1 imply

$$\begin{aligned} \mathrm{Var}_{\xi_q} D_q(X, \xi_q) &= 2 + 4[(\frac{q}{q + |\xi_q|^2})^2 - \frac{q}{q + |\xi_q|^2}] + O(q^{-1}) \\ &= \sigma^2(|\xi_q|^2/q) + O(q^{-1}), \end{aligned} \tag{4.10}$$

as in (2.11).

Proof of Theorem 2.2. Suppose $|\xi_q|^2/q \to a < \infty$. Since $|\hat{\xi}_{CL}|^2 = [|X|^2 - q - 2]_+$, it follows from (4.8) and the weak law of large numbers that

$$q^{-1}|\hat{\xi}_{CL}|^2 \to a \quad \text{in probability.} \tag{4.11}$$

Thus, by Theorem 2.1, $\hat{\sigma}_A$ converges to $\sigma(a)$ in probability and the distribution of $q^{-1/2} D_q(X, \xi_q)/\hat{\sigma}_A$ converges weakly to the standard normal distribution. This and reasoning like that in (2.13) imply (2.17).

From (2.6) and (2.2),

$$G_q(C_{S,A}, \xi_q) = q^{-1/2} E_{\xi_q} |\hat{\xi}_S - \xi_q| + q^{-1/2} E_{\xi_q} \hat{d}_A. \tag{4.12}$$

By the weak law of large numbers, $q^{-1/2}|\hat{\xi}_S - \xi_q| \to r_S(a)/2$ and $q^{-1/2}\hat{d}_A \to r_S(a)/2$, both in probability. Moreover, from (1.5), $q^{-1} E_{\xi_q} |\hat{\xi}_S - \xi_q|^2 \to r_S^2(a)/4$. Dominated convergence reasoning completes the proof of (2.18).

Lemma 4.2. *Suppose that $\{\xi_q \in R^q\}$ is any sequence such that $|\xi_q|^2/q \to a < \infty$ as $q \to \infty$. Then, under P_{ξ_q},*

$$q^{-1/2}(|\hat{\xi}_{CL}|^2 - |\xi_q|^2) \Rightarrow N(0, 2 + 4a). \tag{4.13}$$

Proof. By (4.8),

$$q^{-1/2}(|X|^2 - q - |\xi_q|^2) = q^{-1/2}(|X - \xi_q|^2 - q) + 2q^{-1/2}\xi_q'(X - \xi_q). \tag{4.14}$$

The desired weak convergence follows from the definition (2.14) of $\hat{\xi}_{CL}$ and from Lemma 4.1.

Proof of Theorem 2.3. The lower asymptotic minimax bound (2.23) follows by an extension of the argument for Theorem 3.2 in Beran (1994). Because of Lemma 4.2, the limiting distribution of $q^{-1/2}[f(|\hat{\xi}_{CL}|^2/q) - f(|\xi_q|^2/q)]$ under P_{ξ_q} is $N(0, (2+4a)[f'(a)]^2)$. Because the function u is continuous almost everywhere, conclusion (2.24) follows.

Proof of (3.1). This follows from (4.9), the first three moments of the chi-squared and normal distributions, and a dominated convergence argument.

Proof of Theorem 3.1. Suppose $|\xi_q|^2/q \to a < \infty$ as $q \to \infty$. Then, by Lemma 4.1 and (4.8), $|\hat{\xi}_{CL}|^2/q \to a$, $|X|^2/q \to a$ and $|\hat{\xi}_S|^2/q \to a^2/(1+a)$, all in P_{ξ_q}-probability. These convergences and Theorem 2.1 imply Theorem 3.1.

Proof of Theorem 3.2. A relatively short heuristic argument for (3.17) and for the analogous assertions concerning $C_{S,A}$ or $C_{S,B}$ parallels the discussion in Beran (1988), Section 3. The starting point is the Edgeworth expansion (3.3). A more rigorous argument can be constructed by the methods in Hall (1992).

Acknowledgements

This research was supported in part by National Science Foundation grant DMS 9224868. Conversations with Dave Donoho and Michael Nussbaum were particularly instructive.

References

Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. J. Amer. Statist. Assoc. **83**, 687-697.

Beran, R. (1994). Stein estimation in high dimensions: A retrospective. Unpublished preprint.

Beran, R. and Millar, P. W. (1985). Asymptotic theory of confidence sets. In Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer (Edited by L. M. LeCam and R. A. Olshen), Vol. II, 865-887. Wadsworth, Belmont, California.

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

Brandwein, A. C. and Strawderman, W. E. (1990). Stein estimation: The spherically symmetric case. Statist. Sci. **5**, 356-369.

Casella, G. and Hwang, J. T. (1982). Limit expressions for the risk of James-Stein estimators. Canad. J. Statist. **10**, 305-309.

Casella, G. and Hwang, J. T. (1983). Empirical Bayes confidence sets for the mean of a multivariate normal distribution. J. Amer. Statist. Assoc. **78**, 688-698.

Casella, G. and Hwang, J. T. (1987). Employing vague prior information in the construction of confidence sets. J. Multivariate Anal. **21**, 79-104.

Casella, G., Hwang, J. T. G. and Robert, C. (1993). A paradox in decision-theoretic interval estimation. Statistica Sinica **3**, 141-155.

Cohen, A. and Strawderman, W. E. (1973). Admissibility implications for different criteria in confidence estimation. Ann. Statist. **1**, 363-366.

Donoho, D. L. and Johnstone, I. M. (1994). Minimax risk over l_q-balls for l_p-error. Probab. Theory Related Fields **99**, 277-303.

Hall, P. (1992). The Bootstrap and Edgeworth Expansion. Springer-Verlag, New York.

Hasminski, R. Z. and Nussbaum, M. (1984). An asymptotic minimax bound in a regression problem with an increasing number of nuisance parameters. Proc. 3rd Prague Symp. Asymptotic Statistics (Edited by P. Mandl and M. Hŭsková), 275-283. Elsevier, New York.

James, W. and Stein C. (1961). Estimation with quadratic loss. Proc. 4th Berkeley Symp. Math. Statist. Probab. **1**, 361-380. University California Press.

Li, K.-C. and Hwang, J. T. (1984). The data-smoothing aspect of Stein estimates. Ann. Statist. **12**, 887-897.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. J. Amer. Statist. Assoc. **78**, 47-55.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. Philos. Trans. Roy. Soc. Ser.A **236**, 333-380.

Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian white noise. Problems Inform. Transmission **16**, 120-133.

Rice, J. (1984). Bandwidth choice for nonparametric regression. Ann. Statist. **12**, 1215-1230.

Stein, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proc. 3rd Berkeley Symp. Math. Statist. Probab. **1**, 197-206. University California Press.

Stein, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution (with discussion). J. Roy. Statist. Soc. Ser.B **24**, 265-296.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. Ann. Statist. **9**, 1135-1151.

Department of Statistics, University of California, Berkeley, CA 94720-3860, U.S.A.

(Received April 1993; accepted September 1994)