Bootstrap Variable-Selection and Confidence Sets

Rudolf Beran
University of California, Berkeley

This paper analyzes estimation by bootstrap variable-selection in a simple Gaussian model where the dimension of the unknown parameter may exceed that of the data. A naive use of the bootstrap in this problem produces risk estimators for candidate variable-selections that have a strong upward bias. Resampling from a less overfitted model removes the bias and leads to bootstrap variable-selections that minimize risk asymptotically. A related bootstrap technique generates confidence sets that are centered at the best bootstrap variable-selection and have two further properties: the asymptotic coverage probability for the unknown parameter is as desired; and the confidence set is geometrically smaller than a classical competitor. The results suggest a possible approach to confidence sets in other inverse problems where a regularization technique is used.

Key words and phrases. Coverage probability, geometric loss, C_p-estimator.

1. Introduction

Certain statistical estimation problems, such as curve estimation, signal recovery, or image reconstruction, share two distinctive features: the dimension of the parameter space exceeds that of the data; and each component of the unknown parameter may be important. In such problems, ordinary least squares or maximum likelihood estimation typically overfits the model. One general approach to estimation in such problems has three stages: First, devise a promising class of candidate estimators, such as penalized maximum likelihood estimators corresponding to a family of penalty functions or Bayes estimators generated by a family of prior distributions. This step is sometimes called using a regularization technique. Second, estimate the risk of each candidate estimator. Third, use the candidate estimator with smallest estimated risk.

Largely unresolved to date is the question of constructing accurate confidence sets based on such adaptive, regularized estimators. Even obtaining reliable estimators of risk can be difficult. This paper treats both matters

in the following problem, which is relatively simple to analyze explicitly, yet sufficiently general to indicate potential directions for other problems that involve a regularization technique. Suppose that X_n is an observation on a discretized signal ξ that is measured with error at n time points. The errors are independent, identically distributed, Gaussian random variables with means zero. Thus, X_n is a random vector whose distribution is $N(\xi_n, \sigma_n^2 I_n)$. Both ξ_n and σ_n^2 are unknown. The problem is to estimate the signal ξ_n.

The integrated squared error of an estimator $\hat{\xi_n}$ is

$$L_n(\hat{\xi_n}, \xi_n) = n^{-1}|\hat{\xi_n} - \xi_n|^2, \tag{1.1}$$

where $|\cdot|$ is Euclidean norm. Under this loss, Stein (1956) showed that X_n, the maximum likelihood or least squares estimator of ξ_n, is inadmissible for $n \geq 3$. Better estimators for ξ_n include the James-Stein (1961) estimator, locally smoothed estimators such as the kernel variety treated by Rice (1984), and variable-selection estimators, to be described in the next paragraph. Each of these improved estimators accepts some bias in return for a greater reduction in variance.

A variable-selection approach to estimating ξ_n consists of three steps: first, transform X_n orthogonally to $X_n' = OX_n$; second, replace selected components of X_n' with zero; and third, apply the inverse rotation O^{-1} to the outcome of step two. The vector generated by such a process will be called a *variable-selection* estimator of ξ_n.

How shall we choose the orthogonal matrix O? Ideally, the components of the rotated mean vector $O\xi_n$ would be either very large or very small relative to measurement error. The nature of the experiment that generated X_n may suggest that O be a finite Fourier transform, or an analysis of variance transform, or an orthogonal polynomial transform, or a wavelet transform. Important though it is, we will not deal further, in this paper, with the choice of O.

Having rotated X_n, how shall we choose which components of X_n' to zero out? Thereafter, how shall we construct, around the variable-selection estimator, an accurate confidence set for ξ_n? A plausible answer is to compare candidate variable-selections through their bootstrap risks; and then bootstrap the empirically best candidate estimator to obtain a confidence set for ξ_n. Efron and Tibshirani (1993, Chapter 17) discussed simple bootstrap estimators of mean squared prediction error. However, Freedman et

al. (1988) and Breiman (1992) showed that simple bootstrap estimators of mean-squared prediction error can be untrustworthy for variable-selection.

This paper treats variable-selection for estimation rather than prediction and allows the dimension of the unknown parameter to increase with sample size n. The second point is very important. A stronger model assumption used by Speed and Yu (1993) and others—that the dimension of the parameter space is fixed for all n—restricts the possible bias induced by candidate variable-selections. In such restricted models, variable-selection by C_p does not choose well. On the other hand, C_p can be asymptotically correct when the dimension of the parameter space increases quickly with n and the selection class is not too large (cf. Section 2). Rice (1984, Section 3) and Speed and Yu (1993, Section 4) discuss other instances and aspects of this phenomenon.

Section 2 of this paper proves for our estimation problem that naive bootstrapping—resampling from a $N(X_n, \hat{\sigma_n}^2 I_n)$ model, where $\hat{\sigma_n}^2$ estimates σ_n^2—yields upwardly biased risk estimators for candidate variable-selections. However, resampling from a $N(\tilde{\xi}_n, \hat{\sigma_n}^2)$ distribution, where $\tilde{\xi}_n$ is obtained by suitably shrinking some of the components of X_n' toward zero, corrects the bias and generates a good bootstrap variable-selection estimator $\hat{\xi}_{n,B}$ for ξ_n. Using a related shrinkage bootstrap, Section 3 then constructs confidence sets centered at $\hat{\xi}_{n,B}$ that have correct asymptotic coverage probability for ξ_n and small geometrical error. Here as well, two plausible but naive bootstrap algorithms give wrong answers.

2. Bootstrap Selection Estimators

This section proposes bootstrap selection estimators for ξ_n and analyzes their asymptotic losses (which equal the asymptotic risks). The choice of bootstrap algorithm proves critical to the success of bootstrap selection. Naive bootstrapping does not work.

The signal vector $X_n = (X_{n,1}, \ldots, X_{n,n})'$ has a $N(\xi_n, \sigma_n^2 I_n)$ distribution on R^n. For brevity, write $\theta_n = (\xi_n, \sigma_n^2)$ and let $P_{\theta,n}$ denote the above normal distribution. Because the estimation problem is invariant under rotation of the coordinate system, we will simplify notation by assuming, without any loss of generality, that the orthogonal matrix O is the identity matrix. Then, the variable selection is done directly on the components of X_n. Consider

candidate estimators for ξ_n that have the form

$$\hat{\xi_n}(A) = (a_{n,1}(A)X_{n,1}, \ldots, a_{n,n}(A)X_{n,n})', \tag{2.1}$$

where A ranges over subsets of $[0,1]$ and $a_{n,i}(A) = 1$ if $i/(n+1) \in A$ and vanishes otherwise. The goal is to choose A, on the basis of the data X_n, so as to minimize, at least asymptotically, the loss of the corresponding candidate estimator $\hat{\xi_n}(A)$.

Success of this formulation of variable selection appears to require restrictions on the possible values of A. In the paper, we assume that A is the union of m ordered closed intervals:

$$A = \bigcup_{i=1}^{m} [t_{2i-1}, t_{2i}], \tag{2.2}$$

where $0 \leq t_1 \leq \ldots \leq t_{2m} \leq 1$ and m is fixed. The pseudo-distance between two such sets A and B is defined to be

$$d(A,B) = \mu(A \triangle B), \tag{2.3}$$

where μ is Lebesgue measure. After forming equivalence classes, the collection $\mathcal{S}(m)$ of all subsets having the form (2.2) is a compact metric space under d.

Let $\mathcal{A}$ be a compact subset of $\mathcal{S}(m)$, possibly $\mathcal{S}(m)$ itself, that contains the unit interval $[0,1]$ as an element. Consider the candidate estimators $\hat{\xi}_n(A)$ that are generated as A ranges over $\mathcal{A}$. Since $[0,1]$ is a element of $\mathcal{A}$, the unbiased estimator X_n is among these candidate estimators. Let A^c be the complement of A in $[0,1]$. The quadratic loss of $\hat{\xi}_n(A)$ is then

$$\begin{aligned} L_n(\hat{\xi}_n(A), \xi_n) &= n^{-1} |\hat{\xi}_n(A) - \xi_n|^2 \\ &= n^{-1} \sum_{i/(n+1) \in A} (X_{n,i} - \xi_{n,i})^2 + \nu_n(A^c), \end{aligned} \tag{2.4}$$

where ν_n is the non-negative measure defined by

$$\nu_n(A) = n^{-1} \sum_{i/(n+1) \in A} \xi_{n,i}^2. \tag{2.5}$$

Estimators of this loss or of the associated risk are naturally phrased in terms of the discrete uniform measure

$$\mu_n(A) = n^{-1} \sum_{i/(n+1) \in A} 1 \tag{2.6}$$

and the empirical measure

$$\hat{\lambda}_n(A) = n^{-1} \sum_{i/(n+1)\in A} X_{n,i}^2. \tag{2.7}$$

Consider the following two bootstrap risk estimators:

Naive bootstrap. Suppose $\hat{\sigma}_n^2$ is a consistent estimator of σ_n^2, such as the variance estimators to be discussed in Section 3. Let X_n^* be a random vector such that the conditional distribution of X_n^* given X_n is $N(X_n, \hat{\sigma}_n^2 I_n)$. Let $\xi_n^*(A)$ denote the recalculation from X_n^* of the candidate estimator $\hat{\xi}_n(A)$. Let E_* denote expectation with respect to the conditional distribution of X_n^* given X_n. The *naive bootstrap* risk estimator produced by the scheme is

$$\begin{aligned} \hat{R}_{n,N}(A, \hat{\sigma}_n^2) &= E_* L_n(\xi_n^*(A), X_n) \\ &= E_*[n^{-1} \sum_{i/(n+1)\in A} (X_{n,i}^* - X_{n,i})^2 + n^{-1} \sum_{i/(n+1)\in A^c} X_{n,i}^2] \\ &= \hat{\sigma}_n^2 \mu_n(A) + \hat{\lambda}_n(A^c). \end{aligned} \tag{2.8}$$

Unfortunately, if ν_n converges weakly to ν and σ_n^2 converges to σ^2 as n increases, then $\hat{R}_{n,N}(A)$ converges in probability to

$$\bar{\rho}(A) = \sigma^2 + \nu(A^c). \tag{2.9}$$

The actual asymptotic loss or risk of $\hat{\xi}_n(A)$ is

$$\rho(A) = \sigma^2 \mu(A) + \nu(A^c), \tag{2.10}$$

where μ is Lebesgue measure. Theorem 2.1 below gives details. The upward asymptotic bias in $\hat{R}_{n,N}(A, \hat{\sigma}_n^2)$ renders it useless for selection among the candidate estimators.

Shrink bootstrap. Let $[\,\cdot\,]_+$ denote the positive-part function. The modified estimator

$$\hat{R}_{n,B}(A, \hat{\sigma}_n^2) = \hat{\sigma}_n^2 \mu_n(A) + [\hat{\lambda}_n(A^c) - \hat{\sigma}_n^2 \mu_n(A^c)]_+ \tag{2.11}$$

corrects the asymptotic bias in $\hat{R}_{n,N}$ and converges in probability to $\rho(A)$, the correct asymptotic loss of $\hat{\xi}_n(A)$; see Theorem 2.1. Moreover, the risk estimator $\hat{R}_{n,B}(A, \hat{\sigma}_n^2)$ can also be viewed as a bootstrap estimator:

Let

$$\hat{s}_n(A^c) = [1 - \hat{\sigma}_n^2 \mu_n(A^c)/\hat{\lambda}_n(A^c)]_+ \tag{2.12}$$

and define $\tilde{\xi}_n(A) = (\tilde{\xi}_{n,1}(A), \ldots, \tilde{\xi}_{n,n}(A))'$ by

$$\tilde{\xi}_{n,i}(A) = \begin{cases} X_{n,i} & \text{if } i/(n+1) \in A \\ \hat{s}_n^{1/2}(A^c) X_{n,i} & \text{if } i/(n+1) \in A^c \end{cases} . \tag{2.13}$$

Let X_n^* now be a random vector such that the conditional distribution of X_n^* given X_n is $N(\tilde{\xi}_n(A), \hat{\sigma}_n^2 I_n)$. As before, let $\xi_n^*(A)$ denote the recalculation from X_n^* of the candidate estimator $\hat{\xi}_n(A)$. Now the bootstrap risk is

$$\begin{aligned} E_* L_n(\xi_n^*(A), \tilde{\xi}_n) &= E_*[n^{-1} \sum_{i/(n+1) \in A} (X_{n,i}^* - X_{n,i}^2) + n^{-1} \sum_{i/(n+1) \in A^c} \hat{s}_n(A^c) X_{n,i}^2] \\ &= \hat{\sigma}_n^2 \mu_n(A) + \hat{s}_n(A^c) \hat{\lambda}_n(A^c) \qquad (2.14) \\ &= \hat{R}_{n,B}(A, \hat{\sigma}_n^2). \end{aligned}$$

The *shrink bootstrap* method just described has two notable features: It depends on the candidate set A; and it shrinks some, but not all, of the components of X_n towards the origin. In defining $\tilde{\xi}_n(A)$, we could shrink as well the components of X_n for which $i/(n+1) \in A$ without changing the final evaluation in (2.14). In this sense, $\hat{R}_{n,B}(A, \hat{\sigma}_n^2)$ is the bootstrap risk generated by a family of shrink bootstrap algorithms. The shrinkage factor in (2.13) corrects the overfitting of ξ_n that occurs in the naive bootstrap.

The idea of bootstrap variable selection is to choose the candidate estimator whose estimated loss is smallest. Thus, let $\hat{A}_{n,B}$ be any set in $\mathcal{A}$ such that

$$\hat{R}_{n,B}(\hat{A}_{n,B}, \hat{\sigma}_n^2) = \min_{A \in \mathcal{A}} \hat{R}_{n,B}(A, \hat{\sigma}_n^2). \tag{2.15}$$

The minimum is achieved because, for each n, $\hat{R}_{n,B}(\cdot, \hat{\sigma}_n^2)$ has a finite number of possible values. We will call

$$\hat{\xi}_{n,B} = \hat{\xi}_n(\hat{A}_{n,B}) \tag{2.16}$$

a *bootstrap selection* estimator generated by the candidate estimators $\{\hat{\xi}_n(A) : A \in \mathcal{A}\}$.

Let $\|\cdot\|_{\mathcal{A}}$ denote supremum norm taken over all sets $A \in \mathcal{A}$. To study the locally uniform convergences of $\hat{R}_{n,N}$ and $\hat{R}_{n,B}$, we introduce two conditions:

C1. $\mathcal{A}$ is a compact subset of $\mathcal{S}(m)$, in the metric d, that contains $[0,1]$ as an element. The sequence $\{\theta_n = (\xi_n, \sigma_n^2) : n \geq 1\}$ is such that

$$\lim_{n\to\infty} \|\nu_n - \nu\|_{\mathcal{A}} = 0, \quad \lim_{n\to\infty} \sigma_n^2 = \sigma^2 \tag{2.17}$$

for some bounded, d-continuous, non-negative measure ν on $\mathcal{A}$ and some finite positive σ^2.

C2. For every sequence $\{\theta_n : n \geq 1\}$ that satisfies condition C1,

$$\operatorname*{plim}_{n\to\infty} \hat{\sigma}_n^2 = \sigma^2. \tag{2.18}$$

Here plim stands for the limit in $P_{\theta_n,n}$-probability.

Theorem 2.1 *Suppose that conditions C1 and C2 hold. Then, for every positive* ϵ,

$$\operatorname*{plim}_{n\to\infty} \|\hat{R}_{n,N}(\cdot, \hat{\sigma}_n^2) - \bar{\rho}\|_{\mathcal{A}} = 0 \tag{2.19}$$

where $\bar{\rho}$ *is defined in (2.9). By contrast,*

$$\begin{aligned} \operatorname*{plim}_{n\to\infty} \|L_n(\hat{\xi}_n(\cdot), \xi_n) - \rho\|_{\mathcal{A}} &= 0 \\ \operatorname*{plim}_{n\to\infty} \|\hat{R}_{n,B}(\cdot, \hat{\sigma}_n^2) - \rho\|_{\mathcal{A}} &= 0 \end{aligned} \tag{2.20}$$

where ρ *is defined in (2.10). Consequently,*

$$\operatorname*{plim}_{n\to\infty} L_n(\hat{\xi}_{n,B}, \xi_n) = \min_{A\in\mathcal{A}} \rho(A). \tag{2.21}$$

If the limiting loss ρ *has a unique minimizer* $A_0 \in \mathcal{A}$, *then*

$$\operatorname*{plim}_{n\to\infty} d(\hat{A}_{n,B}, A_0) = 0. \tag{2.22}$$

The theorem proof is in Section 4. By equations (2.20) and (2.21), the limiting loss of $\hat{\xi}_{n,B}$ coincides with the limiting loss of the unrealizable candidate estimator that minimizes loss over all selections A in $\mathcal{A}$. In this sense, the bootstrap selection estimator $\hat{\xi}_{n,B}$ is asymptotically optimal. Because $[0,1]$ is an element of $\mathcal{A}$,

$$\min_{A\in\mathcal{A}} \rho(A) \leq \rho([0,1]) = \sigma^2 \tag{2.23}$$

with equality only in special circumstances (e.g. $\nu = \sigma^2\mu$ or $\mu(A_0) = 1$, $\nu(A_0^c) = 0$). Thus, $\hat{\xi}_{n,B}$ asymptotically dominates the unbiased estimator X_n.

An alternative to the shrink bootstrap risk estimator $\hat{R}_{n,B}$ replaces the positive-part function in (2.11) with the identity function. The result is the risk or loss estimator

$$\begin{aligned} \hat{R}_{n,C}(A, \hat{\sigma}_n^2) &= \hat{\sigma}_n^2[\mu_n(A) - \mu_n(A^c)] + \hat{\lambda}_n(A^c) \\ &= \hat{\lambda}_n(A^c) + \hat{\sigma}_n^2[2\mu_n(A) - 1]. \end{aligned} \tag{2.24}$$

Unlike $\hat{R}_{n,B}$, this risk estimator can assume negative values.

Let $\hat{A}_{n,C}$ be any value of $A \in \mathcal{A}$ that minimizes $\hat{R}_{n,C}(A, \hat{\sigma}_n^2)$. We will call $\hat{\xi}_{n,C} = \hat{\xi}_n(\hat{A}_{n,C})$ a *C_p-estimator* generated by the candidate estimators $\{\hat{\xi}_n(A) : A \in \mathcal{A}\}$. This terminology recognizes the analogy between (2.24) and risk estimators discussed by Mallows (1973) in a different context. Conclusions (2.20), (2.21) and (2.22) remain valid when $\hat{R}_{n,B}$, $\hat{A}_{n,B}, \hat{\xi}_{n,B}$ are replaced by $\hat{R}_{n,C}$, $\hat{A}_{n,C}$, $\hat{\xi}_{n,C}$ respectively.

Other variable selection criteria, such as Akaike's (1974) AIC, Shibata's (1981) method, and several competitors discussed by Rice (1984, Section 3), Speed and Yu (1993, Section 4) might also be used to choose A. Under Conditions C1 and C2, these methods do not minimize asymptotic loss in the sense of (2.21).

3. Bootstrap Confidence Sets

A confidence ball for ξ_n, centered at an estimator $\hat{\xi}_n$ and having radius $\hat{d}_n$, is

$$C_n(\hat{\xi}_n, \hat{d}_n) = \{t \in R^k : |\hat{\xi}_n - t| \le \hat{d}_n\}. \tag{3.1}$$

This section studies confidence balls centered at the bootstrap selection estimator $\hat{\xi}_{n,B}$. The first goal is to devise a bootstrap radius $\hat{d}_{n,B}$ such that the coverage probability $P_{\theta,n}[C_n(\hat{\xi}_{n,B}, \hat{d}_{n,B}) \ni \xi_n]$ converges to α as n increases. The second goal is to determine the *geometric loss* of $C_n(\hat{\xi}_n, \hat{d}_n)$ for various choices of $(\hat{\xi}_n, \hat{d}_n)$:

$$\begin{aligned} GL_n(C_n, \xi_n) &= n^{-1/2} \sup_{t \in C_n} |t - \xi_n| \\ &= n^{-1/2}|\hat{\xi}_n - \xi| + n^{-1/2}\hat{d}_n. \end{aligned} \tag{3.2}$$

Geometric loss measures the error of $C_n(\hat{\xi}_n, \hat{d}_n)$ as a set-valued estimator of ξ_n. It has a projection-pursuit interpretation that stems from the identity $|x| = \sup\{u'x: |u| = 1\}$.

Both the definition of $\hat{\xi}_{n,B}$ and the construction of confidence balls centered at $\hat{\xi}_{n,B}$ require a good estimator of σ_n^2. One possibility, used in Rice (1984), is

$$\hat{\sigma}_{n,1}^2 = [2(n-1)]^{-1} \sum_{i=2}^{n} (X_{n,i} - X_{n,i-1})^2. \tag{3.3}$$

The consistency or asymptotic normality of $\hat{\sigma}_{n,1}^2$ requires that the first-order squared differences $\{(\xi_{n,i} - \xi_{n,i-1})^2\}$ be sufficiently small, in a sense that Condition D1 below makes precise.

A second estimator of σ_n^2 works under the assumption that ξ_n lies in a subspace of dimension $n' < n$. Suppose that n' is the integer part of cn, where c is a fraction strictly between 0 and 1. By making an appropriate orthogonal transformation, assume without loss of generality that $X_n = (X_{n'}, Y_{n-n'})$, where $X_{n'}$ has a $N(\xi_{n'}, \sigma_n^2 I_{n'})$ distribution in n' dimensions, $Y_{n-n'}$ has a $N(0, \sigma_n^2 I_{n-n'})$ distribution in $n - n'$ dimensions, and $X_{n'}$, $Y_{n-n'}$ are independent. In this canonical formulation, a bootstrap selection estimator of $\xi_{n'}$ can be formed from $X_{n'}$ and the variance estimator

$$\hat{\sigma}_{n,2}^2 = (n - n')^{-1} |Y_{n-n'}|^2. \tag{3.4}$$

The distribution of $(n - n')\hat{\sigma}_{n,2}^2/\sigma_n^2$ is chi-squared with $n - n'$ degrees of freedom.

The essential features of $\hat{\sigma}_{n,1}^2$ and $\hat{\sigma}_{n,2}^2$ are expressed in the following two assumptions:

D1. The variance estimator $\hat{\sigma}_{n,1}^2$ is defined by (3.3). The sequence of mean vectors $\{\xi_n : n \geq 1\}$ satisfies

$$\lim_{n\to\infty} n^{-1/2} \sum_{i=2}^{n} (\xi_{n,i} - \xi_{n,i-1})^2 = 0. \tag{3.5}$$

D2. The variance estimator $\hat{\sigma}_{n,2}^2$ and X_n are independent random variables. The distribution of $b_n\hat{\sigma}_{n,2}^2/\sigma_n^2$ is chi-squared, where $\{b_n : n \geq 1\}$ is a sequence of constants such that $\lim_{n\to\infty} b_n/n = b < \infty$.

Under D1 and A1, the asymptotic distribution of $n^{1/2}(\hat{\sigma}^2_{n,1} - \sigma^2_n)$ is $N(0, 3\sigma^4)$, as in Gasser et al. (1986) or by the reasoning in Section 4. Under D2 and A1, the asymptotic distribution of $n^{-1/2}(\hat{\sigma}^2_{n,2} - \sigma^2_n)$ is $N(0, 2b^{-1}\sigma^4)$.

To construct confidence balls, we begin by finding the asymptotic distribution of

$$D_n(\xi_n, X_n, \hat{\sigma}^2_{n,j}) = n^{1/2}[L_n(\hat{\xi}_{n,B}, \xi_n) - \hat{R}_{n,C}(\hat{A}_{n,B}, \hat{\sigma}^2_{n,j})]. \tag{3.6}$$

The quantity D_n compares the loss of $\hat{\xi}_{n,B}$ with the simple estimator (2.24) of its risk. On the one hand, the asymptotic distribution of D_n turns out to be normal with mean zero (Theorem 3.2 below). On the other hand, referring $D_n(\xi_n, X_n, \hat{\sigma}^2_{n,j})$ to the αth quantile of its bootstrap distribution generates a confidence ball centered at $\hat{\xi}_{n,B}$ that has asymptotic coverage probability α for ξ_n (Theorem 3.3 below). There is no apparent advantage to replacing $\hat{R}_{n,C}$ in (3.6) with the more complex bootstrap risk estimator $\hat{R}_{n,B}$.

In the remainder of the paper, the notation Dj stands for either condition D1 or D2, according to the value of j.

Theorem 3.1 *Suppose that Conditions C1 and Dj hold and that the limiting loss ρ has a unique minimum at $A_0 \in \mathcal{A}$. Then*

$$\mathcal{L}[D_n(\xi_n, X_n, \hat{\sigma}^2_{n,j})|P_{\theta_n,n}] \Rightarrow N(0, \tau^2_j(\nu, \sigma^2, A_0)), \tag{3.7}$$

where

$$\tau^2_1(\nu, \sigma^2, A_0) = 2\sigma^4 + \sigma^4[2\mu(A_0) - 1]^2 + 4\sigma^2\nu(A^c_0) \tag{3.8}$$

and

$$\tau^2_2(\nu, \sigma^2, A_0) = 2\sigma^4 + 2b^{-1}\sigma^4[2\mu(A_0) - 1]^2 + 4\sigma^2\nu(A^c_0). \tag{3.9}$$

This result is proved in Section 4. The same asymptotic distributions hold if the bootstrap selection estimator in the definition of $D_n(\xi_n, X_n, \hat{\sigma}^2_{n,j})$ is replaced by the C_p-estimator $\hat{\xi}_{n,C}$. Moreover, comparing the proof of Theorem 3.1 with its counterpart for the C_p-estimator establishes the following asymptotic equivalence in loss:

$$\operatorname*{plim}_{n\to\infty} n^{1/2}|L_n(\hat{\xi}_{n,B}, \xi_n) - L_n(\hat{\xi}_{n,C}, \xi_n)| = 0. \tag{3.10}$$

To successfully bootstrap the sampling distribution of $D_n(\xi_n, X_n, \hat{\sigma}^2_{n,j})$ requires an algorithm that recognizes both the data-based selection $\hat{A}_{n,B}$

and the structure of the variance estimator $\hat{\sigma}^2_{n,j}$. For every $A \in \mathcal{A}$, let

$$\tilde{D}_n(\xi_n, A, X_n, \hat{\sigma}^2_{n,j}) = n^{1/2}[L_n(\hat{\xi}_n(A), \xi_n) - \hat{R}_{n,C}(A, \hat{\sigma}^2_{n,j})]. \tag{3.11}$$

We consider two cases:

Bootstrapping $D_n(\xi_n, X_n, \hat{\sigma}^2_{n,1})$. Let

$$\hat{s}_n = [1 - \hat{\sigma}^2_{n,1}\mu_n(\hat{A}^c_{n,B})/\hat{\lambda}_n(\hat{A}^c_{n,B})]_+ \tag{3.12}$$

and define $\tilde{\xi}_n = (\tilde{\xi}_{n,1}, \ldots, \tilde{\xi}_{n,n})$ by

$$\tilde{\xi}_{n,i} = \begin{cases} X_{n,i} & \text{if } i/(n+1) \in \hat{A}_{n,B} \\ \hat{s}_n^{1/2} X_{n,i} & \text{if } i/(n+1) \in \hat{A}^c_{n,B} \end{cases} . \tag{3.13}$$

Let $E^*_n = (E^*_{n,1}, \ldots, E^*_{n,n})$ be a random vector such that the conditional distribution of E^*_n given X_n is $N(0, \hat{\sigma}^2_{n,1} I_n)$. Define $X^*_n = (X^*_{n,1}, \ldots, X^*_{n,n})'$ and $\sigma^{*2}_{n,1}$ by

$$\begin{aligned} X^*_n &= \tilde{\xi}_n + E^*_n \\ \sigma^{*2}_{n,1} &= \hat{\sigma}^2_{n,1} + [2(n-1)]^{-1} \sum_{i=2}^{n} (E^*_{n,i} - E^*_{n,i-1})^2. \end{aligned} \tag{3.14}$$

The partial bootstrap estimator of $\mathcal{L}[D_n(\xi_n, X_n, \hat{\sigma}^2_{n,1})|P_{\theta_n,n}]$ is then

$$\hat{H}_{n,B,1} = \mathcal{L}[D_n(\tilde{\xi}_n, \hat{A}_{n,B}, X^*_n, \sigma^{*2}_{n,1})|X_n]. \tag{3.15}$$

Bootstrapping $D_n(\xi_n, X_n, \hat{\sigma}^2_{n,2})$. Redefine $\hat{s}_n$ and E^*_n above by replacing $\hat{\sigma}^2_{n,1}$ with $\hat{\sigma}^2_{n,2}$. Define $\tilde{\xi}_n$ by (3.13) and X^*_n as in (3.14). Let $\sigma^{*2}_{n,2}$ be a random variable such that $\mathcal{L}[b_n \sigma^{*2}_{n,2}/\hat{\sigma}^2_{n,2}|X_n]$ is chi-squared with b_n degrees of freedom and such that $\sigma^{*2}_{n,2}$, E^*_n are conditionally independent, given X_n. The actual construction of $\sigma^{*2}_{n,2}$ will normally require a separate bootstrap scheme. The partial bootstrap estimator of $\mathcal{L}[D_n(\xi_n, X_n, \hat{\sigma}^2_{n,2})|P_{\theta_n,n}]$ is then

$$\hat{H}_{n,B,2} = \mathcal{L}[D_n(\tilde{\xi}_n, \hat{A}_{n,B}, X^*_n, \hat{\sigma}^{*2}_{n,2})|X_n] \tag{3.16}$$

Theorem 3.2 *Suppose that Conditions C1 and Dj hold and that the limiting loss ρ has a unique minimum at $A_0 \in \mathcal{A}$. Then*

$$\hat{H}_{n,B,j} \Rightarrow N(0, \tau^2_j(\nu, \sigma^2, A_0)) \tag{3.17}$$

in $P_{\theta_n,n}$-probability, where $\tau^2_j(\nu, \sigma^2, A_0)$ is defined by (3.8) and (3.9).

Both bootstrap algorithms in Theorem 3.2 shrink toward zero these components of X_n that are not selected by $\hat{\xi}_{n,B}$. The construction (3.13) of $\tilde{\xi}_n$ is critical for the weak convergence (3.17). If we took instead $\tilde{\xi}_n = X_n$, overfitting ξ_n, then the asymptotic variance in (3.17) would become $\tau_j^2(\nu + \sigma^2\mu, \sigma^2, A_0)$. If we used $\tilde{\xi}_n = \hat{\xi}_{n,B}$, underfitting ξ_n for bootstrap purposes, the asymptotic variance in (3.17) would become $\tau_j^2(0, \sigma^2, A_0)$. These conclusions follow by the method used to prove Theorem 3.2. Thus, neither of these alternative bootstrap algorithms yield consistent estimators of the sampling distribution of $D_n(\xi_n, X_n, \hat{\sigma}_{n,j}^2)$.

For α strictly between 0 and 1, let $\hat{H}_{n,B,j}^{-1}(\alpha)$ be the αth quantile of the bootstrap distribution $\hat{H}_{n,B,j}$ defined in (3.15) or (3.16). Under Condition Dj, define the *bootstrap selection confidence set* for ξ_n to be $C_{n,B,j} = C_n(\hat{\xi}_{n,B}, \hat{d}_{n,B,j})$, where

$$\hat{d}_{n,B,j} = [n\hat{R}_{n,C}(\hat{A}_{n,B}, \hat{\sigma}_{n,j}^2) + n^{1/2}\hat{H}_{n,B,j}^{-1}(\alpha)]_+^{1/2}. \tag{3.18}$$

The following theorem justifies this confidence set centered at $\hat{\xi}_{n,B}$.

Theorem 3.3 *Suppose that Conditions C1 and Dj hold and that the limiting risk ρ has a unique minimizer $A_0 \in \mathcal{A}$ such that $\rho(A_0) > 0$. Then*

$$\lim_{n\to\infty} P_{\theta_n,n}(C_{n,B,j} \ni \xi_n) = \alpha \tag{3.19}$$

and

$$\operatorname*{plim}_{n\to\infty} GL_n(C_{n,B,j}, \xi_n) = 2\rho^{1/2}(A_0). \tag{3.20}$$

If $\rho(A_0) = 0$ then

$$\liminf_{n\to\infty} P_{\theta_n,n}(C_{n,B,j} \ni \xi_n) \geq \alpha. \tag{3.21}$$

Remarks. The exceptional case $\rho(A_0) = 0$ arises only when $\mu(A_0) = \nu(A_0^c) = 0$. This occurs when all but an asymptotically vanishing fraction of the components of ξ_n are zero.

A more familiar confidence set for ξ_n in the normal model is $C_{n,F} = C_n(X_n, \hat{\sigma}_n\chi_n^{-1/2}(\alpha))$, where $\chi_n^{-1/2}(\alpha)$ is the square root of the αth quantile of the chi-squared distribution with n degrees of freedom. Under Conditions C1 and C2,

$$\begin{aligned} \lim_{n\to\infty} P_{\theta_n,n}(C_{n,F} \ni \xi_n) &= \alpha \\ \operatorname*{plim}_{n\to\infty} GL_n(C_{n,F}, \xi_n) &= 2\sigma, \end{aligned} \tag{3.22}$$

the second convergence relying on (3.2) and the normal approximation to the chi-squared distribution. It follows from (2.23), (3.20) and (3.22) that, at asymptotic coverage probability α, the bootstrap-selection confidence balls $C_{n,B,j}$ are both asymptotically smaller than the confidence ball $C_{n,F}$.

As an alternative to bootstrapping, the asymptotic variances in Theorem 3.1 may be estimated consistently from the sample, using $\hat{\sigma}^2_{n,j}$ for σ^2 and $[\hat{\lambda}_n(\hat{A}^c_{n,B}) - \hat{\sigma}^2_{n,j}\mu(\hat{A}^c_{n,B})]_+$ for $\nu(A_0^c)$. Equation (4.4) below justifies the second of these estimators. The estimated normal limit distributions then yield critical values and confidence sets for which an analog of Theorem 3.3 holds.

4. Proofs

The theorem proofs rely on ideas from Beran (1994) augmented by bootstrap considerations. Let $E_{n,i} = X_{n,i} - \xi_{n,i}$ and, for every set $A \in \mathcal{A}$, define

$$
\begin{aligned}
W_{n,1}(A, \theta_n) &= n^{-1/2} \sum_{i/(n+1) \in A} (E^2_{n,i} - \sigma^2_n) \\
W_{n,2}(A, \theta_n) &= n^{-1/2} \sum_{i/(n+1) \in A} \xi_{n,i} E_{n,i}. \qquad (4.1)
\end{aligned}
$$

Let $D(\mathcal{A})$ denote the set of all bounded functions having at most jump discontinuities on the compact set $\mathcal{A}$. Metrize $D(\mathcal{A})$ by supremum norm $\| \cdot \|_{\mathcal{A}}$. The σ-algebra is that generated by open balls. Under Condition C1, the two processes $W_{n,j}(\theta_n) = \{W_{n,j}(A, \theta_n) : A \in \mathcal{A}\}$ are random elements of $D(\mathcal{A})$.

Let $B_j = \{B_j(A) : A \in \mathcal{A}\}$ be two independent Gaussian processes on $\mathcal{A}$ with mean zero and covariance structure

$$
\begin{aligned}
\mathrm{Cov}[B_1(A), B_1(A')] &= \mu(A \cap A') \\
\mathrm{Cov}[B_2(A), B_2(A')] &= \nu(A \cap A') \qquad (4.2)
\end{aligned}
$$

where μ is Lebesgue measure and ν is the bounded non-negative measure defined in Condition C1. Both processes are random elements of $D(\mathcal{A})$ that have d-continuous sample paths.

Lemma 1 *Suppose that Condition C1 holds. Then the bivariate processes* $\{(W_{n,1}(\theta_n), W_{n,2}(\theta_n)\}$ *converge weakly as random elements of* $D(\mathcal{A}) \times D(\mathcal{A})$ *to the process* $(2^{1/2}\sigma^2 B_1, \sigma B_2)$.

Convergence of the finite-dimensional distributions is straightforward. For tightness, see LeCam (1983, Lemma 4) or Alexander and Pyke (1986, Section 4).

Proof of Theorem 2.1. The definitions (2.5), (2.7) and (4.1) entail

$$\hat{\lambda}_n(A) = \nu_n(A) + \hat{\sigma}_n^2 \mu_n(A) + n^{-1/2} W_{n,1}(A, \theta_n) + 2n^{-1/2} W_{n,2}(A, \theta_n). \tag{4.3}$$

Consequently, by Lemma 1,

$$\operatorname*{plim}_{n\to\infty} \|\hat{\lambda}_n(\cdot) - [\nu(\cdot) + \sigma^2 \mu(\cdot)]\|_{\mathcal{A}} = 0. \tag{4.4}$$

Then (2.19) and the second convergence in (2.20) follow from (4.4), Condition C2, and the definitions (2.8), (2.11) of $\hat{R}_{n,N}$ and $\hat{R}_{n,B}$.

On the other hand, by (2.4) and (4.1),

$$L_n(\hat{\xi}_n(A), \xi_n) = \nu_n(A^c) + \sigma_n^2 \mu_n(A) + n^{-1/2} W_{n,1}(A, \theta_n). \tag{4.5}$$

The first convergence in (2.20) follows from Lemma 4.1.

Definition (2.15) of $\hat{\xi}_{n,B}$, (2.20), and the triangle inequality imply

$$\operatorname*{plim}_{n\to\infty} \hat{R}_{n,B}(\hat{A}_{n,B}, \hat{\sigma}_n^2) = \min_{A\in\mathcal{A}} \rho(A) \tag{4.6}$$

and

$$\operatorname*{plim}_{n\to\infty} [\hat{R}_{n,B}(\hat{A}_{n,B}, \hat{\sigma}_n^2) - L_n(\hat{\xi}_{n,B}, \xi_n)] = 0. \tag{4.7}$$

Conclusion (2.21) thus follows.

Limit (4.6) and the second limit in (2.20) imply that

$$\operatorname*{plim}_{n\to\infty} \rho(\hat{A}_{n,B}) = \rho(A_0), \tag{4.8}$$

where A_0 is the unique minimizer of ρ over $\mathcal{A}$. Suppose that (2.22) does not hold. By considering a subsequence, we may assume without loss of generality that convergence (4.8) occurs almost surely while

$$P_{\theta_n,n}[d(\hat{A}_{n,B}, A_0) > \epsilon] \geq \delta \tag{4.9}$$

for some positive ϵ and δ. Because ρ is d-continuous on the compact $\mathcal{A}$ and A_0 uniquely minimizes ρ over $\mathcal{A}$, the almost sure version of (4.8) implies

that $d(\hat{A}_{n,B}, A_0) \to 0$ with probability one. This contradicts (4.8), thereby proving (2.22).

Proof of Theorem 3.1. As above, the definitions (2.4) and (2.24) of L_n and $\hat{R}_{n,C}$ respectively entail that

$$L_n(\hat{\xi}_{n,B}, \xi_n) = \nu(\hat{A}^c_{n,B}) + \sigma_n^2 \mu_n(\hat{A}_{n,B}) + n^{-1/2} W_{n,1}(\hat{A}_{n,B}, \theta_n) \tag{4.10}$$

and

$$\begin{aligned} \hat{R}_{n,C}(\hat{A}_{n,B}, \hat{\sigma}^2_{n,j}) &= \nu_n(\hat{A}^c_{n,B}) + \sigma_n^2 \mu_n(\hat{A}^c_{n,B}) + n^{-1/2} W_{n,1}(\hat{A}^c_{n,B}, \theta_n) \\ &\quad + 2n^{-1/2} W_{n,2}(\hat{A}^c_{n,B}, \theta_n) + \hat{\sigma}^2_{n,j}[\mu_n(\hat{A}_{n,B}) - \mu_n(\hat{A}^c_{n,B})]. \end{aligned} \tag{4.11}$$

Consequently,

$$\begin{aligned} D_n(\xi_n, X_n, \hat{\sigma}^2_{n,j}) &= n^{1/2}[L_n(\hat{\xi}_{n,B}, \xi_n) - \hat{R}_{n,C}(\hat{A}_{n,B}, \hat{\sigma}^2_{n,j})] \\ &= W_{n,1}(\hat{A}_{n,B}, \theta_n) - W_{n,1}(\hat{A}^c_{n,B}, \theta_n) - 2W_{n,2}(\hat{A}^c_{n,B}) \\ &\quad - n^{-1/2}(\hat{\sigma}^2_{n,j} - \sigma_n^2)[2\mu_n(\hat{A}_{n,B}) - 1]. \end{aligned} \tag{4.12}$$

Under Condition D1,

$$\begin{aligned} \hat{\sigma}^2_{n,1} &= 2^{-1}[(n-1)^{-1} \sum_{i=2}^{n} (E^2_{n,i} + E^2_{n,i-1})] \\ &\quad + (n-1)^{-1} \sum_{i=2}^{n} E_{n,i} E_{n,i-1} + o_p(n^{-1/2}) \\ &= \sigma_n^2 + 2^{-1}\{W_{n,1}([2/(n+1), 1], \theta_n) + W_{n,1}([0, (n-1)/(n+1)], \theta_n)\} \\ &\quad + (n-1)^{-1} \sum_{i=2}^{n} E_{n,i} E_{n,i-1} + o_p(n^{-1/2}). \end{aligned} \tag{4.13}$$

The argument for Lemma 1 and the martingale central limit theorem, applied to the quadratic term in the last line of (4.13), which is uncorrelated with $W_{n,1}, W_{n,2}$, imply that

$$n^{1/2}(\hat{\sigma}^2_{n,1} - \sigma_n^2) \Rightarrow 2^{1/2} \sigma^2 B_1([0,1]) + \sigma^2 Z, \tag{4.14}$$

where Z is a $N(0,1)$ random variable such that B_1, B_2 and Z are independent. Moreover,

$$D_n(\xi_n, X_n, \hat{\sigma}^2_{n,1}) = S_n(E_n, \xi_n, \sigma_n^2) + o_p(1), \tag{4.15}$$

where $S_n(E_n, \xi_n, \sigma_n^2)$ is defined by substituting (4.13) into (4.12) and dropping the remainder term.

The d-continuity of μ and ν together with the convergence (2.22) imply that $\text{plim}_{n\to\infty} \mu(\hat{A}_{n,B}) = \mu(A_0)$ and that $\text{plim}_{n\to\infty} \nu(\hat{A}_{n,B}) = \nu(A_0)$. The foregoing considerations yield

$$\begin{aligned} S_n(E_n, \xi_n, \sigma_n^2) \quad \Rightarrow \quad & 2^{1/2}\sigma^2 B_1(A_0) - 2^{1/2} B_1(A_0^c) - 2\sigma B_2(A_0^c) \\ & -[2\mu(A_0) - 1][2^{1/2} B_1([0,1]) + Z]\sigma^2. \end{aligned} \tag{4.16}$$

For $j = 1$, the weak convergence (3.7) follows from (4.2), (4.15) and (4.16). We will use (4.16) again to prove Theorem 3.2.

Under Condition D2, $\hat{\sigma}_{n,2}^2$ is independent of X_n and

$$n^{1/2}(\hat{\sigma}_{n,2}^2 - \sigma_n^2) \Rightarrow 2^{1/2} b^{-1/2} \sigma^2 Z, \tag{4.17}$$

where Z again is a $N(0,1)$ random variable such that B_1, B_2 and Z are independent. Because of (4.12), (2.22) and Lemma 1,

$$\begin{aligned} D_n(\xi_n, X_n, \hat{\sigma}_{n,2}^2) \quad \Rightarrow \quad & 2^{1/2}\sigma^2 B_1(A_0) - 2^{1/2} B_1(A_0^c) - 2\sigma B_2(A_0^c) \\ & -[2\mu(A_0) - 1] 2^{1/2} b^{-1/2} \sigma^2 Z, \end{aligned} \tag{4.18}$$

which establishes (3.7) for $j = 2$.

Proof of Theorem 3.2. Suppose that $\{\nu_n\}$, $\{\sigma_n^2\}$, and $\{A_n \in \mathcal{A}\}$ are such that

$$\lim_{n\to\infty} \|\nu_n - \tilde{\nu}\|^2, \qquad \lim_{n\to\infty} \sigma_n^2 = \sigma^2, \qquad \lim_{n\to\infty} d(A_n, A_0) = 0 \tag{4.19}$$

and $\tilde{\nu}$ is d-continuous. By the reasoning for Theorem 3.1,

$$\tilde{D}_n(\xi_n, A_n, X_n, \hat{\sigma}_{n,1}^2) = \tilde{S}_n(E_n, A_n, \xi_n, \sigma_n^2) + o_p(1), \tag{4.20}$$

where $\tilde{S}_n(E_n, A_n, \xi_n, \sigma_n^2)$ is obtained from the definition of $S_n(E_n, \xi_n, \sigma_n^2)$ by replacing $\hat{A}_{n,B}$ with A_n. As in Theorem 3.1,

$$\mathcal{L}[\tilde{S}_n(E_n, A_n, \xi_n, \sigma_n^2) | P_{\theta_n,n}] \Rightarrow N(0, \tau_1^2(\tilde{\nu}, \sigma^2, A_0)) \tag{4.21}$$

and

$$\mathcal{L}[\tilde{D}_n(\xi_n, A_n, X_n, \hat{\sigma}_{n,2}^2 | P_{\theta_n,n}] \Rightarrow N(0, \tau_2^2(\tilde{\nu}, \sigma^2, A_0)). \tag{4.22}$$

Next, consider the empirical measure

$$\begin{aligned} \tilde{\nu}_n(A) &= n^{-1} \sum_{i/(n+1)\in A} \tilde{\xi}_{n,i}^2 \\ &= \hat{\lambda}_n(A \cap \hat{A}_{n,B}) + \hat{s}_n \hat{\lambda}_n(A \cap \hat{A}_{n,B}^c). \end{aligned} \tag{4.23}$$

Under either Condition D1 or D2, it follows from (4.4) and (2.22) that

$$\operatorname*{plim}_{n\to\infty} \|\tilde{\nu}_n - \tilde{\nu}\|_{\mathcal{A}} = 0, \quad \operatorname*{plim}_{n\to\infty} \hat{\sigma}_{n,j}^2 = \sigma^2, \tag{4.24}$$

where $\tilde{\nu}$ is the measure on $\mathcal{A}$ defined by

$$\begin{aligned} \tilde{\nu}(A) &= \nu(A \cap A_0) + \sigma^2 \mu(A \cap A_0) \\ &\quad + s[\nu(A \cap A_0^c) + \sigma^2 \mu(A \cap A_0^c)] \\ s &= \nu(A_0^c)/[\nu(A_0^c) + \sigma^2 \mu(A_0^c)]. \end{aligned} \tag{4.25}$$

For the case $j = 1$, (3.14) and the reasoning for (4.20) entail that

$$D_n(\tilde{\xi}_n, \hat{A}_{n,B}, X_n^*, \hat{\sigma}_{n,1}^{*2}) = \tilde{S}_n(E_n^*, \hat{A}_{n,B}, \tilde{\xi}_n, \hat{\sigma}_{n,1}^{*2}). \tag{4.26}$$

Consequently, by (4.24), (2.22) and (4.21),

$$\hat{H}_{n,B,1} \Rightarrow N(0, \tau_1^2(\tilde{\nu}, \sigma^2, A_0)) \tag{4.27}$$

in $P_{\theta_n,n}$-probability. This limit law agrees with (3.17) because $\tilde{\nu}(A_0^c)$ from (4.25) equals $\nu(A_0^c)$ in (3.8).

For the case $j = 2$, (4.24), (2.22) and (4.22) yield

$$\hat{H}_{n,B,2} \Rightarrow N(0, \tau_2^2(\tilde{\nu}, \sigma^2, A_0)), \tag{4.28}$$

which agrees with (3.17) because $\tilde{\nu}(A_0^c)$ from (4.25) again equals $\nu(A_0^c)$ in (3.9).

Proof of Theorem 3.3. This result follows from Theorems 3.1 and 3.2. The argument parallels the proof of Theorem 3.2 in Beran (1994).

Acknowledgements

This paper was written while I was a guest of the Statistics group in the Institut für Angewandte Mathematik at Universität Heidelberg. The research was supported in part by National Science Foundation grant DMS 9224868. Comments by Lutz Dümbgen were especially helpful.

REFERENCES

Akaike, H. (1974). A new look at statistical model identification. *IEEE Trans. Auto. Cont.* **19**, 716-723.

Alexander, K.S. and Pyke, R. (1986). A uniform central limit theorem for set-indexed partial-sum processes with finite variance. *Ann. Probab.* **14**, 582-597.

Beran, R. (1994). Confidence sets centered at C_p-estimators. Unpublished preprint.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Amer. Statist. Assoc.* **87**, 738-754.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

Freedman, D.A., Navidi, W., and Peters, S.C. (1988). On the impact of variable selection in fitting regression equations. *On Model Uncertainty and its Statistical Implications* (T.K. Dijkstra, ed.), 1-16. Springer-Verlag, Berlin.

Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625-633.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* Vol. 1, 361-380. Univ. of California Press, Berkeley.

LeCam, L. (1983). A remark on empirical measures. *Festschrift for Erich Lehmann* (P.J. Bickel, K. Doksum, and J.L. Hodges, eds.), 305-327. Wadsworth, Belmont, California.

Mallows, C. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

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

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.

Speed, T.P. and Yu, B. (1993). Model selection and prediction: normal regression. *Ann. Inst. Statist. Math.* **45**, 35-54.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.* Vol. 1, 197-206. Univ. of California Press, Berkeley.

Rudolf Beran
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720–3860
U.S.A.

29 November 1994