Ann. Inst. Statist. Math.
Vol. 48, No. 1, 1–15 (1996)

CONFIDENCE SETS CENTERED AT C_p-ESTIMATORS

RUDOLF BERAN

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

(Received March 18, 1994; revised July 17, 1995)

Abstract. Suppose X_n is an observation, or average of observations, on a discretized signal ξ_n that is measured at n time points. The random vector X_n has a $N(\xi_n, \sigma_n^2 I)$ distribution, the mean and variance being unknown. Under squared error loss, the unbiased estimator X_n of ξ_n can be improved by variable-selection. Consider the candidate estimator $\hat{\xi}_n(A)$ whose i-th component equals the i-th component of X_n whenever $i/(n+1)$ lies in A and vanishes otherwise. Allow the set A to range over a large collection of possibilities. A C_p-estimator is a candidate estimator that minimizes estimated quadratic loss over A. This paper constructs confidence sets that are centered at a C_p-estimator, have correct asymptotic coverage probability for ξ_n, and are geometrically smaller than or equal to the competing confidence balls centered at X_n. The asymptotics are locally uniform in the parameters (ξ_n, σ_n^2). The results illustrate an approach to inference after variable-selection.

Key words and phrases: Variable-selection, coverage probability, geometrical loss, locally uniform asymptotics.

1. Introduction

Suppose that X_n is an observation, or average of observations, on a discretized signal ξ_n that is measured with error at n time points. The goal is to estimate ξ_n from X_n. Assume that the errors made in measuring the values of the discretized signal are independent, identically distributed, and Gaussian with mean zero. Then $X_n = (X_{n,1}, \dots, X_{n,n})'$ is a random vector with distribution $N(\xi_n, \sigma_n^2 I)$. Both the discretized signal $\xi_n = (\xi_{n,1}, \dots, \xi_{n,n})'$ and the variance σ_n^2 are unknown.

The quadratic loss 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 $|\cdot|$ denotes Euclidean norm. This loss is the discrete analog of the integrated squared error criterion commonly used in continuous signal estimation. Quadratic risk is the expected value of (1.1). By the Cauchy-Schwarz inequality,

$$L_n(\hat{\xi}_n, \xi_n) = n^{-1} \sup\{(u'\hat{\xi}_n - u'\xi_n)^2 : u \in R^n, |u| = 1\}. \tag{1.2}$$

Thus, an estimator $\hat{\xi}_n$ is close to ξ_n in quadratic loss if and only if every normalized linear combination of $\hat{\xi}_n$ is close in squared error to the corresponding linear combination of ξ_n.

Stein (1956) proved that the best unbiased estimator X_n is inadmissible for ξ_n, under quadratic loss, whenever $n \geq 3$. Strategies for bettering X_n, when n is large, accept bias in return for smaller variance. One improvement method is the James-Stein (1961) shrinkage estimator. Stein estimators reduce risk at every point ξ_n in the parameter space. As n tends to infinity, they achieve Pinsker's (1980) asymptotic minimax bound for estimation over large compact balls about the shrinkage point. A second improvement technique—local smoothing of the components of X_n—is the topic of discrete curve estimation (cf. Rice (1984)). Smoothing can reduce quadratic risk far more than uniform shrinkage when the parameter value ξ_n is sufficiently 'smooth'. The cost is increased risk when ξ_n is not smooth.

A third general method for improving on the unbiased estimator X_n is variable-selection. Let A be a subset of the unit interval. Consider the estimator $\hat{\xi}_n(A)$ whose i-th component equals $X_{n,i}$ whenever $i/(n+1)$ lies in A and vanishes otherwise. Estimate the quadratic loss of $\hat{\xi}_n(A)$ for each subset A in a large class of candidates. Let $\hat{A}_n$ denote a subset which minimizes estimated loss. The corresponding variable-selection estimator is then $\hat{\xi}_n(\hat{A}_n)$. This variable-selection approach combines simple shrinkage (the shrinkage factor is either zero or one) with localization (the shrinkage factor depends on the index i). The C_p-method, introduced by Mallows (1973), is a direct way of estimating quadratic loss so as to determine $\hat{A}_n$.

Two refinements to this variable-selection strategy are possible. First, we may reduce risk further by shrinking more cleverly the components of X_n that are indexed by A and A^c respectively. Stein (1966) gave the basic discussion. It follows that the C_p-estimators studied in this paper are not admissible. Appealing nevertheless is the logical simplicity of variable selection as a technique for improving the estimator X_n.

Second, variable-selection may be preceded by an orthogonal transformation O of the observation vector X_n, such as a finite Fourier transform, or an orthogonal polynomial transform, or a wavelet transform, or an analysis-of-variance basis transform. Prior information about the data strongly influences the choice of transform. Ideally, each component of the rotated mean vector $O\xi_n$ would be either very large or very small in magnitude relative to the standard error σ_n. When successful, this prior separation of components into an important subset and a negligible subset enhances the efficiency of variable-selection estimators for the rotated mean. These two improvements to variable selection—adaptive shrinkage and prior orthogonal transformation—will not be treated further here. We refer the reader to Donoho *et al.* (1990) for upper bounds on the reduction in risk achievable by nonlinear estimators.

The framework of this paper lets the dimension of the true mean vector ξ_n increase with n and considers candidate estimators $\hat{\xi}_n(A)$ for which A ranges over certain sets having positive Lebesgue measure. We construct confidence sets that are centered at a C_p-estimator $\hat{\xi}_n(\hat{A}_n)$, have asymptotic coverage probability α

for ξ_n, and are asymptotically closer to ξ_n than the classical level α confidence ball centered at X_n. Both of the asymptotic convergences—for coverage probability and for the geometrical error of the new confidence sets—are locally uniform over the parameter space. This local uniformity is valuable because pointwise asymptotics for variable-selection estimators can create a misleading impression of confidence set performance. For example, a natural asymptotic confidence interval based upon the Hodges estimator behaves badly near the shrinkage point, despite favorable pointwise asymptotics (Beran (1992), Section 3).

Variable-selection techniques can also be studied in settings where the dimension of the true ξ_n is fixed for all n or where the dimension of the candidate estimators is relatively small compared to n. A fixed dimension assumption on ξ_n limits the possible importance of estimator bias (cf. Speed and Yu (1993), Section 4). When the dimension of candidate estimators is small relative to n, variable selection by C_p is asymptotically equivalent to variable selection by Akaike's (1974) AIC or by certain other criteria (cf. Shibata (1981) and Rice (1984), Section 3, for details). Equivalence with C_p in this sense does not hold for the larger parameter space used in this paper.

Pötscher (1995) constructed confidence sets after model-selection for the mean in a bivariate normal model. His approach relied on the fixed dimensionality of the parameter space and on the asymptotic conditional distribution of the model-selection estimator given the selected submodel. This conditional distribution is treated more generally in Pötscher (1991). His confidence sets take values in the parameter space of the selected submodel and are conservative in their asymptotic coverage probability. It is important to note that Pötscher's analysis relies on replication asymptotics rather than the time series asymptotics used in this paper.

Section 2 of this paper defines C_p-estimators for ξ_n and their asymptotic loss. Estimation of σ_n^2 from internal analysis of X_n or from replicated observations on the discretized signal ξ_n is discussed. Section 3 develops locally uniform asymptotic distributions for the centered quadratic loss of C_p-estimators. Consistent estimation of the appropriate limit distribution then yields confidence sets for ξ_n that are centered at C_p-estimators.

2. Convergence of C_p-estimators

This section formally defines C_p-estimators for ξ_n and studies their asymptotic loss. The statistical ideas here, which have a long history, draw in particular on Mallows (1973) and Rice (1984). The technical formulation and results of this section will be used to construct confidence sets in Section 3.

2.1 *Estimating ξ_n*

Let A be a subset of $[0,1]$ formed by the union of m ordered closed intervals. Thus,

$$A = \bigcup_{i=1}^{m} [t_{2i-1}, t_{2i}] \tag{2.1}$$

where $0 \le t_1 \le \cdots \le t_{2m} \le 1$. Let $\mathcal{S}(m)$ denote the class of all sets A of the form (2.1). The value of m is fixed. Define the pseudo-distance between two sets A and B in $\mathcal{S}(m)$ to be

$$d(A,B) = \mu(A \Delta B), \tag{2.2}$$

where μ is Lebesgue measure. After forming equivalence classes, $\mathcal{S}(m)$ is a compact metric space under d.

The observed signal vector $X_n = (X_{n,1}, \ldots, X_{n,n})'$ has a $N(\xi_n, \sigma_n^2 I)$ distribution on R^n. Let $\theta_n = (\xi_n, \sigma_n^2)$ and let $P_{\theta_n,n}$ denote this normal distribution. For every $A \in \mathcal{S}(m)$, let

$$h_{n,i}(A) = \begin{cases} 1 & \text{if } i/(n+1) \in A \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

and let

$$\hat{\xi}_n(A) = (h_{n,1}(A)X_{n,1}, \ldots, h_{n,n}(A)X_{n,n})'. \tag{2.4}$$

Candidate estimators for ξ_n are generated by picking a compact subset $\mathcal{A}$ of $\mathcal{S}(m)$, possibly $\mathcal{S}(m)$ itself, and then considering all estimators of the form (2.4) as A ranges over $\mathcal{A}$. To ensure that X_n is among the candidate estimators, we will require that $\mathcal{A}$ contains the unit interval $[0,1]$ as an element.

Example. Suppose $m = 1$ and

$$\mathcal{A} = \{A(t) \in \mathcal{S}(1) : t \in R^2 \text{ and } t_1 = 0\}. \tag{2.5}$$

A typical set in $\mathcal{A}$ has the form $A(t) = [0, t_2]$, where $0 \le t_2 \le 1$. The candidate estimators $\hat{\xi}_n(A(t))$ can be described explicitly as follows. When $1 \le j \le n$, then

$$\hat{\xi}_n(A(t)) = (X_{n,1}, \ldots, X_{n,j}, 0, \ldots 0)' \qquad \text{if } \ j/(n+1) \le t_2 < (j+1)/(n+1). \tag{2.6}$$

When $t_2 < 1/(n+1)$, then $\hat{\xi}_n(A(t)) = (0, \ldots, 0)'$. When $t_2 = 1$, then $\hat{\xi}_n(A(t)) = X_n$. This example produces a nested class of candidate estimators. The candidate estimators generated by other choices of $\mathcal{A}$ need not be hierarchical.

For every set $A \in \mathcal{A}$, define the non-negative set function ν_n by

$$\nu_n(A) = n^{-1} \sum_{i/(n+1) \in A} \xi_{n,i}^2. \tag{2.7}$$

Let A^c denote the set complement in $[0,1]$ of A. The quadratic loss (1.1) of the candidate estimator $\hat{\xi}_n(A)$ is then

$$\begin{aligned} L_n(\hat{\xi}_n(A), \xi_n) &= n^{-1} |\hat{\xi}_n(A) - \xi_n|^2 \\ &= n^{-1} \sum_{i/(n+1) \in A} (X_{n,i} - \xi_{n,i})^2 + \nu_n(A^c). \end{aligned} \tag{2.8}$$

We will see, in Theorem 2.1 below, that this loss converges in probability to a constant. Consequently, minimizing loss is asymptotically equivalent to minimizing risk.

Estimators of the loss $L_n(\hat{\xi}_n(A), \xi_n)$ will be phrased in terms of two nonnegative set functions, one of which depends on the sample X_n:

$$\mu_n(A) = n^{-1} \sum_{i/(n+1) \in A} 1 \tag{2.9}$$

and

$$\hat{\lambda}_n(A) = n^{-1} \sum_{i/(n+1) \in A} X_{n,i}^2. \tag{2.10}$$

Let $\hat{\sigma}_n^2$ be a consistent estimator of σ_n^2, such as those discussed in Subsection 2.2. Define the *estimated loss* of candidate estimator $\hat{\xi}_n(A)$ by

$$\hat{L}_n(\hat{\xi}_n(A), \hat{\sigma}_n^2) = \hat{\lambda}_n(A^c) + \hat{\sigma}_n^2[2\mu_n(A) - 1]. \tag{2.11}$$

Theorem 2.1 below establishes the uniform consistency of this loss estimator for L_n over all sets in $\mathcal{A}$. If $\hat{\sigma}_n^2$ is unbiased for σ_n^2, then $\hat{L}_n(\hat{\xi}_n(A), \hat{\sigma}_n^2)$ is an unbiased predictor of the loss of $\hat{\xi}_n(A)$ and is an unbiased estimator of the risk of $\hat{\xi}_n(A)$. Both (2.8) and (2.11) define random set functions on $\mathcal{A}$.

The idea of C_p-estimation is to select the candidate estimator whose estimated loss is smallest. In the present context, let $\hat{A}_n$ be a set in $\mathcal{A}$ such that

$$\hat{L}_n(\hat{\xi}_n(\hat{A}_n), \hat{\sigma}_n^2) = \min_{A \in \mathcal{A}} \hat{L}_n(\hat{\xi}_n(A), \hat{\sigma}_n^2). \tag{2.12}$$

This minimum is achieved because, for fixed n, the estimated loss assumes a finite number of values as A ranges over $\mathcal{A}$. The estimator

$$\hat{\xi}_{n,C} = \hat{\xi}_n(\hat{A}_n) \tag{2.13}$$

is called a *C_p-estimator* generated by the candidate estimators $\{\hat{\xi}_n(A) \colon A \in \mathcal{A}\}$. This terminology recognizes that $\hat{A}_n$ minimizes the quantity $\hat{\lambda}_n(A^c) + 2\hat{\sigma}_n^2\mu_n(A)$ over all $A \in \mathcal{A}$, a procedure analogous to one proposed by Mallows (1973) in a different context.

To establish basic locally uniform asymptotic convergences for C_p estimators and their losses, we introduce two assumptions. The notation $\|\cdot\|_{\mathcal{A}}$ stands for supremum norm computed over all sets $A \in \mathcal{A}$ while plim stands for limit in $P_{\theta_n,n}$-probability.

A1. $\mathcal{A}$ is a compact subset of $\mathcal{S}(m)$, in the metric d, which contains the element $[0,1]$. The sequence $\{\theta_n = (\xi_n, \sigma_n^2) \colon n \geq 1\}$ is such that

$$\lim_{n\to\infty} \|\nu_n - \nu\|_{\mathcal{A}} = 0, \qquad \lim_{n\to\infty} \sigma_n^2 = \sigma^2 \tag{2.14}$$

for some finite non-negative measure ν on the Borel sets of $[0,1]$ and some finite positive σ^2. The measure ν is absolutely continuous with respect to Lebesgue measure.

A2. For every sequence $\{\theta_n\}$ that satisfies Assumption A1,

$$\operatorname*{plim}_{n\to\infty} \hat{\sigma}_n^2 = \sigma^2. \tag{2.15}$$

The first limit in (2.14) holds if $\xi_{n,i} = \xi(i/(n+1))$, where ξ is a continuous function of $[0,1]$. Such an assumption is typical in the literature on curve estimation or nonparametric regression. In the special case where $\mathcal{A}$ consists only of the set $[0,1]$, the first half of (2.14) requires only that $n^{-1}|\xi_n|^2$ converge to a finite constant. This convergence is central in the asymptotic theory of Stein estimators (cf. Stein (1956), Casella and Hwang (1982)). In general, Assumption A1 falls between the assumptions used in asymptotics for Stein estimators and those for curve estimators. A1 also implies that ν is continuous on $\mathcal{A}$ in the metric d. Estimators $\hat{\sigma}_n^2$ that satisfy A2 will be described in Subsection 2.2.

Let μ denote Lebesgue measure restricted to $\mathcal{A}$. Define the non-negative set function ρ on $\mathcal{A}$ by

$$\rho(A) = \nu(A^c) + \sigma^2\mu(A). \tag{2.16}$$

As the following result shows, $\rho(A)$ is the asymptotic loss of the candidate estimator $\hat{\xi}_n(A)$.

THEOREM 2.1. *Suppose that Assumptions* A1 *and* A2 *hold. Then,*

$$\begin{aligned} &\operatorname*{plim}_{n\to\infty} \|L_n(\hat{\xi}_n(\cdot), \xi_n) - \rho\|_{\mathcal{A}} = 0 \\ &\operatorname*{plim}_{n\to\infty} \|\hat{L}_n(\hat{\xi}_n(\cdot), \hat{\sigma}_n^2) - \rho\|_{\mathcal{A}} = 0. \end{aligned} \tag{2.17}$$

Consequently,

$$\operatorname*{plim}_{n\to\infty} L_n(\hat{\xi}_{n,C}, \xi_n) = \min_{A\in\mathcal{A}} \rho(A). \tag{2.18}$$

If M denotes the set of minimizers of ρ over $\mathcal{A}$, then

$$\operatorname*{plim}_{n\to\infty} \inf_{A\in M} d(\hat{A}_n, A) = 0. \tag{2.19}$$

This theorem is proved in Section 4. By equations (2.17) and (2.18), the loss of the C_p-estimator $\hat{\xi}_{n,C}$ coincides asymptotically with the loss of the unknown candidate estimators for ξ_n that have smallest loss. In this restricted sense, the C_p estimator is asymptotically efficient (but see the discussion in the Introduction). Moreover, because $\mathcal{A}$ was assumed to contain $[0,1]$,

$$\min_{A\in\mathcal{A}} \rho(A) \leq \rho([0,1]) = \sigma^2. \tag{2.20}$$

Equality in (2.20) holds in special circumstances such as $\nu(A^c) > \sigma^2\mu(A^c)$ for every $A \in \mathcal{A}$ except $A = [0,1]$. This condition on ν holds if each $\xi^2_{n,i}$ is greater than and bounded away from σ^2 for all sufficiently large n. In general, the limiting loss of the C_p estimator is strictly less than that of the unbiased estimator X_n.

The loss of the C_p-estimator $\hat{\xi}_{n,C}$ is bounded above by $n^{-1}\sum_{i=1}^n (X_{n,i}-\xi_{n,i})^2 + \nu_n([0,1])$, whose expectation is $\sigma_n^2 + \nu_n([0,1])$. Thus, by a standard uniform integrability argument, the risk of the C_p estimator also converges to the limiting loss $\min_{A\in\mathcal{A}} \rho(A)$.

2.2 *Estimating* σ_n^2

Consistent estimators of σ_n^2, required to define the C_p-estimator $\hat{\xi}_{n,C}$, may be internal or external. Internal estimators depend only on the observed X_n and require smoothness or dimensionality restrictions of the possible values of ξ_n. External estimators rely on replication or some other source of independent data concerning σ_n^2. One internal estimator of σ_n^2, suggested by Rice (1984), is

$$\hat{\sigma}^2_{n,1} = [2(n-1)]^{-1}\sum_{i=2}^{n}(X_{n,i} - X_{n,i-1})^2. \tag{2.21}$$

If the sequence $\{\theta_n = (\xi_n, \sigma_n^2)\}$ is such that

$$\lim_{n\to\infty} n^{-1}\sum_{i=2}^{n}(\xi_{n,i} - \xi_{n,i-1})^2 = 0, \qquad \lim_{n\to\infty}\sigma_n^2 = \sigma^2 < \infty, \tag{2.22}$$

then $\hat{\sigma}_n^2$ converges in $P_{\theta_n,n}$-probability to σ^2. Variants of this internal variance estimator that require stronger assumptions on ξ_n to achieve faster rates of convergence were discussed by Rice (1984) and by Gasser *et al.* (1986).

A second internal estimator is appropriate under the assumption that the discretized signal ξ_n lies in a subspace of dimension $n' < n$. For simplicity, suppose that n' is the integer part of cn and c is a constant strictly between 0 and 1. By making a suitable orthogonal transformation, assume without loss of generality that $X_n = (X_{n'}, Z_{n-n'})$, where $X_{n'}$ has a $N(\xi_{n'}, \sigma_n^2 I)$ distribution in n' dimensions, $Z_{n-n'}$ has a $N(0, \sigma_n^2 I)$ distribution in $n - n'$ dimensions, and $X_{n'}$, $Z_{n-n'}$ are independent. In this canonical setting, a C_p-estimator of $\xi_{n'}$ can be constructed from $X_{n'}$ and the variance estimator

$$\hat{\sigma}^2_{n,2} = (n-n')^{-1}|Z_{n-n'}|^2. \tag{2.23}$$

If the second half of (2.22) holds, then $\hat{\sigma}^2_{n,2}$ converges in $P_{\theta_n,n}$-probability to σ^2.

Replication makes possible external estimation of variance. Suppose we observe independent random vectors $\{X_n^{(j)} \colon 1 \le j \le J\}$, each of which has a $N(\xi_n, \sigma_n^2)$ distribution. A C_p-estimator would now be based on the average vector

$$X_n = J^{-1}\sum_{j=1}^{J} X_n^{(j)}, \tag{2.24}$$

using the variance estimator

$$(2.25)\qquad \tilde{\sigma}^2_{n,2} = [n(J-1)]^{-1}\sum_{j=1}^{J}|X_n^{(j)} - X_n|^2.$$

If the second half of (2.22) holds, then $\tilde{\sigma}^2_{n,2}$ also converges in $P_{\theta_n,n}$-probability to σ^2.

The confidence set constructions of Section 3 involve more than consistency of the variance estimator used. If $\hat{\sigma}^2_{n,1}$ is the internal estimator defined in (2.21) and the first half of condition (2.22) is strengthened to (2.26) below, then the limiting distribution of $n^{1/2}(\hat{\sigma}^2_{n,1} - \sigma^2_n)$ is $N(0, 3\sigma^4)$. For a proof, see Gasser *et al.* (1986) or the argument in Section 4 for Theorem 3.1. On the other hand, the distribution of $(n-n')\hat{\sigma}^2_{n,2}/\sigma^2_n$ is chi-squared with $n-n'$ degrees of freedom and the distribution of $n(J-1)\tilde{\sigma}^2_{n,2}/\sigma^2_n$ is chi-squared with $n(J-1)$ degrees of freedom.

The essential features of these variance estimators are captured in one or the other of the following assumptions:

B1. The variance estimator $\hat{\sigma}^2_{n,1}$ is defined by (2.21). The sequence $\{\xi_n\}$ satisfies

$$(2.26)\qquad \lim_{n\to\infty} n^{-1/2}\sum_{i=2}^{n}(\xi_{n,i} - \xi_{n,i-1})^2 = 0.$$

B2. The variance estimator $\hat{\sigma}^2_{n,2}$ and X_n are independent random variables. The distribution of $b_n\hat{\sigma}^2_{n,2}/\sigma^2_n$ is chi-squared with b_n degrees of freedom and $\lim_{n\to\infty} b_n/n = b < \infty$.

Note that either B1 or B2 implies the consistency Assumption A2. Under B2, the limiting distribution of $n^{1/2}(\hat{\sigma}^2_{n,2} - \sigma^2_n)$ is $N(0, 2b^{-1}\sigma^4)$.

3. Confidence sets for ξ_n

This section considers confidence balls for ξ_n, centered at an estimator $\hat{\xi}_n$ and having radius $\hat{d}_n$:

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

Once the center $\hat{\xi}_n$ is specified, the radius $\hat{d}_n$ is defined so that the coverage probability $P_{\theta_n,n}[C_n \ni \xi_n]$ converges to α as n increases. The geometrical error in $C_n(\hat{\xi}_n, \hat{d}_n)$ as a set-valued estimator of ξ_n is then measured by the geometrical loss

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

Let $U = \{u \in R^n : |u| = 1\}$ denote the unit sphere. Confidence set C_n is equivalent, by the Cauchy-Schwarz inequality, to the following simultaneous one-sided confidence intervals for linear combinations of ξ_n:

$$\begin{aligned}(3.3)\qquad C_n(\hat{\xi}_n, \hat{d}_n) &= \{t \in R^n : \sup_{u \in U}(u't - u'\hat{\xi}_n) \leq \hat{d}_n\}\\ &= \{t \in R^n : u't \leq u'\hat{\xi}_n + \hat{d}_n \quad \forall u \in U\}.\end{aligned}$$

This is the argument for Scheffé's method of multiple comparisons.

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

$$\begin{aligned}(3.4)\qquad &\max_{u \in U} n^{-1/2} \max\{u'\hat{\xi}_n + \hat{d}_n - u'\xi_n, 0\}\\ &= n^{-1/2} \max_{u \in U} |u'(\hat{\xi}_n - \xi_n)| + n^{-1/2}\hat{d}_n\\ &= GL_n(C_n, \xi_n).\end{aligned}$$

Minimizing the geometrical loss of confidence set C_n is thus the same as minimizing the maximum overshoot of the equivalent simultaneous one-sided confidence intervals for ξ_n.

A standard confidence ball for ξ_n, centered at the unbiased estimator X_n, is

$$(3.5)\qquad C_{n,S} = C_n(X_n, \hat{\sigma}_n \chi_n^{-1/2}(\alpha)).$$

Here $\chi_n^{-1/2}(\alpha)$ denotes the square root of the α-th quantile of the chi-squared distribution with n degrees of freedom. Under Assumptions A1 and A2,

$$(3.6)\qquad \lim_{n \to \infty} P_{\theta_n, n}[C_{n,S} \ni \xi_n] = \alpha$$

and

$$(3.7)\qquad \operatorname*{plim}_{n \to \infty} GL_n(C_{n,S}, \xi_n) = 2\sigma,$$

by (3.2) and the normal approximation to the chi-squared distribution. To check (3.7), note that $n^{-1/2}|X_n - \xi_n|$ and $\hat{\sigma}_n$ both converge in probability to σ while $n^{-1/2}\chi^{-1}(\alpha) = n^{-1/2}[n + O(n^{1/2})]^{1/2}$ converges to 1.

3.1 *Estimated sampling distributions*

To construct confidence balls for ξ_n that are centered at a C_p-estimator $\hat{\xi}_{n,C}$, we will study the sampling distribution of

$$(3.8)\qquad D_n(\xi_n, X_n, \hat{\sigma}_n^2) = n^{1/2}[L_n(\hat{\xi}_{n,C}, \xi_n) - \hat{L}_n(\hat{\xi}_{n,C}, \hat{\sigma}_n^2)].$$

This quantity compares the loss of the C_p-estimator with its estimated loss. The asymptotics for D_n depend upon the choice of variance estimator $\hat{\sigma}_n^2$. The notation Bi will stand for either Assumption B1 or B2 of Section 2, according to the value of i. The following locally uniform weak convergence is proved in Section 4.

THEOREM 3.1. *Suppose that conditions* A1 *and* Bi *hold and that the limiting loss measure* ρ *has a unique minimum at* A_0 *in* $\mathcal{A}$. *Then*

$$\mathcal{L}[D_n(\xi_n, X_n, \hat{\sigma}^2_{n,i})|P_{\theta_n,n}] \Rightarrow N(0, \tau_i^2(\sigma^2, \nu, A_0)) \tag{3.9}$$

where

$$\tau_1^2(\sigma^2, \nu, A_0) = 2\sigma^4 + \sigma^4[2\mu(A_0) - 1]^2 + 4\sigma^2\nu(A_0^c) \tag{3.10}$$

and

$$\tau_2^2(\sigma^2, \nu, A_0) = 2\sigma^4 + 2b^{-1}\sigma^4[2\mu(A_0) - 1]^2 + 4\sigma^2\nu(A_0^c). \tag{3.11}$$

Consistent estimators for the asymptotic distributions in this theorem are easily constructed. Let $\hat{A}_n$ be the set, defined in (2.12), that determines the C_p-estimator (2.13). Let $\hat{\nu}_{n,i}$ be $[\hat{\lambda}_n - \hat{\sigma}^2_{n,i}\mu_n]_+$, where $[\cdot]_+$ denotes the non-negative part function. It follows from (4.7) and Assumption Bi that $\text{plim}_{n\to\infty} \|\hat{\nu}_{n,i} - \nu\|_{\mathcal{A}} = 0$. The natural estimator of the limit distribution (3.9) is then

$$\hat{H}_{n,i} = N(0, \tau_i^2(\hat{\sigma}^2_{n,i}, \hat{\nu}_{n,i}, \hat{A}_n)). \tag{3.12}$$

Under Assumptions A1 and Bi, it follows from (2.19), the convergence of $\hat{\nu}_{n,i}$, and the d-continuity of the measures μ, ν that

$$\hat{H}_{n,i} \Rightarrow N(0, \tau_i^2(\sigma^2, \nu, A_0)) \tag{3.13}$$

in $P_{\theta_n,n}$-probability. The d-continuity of μ stems from definition (2.2) while the d-continuity of ν is implied by A1.

3.2 *Confidence sets*

For α strictly between 0 and 1, let $\hat{H}^{-1}_{n,i}(\alpha)$ denote the α-th quantile of the estimated sampling distribution defined in (3.12). The *asymptotic C_p-confidence set* for ξ_n under condition Bi is then defined to be $C_{n,i} = C_n(\hat{\xi}_{n,C}, \hat{d}_{n,i})$, where

$$\hat{d}_{n,i} = [n\hat{L}_n(\hat{\xi}_{n,C}, \hat{\sigma}^2_{n,i}) + n^{1/2}\hat{H}^{-1}_{n,i}(\alpha)]_+^{1/2}. \tag{3.14}$$

The main asymptotic properties of this confidence set are expressed in the following result.

THEOREM 3.2. *Suppose that conditions* A1 *and* Bi *hold and that the limiting loss measure* ρ *has a unique minimum at* A_0 *in* $\mathcal{A}$. *Then*

$$\operatorname*{plim}_{n\to\infty} GL_n(C_{n,i}, \xi_n) = 2\rho^{1/2}(A_0). \tag{3.15}$$

If $\rho(A_0) > 0$, *then*

$$\lim_{n\to\infty} P_{\theta_n,n}[C_{n,i} \ni \xi_n] = \alpha. \tag{3.16}$$

If $\rho(A_0) = 0$, *then*

$$\liminf_{n\to\infty} P_{\theta_n,n}[C_{n,i} \ni \xi_n] \geq \alpha. \tag{3.17}$$

The exceptional case $\rho(A_0) = 0$ arises only if both $\mu(A_0)$ and $\nu(A_0^c)$ vanish. Roughly speaking, this means that for large n, all but a tiny fraction of the $\{\xi_{n,i}^2\}$ are close to zero. It follows from (2.20) and (3.7) that the asymptotic geometrical risk of the C_p-confidence set $C_{n,i}$ is less than or equal to that of the confidence set $C_{n,S}$ centered at X_n. The construction of this section thus translates the improved asymptotic loss of a C_p-estimator into improved asymptotic geometrical loss for the associated confidence set.

4. Proofs

The theorem proofs rely on the following lemma. Let

$$\begin{aligned} W_{n,1}(A,\theta_n) &= n^{-1/2} \sum_{i/(n+1)\in A} [(X_{n,i} - \xi_{n,i})^2 - \sigma_n^2] \\ W_{n,2}(A,\theta_n) &= n^{-1/2} \sum_{i/(n+1)\in A} \xi_{n,i}(X_{n,i} - \xi_{n,i}) \end{aligned} \tag{4.1}$$

for every set A in $\mathcal{S}(m)$. Let $L_\infty(\mathcal{S})$ denote the set of all bounded measurable functions on $\mathcal{S}(m)$, metrized by supremum norm. Under Assumption A1, the two processes $W_{n,i}(\theta_n) = \{W_{n,i}(A,\theta_n): A \in \mathcal{S}(m)\}$ are random elements of $L_\infty(\mathcal{S})$.

Let $B_i = \{B_i(A): A \in \mathcal{S}(m)\}$ be two independent Gaussian processes on $\mathcal{S}(m)$ with means zero and

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

where μ is Lebesgue measure and ν is the bounded non-negative measure defined in Assumption A1. Both processes are random elements of $L_\infty(\mathcal{S})$, as will be seen in the following proof.

LEMMA 4.1. *Suppose that Assumption* A1 *holds. The processes* $\{(W_{n,1}(\theta_n), W_{n,2}(\theta_n))\}$ *then converge weakly as random elements of* $L_\infty(\mathcal{S}) \times L_\infty(\mathcal{S})$ *to the process* $(2^{1/2}\sigma^2 B_1, \sigma B_2)$.

PROOF. Without loss of generality, suppose that $X_{n,i} = \xi_{n,i} + \sigma_n Z_i$, where the $\{Z_i\}$ are iid standard normal random variables. Consider the two partial sum processes with sample paths in $D[0,1]$ that are given by

$$\begin{aligned} V_{n,1}(t) &= n^{-1/2} \sum_{i/(n+1)\leq t} (Z_i^2 - 1) \\ V_{n,2}(t) &= n^{-1/2} \sum_{i/(n+1)\leq t} \xi_{n,i} Z_i. \end{aligned} \tag{4.3}$$

Define the processes $G_i(t) = B_i([0,t])$ on $[0,1]$. Note that G_1 is a Brownian Bridge process. Then, under A1,

$$(V_{n,1}, V_{n,2}) \Rightarrow (2^{1/2}G_1, G_2) \tag{4.4}$$

as random elements of $D[0,1] \times D[0,1]$.

To verify (4.4), observe that the marginal weak convergence of $V_{n,i}$ that is entailed by (4.4) is known (cf. Section 16 of Billingsley (1968) and Section 4 of Alexander and Pyke (1986)). Consequently, the bivariate processes $\{(V_{n,1}, V_{n,2})\}$ are tight. Because $V_{n,1}$ and $V_{n,2}$ are uncorrelated, the finite dimensional distributions of any linear combination $aV_{n,1}(t)+bV_{n,2}(t)$ converge to those of $2^{1/2}aG_1(t)+bG_2(t)$, by the central limit theorem. The weak convergence (4.4) thus follows.

Next, for any set A of the form (2.1),

$$\begin{aligned} W_{n,1}(A,\theta_n) &= \sigma_n^2 \sum_{i=1}^{m} [V_{n,1}(t_{2i}) - V_{n,1}(t_{2i-1}-)] \\ W_{n,2}(A,\theta_n) &= \sigma_n \sum_{i=1}^{m} [V_{n,2}(t_{2i}) - V_{n,2}(t_{2i-1}-)]. \end{aligned} \tag{4.5}$$

The lemma follows by applying (4.4) to (4.5) and then checking that the covariance structure of the limit process coincides with that given in (4.2).

Proof of Theorem 2.1. From (2.10), (2.7) and the definition of $W_{n,i}(\theta_n)$

$$\hat{\lambda}_n(A) = \nu_n(A) + \sigma_n^2\mu_n(A) + n^{-1/2}W_{n,1}(A,\theta_n) + 2n^{-1/2}W_{n,2}(A,\theta_n). \tag{4.6}$$

Lemma 4.1 and Assumption A1 then imply

$$\|\hat{\lambda}_n(\cdot) - [\nu(\cdot) + \sigma^2\mu(\cdot)]\|_{\mathcal{A}} \to 0 \tag{4.7}$$

in $P_{\theta_n,n}$-probability. The second convergence in (2.17) follows from (2.11), (4.7) and Assumption A2.

Similarly, by (2.8),

$$L_n(\hat{\xi}_n(A), \xi_n) = n^{-1/2}W_{n,1}(A,\theta_n) + \sigma_n^2\mu_n(A) + \nu_n(A^c). \tag{4.8}$$

The first convergence in (2.17) follows from Lemma 4.1 and Assumption A1.

The definition (2.13) of $\hat{\xi}_{n,C}$, (2.17) and the triangle inequality yield

$$\hat{L}_n(\hat{\xi}_{n,C}, \hat{\sigma}_n^2) \to \min_{A\in\mathcal{A}} \rho(A) \tag{4.9}$$

and

$$\hat{L}_n(\hat{\xi}_{n,C}, \hat{\sigma}_n^2) - L_n(\hat{\xi}_{n,C}, \xi_n) \to 0, \tag{4.10}$$

both convergences being in $P_{\theta_n,n}$-probability. Conclusion (2.18) follows immediately.

Suppose that (2.19) does not hold. In view of the second limit in (2.17), we may assume without loss of generality, by going to a subsequence, that

$$\|\hat{L}_n(\hat{\xi}_n(\cdot),\hat{\sigma}_n^2)-\rho\|_{\mathcal{A}}=0 \quad \text{w.p.1} \tag{4.11}$$

while

$$P_{\theta_n,n}[\inf_{A\in M} d(\hat{A}_n,A)>\epsilon]\geq\delta \tag{4.12}$$

for some positive ϵ and δ. Because $\mathcal{A}$ is compact, ρ is d-continuous, and $\hat{A}_n$ minimizes $\hat{L}_n(\hat{\xi}_n(\cdot),\hat{\sigma}_n^2)$ over $\mathcal{A}$, the uniform convergence (4.11) implies that $\inf_{A\in M} d(\hat{A}_n,A)\to 0$ with probability one. This contradicts (4.12) and so establishes (2.19).

Proof of Theorem 3.1. As in the preceding proof, the definitions of L_n and $\hat{L}_n$ entail

$$L_n(\hat{\xi}_{n,C},\xi_n)=\nu_n(\hat{A}_n^c)+\sigma_n^2\mu_n(\hat{A}_n)+n^{-1/2}W_{n,1}(\hat{A}_n,\theta_n) \tag{4.13}$$

and

$$\begin{aligned}\hat{L}_n(\hat{\xi}_{n,C},\hat{\sigma}_{n,i}^2)&=\nu_n(\hat{A}_n^c)+\sigma_n^2\mu_n(\hat{A}_n^c)+n^{-1/2}W_{n,1}(\hat{A}_n^c,\theta_n)\\&\quad+2n^{-1/2}W_{n,2}(\hat{A}_n^c,\theta_n)+\hat{\sigma}_{n,i}^2[\mu_n(\hat{A}_n)-\mu_n(\hat{A}_n^c)].\end{aligned} \tag{4.14}$$

Hence

$$\begin{aligned}D_n(\xi_n,X_n,\hat{\sigma}_{n,i}^2)&=n^{1/2}[L_n(\hat{\xi}_{n,C},\xi_n)-\hat{L}_n(\hat{\xi}_{n,C},\hat{\sigma}_{n,i}^2)]\\&=W_{n,1}(\hat{A}_n,\theta_n)-W_{n,1}(\hat{A}_n^c,\theta_n)-2W_{n,2}(\hat{A}_n^c,\theta_n)\\&\quad-n^{1/2}(\hat{\sigma}_{n,i}^2-\sigma_n^2)[2\mu_n(\hat{A}_n)-1].\end{aligned} \tag{4.15}$$

Suppose Assumption B1 holds and write $X_{n,i}=\xi_{n,i}+\sigma_n Z_i$, as in the proof of Lemma 4.1, where the $\{Z_i\}$ are iid standard normal random variables. Then, under $P_{\theta_n,n}$,

$$\begin{aligned}\hat{\sigma}_{n,1}^2&=2^{-1}\left[(n-1)^{-1}\sigma_n^2\sum_{i=2}^{n}(Z_i^2+Z_{i-1}^2)\right]\\&\quad+(n-1)^{-1}\sigma_n^2\sum_{i=2}^{n}Z_iZ_{i-1}+o_p(n^{-1/2}).\end{aligned} \tag{4.16}$$

Consequently,

$$\begin{aligned}n^{1/2}(\hat{\sigma}_{n,1}^2-\sigma_n^2)&=2^{-1}[W_{n,1}([2/(n+1),1],\theta_n)\\&\qquad+W_{n,1}([0,(n-1)/(n+1)],\theta_n)]\\&\quad+n^{1/2}(n-1)^{-1}\sigma_n^2\sum_{i=2}^{n}Z_iZ_{i-1}+o_p(1).\end{aligned} \tag{4.17}$$

The quadratic term in (4.17) converges, by the martingale central limit theorem, to a Gaussian limit. Moreover, let Z be a standard normal variable such that B_1, B_2 and Z are mutually independent. By extension of the proof for Lemma 4.1, the processes $\{(W_{n,1}, W_{n,2}, n^{-1/2}\sigma_n^2 \sum_{i=2}^n Z_i Z_{i-1})\}$ converge weakly to $(2^{1/2}\sigma^2 B_1, \sigma B_2, \sigma^2 Z)$, as random elements of $L_\infty(\mathcal{S}) \times L_\infty(\mathcal{S}) \times R$.

It follows from (4.17) that $n^{1/2}(\hat{\sigma}_{n,1}^2 - \sigma_n^2)$ converges weakly to $2^{1/2}\sigma^2 B_1([0,1]) + \sigma^2 Z$, whose distribution is $N(0, 3\sigma^4)$. Moreover, (2.19) and the d-continuity of the measures μ, ν imply that

$$(4.18) \qquad \mu(\hat{A}_n) \to \mu(A_0) \quad \text{and} \quad \nu(\hat{A}_n) \to \nu(A_0)$$

in $P_{\theta_n,n}$-probability. The foregoing considerations establish

$$(4.19) \quad D_n(\xi_n, X_n, \hat{\sigma}_{n,1}^2) \Rightarrow 2^{1/2}\sigma^2 B_1(A_0) - 2^{1/2} B_1(A_0^c) - 2\sigma B_2(A_0^c) - [2\mu(A_0) - 1][2^{1/2} B_1([0,1]) + Z]\sigma^2.$$

For any set A in $\mathcal{A}$, (4.2) entails

$$(4.20) \qquad \operatorname{Cov}[B_1(A), B_1([0,1])] = \mu(A).$$

Thus, after simplification of the variance, the right side of (4.19) has a $N(0, \tau_1(\sigma^2, \nu, A_0))$ distribution.

Suppose Assumption B2 holds. Then $\hat{\sigma}_{n,2}^2$ is independent of X_n and

$$(4.21) \qquad n^{1/2}(\hat{\sigma}_{n,2}^2 - \sigma_n^2) \Rightarrow 2^{1/2} b^{-1/2} \sigma^2 Z,$$

where again Z is a standard normal random variable such that B_1, B_2 and Z are independent. In view of (4.15), (2.19) and Lemma 4.1,

$$(4.22) \quad D_n(\xi_n, X_n, \hat{\sigma}_{n,2}^2) \Rightarrow 2^{1/2}\sigma^2 B_1(A_0) - 2^{1/2} B_1(A_0^c) - 2\sigma B_2(A_0^c) - [2\mu(A_0) - 1] 2^{1/2} b^{-1/2} \sigma^2 Z.$$

The right side of (4.22) has a $N(0, \tau_2(\sigma^2, \nu, A_0))$ distribution.

PROOF OF THEOREM 3.2. By the definition of $C_{n,i}$ and the non-negativity of the loss $L_n(\hat{\xi}_{n,C}, \xi_n)$,

$$\begin{aligned}(4.23) \quad & P_{\theta_n,n}[C_{n,i} \ni \xi_n] \\ & = P_{\theta_n,n}\{L_n(\hat{\xi}_{n,C}, \xi_n) \le [\hat{L}_n(\hat{\xi}_{n,C}, \hat{\sigma}_{n,i}^2) + n^{-1/2}\hat{H}_{n,i}^{-1}(\alpha)]_+\} \\ & \ge P_{\theta_n,n}[D_n(\xi_n, X_n, \hat{\sigma}_{n,i}^2) \le \hat{H}_{n,i}^{-1}(\alpha)].\end{aligned}$$

This and (3.13) imply (3.17). By Theorem 2.1, the estimated loss $\hat{L}_n(\hat{\xi}_{n,C}, \hat{\sigma}_{n,i}^2)$ converges in probability to $\rho(A_0)$. Thus, when $\rho(A_0) > 0$, the greater than or equal in (4.23) may be replaced asymptotically by equality, proving (3.16).

From (3.2) and (3.14),

$$(4.24) \quad GL_n(C_{n,i}, \xi_n) = L_n^{1/2}(\hat{\xi}_{n,C}, \xi_n) + [\hat{L}_n(\hat{\xi}_{n,C}, \hat{\sigma}_{n,i}^2) + n^{-1/2}\hat{H}_{n,i}^{-1}(\alpha)]_+^{1/2}.$$

The right side of (4.24) converges to $2\rho^{1/2}(A_0)$ because of Theorem 2.1 and (3.13).

Acknowledgements

An anonymous referee and Lutz Dümbgen made valuable comments on the original version of this paper. The research was supported in part by National Science Foundation grant DMS 9224868.

References

Akaike, H. (1974). A new look at statistical model identification, *IEEE Trans. Automat. Control*, **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**, 582–587.

Beran, R. (1992). The radial process for confidence sets, *Probability in Banach Spaces*, 8 (eds. R. M. Dudley, M. G. Hahn and J. Kuelbs), 479–496, Birkhäuser, Boston.

Billingsley, P. (1968). *Convergence of Probability Measures*, Wiley, New York.

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

Donoho, D. L., Liu, R. C. and MacGibbon, B. (1990). Minimax risk over hyperrectangles and implications, *Ann. Statist.*, **18**, 1416–1437.

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. on Math. Statist. Prob.*, Vol. 1, 361–380, University of California Press, Berkeley.

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

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

Pötscher, B. M. (1991). Effects of model selection on inference, *Econom. Theory*, **7**, 163–185.

Pötscher, B. M. (1995). Comment on "The effect of model selection on confidence regions and prediction regions" by P. Kabaila, *Econom. Theory*, **11**, 550–559.

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. on Math. Statist. Prob.*, Vol. 1 (ed. J. Neyman), 197–206, University of California Press, Berkeley.

Stein, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs, *Research Papers in Statistics: Festschrift for J. Neyman* (ed. F. N. David), 351–366, Wiley, New York.