Calibrating Prediction Regions

Rudolf Beran*

University of California, Berkeley

Technical Report No. 183
November 1988

*This research was supported in part by NSF Grant DMS 87-01426.
Part of the work was done while the author was a guest of Sonderforschungsbereich 123
at Universität Heidelberg.
The author thanks G. Sawitzki and F. Seillier for helpful comments.

Department of Statistics
University of California
Berkeley, California
Calibrating Prediction Regions

Rudolf Beran*
University of California, Berkeley

ABSTRACT

Suppose the variable $X$ to be predicted and the learning sample $Y_n$ that was observed have a joint distribution, which depends on an unknown parameter $\theta$. The parameter $\theta$ can be finite or infinite dimensional. A prediction region $D_n$ for $X$ is a random set, depending on $Y_n$, that contains $X$ with prescribed probability $\alpha$. This paper studies methods for controlling simultaneously the conditional coverage probability of $D_n$, given $Y_n$, and the overall (unconditional) coverage probability of $D_n$. The basic construction yields a prediction region $D_n$ which has the following properties in regular models: Both the conditional and overall coverage probabilities of $D_n$ converge to $\alpha$ as the size $n$ of the learning sample increases. The convergence of the former is in probability. Moreover, the asymptotic distribution of the conditional coverage probability about $\alpha$ is typically normal; and the overall coverage probability tends to $\alpha$ at rate $n^{-1}$. Can one reduce the dispersion of the conditional coverage probability about $\alpha$ and increase the rate at which overall coverage probability converges to $\alpha$? Both issues are addressed. The paper establishes a lower bound for the asymptotic dispersion of conditional coverage probability. The paper also shows how to calibrate $D_n$ so as to make its overall coverage probability converge to $\alpha$ at the faster rate $n^{-2}$. This calibration adjustment does not affect the asymptotic distribution or dispersion of the conditional coverage probability, in a first-order analysis. In general, a bootstrap Monte Carlo algorithm accomplishes the calibration of $D_n$. In special cases, analytical calibration is possible.

* This research was supported in part by NSF Grant DMS 87-01426. Part of the work was done while the author was a guest of Sonderforschungsbereich 123 at Universität Heidelberg. The author thanks G. Sawitzki and F. Seillier for helpful comments.
1. INTRODUCTION

Prediction regions are of interest in the following context. The variable $X$ to be predicted and the learning sample $Y_n$ that was observed have a joint distribution $P_{\theta,n}$. The parameter $\theta$ is unknown but is restricted to a finite or infinite dimensional parameter space. A prediction region for $X$ is a random set $D_n = D_n(\alpha, Y_n)$, depending on the learning sample, that contains $X$ with prescribed probability $\alpha$.

Let $P_\theta(\cdot | Y_n)$ denote the conditional distribution of $X$ given $Y_n$. The conditional coverage probability of $D_n$ given $Y_n$ is

$$CP(D_n | Y_n, \theta) = P_\theta \{ X \in D_n | Y_n \}. \quad (1.1)$$

The overall coverage probability of $D_n$ is

$$CP(D_n | \theta) = E_\theta CP(D_n | Y_n, \theta) = P_{\theta,n} \{ X \in D_n \}, \quad (1.2)$$

where the expectation is with respect to the distribution $Q_{\theta,n}$ of $Y_n$. If $CP(D_n | Y_n, \theta)$ equals $\alpha$, exactly or asymptotically, then so does $CP(D_n | \theta)$. The converse may not be true, as shown by the following example.

**Example 1.** Consider the stationary first-order autoregressive model

$$X_i = \theta X_{i-1} + E_i \quad (1.3)$$

where the $\{E_i\}$ are i.i.d. standard normal random variables and $|\theta|$ is no larger than $1 - \varepsilon$, for some small positive $\varepsilon$. Suppose the learning sample is $Y_n = (X_1, \ldots, X_n)$ and the goal is to predict $X = X_{n+1}$. Let $\hat{\theta}_n$ denote the least squares estimate of $\theta$ based on $Y_n$, clipped so that $|\hat{\theta}_n|$ does not exceed $1 - \varepsilon$:

$$\hat{\theta}_n = \max \{ \min \{ \sum_{i=1}^{n-1} X_i X_{i+1}/\sum_{i=1}^{n-1} X_i^2, 1 - \varepsilon \}, -1 + \varepsilon \}. \quad (1.4)$$

Let $\Phi$ denote the standard normal cdf and let $z_\alpha = \Phi^{-1}(\alpha)$. For the one-sided prediction interval

$$D_n = (-\infty, z_\alpha + \hat{\theta}_n X_n], \quad (1.5)$$

the conditional coverage probability is

$$CP(D_n | Y_n, \theta) = \Phi [ z_\alpha + (\hat{\theta}_n - \theta) X_n ] \quad (1.6)$$

because the conditional distribution of $X$ given $Y_n$ is $N(\theta X_n, 1)$ here. Since $\hat{\theta}_n$ is a consistent estimate of $\theta$, both $CP(D_n | Y_n, \theta)$ and $CP(D_n | \theta)$ converge to $\alpha$ as $n$ increases, the former in probability.
On the other hand, for the alternative one-sided prediction interval
\[ D_n = (-\infty, z_{\alpha}(1 - \hat{\theta}_n^2)^{-1/2}], \]  
the overall coverage probability tends to \( \alpha \) as \( n \) increases, because \( X \) has a \( N(0, (1 - \theta^2)^{-1}) \) distribution. However
\[ \text{CP}(D_n|Y_n, \theta) = \Phi[z_{\alpha}(1 - \hat{\theta}_n^2)^{-1/2} - \theta X_n] \]
does not converge to \( \alpha \), because in the limit it is distributed as \( \Phi[(z_\alpha - \theta Z)(1 - \theta^2)^{-1/2}] \), where \( Z \) is a standard normal random variable.

This paper presents methods for controlling simultaneously the overall coverage probability and the conditional coverage probability of a prediction region. Both goals are important. Several recent authors, including Cox and Hinkley (1974), Cox (1975, 1986), Atwood (1984), Bai and Olshen (1988) have emphasized overall coverage probability. Other recent authors, including Guttman (1970), Butler and Rothman (1980), Butler (1982), Stine (1985), have discussed conditional coverage probability as well. Terminology varies greatly in the works just cited.

The main findings of this paper are as follows. For a natural construction of the prediction region \( D_n \), the overall coverage probability \( \text{CP}(D_n|\theta) \) converges to \( \alpha \) at rate \( n^{-1} \) while \( n^{1/2} \{ \text{CP}(D_n|Y_n, \theta) - \alpha \} \) has a normal limit law with mean zero and variance \( \sigma^2(\theta) \). Appropriate calibration of \( D_n \) makes the overall coverage probability tend to \( \alpha \) at rate \( n^{-2} \), without affecting the normal limit law for the centered conditional coverage probability. At the same time, the asymptotic variance \( \sigma^2(\theta) \) can be minimized by the construction, without affecting the \( n^{-1} \) or \( n^{-2} \) rate of convergence for overall coverage probability. Sections 3 and 4 give details.

The calibration operation mentioned above is introduced in section 2. In the simplest examples, the calibration of \( D_n \) can be done exactly. In a larger class of examples, an asymptotic approximation is available for the calibration operation. This approximation is linked to ideas in Cox (1975). In general, a bootstrap Monte Carlo algorithm accomplishes the calibration, without bogging down in complicated algebra.

2. CONSTRUCTIONS

An extension of the classical pivotal method generates a prediction region \( D_n \) whose conditional coverage probability and overall coverage probability both converge to \( \alpha \) as the size of the learning sample increases. Calibration of the critical value of \( D_n \) seeks to reduce error in overall coverage probability without increasing the dispersion of the conditional coverage probability. Monte Carlo approximations are available for the critical values and calibrated critical values of \( D_n \).
2.1 Constructing Dn

Let \( R_n = R_n(X, Y_n) \) be a root for the prediction region — a function of \( X \) and \( Y_n \) which will be referred to a critical value in order to generate the desired prediction region \( D_n \) for \( X \). Let \( A_n(\cdot, \theta, Y_n) \) be the conditional cdf of \( R_n \) given \( Y_n \). Assume this cdf to be continuous. Suppose \( \hat{\theta}_n = \hat{\theta}_n(Y_n) \) is a consistent estimate of \( \theta \) based on the learning sample. The plug-in estimate of \( A_n(\cdot, \theta, Y_n) \) is then \( A_n(\cdot, \hat{\theta}_n, Y_n) \). Define the prediction region \( D_n \) by referring \( R_n \) to the largest \( \alpha \)th quantile of \( A_n(\cdot, \hat{\theta}_n, Y_n) \):

\[
D_n = \{ x : R_n(x, Y_n) \leq A_n^{-1}(\alpha, \hat{\theta}_n, Y_n) \} = \{ x : A_n[R_n(x, Y_n), \hat{\theta}_n, Y_n] \leq \alpha \}. \tag{2.1}
\]

This construction is motivated by prediction interval (1.5) for Example 1.

The conditional coverage probability of \( D_n \) for \( X \) is

\[
CP(D_n|Y_n, \theta) = A_n[A_n^{-1}(\alpha, \hat{\theta}_n, Y_n), \theta, Y_n]. \tag{2.2}
\]

Under regularity conditions, such as those to be discussed in section 3, both \( CP(D_n|Y_n, \theta) \) and \( CP(D_n|\theta) \) converge to \( \alpha \) as \( n \) increases, the former converging in probability.

Example 2. Suppose \( X \) and the elements of \( Y_n = (X_1, \ldots, X_n) \) are i.i.d. \( N(\mu, \sigma^2) \) random variables, the parameter \( \theta = (\mu, \sigma^2) \) being unknown. Let \( \bar{\theta}_n = (\bar{X}_n, s^2_n) \) denote the usual unbiased estimate of \( \theta \) based on \( Y_n \). A classical root for this problem,

\[
R_n(X, Y_n) = (X - \bar{X}_n) / [s_n(1 + n^{-1})^{1/2}], \tag{2.3}
\]

generates, through (2.1), the one-sided prediction interval

\[
D_n = (-\infty, \bar{X}_n + s_n z_\alpha]. \tag{2.4}
\]

It is easily seen, in this case, that both \( CP(D_n|Y_n, \theta) \) and \( CP(D_n|\theta) \) tend to \( \alpha \) as \( n \) increases, the first convergence being in probability. Nevertheless, prediction interval (2.4) is disappointing in two ways. First, the simpler root \( R_n(X, Y_n) = X \) yields the same prediction interval by construction (2.1). Secondly, the classical interval for this problem replaces \( z_\alpha \) in (2.4) with (1 + \( n^{-1} \))\(^{1/2} \) times the \( \alpha \)th quantile of the \( t \)-distribution with \( n - 1 \) degrees of freedom. This substitution produces a prediction interval whose overall coverage probability is exactly \( \alpha \) and whose conditional coverage probability still converges to \( \alpha \) as \( n \) increases. The possibility of reducing the coverage probability error of \( D_n \) in general models is the subject of the next section.
2.2 Calibrating $D_n$

Let $H_n(\cdot, \theta)$ denote the cdf of the transformed root $A_n(R_n, \hat{\theta}_n, Y_n)$. By (1.2) and the second line of (2.1) respectively,

$$\text{CP}(D_n | \theta) = E_{\theta} \text{CP}(D_n | Y_n, \theta)$$

$$= H_n(\alpha, \theta) \quad (2.5)$$

Both expressions for overall coverage probability of $D_n = D_n(\alpha)$ will be used in the sequel. Suppose $H_n(\cdot, \theta)$ is continuous at its $\alpha$th quantiles and $H_n^{-1}(\alpha, \theta)$ denotes the largest $\alpha$th quantile. If $\theta$ were known, replacing prediction region $D_n(\alpha)$ with $D_n[H_n^{-1}(\alpha, \theta)]$ would yield a prediction region whose coverage probability is exactly $\alpha$. Since $\theta$ is unknown but has an estimate $\hat{\theta}_n$, it is natural to consider the calibrated prediction set

$$D_{n,1}(\alpha) = D_n[H_n^{-1}(\alpha, \hat{\theta}_n)]. \quad (2.6)$$

It will be shown in section 3 that $\text{CP}(D_{n,1} | \theta)$ typically converges to $\alpha$ at a faster rate than $\text{CP}(D_n | \theta)$. On the other hand, $\text{CP}(D_{n,1} | Y_n, \theta)$ converges in probability to $\alpha$ at the same rate as $\text{CP}(D_n | Y_n, \theta)$.

**Example 2 (continued).** Both the root $R_n = X$ and the root (2.3) yield the one-sided prediction interval (2.4). By (2.2),

$$\text{CP}(D_n | Y_n, \theta) = \Phi[z_\alpha + \sigma^{-1} \{(s_n - \sigma)z_\alpha + (\bar{X}_n - \mu)\}]$$

$$= \alpha + O_p(n^{-1/2}). \quad (2.7)$$

Let $J_r$ denote the cdf of the t-distribution with $r$ degrees of freedom. Since

$$H_n(\alpha, \theta) = J_{n-1}[z_\alpha (1 + n^{-1})^{-1/2}], \quad (2.8)$$

it follows by the second line of (2.5) that

$$\text{CP}(D_n | \theta) = J_{n-1}[z_\alpha (1 + n^{-1})^{-1/2}]$$

$$= \alpha - (4n)^{-1}(z_\alpha^3 + 3z_\alpha)\phi(z_\alpha) + O(n^{-2}). \quad (2.9)$$

From (2.8),

$$H_n^{-1}(\alpha, \theta) = \Phi[J_{n-1}^{-1}(\alpha) (1 + n^{-1})^{1/2}]. \quad (2.10)$$

Consequently, the calibrated prediction interval is

$$D_{n,1}(\alpha) = (-\infty, \bar{X}_n + s_n (1 + n^{-1})^{1/2} J_{n-1}^{-1}(\alpha)], \quad (2.11)$$

the classical answer in this situation. Clearly $\text{CP}(D_{n,1} | \theta) = \alpha$ and

$$\text{CP}(D_{n,1} | Y_n, \theta) = \alpha + O_p(n^{-1/2}) \quad (2.12)$$
by extending the reasoning for (2.7). In fact, both \( n^{1/2} \left[ CP(D_n|Y_n, \theta) - \alpha \right] \) and \( n^{1/2} \left[ CP(D_{n,1}|Y_n, \theta) - \alpha \right] \) converge weakly to the same normal limit distribution. This phenomenon will be explained more generally in section 4.

2.3 Monte Carlo Approximations

In a few cases, such as Example 2, \( D_{n,1} \) can be constructed analytically. In a larger class of problems, it is possible to approximate \( D_{n,1} \) by using asymptotic expansions for \( H_n(\cdot, \hat{\theta}_n) \) and, if necessary, for \( A_n(\cdot, \hat{\theta}_n, Y_n) \). The most general approach approximates the first or both of these cdf's by Monte Carlo methods. Algorithm 1 below assumes that \( A_n(\cdot, \hat{\theta}_n, Y_n) \) can be found analytically and gives an approximation for \( H_n(\cdot, \hat{\theta}_n) \). Algorithm 2 provides approximations for both cdf's. Both algorithms are bootstrap algorithms in the sense that they rely on sampling from fitted models.

The following representations for \( H_n(\cdot, \hat{\theta}_n) \) and \( A_n(\cdot, \hat{\theta}_n, Y_n) \) underlie the two algorithms. Let \( \bar{X} \) be a random variable whose conditional distribution, given \( Y_n \), is \( P_{\theta_n}(\cdot|Y_n) \). Then

\[
A_n(x, \hat{\theta}_n, Y_n) = \Pr[R_n(\bar{X}, Y_n) \leq x|Y_n].
\] (2.13)

Let \((X^*, Y_n^*)\) be random variables whose conditional joint distribution, given \( Y_n \), is \( P_{\theta_n^n} \). Let \( \theta_n^* = \hat{\theta}_n(Y_n^*) \). Then

\[
H_n(x, \hat{\theta}_n) = \Pr[A_n(R_n(X^*, Y_n^*), \theta_n^*, Y_n^*) \leq x|Y_n].
\] (2.14)

Let \( \bar{X}^* \) be a random variable whose conditional distribution, given \( X^* \), \( Y_n^* \) and \( Y_n \), is \( P_{\theta_n^*}(\cdot|Y_n^*) \). In view of (2.13),

\[
A_n[R_n(X^*, Y_n^*), \theta_n^*, Y_n^*] = \Pr[R_n(\bar{X}^*, Y_n^*) \leq R_n(X^*, Y_n^*)|X^*, Y_n^*, Y_n].
\] (2.15)

Representation (2.14) is the basis for the bootstrap Algorithm 1 below. Representations (2.13), (2.14) and (2.15) are the foundation for the double bootstrap Algorithm 2.

**Algorithm 1.** \( (A_n \text{ is known, } H_n \text{ to be found}) \). Draw \( J \) bootstrap samples \( \{(X^*_j, Y_n^*_j): 1 \leq j \leq J\} \) from the fitted distribution \( P_{\theta_n^n} \) for \((X, Y_n)\). These bootstrap samples are conditionally independent, given \( Y_n \). For each \( j \), calculate \( \theta_n^*_j = \hat{\theta}_n(Y_n^*_j) \) and \( R_n^*_j = R_n(X^*_j, Y_n^*_j) \). The empirical cdf of the values \( \{A_n(R_{n,j}^*, \theta_n^*_j, Y_n^*_j): 1 \leq j \leq J\} \) approximates \( H_n(\cdot, \hat{\theta}_n) \) for sufficiently large \( J \).

**Algorithm 2.** \( (\text{Both } A_n \text{ and } H_n \text{ to be found}) \). Draw \( K \) bootstrap variables \( \{\bar{X}_k: 1 \leq k \leq K\} \) from the fitted conditional distribution \( P_{\theta_n}(\cdot|Y_n) \) for \( X \). These bootstrap variables are conditionally independent, given \( Y_n \). The empirical cdf of the values
\{R_n(\tilde{X}_k, Y_n) : 1 \leq k \leq K\} approximates A_n(\cdot, \hat{\theta}_n, Y_n) for sufficiently large K.

Draw J bootstrap samples \{(X^*_j, Y^*_j) : 1 \leq j \leq J\} as in Algorithm 1 and calculate \hat{\theta}^*_n = \hat{\theta}_n(Y^*_n) for each j. Then, for each j, draw K bootstrap variables \{\tilde{X}^*_{k,j} : 1 \leq k \leq K\} from the fitted conditional distribution \(P_{\hat{\theta}^*_n}(\cdot|Y^*_n)\). These bootstrap variables are conditionally independent given \(Y_n\) and the \{(X^*_j, Y^*_j)\}. Let \(Z_j\) be the proportion of the values \(\{R_n(\tilde{X}^*_{k,j}, Y^*_n) : 1 \leq k \leq K\}\) which are less than or equal to \(R_n(X^*_j, Y^*_n)\). The empirical cdf of the \(\{Z_j : 1 \leq j \leq J\}\) approximates \(H_n(\cdot, \hat{\theta}_n)\) for sufficiently large \(J\) and \(K\).

A SUN 3/140 workstation suffices to carry out Algorithms 1 and 2 for \(J\) and \(K\) near 1000 and modest sample size \(n\). Both algorithms rely on simple random sampling with replacement. More efficient algorithm based on importance sampling are likely future developments.

**Example 1 (continued).** Table 1 reports some results from a simulation study of three prediction intervals in the gaussian autoregressive model (1.3). For the root \(R_n = X\), the prediction region \(D_n\) defined by (2.1) is simply the one-sided prediction interval (1.5). The corresponding conditional cdf \(A_n(x, \hat{\theta}_n, Y_n)\) has the analytical expression (3.5). Algorithm 1 produced the bootstrap approximation to \(H_n^{-1}(\alpha, \hat{\theta}_n)\) that is needed to construct the calibrated prediction interval \(D_{n,1}\). The third prediction interval considered was

\[
D_{n}^{\prime}(\alpha) = D_n[\alpha + (2n)^{-1}z_{\alpha} \phi(z_{\alpha})].
\] (2.15)

This interval is an analytical approximation to \(D_{n,1}\) for this example, as will be explained in section 4.

In the study comparing \(D_n\), \(D_{n,1}\) and \(D_{n}^{\prime}\), the value of \(\alpha\) was .90 and \(\hat{\theta}_n\) was defined by (1.4) with \(\varepsilon\) being \(10^{-6}\). The numerical entries in Table 1 support several conclusions:

(a) For each \(\theta\), the overall coverage probability of \(D_{n,1}\) or \(D_{n}^{\prime}\) is closer to \(\alpha\) than is the overall coverage probability of \(D_n\). All three overall coverage probabilities converge rapidly to \(\alpha\) as the size \(n\) of the learning sample increases.

(b) For each \(\theta\), the standard deviation of the conditional coverage probability is nearly the same for \(D_n\), \(D_{n,1}\) and \(D_{n}^{\prime}\). All three standard deviations tend to zero as \(n\) increases.

Asymptotics supporting these conclusions are developed in the next two sections. Of course, the asymptotics do not fully explain the remarkably good small sample performance of \(D_{n}^{\prime}\) and \(D_{n,1}\) in Example 1.
### Table 1. Overall Coverage Probabilities and Standard Deviations of Conditional Coverage Probabilities for $D_n$, $D_{n,1}$ and $D_n'$ in Example 1.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\theta$</th>
<th>$D_n$</th>
<th>$D_{n,1}$</th>
<th>$D_n'$</th>
<th>S.D. of CP ($\cdot\mid Y_n, \theta$) for $D_n$</th>
<th>S.D. of CP ($\cdot\mid Y_n, \theta$) for $D_{n,1}$</th>
<th>S.D. of CP ($\cdot\mid Y_n, \theta$) for $D_n'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.</td>
<td>.850</td>
<td>.881</td>
<td>.891</td>
<td>.169</td>
<td>.164</td>
<td>.149</td>
</tr>
<tr>
<td></td>
<td>-.1</td>
<td>.850</td>
<td>.881</td>
<td>.891</td>
<td>.169</td>
<td>.164</td>
<td>.149</td>
</tr>
<tr>
<td></td>
<td>-.5</td>
<td>.856</td>
<td>.887</td>
<td>.897</td>
<td>.150</td>
<td>.144</td>
<td>.131</td>
</tr>
<tr>
<td></td>
<td>-.9</td>
<td>.878</td>
<td>.904</td>
<td>.918</td>
<td>.102</td>
<td>.092</td>
<td>.088</td>
</tr>
<tr>
<td>5</td>
<td>0.</td>
<td>.874</td>
<td>.893</td>
<td>.897</td>
<td>.122</td>
<td>.119</td>
<td>.112</td>
</tr>
<tr>
<td></td>
<td>-.1</td>
<td>.874</td>
<td>.894</td>
<td>.897</td>
<td>.121</td>
<td>.118</td>
<td>.112</td>
</tr>
<tr>
<td></td>
<td>-.5</td>
<td>.875</td>
<td>.893</td>
<td>.899</td>
<td>.113</td>
<td>.110</td>
<td>.103</td>
</tr>
<tr>
<td></td>
<td>-.9</td>
<td>.885</td>
<td>.898</td>
<td>.908</td>
<td>.081</td>
<td>.076</td>
<td>.072</td>
</tr>
</tbody>
</table>

**NOTE:** The intended coverage probability is .90. $D_{n,1}$ uses 999 bootstrap samples. The table entries are calculated from 2500 Monte Carlo trials.
3. CONVERGENCE OF COVERAGE PROBABILITIES

This section establishes conditions under which the conditional and overall coverage probabilities of $D_n$ and $D_{n,1}$ converge to $\alpha$, as the size of the learning sample increases. The first restriction is to assume that

$$\text{CP} (D_n | Y_n, \theta) = C(\alpha, \theta, \hat{\theta}_n, U_n), \quad (3.1)$$

where the function $C$ does not depend on $n$ and $U_n = U_n(Y_n)$ is a statistic. This situation arises frequently in examples, as will be shown later in the section. Suppose the parameter space is metric. Let $C(\theta)$ denote the set of sequences $\{\theta_n: n \geq 1\}$ in the parameter space such that $\theta_n$ converges to $\theta$. Let $L(U_n|\theta)$ stand for the distribution of $U_n$ under $Q_\theta$.

**Proposition 1.** Suppose (3.1) holds, $C(\alpha, \theta, t, u)$ is continuous in $(\theta, t, u)$ at points where $t = \theta$, and $C(\alpha, \theta, \theta, u) = \alpha$ for all possible values of $(\theta, u)$. Suppose that for every sequence $\{\theta_n\}$ in $C(\theta)$, $\hat{\theta}_n \to \theta$ in $Q_{\theta_n}$ probability and $\{L(U_n|\theta_n)\}$ is tight. Then, for every sequence $\{\theta_n\}$ in $C(\theta)$,

$$\text{CP} (D_n | Y_n, \theta_n) \to \alpha \text{ in } Q_{\theta_n} \text{ probability} \quad (3.2)$$

and

$$\text{CP} (D_n | \theta_n) \to \alpha. \quad (3.3)$$

These conclusions also extend to $D_{n,1}$ if $C(\alpha, \theta, t, u)$ is continuous in $(\alpha, \theta, t, u)$ at points where $t = \theta$.

The proof of this result is in section 6. The uniformity of convergences (3.2) and (3.3) over compact subsets of the parameter space enhances their trustworthiness. The following examples illustrate the scope of Proposition 1.

**Example 1 (continued).** Suppose the root for prediction in this autoregressive model is

$$R_n(X, Y_n) = X. \quad (3.4)$$

The conditional cdf of $R_n$ given $X_n$ is

$$A_n(x, \theta, Y_n) = \Phi(x - \theta X_n). \quad (3.5)$$

Let $\hat{\theta}_n$ be the consistent estimate of $\theta$ given in (1.4). The prediction region $D_n$ defined by (2.1) coincides with the one-sided prediction interval already described in (1.5). Proposition 1 is applicable with $U_n = X_n$ and
\[ C(x, \theta, t, u) = \Phi[ \Phi^{-1}(x) + (t - \theta) u ]. \quad (3.6) \]

A two-sided prediction interval for \( X \) can be obtained in two ways: Combine the one-sided interval (1.5) with the analogous one-sided interval based on the root \( -X \), each of nominal coverage probability \( (1 + \alpha) / 2 \); or use in (2.1) the root
\[ R_n(X, Y_n) = |X - \hat{\theta}_n X_n| \quad (3.7) \]
where \( \hat{\theta}_n X_n \) is the usual point predictor of \( X \). Let \( \Psi(x) = \Phi(x) - \Phi(-x) \) be the cdf of the folded-over standard normal distribution. Because the model is gaussian, both methods just described yield the prediction interval
\[ D_n = [ \hat{\theta}_n X_n - \Psi^{-1}(\alpha), \hat{\theta}_n X_n + \Psi^{-1}(\alpha) ] \quad (3.8) \]
Proposition 1 is again applicable, this time with \( U_n = X_n \) and
\[ C(x, \theta, t, u) = \Phi[ \Psi^{-1}(x) + (t - \theta) u ] - \Phi[ -\Psi^{-1}(x) + (t - \theta) u ]. \quad (3.9) \]

**Example 3.** Suppose the learning sample \( Y_n = (X_1, \ldots, X_n) \) follows the linear model
\[ X_i = \beta c_i + E_i, \quad (3.10) \]
where the \( \{c_i\} \) are known constants. The goal is to predict
\[ X = \beta c + E, \quad (3.11) \]
given \( Y_n \) and the assumption that the errors \( E \), \( \{E_i\} \) are i.i.d. with continuous cdf \( F \), which has mean zero and finite variance. The unknown parameter \( \theta = (\beta, F) \) is estimated by \( \hat{\theta}_n = (\hat{\beta}_n, \hat{F}_n) \), where \( \hat{\beta}_n \) is the least squares estimate of \( \beta \) and \( \hat{F}_n \) is the empirical cdf of the residuals \( \{X_i - \hat{\beta}_n c_i; 1 \leq i \leq n\} \)

The root
\[ R_n(X, Y_n) = X \quad (3.12) \]
in (2.1) generates the one-sided prediction interval
\[ D_n = (-\infty, \hat{F}_n^{-1}(\alpha) + \hat{\beta}_n c] \quad (3.13) \]
for \( X \). Define the distance between \( \theta = (\beta, F) \) and the element \( t = (b, G) \) of the parameter space by
\[ d(t, \theta) = |b - \beta| + m(G, F), \quad (3.14) \]
where \( m \) is bounded Lipschitz distance. Assume \( \max_i c_i^2 / \Sigma_i c_i^2 \) tends to zero and \( \Sigma_i c_i^2 \) tends to infinity as \( n \) increases. Then the estimate \( \hat{\theta}_n \) described above is consistent for \( \theta \). Proposition 1 is applicable with
10

\[ C(x, \theta, t) = F[G^{-1}(x) + (b - \beta)c]. \quad (3.15) \]

The continuity required of the function \( C \) is immediate here because weak convergence of cdf's to a continuous cdf implies uniform convergence.

The root

\[ R_n(X, Y_n) = |X - \hat{\beta}_n c| \quad (3.16) \]

in (2.1) yields the two-sided prediction interval

\[ D_n = [\hat{\beta}_n c - \hat{G}_n^{-1}(\alpha), \hat{\beta}_n c + \hat{G}_n^{-1}(\alpha)], \quad (3.17) \]

where \( \hat{G}_n(x) = \hat{F}_n(x) - \hat{F}_n(-x) \) is the estimated cdf of \( |E| \). The application of Proposition 1 to this case is similar to that of the previous paragraph.

4. FURTHER ASYMPTOTICS

The asymptotic distributions of \( CP(D_n | Y_n, \theta) \) or \( CP(D_n,1 | Y_n, \theta) \) and the rates of convergence to \( \alpha \) of \( CP(D_n | \theta) \) or \( CP(D_n,1 | \theta) \) are the topics of this section. Attention is focused on the simplest case: the learning sample \( Y_n \) and the variable \( X \) to be predicted are independent. The distribution of \( Y_n \) is \( Q_{\theta,n} \) and the distribution of \( X \) is \( P_\theta \). The prediction root has the form \( R(X, \hat{\theta}_n) \), where \( \hat{\theta}_n = \hat{\theta}_n(Y_n) \) is a consistent estimate of \( \theta \). The parameter space in the exposition will be an open subset of the real line. The extension to Euclidean parameter spaces is straightforward, but requires heavier notation.

In this setting, the conditional cdf of the root \( R \) given \( Y_n \) is

\[ A(x, \theta, \hat{\theta}_n) = P_\theta[R(X, \hat{\theta}_n) \leq x], \quad (4.1) \]

where \( \hat{\theta}_n \) is held fixed on the right side. The conditional coverage probability, given \( Y_n \), of the prediction region \( D_n \) generated by (2.1) is thus a function \( C(\alpha, \theta, \hat{\theta}_n) \), which is explicitly \( A[A^{-1}(\alpha, \hat{\theta}_n, \hat{\theta}_n), \theta, \hat{\theta}_n] \). Note that \( C(\alpha, \theta, t) \) is a cdf in its first argument. It will be assumed throughout that

\[ C(\alpha, \theta, \theta) = \alpha \quad (4.2) \]

for every possible \( \theta \). This occurs, in particular, when \( A(x, \theta, t) \) is continuous in \( x \).

4.1 Coverage Probabilities of \( D_n \)

The asymptotic behavior of \( CP(D_n | Y_n, \theta) \) and \( CP(D_n | \theta) \) can be studied by developing expansions for these two quantities. In the discussion, notation like \( f^{(i,j,k)}(x, \theta, t) \) represents the partial derivative \( \partial^{i+j+k}f(x, \theta, t)/\partial x^i \partial \theta^j \partial t^k \). Let \( C_1(\theta) \) denote the class of all sequences \( \{\theta_n\} \) in the parameter space such that \( \{n^{1/2}(\theta_n - \theta)\} \) converges to a finite limit. Let \( <x> \) denote the integer part of \( x \). The following
assumptions describe the leading case.

**Assumption A** \((r)\). For \(1 \leq j \leq r\) and \(r\) an even integer, there exist functions \(\{a_{j,k}(\theta)\}\) such that, as \(n\) increases,

\[
n^{r/2}[E_{\theta_n}(\hat{\theta}_n - \theta_n)^j - \sum_{k=<(j+1)/2>}^{r/2} n^{-k}a_{j,k}(\theta_n)] \to 0 \quad (4.3)
\]

whenever \(\{\theta_n\}\) belongs to \(C_1(\theta)\).

**Assumption B.** If \(\{\theta_n\}\) belongs to \(C_1(\theta)\), then the distribution of \(n^{1/2}(\hat{\theta}_n - \theta_n)\) under \(Q_{\theta_n}\) converges weakly to the \(N(0, a_{2,1}(\theta))\) distribution.

These assumptions are satisfied, for example, when \(\{Q_{\theta_n}\}\) is a smoothly parametrized exponential family and \(\hat{\theta}_n\) is the maximum likelihood estimate of \(\theta\). In this situation, \(\hat{\theta}_n\) is a smooth function of a sample mean and (4.3) follows by a Taylor expansion argument (compare Theorem 5.1 on p. 101 of Lehmann, 1983). The asymptotic normality of Assumption B is a consequence of the central limit theorem.

**Proposition 2A.** Suppose Assumption B and (4.2) hold. Suppose \(C^{(0,0,1)}(\alpha, \theta, t)\) exists and is continuous in \((\theta, t)\) at points where \(t = \theta\). Then, for every sequence \(\{\theta_n\}\) in \(C_1(\theta)\),

\[
L \left[ n^{1/2} \left\{ CP(D_n|Y_n, \theta_n) - \alpha \right\} | \theta_n \right] \Rightarrow N(0, \sigma^2(\theta)) \quad (4.4)
\]

where

\[
\sigma^2(\theta) = [C^{(0,0,1)}(\alpha, \theta, \theta)]^2 a_{2,1}(\theta). \quad (4.5)
\]

In the extension of Proposition 2A to vector parameter \(\theta\), \(C^{(0,0,1)}(\alpha, \theta, t)\) is a column vector, \(a_{2,1}(\theta)\) is a matrix, and

\[
\sigma^2(\theta) = [C^{(0,0,1)}(\alpha, \theta, \theta)]' a_{2,1}(\theta) [C^{(0,0,1)}(\alpha, \theta, \theta)] \quad (4.6)
\]

**Proposition 2B.** Suppose Assumption A(2) and (4.2) hold, and \(a_{2,1}(\theta)\) is continuous in \(\theta\). Suppose \(C^{(0,0,2)}(\alpha, \theta, t)\) exists, is bounded in \((\theta, t)\), and is continuous in \((\theta, t)\) at points where \(t = \theta\). Then, for every sequence \(\{\theta_n\}\) in \(C_1(\theta)\),

\[
n \left[ CP(D_n|\theta_n) - \alpha - n^{-1} b_1(\alpha, \theta_n) \right] \to 0 \quad (4.7)
\]

where
Propositions 2A and 2B are proved in Section 6. The following example illustrates their content.

**Example 2 (continued).** In this example, the parameter $\theta = (\mu, \sigma^2)$ is two-dimensional, $t = (m, s^2)$ and $C(\alpha, \theta, t)$ is

$$C(\alpha, \theta, t) = \Phi \left[ z_{\alpha} + \sigma^{-1} \left( (s - \sigma) z_{\alpha} + (m - \mu) \right) \right]$$

where

$$\delta = \sigma^{-1} \left[ (s - \sigma) z_{\alpha} + (m - \mu) \right]$$

and $\bar{z}$ lies between $z_{\alpha}$ and $z_{\alpha} + \delta$. The reasoning for Proposition 2A and 2B applies, with $\hat{\theta}_n = (\bar{X}_n, \bar{s}_n^2)$. Thus, $CP(D_n \mid Y_n, \theta_n) = C(\alpha, \theta_n, \hat{\theta}_n)$ is asymptotically normal as in (4.4) with

$$\sigma^2(\theta) = (2^{-1} z_{\alpha}^2 + 1) \phi^2(z_{\alpha}).$$

Moreover, taking expectations through the expansion for $CP(D_n \mid Y_n, \theta_n)$ implied by (4.9) establishes conclusion (4.7) with

$$b_1(\alpha, \theta) = -4^{-1} z_{\alpha} \phi(z_{\alpha}) + 2^{-1} (2^{-1} z_{\alpha}^3 + 1) \phi'(z_{\alpha})$$

$$= -4^{-1} (z_{\alpha}^3 + z_{\alpha}) \phi(z_{\alpha}).$$

This argument is an alternative derivation for (2.9).

The next two examples illustrate what can happen in situations where the reasoning for Proposition 2A and 2B does not apply, in some respect.

**Example 1 (continued).** For root (3.4) in this AR(1) model, the expression (3.6) entails

$$CP(D_n \mid Y_n, \theta_n) = \Phi \left[ z_{\alpha} + (\hat{\theta}_n - \theta_n) X_n \right]$$

where $\delta_n$ represents $(\hat{\theta}_n - \theta_n) X_n$ and $z_n$ lies between $z_{\alpha}$ and $z_{\alpha} + \delta_n$. Proposition 2A does not apply to this example. Indeed, if $\{\theta_n\}$ lies in $C_1(\theta)$, then

$$n^{1/2} \left[ CP(D_n \mid Y_n, \theta_n) - \alpha \right]$$

converges weakly to the product of two independent $N(0, 1)$ random variables and $\phi(z_{\alpha})$. 

\[
b_1(\alpha, \theta) = \sum_{j=1}^{2} a_{j,1}(\theta) C^{(0,0,j)}(\alpha, \theta, \theta) / j!
\]

(4.8)
On the other hand, taking expectations through (4.13) establishes

$$CP(D_n|\theta_n) = \alpha - (2n)^{-1} z_\alpha \phi(z_\alpha) + o(n^{-1})$$  \hspace{1cm} (4.14)

By symmetry, the expectation of $\delta_n$ is zero in this gaussian model. While Proposition 2B is not applicable to this example, conclusion (4.14) is analogous to (4.7).

**Example 4.** Suppose $X$ and the elements of $Y = (X_1, \ldots, X_n)$ are iid random variables with unknown continuous cdf $F$. This cdf, which is the parameter $\theta$ here, is estimated by the empirical cdf. Let $X_{(1)} \leq \cdots \leq X_{(n)}$ denote the order statistics of the learning sample. Define $X_{(n+1)}$ to be $\infty$. From the root $R_n = X$, definition (2.1) generates the one-sided prediction interval

$$D_n = (-\infty, X_{<\alpha n+1>}]$$ \hspace{1cm} (4.15)

where $<\cdot>$ is the integer part function. Evidently, $CP(D_n|Y_n, \theta)$ equals $F(X_{<\alpha n+1>})$ for every continuous $F$. Since the $i$th uniform order statistic has a Beta(i, n - i + 1) distribution, it follows that

$$n^{1/2} [CP(D_n|Y_n, \theta) - \alpha] \Rightarrow N(0, \alpha(1 - \alpha))$$ \hspace{1cm} (4.16)

and

$$CP(D_n|\theta) = (n + 1)^{-1} <\alpha n + 1> = \alpha + O(n^{-1})$$ \hspace{1cm} (4.17)

Both convergences are uniform over all continuous $F$. These conclusions parallel Propositions 2A and 2B, though for different reasons.

Proposition 2B has several implications and extensions:

(a) Two plausible estimates for $CP(D_n|\theta)$ are the plug-in estimate $CP(D_n|\hat{\theta}_n)$ and the naive estimate $\hat{CP} = \alpha$. Under the assumptions of Proposition 2B,

$$n [CP(D_n|\hat{\theta}_n) - \alpha - n^{-1} b_1(\alpha, \hat{\theta}_n)] \rightarrow 0$$ \hspace{1cm} (4.18)

in $Q_{\theta_n}$ probability. Consequently,

$$n [CP(D_n|\hat{\theta}_n) - CP(D_n|\theta_n)] \rightarrow 0$$ \hspace{1cm} (4.19)

in $Q_{\theta_n}$ probability, provided $b_1(\alpha, \theta)$ is continuous in $\theta$. By contrast, the naive estimate satisfies

$$n [\hat{CP} - CP(D_n|\theta_n)] \rightarrow -b_1(\alpha, \theta),$$ \hspace{1cm} (4.20)

and so is less efficient than the plug-in estimate. This fact and (2.5) are the bases for the calibration procedure which generates $D_{n,1}$ from $D_n$. 

(b) Suppose $C^{(0,0,2)}(\alpha, \theta, t)$ is bounded in all three arguments and continuous at points where $t = \theta$, in addition to the assumptions of Proposition 2B. Suppose also that $b_1(\alpha, \theta)$ is continuous in both arguments. Then

$$n \left[ H_n^{-1}(\alpha, \hat{\theta}_n) - \alpha + n^{-1} b_1(\alpha, \hat{\theta}_n) \right] \to 0$$

(4.21)

in $Q_{\theta_m}$ probability, for every sequence $\{\theta_n\}$ in $C_1(\theta)$.

(c) Higher order expansions for $H_n(\alpha, \theta)$ and $H_n^{-1}(\alpha, \theta)$ exist under stronger regularity conditions. For example, suppose Assumption A(4) and (4.2) hold, $a_{4,2}(\theta)$ is continuous, and $C^{(0,0,4)}(\alpha, \theta, t)$ is bounded in all three arguments and is continuous at points where $t = \theta$. Then, for every convergent sequence $\{\alpha_n\}$ in the unit interval and every sequence $\{\theta_n\}$ in $C_1(\theta)$,

$$n^2 \left[ H_n(\alpha_n, \theta_n) - \alpha_n - \sum_{i=1}^{2} n^{-i} b_1(\alpha_n, \theta_n) \right] \to 0$$

(4.22)

where $b_1$ is given by (4.8) and

$$b_2(\alpha, \theta) = \sum_{j=1}^{4} a_{j,2}(\theta) C^{(0,0,j)}(\alpha, \theta, \theta) / j!$$

(4.23)

Moreover, if $b_2(\alpha, \theta)$ and the derivative $b_1^{(1,0)}(\alpha, \theta)$ are continuous in both arguments, then

$$n^2 \left[ H_n^{-1}(\alpha, \hat{\theta}_n) - \alpha - \sum_{i=1}^{2} n^{-i} c_i(\alpha, \hat{\theta}_n) \right] \to 0$$

(4.24)

in $Q_{\theta_m}$ probability, where

$$c_1(\alpha, \theta) = -b_1(\alpha, \theta)$$

$$c_2(\alpha, \theta) = b_1(\alpha, \theta) b_1^{(1,0)}(\alpha, \theta) - b_2(\alpha, \theta).$$

(4.25)

4.2 Coverage Probabilities of $D_{n,1}$

From the definition (2.6) of $D_{n,1}$, it is immediate that

$$CP(D_{n,1} | Y_n, \theta) = C \left[ H_n^{-1}(\alpha, \hat{\theta}_n), \theta, \hat{\theta}_n \right].$$

(4.26)

Consequently, the behavior of $H_n^{-1}(\alpha, \hat{\theta}_n)$ strongly influences the conditional and overall coverage probabilities of $D_{n,1}$. The discussion surrounding (4.21) and (4.24) motivates the assumptions made about $H_n^{-1}(\alpha, \hat{\theta}_n)$ in this section.

**Proposition 3A.** Suppose Assumption B and (4.2) hold. Suppose $C^{(0,0,1)}(\alpha, \theta, t)$ exists and is continuous in all three arguments at points where $t = \theta$. Suppose that

$$n \left[ H_n^{-1}(\alpha, \hat{\theta}_n) - \alpha \right] \to 0$$

(4.27)
in $Q_{\theta_n}$ probability for every sequence $\{\theta_n\}$ in $C_1(\theta)$. Then

$$L_n \{ n^{1/2} \{ CP(D_{n,1} | Y_n, \theta_n) - \alpha \} | \theta_n \} \Rightarrow N(0, \sigma^2(\theta))$$ (4.28)

where $\sigma^2(\theta)$ is defined by (4.5).

Comparing Propositions 2A and 3A reveals that both $CP(D_n | Y_n, \theta_n)$ and $CP(D_{n,1} | Y_n, \theta_n)$ have the same normal limiting distribution whose asymptotic variance is proportional to the asymptotic variance of $\hat{\theta}_n$. Thus, the estimate $\hat{\theta}_n$ should be asymptotically efficient to minimize the asymptotic dispersion about $\alpha$ of $CP(D_n | Y_n, \theta)$ or $CP(D_{n,1} | Y_n, \theta)$. This statement can be formalized in a lower asymptotic minimax bound and in a convolution theorem, which will be presented elsewhere.

As the next result shows, calibration of $D_n$ achieves a bias correction of order $n^{-1}$ in the conditional coverage probability of $D_n$. Let

$$r_n(\alpha, \theta) = n^2 \{ H_n^{-1}(\alpha, \theta) - \alpha + n^{-1} b_1(\alpha, \theta) \}$$ (4.29)

for $b_1$ defined as in (4.8).

**Proposition 3B.** Suppose Assumption A(2) and (4.2) hold. Suppose $C^{(1,0,0)}(\alpha, \theta, t)$, $C^{(1,0,1)}(\alpha, \theta, t)$, $C^{(2,0,0)}(\alpha, \theta, t)$, $b^{(0,1)}(\alpha, \theta)$, $b^{(0,2)}(\alpha, \theta)$ exist and are bounded in all of their arguments at points where $t = \theta$. Suppose that for every sequence $\{\theta_n\}$ in $C_1(\theta)$,

$$\limsup_{n \to \infty} r_n(\alpha, \theta_n) < \infty$$ (4.30)

and

$$\limsup_{n \to \infty} E_{\theta_n} | r_n(\alpha, \hat{\theta}_n) | < \infty.$$ (4.31)

Then

$$CP(D_{n,1} | \theta_n) = \alpha + O(n^{-2}).$$ (4.32)

Propositions 3A and 3B are proved in Section 6. Some extensions are possible:

(a) A plausible analytical approximation to $D_{n,1}$ is the prediction region

$$D_n'(\alpha) = D_n[\alpha - n^{-1} b_1(\alpha, \hat{\theta}_n)].$$ (4.33)

If $\{\theta_n\}$ is any sequence in $C_1(\theta)$, it follows from the proofs for Propositions 3B and 3A that

$$CP(D_n' | \theta_n) = \alpha + O(n^{-2})$$ (4.34)
and

\[ L \left[ n^{1/2} \{ \text{CP} (D_n' | Y_n, \theta_n) - \alpha \} | \theta_n \right] \Rightarrow N(0, \sigma^2(\theta)), \quad (4.35) \]

like prediction region \( D_{n,1} \). Nevertheless, as shown below in the discussion of example 2, prediction regions \( D_{n,1} \) and \( D_n' \) can still differ significantly. The analytical calibration idea that underlies \( D_n' \) is related to Cox (1975).

(b) Calibration of \( D_n \) can be iterated. Let \( H_n(a, \theta) = \text{CP} \left[ D_{n,1} (\alpha) | \theta \right] \) and define

\[
D_{n,2}(\alpha) = D_{n,1} \left[ H_n^{-1}(\alpha, \hat{\theta}_n) \right] = D_n \left[ H_n^{-1}(H_n^{-1}(\alpha, \hat{\theta}_n), \hat{\theta}_n) \right].
\]

Under stronger regularity conditions, the conclusion of Proposition 3A extends to \( \text{CP} \left( D_{n,2} | Y_n, \theta_n \right) \) and

\[ \text{CP} (D_{n,2} | \theta_n) = \alpha + O(n^{-3}). \quad (4.37) \]

In other words, \( D_{n,2} \) achieves a second-order bias correction to \( \text{CP} (D_n | Y_n, \theta_n) \). A nested double bootstrap Monte Carlo algorithm can be used to approximate \( H_n(\alpha, \hat{\theta}_n) \) and so the critical value of \( D_{n,2} \).

Thus, for all sufficiently large \( n \), the coverage probability of \( D_{n,2} \) is more accurate than that of \( D_{n,1} \), which in turn is more accurate than the coverage probability of \( D_n \). For \textit{fixed} \( n \), calibration and iterated calibration may or may not improve coverage probability. Examples 1 and 2, the numerical study reported in section 2.3, and experience with asymptotic corrections in other problems suggest that at least one round of calibration is often worthwhile. On the other hand, the discussion below of example 4 shows that regularity conditions are essential for the success of calibration.

\textbf{Example 2 (continued).} In this normal location-scale model, the function \( b_1(\alpha, \theta) \) is given by (4.12). Let

\[ \alpha_n = \alpha + (4n)^{-1} (z_\alpha^2 + 3z_\alpha) \phi(z_\alpha) \]

so that \( D_n' \) is just \( D_n(\alpha_n) \). From (2.9) it is immediate that

\[
\text{CP} (D_n' | \theta) = \alpha_n - 4^{-1} n^{-1} (z_{\alpha_n}^3 + 3z_{\alpha_n}) \phi(z_{\alpha_n}) + O(n^{-2})
\]

\[ = \alpha + O(n^{-2}) \quad (4.39) \]

uniformly in \( \theta \), as expected. However, in this example \( \text{CP} (D_{n,1} | \theta) \) equals \( \alpha \) exactly, as discussed after (2.11). Both \( n^{1/2} \left[ \text{CP} (D_n' | Y_n, \theta) - \alpha \right] \) and \( n^{1/2} \left[ \text{CP} (D_{n,1} | Y_n, \theta) - \alpha \right] \) converge weakly to a \( N(0, \sigma^2(\theta)) \) distribution, the variance being given by (4.11). This example illustrates the extension of Propositions 3A and 3B to vector \( \theta \) and to \( D_n' \).
Example 4 (continued). Success of the calibration adjustment requires continuity of \( H_n(\alpha, \theta) \) in \( \alpha \). Such continuity is missing in example 4, where by (4.17),
\[
H_n(\alpha, \theta) = (n + 1)^{-1} < \alpha n + 1 > ,
\]
the discrete uniform distribution supported on the values \( \{n^{-1} j : 0 \leq j \leq n\} \). In this instance,
\[
H_n^{-1}(\alpha, \theta) = n^{-1} < \alpha (n + 1) > \quad \text{if} \quad 0 \leq \alpha < 1
\]
\[
= \infty \quad \text{if} \quad \alpha = 1 .
\]
Consequently, for \( \alpha \) less than 1,
\[
D_{n,1} = (-\infty, X(\beta)],
\]
where \( \beta = \alpha (n + 1) > + 1 \). It follows that
\[
n^{1/2} \{ \text{CP} (D_{n,1} | Y_n, \theta) - \alpha \} \Rightarrow N(0, \alpha (1 - \alpha))
\]
and
\[
\text{CP} (D_{n,1} | \theta) = \alpha + O(n^{-1})
\]
uniformly over all continuous \( F \). In this nonparametric case, \( D_{n,1} \) is no better than \( D_n \), in either conditional or overall coverage probability.

Example 1 (continued). For the gaussian AR(1) model, with prediction root \( R_n = X \), the coverage probability calculation (4.14) and the reasoning behind (4.33) yield the analytically calibrated prediction interval.
\[
D_n' (\alpha) = D_n [\alpha + (2n)^{-1} z_\alpha \phi(z_\alpha)],
\]
where \( D_n \) is given by (1.5). By direct arguments based on (4.14), both \( \text{CP} (D_{n,1} | \theta_n) \) and \( \text{CP} (D_n' | \theta_n) \) equal \( \alpha + o(n^{-1}) \) while \( \text{CP} (D_n | \theta_n) \) is \( \alpha + O(n^{-1}) \). Moreover, the asymptotic variance of \( n^{1/2} [\text{CP} (\cdot | Y_n^{\neq}, \theta) - \alpha] \) is \( \phi^2(z_\alpha) \) for each of the prediction intervals \( D_n, D_{n,1} \) and \( D_n' \). These asymptotics are consistent with the simulation results reported in section 2.2. They indicate that calibration works beyond the setting of Propositions 3B and 3A.

5. PROOFS

This section proves the Propositions stated earlier in the paper.

Proof of Proposition 1. Suppose (3.2) does not hold. By going to a subsequence assume without loss of generality that
\[
Q_{\theta_n,n} [\{ \text{CP} (D_n | Y_n, \theta_n) - \alpha \} > \varepsilon ] > \delta
\]
for some sequence \( \{\theta_n\} \in C(\theta) \) and some positive \((\epsilon, \delta)\). By going to a further subsequence, assume without loss of generality that \( \hat{\theta}_n \to \theta \) in \( Q_{\theta_0} \) probability and that \( U_n \) converges weakly under \( Q_{\theta_0} \) to a random variable \( U \). For Skorokhod versions of the random variables involved, \( (\hat{\theta}_n, U_n) \) converges almost surely to \((\theta, U)\), by Wichura (1970). From this and the assumed properties of \( C \), it follows that \( C(\alpha, \theta_n, \hat{\theta}_n, U_n) \to \alpha \) with probability one. This contradicts (5.1) because of (3.1). Hence (3.2) and (3.3) hold.

The argument for the coverage probabilities of \( D_{n1} \) is analogous and uses two additional facts: In view of (3.3), which holds for every possible \( \alpha \), and (2.5), it follows that \( H_n^{-1}(\alpha, \theta_n) \) converges to \( \alpha \) whenever \( \{\theta_n\} \) belongs to \( C(\theta) \). Moreover,

\[
\text{CP}(D_{n1} | Y_n, \theta_n) = C[H_n^{-1}(\alpha, \hat{\theta}_n), \theta_n, \hat{\theta}_n, U_n]. \tag{5.2}
\]

**Proof of Proposition 2A.** It follows from (4.2) that

\[
\text{CP}(D_n | Y_n, \theta_n) = C(\alpha, \theta_n, \hat{\theta}_n)
\]

\[
= \alpha + (\hat{\theta}_n - \theta_n) C^{(0,0,1)}(\alpha, \theta_n, \bar{\theta}_n) \tag{5.3}
\]

where \( \bar{\theta}_n \) lies between \( \theta_n \) and \( \hat{\theta}_n \). Assumption B implies that \((\theta_n, \bar{\theta}_n)\) converges to \((\theta, \theta)\) in \( Q_{\theta_0} \) probability. The asymptotic normality (4.4) thus follows from (5.3), Assumption B, and the continuity assumption on \( C^{(0,0,1)}(\alpha, \theta, \tau) \).

**Proof of Proposition 2B.** In this case, it follows from (4.2) that

\[
\text{CP}(D_n | Y_n, \theta_n) = \alpha + (\hat{\theta}_n - \theta_n) C^{(0,0,1)}(\alpha, \theta_n, \theta_n)
\]

\[
+ 2^{-1} (\hat{\theta}_n - \theta_n)^2 C^{(0,0,2)}(\alpha, \theta_n, \bar{\theta}_n) \tag{5.4}
\]

where \( \bar{\theta}_n \) lies between \( \theta \) and \( \hat{\theta}_n \). By Assumption A(2),

\[
E_{\theta_n} (\hat{\theta}_n - \theta_n) = n^{-1} a_{1,1}(\theta_n) + o(n^{-1}) \tag{5.5}
\]

\[
E_{\theta_n} (\hat{\theta}_n - \theta_n)^2 = n^{-1} a_{2,1}(\theta_n) + o(n^{-1}).
\]

The second line in (5.5) and the continuity assumptions on \( C^{(0,0,2)}(\alpha, \theta, \tau) \) and \( a_{2,1}(\theta) \) entail

\[
n (\hat{\theta}_n - \theta_n)^2 [C^{(0,0,2)}(\alpha, \theta_n, \bar{\theta}_n) - C^{(0,0,2)}(\alpha, \theta_n, \theta_n)] \to 0 \tag{5.6}
\]

in \( Q_{\theta_0} \) probability. A uniform integrability argument using the boundedness of \( C^{(0,0,2)} \) and the continuity of \( a_{2,1}(\theta) \) shows that convergence (5.6) also occurs in expectation. The conclusion (4.7) follows from the facts just established, by taking expectations through (5.4).
Proof of Proposition 3A. It follows from (4.2) and (4.26) that
\[
\text{CP}(D_{n,1} | Y_n, \theta_n) = H_n^{-1}(\alpha, \hat{\theta}_n) + (\hat{\theta}_n - \theta_n) C^{(0,0,1)}[H_n^{-1}(\alpha, \hat{\theta}_n), \theta_n, \bar{\theta}_n],
\]
where \(\bar{\theta}_n\) lies between \(\theta_n\) and \(\hat{\theta}_n\). In view of (4.27), the result follows by the reasoning for Proposition 2A.

Proof of Proposition 3B. Let \(\{\theta_n\}\) be any sequence in \(C_1(\theta)\). Write \(K_n\) for the quantile function \(H_n^{-1}\) and define
\[
K_{n,o}(\alpha, \theta) = \alpha - n^{-1} b_1(\alpha, \theta).
\]
By Taylor expansion,
\[
C[K_n(\alpha, \hat{\theta}_n), \theta_n, \hat{\theta}_n] = C[K_{n,o}(\alpha, \hat{\theta}_n), \theta_n, \hat{\theta}_n] + n^{-2} r_n(\alpha, \hat{\theta}_n) C^{(1,0,0)}(k_n, \theta_n, \hat{\theta}_n)
\]
where \(k_n\) lies between \(K_n(\alpha, \hat{\theta}_n)\) and \(K_{n,o}(\alpha, \hat{\theta}_n)\). Since \(C^{(1,0,0)}\) is bounded, it follows from (4.26) and (4.31) that
\[
\text{CP}(D_{n,1} | \theta_n) = E_{\hat{\theta}_n} C[K_{n,o}(\alpha, \hat{\theta}_n), \theta_n, \hat{\theta}_n] + O(n^{-2}).
\]
The following three Taylor expansions hold:
\[
\delta_n = K_{n,o}(\alpha, \hat{\theta}_n) - K_{n,o}(\alpha, \theta_n)
\]
\[
= -n^{-1}(\hat{\theta}_n - \theta_n) b_1^{(0,1)}(\alpha, \bar{\theta}_{n,1})
\]
\[
= -n^{-1} [(\hat{\theta}_n - \theta_n) b_1^{(0,1)}(\alpha, \theta_n) + 2^{-1} (\hat{\theta}_n - \theta_n)^2 b_1^{(0,2)}(\alpha, \bar{\theta}_{n,2})]
\]
where \(\bar{\theta}_{n,1}\) and \(\bar{\theta}_{n,2}\) lie between \(\hat{\theta}_n\) and \(\theta_n\). Also
\[
C[K_{n,o}(\alpha, \hat{\theta}_n), \theta_n, \hat{\theta}_n] = C[K_{n,o}(\alpha, \theta_n), \theta_n, \hat{\theta}_n]
\]
\[\quad + \delta_n C^{(1,0,0)}[K_{n,o}(\alpha, \theta_n), \theta_n, \theta_n] + (\hat{\theta}_n - \theta_n) C^{(1,0,1)}[K_{n,o}(\alpha, \theta_n), \theta_n, \bar{\theta}_{n,3}]
\]
\[\quad + 2^{-1} \delta_n^2 C^{(2,0,0)}(\bar{k}_n, \theta_n, \hat{\theta}_n)
\]
where \(\bar{k}_n\) lies between \(K_{n,o}(\alpha, \hat{\theta}_n)\) and \(K_{n,o}(\alpha, \theta_n)\) while \(\bar{\theta}_{n,3}\) lies between \(\hat{\theta}_n\) and \(\theta_n\).

From Assumption A2, (5.10), and the boundedness of the derivatives in (5.11), (5.12), it is apparent that
\[
\text{CP}(D_{n,1} | \theta_n) = E_{\hat{\theta}_n} C[K_{n,o}(\alpha, \theta_n), \theta_n, \hat{\theta}_n] + O(n^{-2}).
\]
On the other hand, (4.30) and an expansion analogous to (5.9) establish
\[
E_{\hat{\theta}_n} C[K_{n,o}(\alpha, \theta_n), \theta_n, \hat{\theta}_n] = E_{\hat{\theta}_n} C[K_n(\alpha, \theta_n), \theta_n, \hat{\theta}_n] + O(n^{-2})
\]
\[\quad = \alpha + O(n^{-2}).
\]
The second line in (5.14) relies on the continuity of \(H_n(\alpha, \theta)\) in \(\alpha\), itself a consequence of (2.5), the first line in (5.3), and the boundedness of \(C^{(1,0,0)}\). Combining
(5.13) with (5.14) completes the proof.
REFERENCES


TECHNICAL REPORTS
Statistics Department
University of California, Berkeley


17. DRAPER, D. C. (Feb 1983). Rank-based robust inference in regression models with several observations per cell.


32. BREIMAN, L. (May 1984). Nail finders, edifices, and Oz.


43. BICKEL, P. J. and YAHAV, J. A. (1985). On estimating the number of unseen species: how many executions were there?


56. LE CAM, L. and YANG, G. L. (January 1986). Replaced by No. 68.


64. O’SULLIVAN, F. (May 1986). Nonparametric Estimation in the Cox Proportional Hazards Model.


71. BICKEL, P.J. and YAHAV, J.A. (July 1986). Richardson Extrapolation and the Bootstrap.


77. O’SULLIVAN, F. (September 1986). Deconvolution of episodic hormone data.


90. DABROWSKA, D.M. (March 1987). Rank tests for matched pair experiments with censored data.


95. CANCELLED


114. RITOV, Y. (September 1987). Estimation in a linear regression model with censored data.
118. ALDOUS, D.J. (October 1987). Meeting times for independent Markov chains.
127. Same as No. 133.
128. DONOHO, D. and GASKO, M. (December 1987). Multivariate generalizations of the median and trimmed mean II.
131. Same as No. 140
134. CANCELLED
166. FAN, JIANQING (July 1988). Nonparametric estimation of quadratic functionals in Gaussian white noise.
172. ADLER, R.J. and EPSTEIN, R. (September 1988). Intersection local times for infinite systems of planar brownian motions and for the brownian density process.


179. DONOHO, D.L. and STARK, P.B. (October 1988). Recovery of sparse signal when the low frequency information is missing.


209. BREIMAN, L. (July 1989). Fitting additive models to regression data.


Copies of these Reports plus the most recent additions to the Technical Report series are available from the Statistics Department technical typist in room 379 Evans Hall or may be requested by mail from:

Department of Statistics
University of California
Berkeley, California 94720

Cost: $1 per copy.