The Annals of Statistics
1992, Vol. 20, No. 2, 1110–1119

CONTROLLING CONDITIONAL COVERAGE PROBABILITY IN PREDICTION[1]

BY RUDOLF BERAN

University of California, Berkeley

Suppose the variable X to be predicted and the learning sample Y_n that was observed are independent, with a joint distribution that depends on an unknown parameter θ. A prediction region D_n for X is a random set, depending on Y_n, that contains X with prescribed probability α. In sufficiently regular models, D_n can be constructed so that overall coverage probability converges to α at rate n^{-r}, where r is any positive integer. This paper shows that the *conditional* coverage probability of D_n, given Y_n, converges in probability to α at a rate which usually cannot exceed $n^{-1/2}$.

1. Introduction. A random variable X that is to be predicted and a learning sample Y_n that was observed have a joint distribution $P_{\theta,n}$. The parameter θ is unknown. A prediction region for X is a random set D_n, depending on the learning sample Y_n, that contains X with prescribed probability α.

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 the random variable

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

A basic problem is to construct D_n so that $CP(D_n|Y_n,\theta)$ converges to α in probability. Controlling conditional coverage probability, at least asymptotically, is a natural goal in predicting time series [Box and Jenkins (1976), Section 5.2.4] and in establishing tolerance regions [Guttman (1970), Butler (1982)]. Other recent discussions of conditional coverage probability in prediction appear in Butler and Rothman (1980), Stine (1985) and Beran (1990).

The expectation of the conditional coverage probability $CP(D_n|Y_n,\theta)$ is the *overall coverage probability* of D_n:

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

This expectation is taken with respect to the distribution $Q_{\theta,n}$ of Y_n. In view of (1.2), the bias in $CP(D_n|Y_n,\theta)$ as an estimator of α is the same as the error in the overall coverage probability of D_n. Cox (1975) develops an algebraic adjustment to D_n which reduces this bias to asymptotic order n^{-2} in regular models. Beran (1990) gives a bootstrap adjustment to D_n which has the same

Received September 1989; revised January 1991.
[1]Research supported in part by NSF Grant DMS-87-01426.
AMS 1980 *subject classifications*. Primary 62M20; secondary 62E20.
Key words and phrases. Prediction region, conditional coverage probability, local asymptotic minimax, convolution representation.

effect. In sufficiently regular cases, iteration of such bias adjustments reduces the error in $CP(D_n|\theta)$ to asymptotic order n^{-r}, where r is any positive integer.

The use of the word estimator in the previous paragraph is imprecise because α is known and the conditional coverage probability (1.2) depends on the unknown parameter θ. However, the normalized error in conditional coverage probability

$$T_n(\theta) = n^{1/2}[CP(D_n|Y_n, \theta) - \alpha] \tag{1.3}$$

often has a limiting distribution as n increases [Butler (1982), Beran (1990)]. Think of $T_n(\theta)$ as analogous to the normalized error $\tilde{T}_n(\theta) = n^{1/2}(\hat{\theta}_n - \theta)$, where $\hat{\theta}_n$ is some estimator of θ. If the analogy has substance, the Hájek convolution representation for the limiting distribution of $\tilde{T}_n(\theta)$ and the local asymptotic minimax bound on the dispersion of $\tilde{T}_n(\theta)$ should have counterparts for $T_n(\theta)$.

This turns out to be the case. Section 2 establishes a sharp asymptotic lower bound on the dispersion of $T_n(\theta)$ and also gives a convolution representation for the limiting distribution of $T_n(\theta)$. Consequently, unlike overall coverage probability, conditional coverage probability converges to α at a rate which cannot exceed $n^{-1/2}$ in classically regular models. This circumstance limits the statistician's ability to control conditional coverage probability when designing a prediction region.

2. Dispersion of conditional coverage probability. The analysis in this paper is directed at the simplest case, where the learning sample Y_n and the sample X to be predicted are independent. The distribution of Y_n is $Q_{\theta,n}$ and the distribution of X is P_θ. The parameter space is an open subset of the real line. The extension to Euclidean parameter spaces is straightforward and will be sketched in the exposition.

Treatment of infinite-dimensional θ is harder because Fréchet differentiability conditions analogous to those in Proposition 1 below are now too strong to be useful. On a case-by-case basis, some infinite-dimensional extensions of Proposition 1 are possible, as shown by Example 4 below. For time series models, Assumption B—local asymptotic normality—usually does not hold. Thus, a substantial further development is needed to handle these models.

Suppose $\hat{\theta}_n = \hat{\theta}_n(Y_n)$ is an estimate of θ. Let $R_n = R(X, \hat{\theta}_n)$ be a root for the prediction region— a function of X and $\hat{\theta}_n$ which is referred to a critical value in order to generate the desired prediction region for X. Consider in this setting the following design question. Let $\{D_n(c)\}$ be a sequence of prediction regions for X of the form

$$D_n(c) = \left\{x\colon R\left(x, \hat{\theta}_n\right) \le d_n\right\}, \tag{2.1}$$

where $c = \{(d_n, \hat{\theta}_n)\}$ is a sequence of critical values and estimates. Both d_n and $\hat{\theta}_n$ are computed from the learning sample Y_n. Examples of construction

(2.1) are given later in this section. Impose on the sequence c the following constraint:

ASSUMPTION A. As n increases,

$$CP[D_n(c)|Y_n, \theta] \to_p \alpha, \tag{2.2}$$
$$\hat{\theta}_n \to_p \theta,$$

in $Q_{\theta,n}$ probability.

How should the sequence c be chosen so as to minimize the dispersion of

$$T_n(c, \theta) = n^{1/2}\{CP[D_n(c)|Y_n, \theta] - \alpha\}? \tag{2.3}$$

To answer this question, introduce the following local asymptotic normality assumption on the distribution $Q_{\theta,n}$ of the learning sample. In classical parametric models, the variance $I(\theta)$ in (2.5) below is the Fisher information or a related limit.

ASSUMPTION B. For $\theta_n = \theta + n^{-1/2}h$, where h is real, let $Q^c_{\theta_n,n}$ and $Q^s_{\theta_n,n}$ denote, respectively, the absolutely continuous and the singular parts of $Q_{\theta_n,n}$ with respect to $Q_{\theta,n}$. Let $L_n(h, \theta)$ denote the log-likelihood ratio of $Q^c_{\theta_n,n}$ with respect to $Q_{\theta,n}$. There exist random variables $\xi_n(\theta)$ depending on Y_n and on θ and a positive constant $I(\theta)$ such that

$$L_n(h, \theta) - h\xi_n(\theta) + 2^{-1}h^2 I(\theta) \to_p 0 \tag{2.4}$$

in $Q_{\theta,n}$ probability, for every real h, and

$$\mathscr{L}[\xi_n(\theta)|\theta] \Rightarrow N(0, I(\theta)). \tag{2.5}$$

Without loss of generality, we may assume that $\xi_n(\theta)$ is constructed so that

$$\xi_n(\theta_n) - \xi_n(\theta) + hI(\theta) \to_p 0 \tag{2.6}$$

in $Q_{\theta,n}$ probability for every real h [Le Cam (1969), page 68].

This section presents two Propositions that bound from below the dispersion of $T_n(c, \theta)$ when n is large. The first result is related to the Hájek–Le Cam local asymptotic minimax theory; the other is linked to the Hájek convolution representation for limiting distributions of regular estimates. Proofs are deferred to Section 3.

2.1. *The local asymptotic minimax approach.* Let w be a monotone function on the nonnegative reals, with $w(0) = 0$. Measure the dispersion of $CP[D_n(c)|Y_n, \theta]$ about α through the risk

$$\rho_n(c, \theta) = E_\theta w[|T_n(c, \theta)|]. \tag{2.7}$$

The conditional cdf of the root R_n, given Y_n, is

$$A(x,\theta,\hat{\theta}_n) = P_\theta\big[R(X,\hat{\theta}_n) \le x \big| Y_n\big], \tag{2.8}$$

where $\hat{\theta}_n$ is held fixed on the right side. The conditional coverage probability of $D_n(c)$ is thus

$$CP\big[D_n(c)|Y_n,\theta\big] = A(d_n,\theta,\hat{\theta}_n). \tag{2.9}$$

For use in the sequel, define the function

$$C(\alpha,\theta,t) = A\big[A^{-1}(\alpha,t,t),\theta,t\big], \tag{2.10}$$

where $A^{-1}(\alpha,t,t)$ is the largest αth quantile of the conditional cdf $A(x,t,t)$. Notation like $f^{(i,j,k)}(x,\theta,t)$ will represent the partial derivative $\partial^{i+j+k}f(x,\theta,t)/\partial x^i\,\partial\theta^j\,\partial t^k$.

PROPOSITION 1. *Suppose Assumptions A and B hold; and the cdf $A(x,\theta,t)$ is strictly monotone in x and is continuous in all three arguments. Suppose also that $A^{(1,0,0)}(x,\theta,t)$, $A^{(0,0,1)}(x,\theta,t)$, $C^{(0,0,1)}(x,\theta,t)$ exist and are continuous in (x,θ), (x,θ,t), (θ,t), respectively, at points where $t=\theta$. Then, for $\theta_n = \theta + n^{-1/2}h$,*

$$\lim_{b\to\infty}\liminf_{n\to\infty}\inf_{c}\sup_{|h|\le b}\rho_n(c,\theta_n) \ge Ew\big[|\tau(\theta)Z|\big], \tag{2.11}$$

where Z is a standard normal random variable and

$$\tau^2(\theta) = \big[C^{(0,0,1)}(\alpha,\theta,\theta)\big]^2 I^{-1}(\theta). \tag{2.12}$$

If w is bounded and c is such that

$$T_n(c,\theta_n) - C^{(0,0,1)}(\alpha,\theta,\theta)I^{-1}(\theta)\xi_n(\theta_n) \to_p 0 \tag{2.13}$$

in $Q_{\theta_n,n}$ probability, for every real h, then

$$\lim_{n\to\infty}\sup_{|h|\le b}\rho_n(c,\theta_n) = Ew\big[|\tau(\theta)Z|\big] \tag{2.14}$$

for every positive b.

Two remarks concerning Proposition 1 are:

REMARK A. The following condition on the sequence $c = \{(d_n,\hat{\theta}_n)\}$ ensures that the lower bound (2.13) is attained in the sense (2.14), under the assumptions of the proposition:

$$\begin{aligned} &n^{1/2}(\hat{\theta}_n - \theta) - I^{-1}(\theta)\xi_n(\theta) \to_p 0,\\ &n^{1/2}\big[d_n - A^{-1}(\alpha,\hat{\theta}_n,\hat{\theta}_n)\big] \to_p 0 \end{aligned} \tag{2.15}$$

in $Q_{\theta,n}$ probability. Indeed, by (2.6) and the contiguity entailed by Assumption B, (2.15) implies

$$n^{1/2}(\hat{\theta}_n - \theta_n) - I^{-1}(\theta)\xi_n(\theta_n) \to_P 0 \tag{2.16}$$

in $Q_{\theta_n, n}$ probability. Moreover $d_n \to_P A^{-1}(\alpha, \theta, \theta)$ under $Q_{\theta_n, n}$ and $C(\alpha, t, t) = \alpha$ for every possible t. Hence, by several first-order Taylor expansions,

$$\begin{aligned} CP[D_n(c)|Y_n, \theta_n] &= A(d_n, \theta_n, \hat{\theta}_n) \\ &= C(\alpha, \theta_n \hat{\theta}_n) + o_p(n^{-1/2}) \\ &= \alpha + C^{(0,0,1)}(\alpha, \theta, \theta)(\hat{\theta}_n - \theta_n) + o_p(n^{-1/2}) \end{aligned} \tag{2.17}$$

in $Q_{\theta_n, n}$ probability. Property (2.13) follows from (2.17) and (2.16). Note that the first line in (2.15) is the classical condition that $\hat{\theta}_n$ be an asymptotically efficient estimate of θ.

REMARK B. In the vector parameter extension of Proposition 1, $C^{(0,0,1)}(x, \theta, t)$ is a column vector and $I(\theta)$ is a positive definite matrix. The expression for $\tau^2(\theta)$ becomes

$$\tau^2(\theta) = [C^{(0,0,1)}(\alpha, \theta, \theta)]' I^{-1}(\theta) [C^{(0,0,1)}(\alpha, \theta, \theta)]. \tag{2.18}$$

EXAMPLE 1. The simplest good choice for $c = \{(d_n, \hat{\theta}_n)\}$ is $\hat{\theta}_n$ asymptotically efficient in the sense of (2.15) and

$$d_n = A^{-1}(\alpha, \hat{\theta}_n, \hat{\theta}_n). \tag{2.19}$$

Remark A above applies. Moreover, the overall coverage probability of $D_n(c)$ satisfies

$$CP[D_n(c)|\theta] = \alpha + O(n^{-1}) \tag{2.20}$$

under the assumptions of Proposition 2B in Beran (1990).

A better choice of c is often possible. Let $H_n(\cdot, \theta)$ denote the cdf of the transformed root $A(R_n, \hat{\theta}_n, Y_n)$. Replace (2.19) with the refinement

$$d_n = A^{-1}[H_n^{-1}(\alpha, \hat{\theta}_n), \hat{\theta}_n, \hat{\theta}_n], \tag{2.21}$$

where $\hat{\theta}_n$ is still asymptotically efficient. For sufficiently regular models,

$$H_n^{-1}(\alpha, \hat{\theta}_n) = \alpha + O_p(n^{-1}) \tag{2.22}$$

under $Q_{\theta, n}$ [see equation (4.21) in Beran (1990)]. Consequently, Remark A above still applies. However, for this refined choice of c,

$$CP[D_n(c)|\theta] = \alpha + O(n^{-2}) \tag{2.23}$$

under the assumptions of Proposition 3B in Beran (1990).

The adjustment (2.21) to the critical value d_n thus improves the rate of convergence to α of overall coverage probability; it does not affect the first-order asymptotics of the conditional coverage probability. For both choices (2.19) and (2.21) of d_n, the dispersion of the conditional coverage probability of $D_n(c)$ achieves the local asymptotic minimax bound (2.11), provided $\hat{\theta}_n$ is asymptotically efficient.

EXAMPLE 2. To illustrate the extension of Proposition 1 to vector parameters, suppose X and Y_n are independent, X has a $N(\beta c, \sigma^2)$ distribution and the elements of $Y_n = (X_1, \dots, X_n)$ are iid $N(\beta c_i, \sigma^2)$ random variables. The parameter $\theta = (\beta, \sigma^2)$ is unknown; the $\{c_i\}$ and c are known constants. Let $\hat{\theta}_n = (\hat{\beta}_n, s_n^2)$ denote the least squares estimate of θ based on Y_n. As root for the prediction region, take the function

$$R(X, \hat{\theta}_n) = X. \tag{2.24}$$

The more elaborate roots $X - \hat{\beta}_n c$ or $(X - \hat{\beta}_n c)/s_n$ yield the same prediction intervals as (2.24) in the following discussion.

The choice (2.19) for critical value d_n generates the one-sided prediction interval $(-\infty, \hat{\beta}_n c + s_n z_\alpha]$, where z_α is the αth quantile of the standard normal distribution. Suppose $n^{-1}\sum_{i=1}^n c_i^2 \to a^2$, a finite limit. The overall coverage probability of the interval then differs from α by $O(n^{-1})$. The refined choice (2.21) for d_n generates the classical exact prediction interval $(-\infty, \hat{\beta}_n c + s_n\{1 + (\sum_{i=1}^n c_i^2)^{-1}\}^{1/2} t_{n-1,\alpha}]$, where $t_{r,\alpha}$ is the αth quantile of the t distribution with r degrees of freedom. Let φ denote the standard normal density. By the reasoning for remark (a) above, both of these prediction intervals have the property that (2.14) holds, with

$$I(\theta) = \begin{pmatrix} a^2\sigma^{-2} & 0 \\ 0 & (2\sigma^4)^{-1} \end{pmatrix}, \tag{2.25}$$

$$C^{(1,0,0)}(\alpha, \theta, \theta) = \begin{pmatrix} \sigma^{-1}\varphi(z_\alpha) \\ (2\sigma^2)^{-1} z_\alpha \varphi(z_\alpha) \end{pmatrix} \tag{2.26}$$

and

$$\tau^2(\theta) = (2^{-1}z_\alpha^2 + a^{-2})\varphi^2(z_\alpha) \tag{2.27}$$

in accordance with remark (b) above.

EXAMPLE 3. As an interesting special case of Example 1, suppose that the $\{X_i: i \ge 1\}$ are iid unit vectors from a Fisher (μ, κ) distribution, where μ is a unit vector and κ is positive. The parameter $\theta = (\mu, \kappa)$ is unknown. Suppose the learning sample is $Y_n = (X_1, \dots, X_n)$ and the variable to be predicted is the unit vector $X = X_{n+1}$. Since the Fisher model can be rewritten in canonical exponential form [Beran (1979)], Assumption B holds and the maximum likelihood estimate $\hat{\theta}_n = (\hat{\mu}_n, \hat{\kappa}_n)$ of θ satisfies the first line in (2.15). Moreover, $\hat{\mu}_n$ is the sample mean vector rescaled to unit length.

To generate a prediction cone for X, consider the root

$$R(X, \hat{\theta}_n) = 2^{-1}|X - \hat{\mu}_n|^2 = 1 - \hat{\mu}_n' X. \tag{2.28}$$

By straightforward calculation,

$$A(x, \hat{\theta}_n, \hat{\theta}_n) = \frac{1 - \exp(-\hat{\kappa}_n x)}{1 - \exp(-2\hat{\kappa}_n)}, \qquad 0 \le x \le 2. \tag{2.29}$$

Consequently, the efficient critical value (2.19) is

$$d_n = -\hat{\kappa}_n^{-1} \log\{1 - \alpha[1 - \exp(-2\hat{\kappa}_n)]\}. \tag{2.30}$$

The prediction region (2.1) for X determined by root (2.28) and critical value (2.30) is a cone with axis $\hat{\mu}_n$. By Proposition 1, the conditional coverage probability of this prediction cone is minimally dispersed about α, for large n, because $\{d_n\}$ satisfies (2.15).

Comparing this example with Example 2, we see that the optimality property isolated in Proposition 1 has nothing to do with one-sidedness or multisidedness of a prediction region.

The refined critical value (2.21) satisfies (2.23) and Proposition 1 in this example. A bootstrap algorithm for computing (2.21) is given in Beran (1990).

EXAMPLE 4. Suppose that X and the elements of $Y_n = (X_1, \dots, X_n)$ are iid random variables with unknown continuous cdf F. The parameter $\theta = F$ is estimated by the empirical cdf $\hat{\theta}_n = \hat{F}_n$. From the root $R_n = X$, definition (2.1) and critical value (2.19) generate the one-sided prediction interval

$$D_n = \left(-\infty, \hat{F}_n^{-1}(\alpha)\right], \tag{2.31}$$

where $\hat{F}_n^{-1}(\alpha)$ is the largest αth sample quantile, say.

The conditions for Proposition 1 are not satisfied in this example. However, under regularity conditions on F, an argument using ideas in Koshevnik and Levit (1976) establishes an analog of the lower bound (2.11), with

$$\tau^2(F) = \alpha(1 - \alpha). \tag{2.32}$$

Moreover, the prediction interval (2.31) attains this asymptotic lower bound in the sense of (2.14). This example illustrates the possibility of case-by-case extensions of Proposition 1 when θ is infinite-dimensional.

2.2. *The convolution representation.* A sequence of prediction regions $\{D_n(c)\}$ having the form (2.1) will be called *Hájek-regular* if, for $\theta_n = \theta + n^{-1/2}h$ and for $T_n(c, \theta_n)$ defined by (2.3),

$$\mathscr{L}[T_n(c, \theta_n)|\theta_n] \Rightarrow \mu_\theta(c), \tag{2.33}$$

the limit law $\mu_\theta(c)$ depending on c but not on h.

PROPOSITION 2. *Suppose the assumptions for Proposition* 1 *hold and the prediction regions* $\{D_n(c)\}$ *are Hájek-regular. Then*

$$\mu_\theta(c) = N(0, \tau^2(\theta)) * v_\theta(c), \tag{2.34}$$

where $\tau^2(\theta)$ *is defined by* (2.12) *and* $(v_\theta(c))$ *is a probability measure on the real line. Moreover,* $\{D_n(c)\}$ *is Hájek-regular and* $\mu_\theta(c) = N(0, \tau^2(\theta))$ *if and only if* (2.13) *holds.*

Remarks A and B that follow Proposition 1 also carry over to Proposition 2. Examples 1, 2 and 3 illustrate how to choose the sequence c so that $v_\theta(c)$ in

(2.34) is the point mass at zero—the situation when the limit law of $T_n(c,\theta)$ is least dispersed. An analogous result for Example 4 can be proved using ideas in Beran (1977).

3. Proofs.

This section proves the two propositions stated in Section 2.

PROOF OF PROPOSITION 1. Assumption A and the conditions on the cdf A imply that

$$d_n \to_p A^{-1}(\alpha,\theta,\theta) \tag{3.1}$$

in $Q_{\theta,n}$ probability. Write ρ_0 for the right side of (2.11). Suppose the proposition is false. Then, there exists positive ε such that

$$\liminf_{n\to\infty} \inf_{c} \sup_{|h|\le b} \rho_n(c,\theta_n) \le \rho_0 - \varepsilon \tag{3.2}$$

for every positive b. By extracting a suitable subsequence, assume without loss of generality that there exists a sequence c such that (3.1) holds and

$$\rho_n(c,\theta_n) \le \rho_0 - \varepsilon/4 \tag{3.3}$$

for every $|h| \le b$ and every n.

On the other hand, for every fixed h,

$$\rho_n(c,\theta_n) \ge E_\theta\{u(T_n)\exp[L_n(h,\theta)]\}, \tag{3.4}$$

where

$$T_n = T_n(c,\theta_n) = n^{1/2}\big[A(d_n,\theta_n,\hat\theta_n) - \alpha\big] \tag{3.5}$$

can be written as the sum of two terms $T_{n,1}$ and $T_{n,2}$ as follows:

$$\begin{aligned} T_{n,1} &= n^{1/2}[A(d_n,\theta_n,\theta) - \alpha] \\ &= n^{1/2}[C(\alpha,\theta_n,\theta) - \alpha] + W_n \\ &= -hC^{(0,0,1)}(\alpha,\theta_n,\bar\theta_{n,1}) + W_n, \end{aligned} \tag{3.6}$$

where $\bar\theta_{n,1}$ lies between θ_n and θ and

$$W_n = n^{1/2}\big[d_n - A^{-1}(\alpha,\theta,\theta)\big]A^{(1,0,0)}(\bar d_n,\theta_n,\theta) \tag{3.7}$$

for $\bar d_n$ between d_n and $A^{-1}(\alpha,\theta,\theta)$. Moreover,

$$\begin{aligned} T_{n,2} &= n^{1/2}\big[A(d_n,\theta_n,\hat\theta_n) - A(d_n,\theta_n,\theta)\big] \\ &= n^{1/2}(\hat\theta_n - \theta)A^{(0,0,1)}(d_n,\theta_n,\bar\theta_{n,2}), \end{aligned} \tag{3.8}$$

where $\bar\theta_{n,2}$ lies between $\hat\theta_n$ and θ.

In view of (3.5) through (3.8) and the assumptions of the proposition, assume without loss of generality, by going to a subsequence, that

$$(T_n, \xi_n(\theta)) \Rightarrow (V - hC^{(0,0,1)}(\alpha,\theta,\theta), I^{1/2}(\theta)Z) \tag{3.9}$$

under $Q_{\theta,n}$. Here V is a random variable on the extended real line whose distribution does not depend on h and Z has a standard normal distribution.

Set $t = I^{1/2}(\theta)h$ and $b(\theta) = C^{(0,0,1)}(\alpha, \theta, \theta)I^{-1/2}(\theta)$. From (3.4), (3.9), Assumption B and Fatou's lemma,

$$\begin{aligned} \lim_{n\to\infty} \rho_n(c, \theta_n) &\geq E\{u[|V - b(\theta)t|]\exp(tZ - 2^{-1}t^2)\} \\ &= \int E\{u[|V - b(\theta)t|]Z = z\}\phi(z - t)\,dz \\ &= \int\int u[|v - b(\theta)t|]M(dv, z)\phi(z - t)\,dz \\ &= J(M, t), \quad \text{say,} \end{aligned} \tag{3.10}$$

where $M(dv, z)$ is the probability element of the conditional distribution of V given $Z = z$. Combining (3.10) with (3.3) establishes

$$J(M, t) \leq \rho_0 - \varepsilon/4 \tag{3.11}$$

for every $|t| \leq I^{1/2}(\theta)b$. The argument works for every positive b.

Inequality (3.11) thus contradicts the classical minimax bound in the normal location model:

$$\lim_{a\to\infty} \inf_M \sup_{|t|\leq a} J(M, t) = \rho_0. \tag{3.12}$$

Hence Proposition 1 is true. □

Proof of Proposition 2. Because of Assumption B, $Q^s_{\theta_n, n}(R^1) \to 0$ as n increases. Hence the characteristic function of $\mathscr{L}[T_n|Q_{\theta_n, n}]$ satisfies

$$E_{\theta_n} \exp(iuT_n) = E_\theta[iuT_n + L_n(h, \theta)] + o(1). \tag{3.13}$$

By going to a subsequence, assume without loss of generality that (3.9) holds. In view of (2.33), specializing to $h = 0$ shows that V in (3.9) an ordinary random variable with distribution $\mu_\theta(c)$. From (3.9) and a uniform integrability argument, passing to the limit in (3.13) as n increases yields

$$\begin{aligned} E\exp(iuV) = E\exp&[iu\{V - hC^{(0,0,1)}(\alpha, \theta, \theta)\}] \\ &\times \exp[hI^{1/2}(\theta)Z - 2^{-1}h^2I(\theta)]. \end{aligned} \tag{3.14}$$

Since the right side of (3.14) is analytic in h and is constant for all real h, the relation (3.14) must be valid for all complex h. In particular, setting $h = -iI^{-1}(\theta)C^{(0,0,1)}(\alpha, \theta, \theta)u$ in (3.14) gives

$$\begin{aligned} E\exp(iuV) &= \exp[-2^{-1}\tau^2(\theta)u^2] \\ &\quad \times E\exp[iu\{V - C^{(0,0,1)}(\alpha, \theta, \theta)I^{-1/2}(\theta)Z\}]. \end{aligned} \tag{3.15}$$

This proves (2.34).

The if and only if part: Suppose (2.13) holds. By contiguity reasoning and Assumption B,

$$\mathscr{L}[\xi_n(\theta)|\theta_n] \Rightarrow N(hI(\theta), I(\theta)). \tag{3.16}$$

From this, (2.6) and (2.13),

$$\mathscr{L}[T_n|\theta_n] \Rightarrow N(0, \tau^2(\theta)) \tag{3.17}$$

for every real h, as asserted in Proposition 2.

Conversely, suppose that (3.17) holds for every real h while convergence (2.13) does not occur under $Q_{\theta_n, n}$, and hence under $Q_{\theta, n}$ by contiguity. By going to a subsequence, assume without loss of generality that

$$Q_{\theta, n}\left[\left|T_n - C^{(0,0,1)}(\alpha, \theta, \theta) I^{-1}(\theta)\xi_n(\theta_n)\right| \geq \varepsilon\right] > \delta \tag{3.18}$$

for every n and some positive ε and δ. By going to a further subsequence, as in the first part of the proof, assume without loss of generality that (3.9) holds under $Q_{\theta, n}$, with V having a $N(0, \tau^2(\theta))$ distribution in view of (3.17). From this, (2.6) and (3.18),

$$\Pr\left[\left|V - C^{(0,0,1)}(\alpha, \theta, \theta) I^{-1/2}(\theta) Z\right| \geq \varepsilon\right] > \delta. \tag{3.19}$$

At the same time, (3.15) also holds and here entails

$$V = C^{(0,0,1)}(\alpha, \theta, \theta) I^{-1/2}(\theta) Z \quad \text{w.p.1.} \tag{3.20}$$

The contradiction between (3.19) and (3.20) establishes that (2.13) must hold.

This argument draws in part on Bickel's unpublished proof of the Hájek convolution representation for regular estimates. □

REFERENCES

BERAN, R. (1977). Estimating a distribution function. *Ann. Statist.* **5** 400–404.

BERAN, R. (1979). Exponential models for directional data. *Ann. Statist.* **7** 1162–1178.

BERAN, R. (1990). Calibrating prediction regions. *J. Amer. Statist. Assoc.* **85** 715–723.

BOX, G. E. P. and JENKINS, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, Oakland.

BUTLER, R. W. (1982). Nonparametric interval and point prediction using data trimmed by a Grubbs type outlier. *Ann. Statist.* **10** 197–204.

BUTLER, R. and ROTHMAN, E. D. (1980). Predictive intervals based on reuse of the sample. *J. Amer. Statist. Assoc.* **75** 881–889.

COX, D. R. (1975). Prediction intervals and empirical Bayes confidence intervals. In *Perspectives in Probability and Statistics* (J. Gani, ed.) 47–55. Academic, London.

GUTTMAN, I. (1970). *Statistical Tolerance Regions: Classical and Bayesian*. Griffin, London.

KOSHEVNIK, YU. A. and LEVIT, B. YA. (1976). On a nonparametric analog of the information matrix. *Theory Probab. Appl.* **21** 738–753.

LE CAM, L. M. (1969). *Théorie Asymptotique de la Décision Statistique*. Univ. Montreal Press.

STINE, R. A. (1985). Bootstrap prediction intervals for regression. *J. Amer. Statist. Assoc.* **80** 1026–1031.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720