The Annals of Statistics
1993, Vol. 21, No. 4, 1967–1981

PROBABILITY-CENTERED PREDICTION REGIONS[1]

BY RUDOLF BERAN

University of California, Berkeley

Consider the problem of constructing a prediction region D_n for a potentially observable variable X on the basis of a learning sample of size n. Usually, the requirement that D_n contain X with probability α, conditionally on the learning sample, does not uniquely determine D_n. This paper develops a general probability-centering concept for prediction regions that extends to vector-valued or function-valued X the classical notion of an equal-tailed prediction interval. The dual requirements of probability centering and specified coverage probability determine D_n uniquely. Several examples illustrate the scope and consequences of the proposed centering concept.

1. Introduction. Suppose a potentially observable variable X and a learning sample Y_n of size n have a joint distribution that depends upon an unknown parameter θ. The future variable X can be real-valued, vector-valued or function-valued, and the parameter θ can be finite- or infinite-dimensional. The problem is to construct a good prediction region D_n for X on the basis of the learning sample Y_n.

A basic requirement on D_n is that the conditional coverage probability for X, given Y_n, should converge in probability to a preselected value α as n increases. Authors who have expressed this design goal include Box and Jenkins (1976), Butler and Rothman (1980), Butler (1982), Stine (1985) and Beran (1990). Such convergence in conditional coverage probability determines the asymptotic form of a one-sided prediction interval D_n for a real-valued X: the critical value must estimate consistently an appropriate quantile of the conditional distribution of X given Y_n.

On the other hand, a two-sided prediction interval for real-valued X is not determined by its coverage probability alone. An additional "centering" condition is needed. A familiar probability-centering concept for two-sided prediction intervals specifies that the conditional probability, given Y_n, of X exceeding the upper endpoint of D_n should equal the conditional probability of X falling below the lower endpoint. This type of centering is relatively easy to achieve, at least asymptotically, and can be generalized to prediction regions for vector-valued or function-valued X. Such generalizations are the topic of this paper.

Alternative centering concepts for two-sided prediction intervals rely on notions of shortest expected length or on notions of most concentrated support

Received February 1991; revised August 1992.
[1]Supported in part by NSF Grant DMS-90-01710.
AMS 1991 *subject classifications*. Primary 62M20; secondary 62G09.
Key words and phrases. Simultaneous prediction intervals, bootstrap, design goals.

for the conditional distribution of X given the learning sample. Butler (1982) gives an instructive example of the latter approach; this is extended to a multivariate setting in Butler, Davies and Jhun (1993). Another multivariate approach, the method of cuts of J. Tukey and D. A. S. Fraser, is described in Section 2 of Guttman (1970).

The basic idea in this paper is to construct prediction region D_n as a collection of simultaneous one-sided prediction intervals for suitably chosen real-valued functions of X, such as linear functionals of X. Requiring the marginal coverage probabilities of these one-sided intervals to be equal defines the probability-centering of D_n. Section 2 gives the main theoretical results and several illustrative examples for vector-valued and function-valued X. Section 3 discusses numerical algorithms needed to construct probability-centered prediction regions. Proofs are gathered in Section 4.

2. Probability centering. This section defines probability-centered prediction regions quite generally and shows how to construct them when the learning sample is moderately large. Several examples illustrate the scope of the theory. Numerical aspects are discussed in Section 3.

2.1. *Design goals.* Suppose the variable X to be predicted and the learning sample Y_n have a joint distribution $P_{\theta,n}$. The unknown parameter θ lies in a metric space Θ. Associate with X a random process $Z = \{Z(u, X)\colon u \in U\}$ whose index set U is also a metric space. The prediction region D_n for X is assumed to take the form

$$D_n = \{x\colon Z(u, x) \le c_n(u, Y_n), \forall\, u \in U\}, \tag{2.1}$$

where the critical values $c_n(u, Y_n)$ can depend on the learning sample Y_n. Evidently, D_n is equivalent to simultaneously asserting the prediction regions

$$D_{n,u} = \{x\colon Z(u, x) \le c_n(u, Y_n)\}, \qquad u \in U. \tag{2.2}$$

The problem is to select the critical values $\{c_n(u, Y_n)\}$ in a reasonable way.

Structure (2.1) for prediction regions and possible choices for the process Z are illustrated by Example 1 and by the additional examples in subsection 2.3. The reader may wish to glance at these before continuing any further.

The performance of prediction region D_n can be assessed through the coverage probabilities of D_n and the component regions $\{D_{n,u}\colon u \in U\}$. 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

$$\mathrm{CP}(D_n|Y_n, \theta) = P_\theta(X \in D_n|Y_n) \tag{2.3}$$

and the (unconditional) *coverage probability of* D_n is

$$\mathrm{CP}(D_n|\theta) = E_\theta\, \mathrm{CP}(D_n|Y_n, \theta), \tag{2.4}$$

where the expectation is calculated with respect to the distribution $Q_{\theta,n}$ of the learning sample Y_n. The notation $\mathrm{CP}(D_{n,u}|Y_n, \theta)$ and $\mathrm{CP}(D_{n,u}|\theta)$ will similarly

denote the conditional and unconditional coverage probabilities of the component prediction region $D_{n,u}$.

The two design goals for a prediction region D_n that were stated in the Introduction can now be formulated technically:

DG1 (Conditional coverage probability of D_n). Choose the critical values $\{c_n(u, Y_n)\}$ in (2.1) so that

$$\text{CP}(D_n|Y_n, \theta) \to \alpha \tag{2.5}$$

in $Q_{\theta,n}$-probability as n increases. The value of $\alpha \in (0, 1)$ is fixed in advance.

DG2 (Conditional probability centering of D_n). Subject to DG1, choose the critical values $\{c_n(u, Y_n)\}$ in (2.1) so that

$$\sup_{u \in U} \left|\text{CP}(D_{n,u}|Y_n, \theta) - \beta(\alpha, \theta)\right| \to 0 \tag{2.6}$$

in $Q_{\theta,n}$-probability for some constant $\beta(\alpha, \theta)$ which does not depend on u but can depend on (α, θ).

Note that DG1 and DG2 imply the analogous convergences for unconditional coverage probabilities. The interpretation of DG2 as a probability-centering condition requires a suitable choice of the process Z that appears in (2.1) and (2.2). Basic is the requirement that the variables $\{Z(u, X)\colon u \in U\}$ each measure logically similar attributes of the variable X to be predicted. The following example illustrates this point and the link with classical normal-model prediction ellipsoids.

EXAMPLE 1 (Multivariate normal model). Suppose that the $\{X_i\colon i \geq 1\}$ are iid r-variate $N(\mu, \Sigma)$ random column vectors, the parameter $\theta = (\mu, \Sigma)$ being unknown, with Σ positive-definite. The learning sample is $Y_n = (X_1, \ldots, X_n)$ and the vector to be predicted is $X = X_{n+1}$. Let $\hat{\theta}_n = (\bar{X}_n, S_n)$ denote the usual estimate of $\theta = (\mu, \Sigma)$. A classical prediction ellipsoid for X is

$$D_n = \left\{x\colon \left(x - \bar{X}_n\right)' S_n^{-1}\left(x - \bar{X}_n\right) \leq d_n(\alpha)\right\}, \tag{2.7}$$

where the critical value $d_n(\alpha)$ is the αth quantile of the F distribution with r and $n - r$ degrees of freedom multiplied by the factor $(1 + 1/n)r(n - 1)/(n - r)$. For the underlying distribution theory, see Theorem 5.2.2 in Anderson (1958).

This prediction ellipsoid for X can be rewritten in the form (2.1) as follows. Let $U = \{u \in R^r\colon |u| = 1\}$ be the unit sphere, and define the process Z on U by

$$Z(u, X) = u'X, \qquad u \in U. \tag{2.8}$$

In this example, Z is a Gaussian process on the sphere U. By the derivation of Scheffé's method for simultaneous comparison of linear combinations [cf.

Miller (1981), Section 2 of Chapter 2], Definition (2.7) is equivalent to

$$D_n = \{x: Z(u, X) \le u'\bar{X}_n + s_{n,u} d_n^{1/2}(\alpha), \forall\, u \in U\}, \tag{2.9}$$

where $s_{n,u}^2 = u'S_n u$. Thus, the prediction ellipsoid (2.7) for X is the intersection of the uncountably many prediction half-spaces

$$D_{n,u} = \{x: u'x \le u'\bar{X}_n + s_{n,u} d_n^{1/2}(\alpha)\}, \qquad u \in U. \tag{2.10}$$

Let Φ denote the standard normal cdf and let χ_r denote the cdf of the chi-squared distribution with r degrees of freedom. By direct analysis of (2.7) and (2.10),

$$\begin{aligned} &\mathrm{CP}(D_n|Y_n, \theta) \to \alpha, \\ &\sup_{u \in U} \left|\mathrm{CP}(D_{n,u}|Y_n, \theta) - \Phi\left\{\left[\chi_r^{-1}(\alpha)\right]^{1/2}\right\}\right| \to 0 \end{aligned} \tag{2.11}$$

in $Q_{\theta,n}$-probability. Thus, the prediction ellipsoid (2.7) meets both design goals DG1 and DG2, with Z given by (2.8) and with

$$\beta(\alpha, \theta) = \Phi\left\{\left[\chi_r^{-1}(\alpha)\right]^{1/2}\right\} \tag{2.12}$$

in (2.6).

In what way is the prediction ellipsoid D_n probability-centered? Consider $\bar{D}_{n,u}$, the half-space complementary to $D_{n,u}$. The boundary of $\bar{D}_{n,u}$ is a hyperplane tangent to the ellipsoid D_n, and $\bar{D}_{n,u}$ is thus a supporting half-space for the ellipsoid D_n. From (2.11), $\mathrm{CP}(\bar{D}_{n,u}|Y_n, \theta)$ converges in probability to a limit which stays the same for every u (i.e., for every supporting half-space). It is in this sense that prediction ellipsoid D_n is asymptotically probability-centered.

For the special case of real-valued X, D_n reduces to a prediction interval, probability-centered in large samples as described in the Introduction. For two-dimensional X, Figure 1 illustrates how the prediction ellipse D_n is probability-centered.

2.2. *Characterizing critical values.* The explicit calculation of prediction region D_n in Example 1 drew on the independence and multivariate normality of X and Y_n. The Proposition stated in this section is much more general. It characterizes the asymptotic form of critical values $\{c_n(u, Y_n)\}$ for D_n that achieve design goals DG1 and DG2.

Let $L(U)$ denote the space of all functions on U which take values in $[0, 1]$, metrized by the supremum norm $\|\cdot\|$ taken over U. Without loss of generality, assume that the sample paths of the process $Z(\cdot, X)$ belong to $L(U)$. This property can always be achieved by a strictly monotone transformation of the $\{Z(u, X):\ u \in U\}$ without changing the form (2.1) of D_n. The following regularity assumptions are made on the joint distribution of (X, Y_n). For ways to handle measurability issues concerning infima, see Pollard (1984).

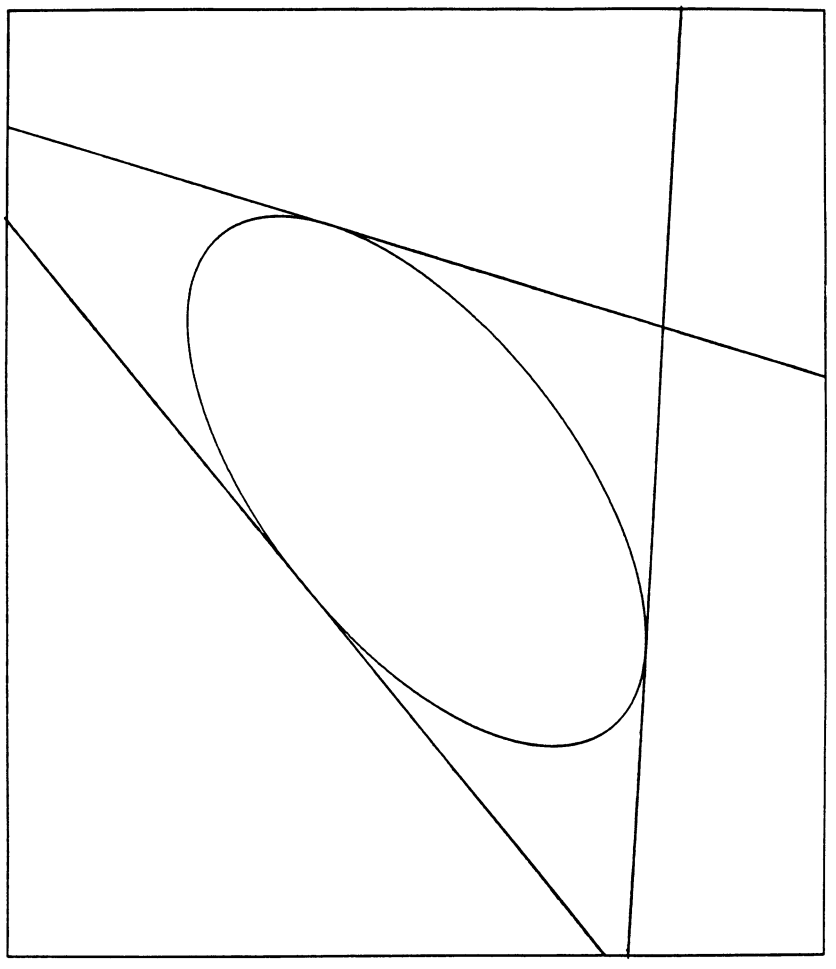

FIG. 1. *A probability-centered prediction ellipse. Each supporting half-plane has equal probability content.*

ASSUMPTION A. There exist statistics $\{V_n = V_n(Y_n): n \geq 1\}$, cdf's $\{A_u(\cdot, \theta, v): u \in U\}$ and a functional $B(\cdot, \theta, v)$ on $L(U)$ such that, for every θ, the following hold:

(i) The $\{V_n\}$ are tight under $Q_{\theta,n}$, as elements of some metric space.
(ii) For every $x \in [0, 1]$ and every $u \in U$,

$$P_\theta[Z(u, X) \leq x | Y_n] = A_u(x, \theta, V_n). \tag{2.13}$$

(iii) For every $f \in L(U)$,

$$P_\theta[Z(u, X) \leq f(u), \forall\, u \in U | Y_n] = B(f, \theta, V_n). \tag{2.14}$$

ASSUMPTION B. For every θ, the cdf's $\{A_u(x,\theta,v)\colon u \in U\}$ on $[0,1]$ are strictly monotone increasing in x and are continuous in (x,v) as elements of $L(U)$. The quantiles $\{A_u^{-1}(x,\theta,v)\colon u \in U\}$ are continuous in (x,v) as elements of $L(U)$.

ASSUMPTION C. For every θ, the functional $B(f,\theta,v)$ is continuous in (f,v) and is strictly monotone increasing in f in the following sense: If $f, g \in L(U)$ and $f(u) > g(u)$ for every $u \in U$, then $B(f,\theta,v) > B(g,\theta,v)$.

Section 2.3 gives four diverse examples that satisfy these assumptions. In view of (2.3) and Assumption A,

$$\begin{aligned} \mathrm{CP}(D_{n,u}|Y_n,\theta) &= A_u\big[c_n(u),\theta,V_n\big], \\ \mathrm{CP}(D_n|Y_n,\theta) &= B\big[c_n(\cdot),\theta,V_n\big]. \end{aligned} \tag{2.15}$$

Define the cdf $A(\cdot,\theta,v)$ on $[0,1]$ by

$$A(x,\theta,v) = B\big[A_\cdot^{-1}(x,\theta,v),\theta,v\big], \tag{2.16}$$

where $A_\cdot^{-1}(x,\theta,v)$ denotes the function on U whose value at $u \in U$ is $A_u^{-1}(x,\theta,v)$. It follows from (2.14) and (2.16) that

$$\begin{aligned} A(x,\theta,V_n) &= P_\theta\big[Z(u,X) \le A_u^{-1}(x,\theta,V_n), \forall\, u \in U|Y_n\big] \\ &= P_\theta\Big[\sup_u A_u\{Z(u,X),\theta,V_n\} \le x|Y_n\Big]. \end{aligned} \tag{2.17}$$

By Assumptions B and C, the cdf $A(x,\theta,v)$ is strictly increasing in x and is continuous in (x,v).

PROPOSITION 1. *Suppose Assumptions* A, B *and* C *hold. Design goals DG*1 *and DG*2 *are met by prediction region* D_n *if and only if the following hold*:

(i) *there exists a cdf* $\tilde{A}(x,\theta)$ *on* $[0,1]$, *strictly monotone increasing and continuous in* x, *such that, for every* (x,θ),

$$A(x,\theta,V_n) \to \tilde{A}(x,\theta) \tag{2.18}$$

in $Q_{\theta,n}$*-probability*; *and*

(ii) *the critical values in* (2.1) *satisfy*

$$\big\|c_n(\cdot,Y_n) - A_\cdot^{-1}\big[\tilde{A}^{-1}(\alpha,\theta),\theta,V_n\big]\big\| \to 0 \tag{2.19}$$

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

REMARKS. Proposition 1 is proved in Section 4. The argument shows that (2.18) and the probability-centering requirement DG2 are linked through the equation

$$\beta(x,\theta) = \tilde{A}^{-1}(x,\theta) \tag{2.20}$$

for every $x \in [0,1]$.

Suppose $\hat{\theta}_n = \hat{\theta}_n(Y_n)$ is a consistent estimator of θ. Then, the plug-in critical values

$$c_n(u, Y_n) = A_u^{-1}\left[A^{-1}\left(\alpha, \hat{\theta}_n, V_n\right), \hat{\theta}_n, V_n\right] \tag{2.21}$$

have property (2.19), under assumptions slightly stronger than those for Proposition 1. Section 3 gives a general bootstrap algorithm for evaluating these critical values.

To see heuristically why the critical values (2.21) achieve design goals DG1 and DG2 under assumption (2.18), note that

$$\begin{aligned} \mathrm{CP}(D_{n,u}|Y_n, \theta) &= P_\theta[Z(u, X) \le c_n(u, Y_n)|Y_n] \\ &\cong A^{-1}\left(\alpha, \hat{\theta}_n, V_n\right) \\ &\cong \tilde{A}^{-1}(\alpha, \theta), \end{aligned}$$

for every u, and that

$$\begin{aligned} \mathrm{CP}(D_n|Y_n, \theta) &= P_\theta[Z(u, X) \le c_n(u, Y_n), \forall\, u \in U|Y_n] \\ &\cong A\left[A^{-1}\left(\alpha, \hat{\theta}_n, V_n\right), \theta, V_n\right] \\ &\cong \alpha. \end{aligned}$$

2.3. *Examples.* We apply Proposition 1 to Example 1 and to three harder examples which do not have closed-form analytical solutions. Technical details are sketched in Section 4. For simplicity, we will write $c_n(u)$, instead of $c_n(u, Y_n)$.

EXAMPLE 1 (Continued). The process Z is defined in (2.8) and V_n is trivially constant because X and Y_n are independent. Explicitly,

$$\begin{aligned} A_u(x, \theta, v) &= \Phi\left[(x - u'\mu)/\sigma_u\right], \\ A(x, \theta, v) &= \chi_r\left[\left\{\Phi^{-1}(x)\right\}^2\right], \end{aligned} \tag{2.22}$$

where $\sigma_u^2 = u'\Sigma u$. The assumptions for Proposition 1 hold, by the argument in Section 4. Consequently, the prediction region

$$D_n = \{x: u'x \le c_n(u), \forall\, u \in U\} \tag{2.23}$$

meets design goals DG1 and DG2 if and only if

$$c_n(u) \to u'\mu + \sigma_u\left\{\chi_r^{-1}(\alpha)\right\}^{1/2} \tag{2.24}$$

in $Q_{\theta,n}$-probability, uniformly in u.

The obvious choice of critical values,

$$c_n(u) = u'\bar{X}_n + s_{n,u}\left\{\chi_r^{-1}(\alpha)\right\}^{1/2}, \tag{2.25}$$

where $s_{n,u}^2$ is the usual estimate of σ_u^2, satisfies (2.24) and yields the prediction ellipsoid

$$D_n = \left\{x: \left(x - \bar{X}_n\right)' S_n^{-1}\left(x - \bar{X}_n\right) \le \chi_r^{-1}(\alpha)\right\} \tag{2.26}$$

from (2.23). The refined critical values, which replace $\chi_r^{-1}(\alpha)$ in (2.25) and (2.26) with the $d_n(\alpha)$ defined following (2.7), still satisfy (2.24) and generate the classical prediction ellipsoid discussed earlier in subsection 2.1. The point of the refinement is to make the unconditional coverage probability of D_n exactly α.

EXAMPLE 2 (Multivariate nonparametric model). Suppose the $\{X_i\colon i \geq 1\}$ are iid r-variate random vectors with unknown absolutely continuous cdf F. The support of F is $\mathbb{R}^r$. The learning sample is $Y_n = (X_1, \ldots, X_n)$ and the vector to be predicted is $X = X_{n+1}$. If the process Z is again defined by (2.8) and U is the unit sphere or a subset thereof, this example is a nonparametric version of Example 1, with $\theta = F$.

Let $J_u(x, F)$ denote the cdf of $u'X$ and let $J(x, F)$ denote the cdf of $\sup\{J_u(u'X)\colon |u| = 1\}$. Evidently,

$$\begin{aligned} A_u(x, F, v) &= J_u(x, F), \\ A(x, F, v) &= J(x, F). \end{aligned} \tag{2.27}$$

By the reasoning in Section 4, the assumptions for Proposition 1 hold. Consequently, DG1 and DG2 are met if and only if the critical values of D_n satisfy

$$\sup_{|u|=1} \left| c_n(u) - J_u^{-1}\left[J^{-1}(\alpha, F), F\right] \right| \to 0 \tag{2.28}$$

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

Let $\hat{F}_n$ denote the empirical cdf of the learning sample. The plug-in critical values

$$c_n(u) = J_u^{-1}\left[J^{-1}\left(\alpha, \hat{F}_n\right), \hat{F}_n\right] \tag{2.29}$$

satisfy (2.28), by a direct argument. The corresponding prediction region for X,

$$D_n = \left\{x\colon u'x \leq J_u^{-1}\left[J^{-1}\left(\alpha, \hat{F}_n\right), \hat{F}_n\right], \forall\, u \in U\right\}, \tag{2.30}$$

is a convex set whose shape depends on the learning sample.

Note that $J_u(x, \hat{F}_n)$ is the empirical cdf of the $\{u'X_i\colon 1 \leq i \leq n\}$ while $J(x, \hat{F}_n)$ is the empirical cdf of the values $\{\sup\{n^{-1} \operatorname{rank}(u'X_i)\colon u \in U\}\colon 1 \leq i \leq n\}$. The interpretation of probability centering parallels that in Example 1, for any reasonable choice of the subset U of the unit sphere. The following numerical example illustrates these points.

Mardia, Kent and Bibby (1979) reported test scores for 88 college students, each of whom took two closed-book and three open-book tests. Earlier analyses of this data by several authors [cf. Example 2 in subsection 2.2 of Beran (1988) for references] have established two points: (a) the data resembles a multivariate normal sample in five dimensions but exhibits certain departures from normality; (b) the closed-book test scores contain information about the students that is not present in the open-book scores.

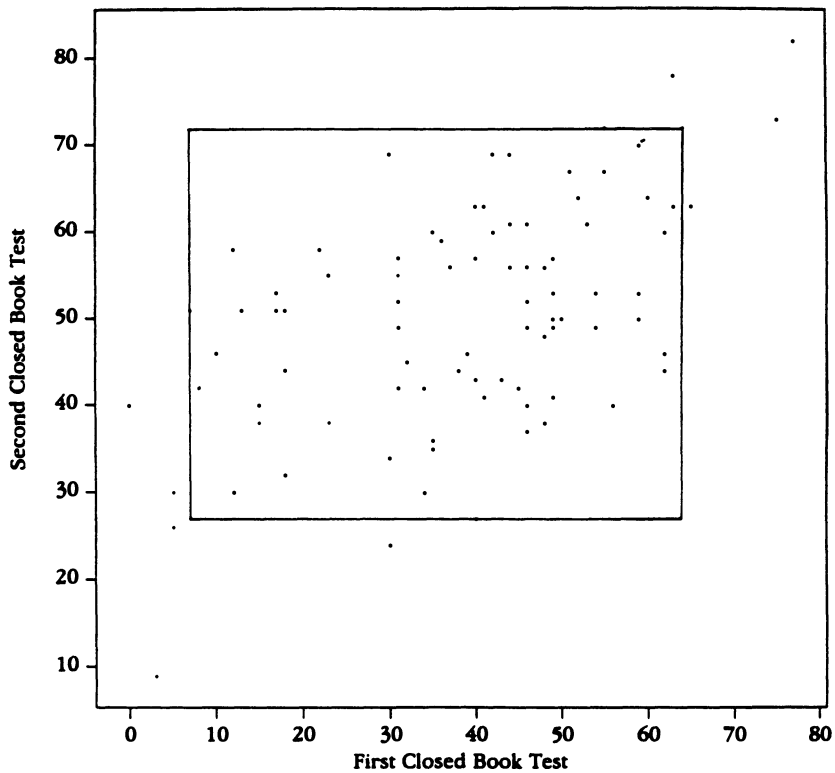

FIG. 2. *Probability-centered nonparametric prediction box for the closed-book test score data. Here coverage probability* $\alpha = 0.90$, *the centering is in directions parallel to the coordinate axes and* $\beta(\alpha, \theta) \doteq 85/88 = 0.97$.

We will consider a different aspect of the data. Figure 2 displays the scatterplot of the closed-book scores. Taking these closed-book scores as the learning sample, suppose that U consists of the four vectors $(\pm 1, 0)$ and $(0, \pm 1)$ and that $\alpha = 0.90$. Then, the prediction region D_n defined in (2.30) is the box $[7, 64] \times [27, 72]$. From Figure 2, it is evident that the plug-in estimate of $\beta(\alpha, F)$ for this D_n is $\beta(\alpha, \hat{F}_n) = 85/88 = 0.966$. This nonparametric prediction box is probability-centered in the horizontal and vertical directions only. It is easy to construct and easy to use in predicting how future students may score on the same tests.

To devise a nonparametric analog to the normal-model prediction ellipse for the closed-book test scores, take U to be all unit vectors in $\mathbb{R}^2$ and $\alpha = 0.90$. Figure 3 displays the convex prediction set D_n defined by (2.30) for this U. From Figure 3, the estimate of $\beta(\alpha, F)$ is now $\beta(\alpha, \hat{F}_n) = 87/88 = 0.989$—larger, as expected, than for the prediction box. This nonparametric prediction set is probability-centered, asymptotically, in every direction. Consequently, it reflects the shape of the scatterplot in a natural way. Constructing

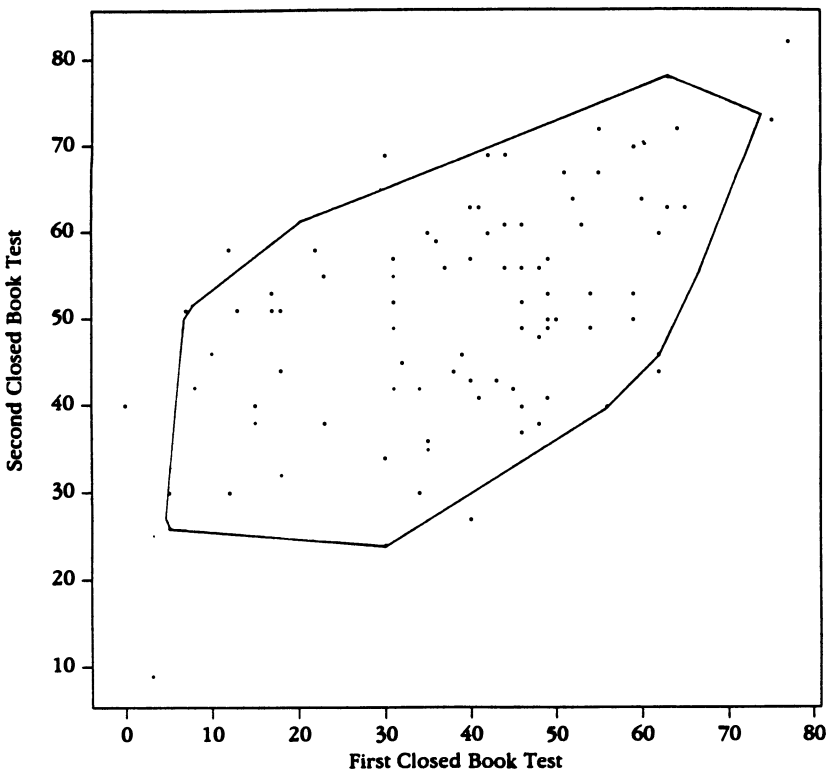

FIG. 3. *Probability-centered nonparametric prediction set for the closed-book test score data. Here coverage probability* $\alpha = 0.90$, *the centering is in all directions and* $\beta(\alpha, \theta) \doteq 87/88 = 0.99$.

this prediction region D_n amounts to the convex hull peeling of Tukey [cf. Green (1985)].

EXAMPLE 3 (Gaussian autoregression). Suppose that $\{X_i\colon i \geq 1\}$ is a stationary sequence of random variables that satisfy the Gaussian AR(1) model

$$X_{i+1} = \theta X_i + E_i, \tag{2.31}$$

where $|\theta| < 1$, θ is otherwise unknown and the $\{E_i\}$ are independent standard normal random variables. The learning sample is $Y_n = (X_1, \ldots, X_n)$. To be predicted are the next two observations in the sequence, $X = (X_{n+1}, X_{n+2})$.

To obtain a probability-centered prediction box for X, define the process $Z = \{Z(u, X)\}$ by setting $U = \{-2, -1, 1, 2\}$ and

$$Z(u, X) = \operatorname{sgn}(u) X_{n+|u|}, \qquad u \in U. \tag{2.32}$$

Then $V_n = X_n$ and

$$
\begin{aligned}
A_1(x, \theta, V_n) &= \Phi(x - \theta X_n), \\
A_{-1}(x, \theta, V_n) &= 1 - \Phi(-x - \theta X_n), \\
A_2(x, \theta, V_n) &= \Phi\left[(1 + \theta^2)^{-1/2}(x - \theta^2 X_n)\right], \\
A_{-2}(x, \theta, V_n) &= 1 - \Phi\left[(1 + \theta^2)^{-1/2}(-x - \theta^2 X_n)\right].
\end{aligned} \tag{2.33}
$$

Let (W_1, W_2) denote a bivariate normal random vector with means 0, variances 1 and covariance $\theta(1 + \theta^2)^{-1/2}$. Then, from (2.33) and (2.17),

$$
\begin{aligned}
A(x, \theta, V_n) &= P\left[\Phi^{-1}(1 - x) \le W_i \le \Phi^{-1}(x), \text{ for } i = 1, 2\right] \\
&= J(x, \theta), \quad \text{say.}
\end{aligned} \tag{2.34}
$$

Checking the assumptions for Proposition 1 is straightforward by the arguments in Section 4. Let $\hat{\theta}_n$ be the least squares estimator of θ, clipped so that $|\hat{\theta}_n| < 1$. The critical values

$$
c_n(u) = \begin{cases}
\hat{\theta}_n X_n + \Phi^{-1}\left[J^{-1}(\alpha, \hat{\theta}_n)\right], & u = 1, \\
-\hat{\theta}_n X_n - \Phi^{-1}\left[1 - J^{-1}(\alpha, \hat{\theta}_n)\right], & u = -1, \\
\hat{\theta}_n^2 X_n + \left(1 + \hat{\theta}_n^2\right)^{1/2} \Phi^{-1}\left[J^{-1}(\alpha, \hat{\theta}_n)\right], & u = 2, \\
-\hat{\theta}_n^2 X_n - \left(1 + \hat{\theta}_n^2\right)^{1/2} \Phi^{-1}\left[1 - J^{-1}(\alpha, \hat{\theta}_n)\right], & u = -2,
\end{cases} \tag{2.35}
$$

thus yield a probability-centered prediction box for $X = (X_{n+1}, X_{n+2})$ that satisfies DG1 and DG2. Note that this box is geometrically centered at $(\hat{\theta}_n X_n, \hat{\theta}_n^2 X_n)$, the usual point predictor for X, and has sides of lengths $2\Phi^{-1}[J^{-1}(\alpha, \hat{\theta}_n)]$ and $2(1 + \hat{\theta}_n^2)^{1/2}\Phi^{-1}[J^{-1}(\alpha, \hat{\theta}_n)]$, respectively. This is a consequence of DG2 and the symmetry of the normal distribution.

A nonparametric version of this example can also be worked through, using ideas from Example 2.

EXAMPLE 4 (Nonparametric prediction band). The fine discussion of gait analysis by Olshen, Biden, Wyatt and Sutherland (1989) suggests the following simplified model. Suppose that $\{X_i\colon i \ge 1\}$ are iid random processes with continuous sample paths on the compact interval $[a, b]$. The distribution P of X_i is unknown. On the basis of the learning sample $Y_n = (X_1, \ldots, X_n)$, the problem is to construct a centered prediction band for the process $X = X_{n+1}$.

For notational convenience, assume without loss of generality that $a \ne 0$. Define the process

$$
Z(u, X) = \begin{cases}
X(u), & \text{if } a \le u \le b, \\
-X(-u), & \text{if } -b \le u \le -a,
\end{cases} \tag{2.36}
$$

on the set $U = [-b, -a] \cup [a, b]$. In this example, the unknown parameter θ is the distribution P of the process X. Let $F_u(\cdot, P)$ denote the cdf of $X(u)$. Then

$$A_u(x, P, v) = \begin{cases} F_u(x, P), & \text{if } a \le u \le b, \\ 1 - F_{-u}(-x, P), & \text{if } -b \le u \le -a, \end{cases} \tag{2.37}$$
$$= J_u(x, P), \quad \text{say,}$$

and

$$A(x, P, v) = P\Big[\sup_{a \le u \le b} \max\{F_u[X(u), P], 1 - F_u[X(u), P]\} \le x\Big] \tag{2.38}$$
$$= J(x, P), \quad \text{say.}$$

Suppose that the cdf's $\{J_u(x, P): u \in U\}$ are equicontinuous in x and the cdf $J(x, P)$ is continuous and strictly monotone in x. Let $\hat{P}_n$ be the empirical distribution of the learning sample. Let $X_{(1)}(u) \le \cdots \le X_{(n)}(u)$ denote the order statistic of the observed processes at time u. Let $r_{n,1}$ and $r_{n,2}$ denote the integer parts of $n[1 - J^{-1}(\alpha, \hat{P}_n)]$ and $nJ^{-1}(\alpha, \hat{P}_n)$, respectively. From the argument in Section 4, the prediction band

$$D_n = \{x: X_{(r_{n,1})}(u) \le x(u) \le X_{(r_{n,2})}(u), \forall\, u \in [a, b]\} \tag{2.39}$$

satisfies design goals DG1 and DG2. The probability-centering of this band is pointwise in u. The width of the prediction band varies with u, to reflect the changing distribution of $X(u)$. The boundaries of D_n are themselves continuous in u.

3. Computational remarks. This section considers two practical points that are related to Proposition 1.

3.1. *Computing critical values.* In general, a bootstrap algorithm can be used to approximate the plug-in critical values (2.21) that satisfy the necessary and sufficient condition of Proposition 1. Let $P_\theta(\cdot|Y_n)$ denote the conditional distribution of X given Y_n. Let $\tilde{X}$ be a random variable whose conditional distribution given Y_n is $P_{\hat{\theta}_n}(\cdot|Y_n)$. From (2.13) and (2.17),

$$A_u(x, \hat{\theta}_n, V_n) = \Pr[Z(u, \tilde{X}) \le x|Y_n] \tag{3.1}$$

and

$$A(x, \hat{\theta}_n, V_n) = \Pr\Big[\sup_u A_u\{Z(u, \tilde{X}), \hat{\theta}_n, V_n\} \le x|Y_n\Big]. \tag{3.2}$$

By drawing independent realizations of $\tilde{X}$, we can construct Monte Carlo approximations to the cdf's A_u and A and hence to the critical values (2.21). When U is infinite, the supremum in (3.2) over all u in U may have to be approximated numerically.

3.2. *Computing* D_n. When the process Z is given by (2.8), D_n is a convex set defined by the intersection of the half-spaces $\{x\colon u'x \le c_n(u),\ \forall\ u \in U\}$. Suppose X is two-dimensional. Graphing D_n with ruler and pencil is straightforward for learning samples of moderate size. Figures 2 and 3 were drawn in this manner. A more systematic graphing method, suitable for routine use and for larger learning samples, would use an efficient algorithm to determine the set intersection of an ordered set of half-spaces. See Middleditch [(1988), pages 228–231] for discussion of the problem and an algorithm. Algorithms for identifying D_n efficiently when X is high-dimensional and U is large are a challenging problem.

4. Proofs. In this section I prove Proposition 1 and sketch how it applies to the examples in subsection 2.3.

PROOF OF PROPOSITION 1. Suppose DG1 and DG2 are both met but (2.18) or (2.19) fails. Because of Assumption A, assume without loss of generality, by going to a subsequence, that $V_n \Rightarrow V$. From Assumption B, if $v_n \to v$, then

$$\sup_x \sup_u \left| A_u(x,\theta,v_n) - A_u(x,\theta,v) \right| \to 0. \tag{4.1}$$

Consequently,

$$\sup_u \left| A_u\left[c_n(u),\theta,V_n\right] - A_u\left[c_n(u),\theta,V\right] \right| \to 0 \tag{4.2}$$

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

From DG2, (2.15) and (4.2),

$$\sup_u \left| A_u\left[c_n(u),\theta,V\right] - \beta(\alpha,\theta) \right| \to 0 \tag{4.3}$$

in $Q_{\theta,n}$-probability. Hence, using the second part of Assumption B,

$$\sup_u \left| c_n(u) - A_u^{-1}\left[\beta(\alpha,\theta),\theta,V\right] \right| \to 0 \tag{4.4}$$

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

On the other hand, from DG1 and (2.15),

$$B\left[c_n(\cdot),\theta,V_n\right] \to \alpha \tag{4.5}$$

in $Q_{\theta,n}$-probability. By (4.3), Assumption C and (2.16),

$$\begin{aligned} B\left[c_n(\cdot),\theta,V_n\right] &\Rightarrow B\left[A^{-1}\{\beta(\alpha,\theta),\theta,V\},\theta,V\right] \\ &= A\left[\beta(\alpha,\theta),\theta,V\right]. \end{aligned} \tag{4.6}$$

Thus, with probability 1,

$$A\left[\beta(\alpha,\theta),\theta,V\right] = \alpha, \tag{4.7}$$

for every α in $[0,1]$. Since the cdf $A(x,\theta,V)$ is strictly monotone and continuous in x, so is $\beta(x,\theta)$ and

$$A(x,\theta,V) = \beta^{-1}(x,\theta). \tag{4.8}$$

Define $\tilde{A}(x, \theta)$ to be $\beta^{-1}(x, \theta)$. Then (2.18) is immediate, while (2.19) follows from (4.4), the definition of $\tilde{A}(x, \theta)$ and the second part of Assumption B. These conclusions contradict the opening supposition at the start of the proof.

The other half of the proposition is easily verified at this point. □

4.1. *Argument for Example* 2. We will use repeatedly the following result: If $\{G_n\}$ and G are continuous cdf's such that G_n converges weakly to G as n increases and G is strictly monotone, then the quantiles $G_n^{-1}(\alpha)$ converge to $G^{-1}(\alpha)$ for every α in $(0, 1)$.

Assumption A is obvious in this example.

The function $\{J_u(x, F)\colon u \in U\}$ is continuous in x as an element of $L(U)$. Suppose not. Recall that U is compact. Then, without loss of generality, there exist $\delta > 0$ and sequences x_n converging to x and u_n converging to u such that

$$\left|J_{u_n}(x_n, F) - J_{u_n}(x, F)\right| \geq \delta, \qquad \forall\, n. \tag{4.9}$$

At the same time, $u'_n X - x_n$ converges weakly to $u'X - x$, which has a continuous distribution because F is absolutely continuous. Thus, both $J_{u_n}(x_n, F)$ and $J_{u_n}(x, F)$ converge to $J_u(x, F)$, contradicting (4.9) and thereby proving the first sentence in this paragraph.

The strict monotonicity in x of $J_u(x, F)$ follows from the full support assumption on F.

The quantile function $\{J_u^{-1}(x, F)\colon u \in U\}$ is continuous in x as an element of $L(U)$. Suppose not. Then, without loss of generality, there exist $\delta > 0$ and sequences x_n converging to x and u_n converging to u such that

$$\left|J_{u_n}^{-1}(x_n, F) - J_{u_n}^{-1}(x, F)\right| \geq \delta, \qquad \forall\, n. \tag{4.10}$$

On the other hand, both $J_{u_n}^{-1}(x_n, F)$ and $J_{u_n}^{-1}(x, F)$ converge to $J_u^{-1}(x, F)$, because of the preceding paragraphs. The contradiction to (4.10) establishes the first sentence in this paragraph.

The last three paragraphs show that Assumption B is satisfied.

To verify Assumption C, observe that

$$B(f, F) = P_F[h(X, f) \leq 0], \tag{4.11}$$

where

$$h(x, f) = \sup_{|u|=1} [u'x - f(u)] \tag{4.12}$$

and f is an element of $L(U)$. Since $h(x, f)$ is convex in x and F is absolutely continuous,

$$P_F[h(X, f) = 0] = 0. \tag{4.13}$$

Continuity of $B(f, F)$ in f now follows from the continuity of $h(x, f)$ in f. Strict monotonicity of $B(f, F)$ in F is ensured by the full support of F.

Proposition 1 is thus applicable to Example 2. A triangular array extension in F of the reasoning above shows that the critical values (2.29) satisfy (2.28), as claimed.

4.2. *Arguments for Examples* 1, 3 *and* 4. The reasoning for Examples 1 and 3 parallels that for Example 2. The conclusions in Example 4 follow by minor modification of Theorem 4.1 in Beran (1988).

REFERENCES

ANDERSON, T. W. (1958). *Introduction to Multivariate Analysis*. Wiley, New York.

BERAN, R. (1988). Balanced simultaneous confidence sets. *J. Amer. Statist. Assoc.* **83** 679–686.

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*, revised ed. Holden-Day, Oakland, CA.

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

BUTLER, R. W., DAVIES, P. L. and JHUN, M. (1993). Asymptotics for the minimum covariance determinant estimator. *Ann. Statist.* **21** 1385–1400.

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

GREEN, P. J. (1985). Peeling data. In *Encyclopedia of Statistical Sciences* **6** 660–664. Wiley, New York.

GUTTMAN, I. (1970). *Statistical Tolerance Regions*. Griffin, London.

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic, New York.

MIDDLEDITCH, A. E. (1988). The representation and manipulation of convex polygons. In *Theoretical Foundations of Computer Graphics and CAD* (R. A. Earnshaw, ed.) 211–252. Springer, New York.

MILLER, R. G., JR. (1981). *Simultaneous Statistical Inference*, 2nd ed. Springer, New York.

OLSHEN, R. A., BIDEN, E. N., WYATT, M. P. and SUTHERLAND, D. H. (1989). Gait analysis and the bootstrap. *Ann. Statist.* **17** 1419–1440.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

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

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720