CHAPTER 8.

BOOTSTRAP AND JACKKNIFE ESTIMATION OF SAMPLING DISTRIBUTIONS

1. A General View of the Bootstrap

We begin with a general approach to bootstrap methods. The goal is to formulate the ideas in a context which is free of particular model assumptions.

Suppose that the data $\underline{X} \sim P_\theta \in \mathbf{P} = \{P_\theta : \theta \in \Theta\}$. The parameter space Θ is allowed to be very general; it could be a subset of R^k (in which case the model $\mathbf{P}$ is a parametric model), or it could the distributions of all i.i.d sequences on some measurable space $(\mathbf{X}, \mathbf{A})$ (in which case the model $\mathbf{P}$ is the "nonparametric i.i.d." model).

Suppose that we have an estimator $\hat{\theta}$ of $\theta \in \Theta$, and thereby an estimator $P_{\hat{\theta}}$ of P_θ. Consider estimation of:

A. The distribution of $\hat{\theta}$: e.g. $P_\theta(\hat{\theta} \in A) = P_\theta(\hat{\theta}(\underline{X}) \in A)$ for a measurable subset A of Θ ;

B. If $\Theta \subset R^k$, $Var_\theta(\underline{a}^T \hat{\theta}(\underline{X}))$ for a fixed vector a .

Natural *(ideal) bootstrap* estimators of these parameters are provided by:

A′. $P_{\hat{\theta}}(\hat{\theta}(\underline{X}^*) \in A)$;

B′. $Var_{\hat{\theta}}(\underline{a}^T \hat{\theta}(\underline{X}^*))$.

While these are often difficult to compute exactly, we can often obtain Monte-Carlo estimates thereof by sampling from $P_{\hat{\theta}}$: Let $\underline{X}_1^*, \cdots, \underline{X}_B^*$ be i.i.d. with common distribution $P_{\hat{\theta}}$, and calculate $\hat{\theta}(\underline{X}_j^*)$ for $j = 1, \cdots, B$. Then *Monte-Carlo approximations (or implementations)* of the bootstrap estimators in A' and B' are given by

A″. $B^{-1} \sum_{j=1}^{B} 1_{[\hat{\theta}(\underline{X}_j^*) \in A]}$;

B″. $B^{-1} \sum_{j=1}^{B} (\hat{\theta}(\underline{X}_j^*) - B^{-1} \sum_{j=1}^{B} \hat{\theta}(\underline{X}_j^*))^2$.

If $\mathbf{P}$ is a parametric model, the above approach yields a *parametric bootstrap*. If $\mathbf{P}$ is a nonparametric model, then this yields a *nonparametric bootstrap*. In the following section, we try to make these ideas more concrete first in the context of $\underline{X} = (X_1, \cdots, X_n)$ are i.i.d. F or P with $\mathbf{P}$ nonparametric so that $P_\theta = F \times \cdots \times F$ and $P_{\hat{\theta}} = IF_n \times \cdots \times IF_n$. [Or if the basic underlying sample space for each X_i is not R, $P_\theta = P \times \cdots \times P$ and $P_{\hat{\theta}} = IP_n \times \cdots \times IP_n$.]

2. Bootstrap Methods

We begin with a discussion of Efron's nonparametric bootstrap; we will then discuss some of the many alternatives.

Efron's Nonparametric Bootstrap

Suppose that $T(F)$ is some functional of F If $X_1, \cdots, X_n$ are iid F then we estimate $T(F)$ by $T(I\!F_n) \equiv T_n$ where $I\!F_n$ is the empirical df, $I\!F_n \equiv n^{-1} \sum_{i=1}^{n} 1[X_i \le x]$. More generally, if $T(P)$ is some functional of P and $X_1, \cdots, X_n$ are i.i.d. P, then a natural estimator of $T(P)$ is just $T(I\!P_n)$ where $I\!P_n$ is the empirical measure $I\!P_n = n^{-1} \sum_1^n \delta_{X_i}$.

Consider estimation of:

A. $b_n(F) \equiv n\{E_F(T_n) - T(F)\}$.

B. $n\sigma_n^2(F) \equiv n Var_F(T_n)$.

C. $\kappa_{3,n}(F) \equiv E_F[T_n - E_F(T_n)]^3 / \sigma_n^3(F)$.

D. $H_n(x, F) \equiv P_F(\sqrt{n}(T_n - T(F)) \le x)$.

E. $K_n(x, F) \equiv P_F(\sqrt{n}||I\!F_n - F||_\infty \le x)$.

F. $L_n(x, P) \equiv Pr_P(\sqrt{n}||I\!P_n - P||_{\mathbf{F}} \le x)$ where $\mathbf{F}$ is a class of functions for which the central limit theorem holds uniformly over $\mathbf{F}$ (i.e. a *Donsker class*).

The *(ideal) nonparametric bootstrap estimates* of these quantities are obtained simply via the *substitution principle*: if F (or P) is unknown, estimate it by the empirical distribution function $I\!F_n$ (or the empirical measure $I\!P_n$). This yields the following nonparametric bootstrap estimates in examples A - F:

A′. $b_n(I\!F_n) \equiv n\{E_{I\!F_n}(T_n) - T(I\!F_n)\}$.

B′. $n\sigma_n^2(I\!F_n) \equiv n Var_{I\!F_n}(T_n)$.

C′. $\kappa_{3,n}(I\!F_n) \equiv E_{I\!F_n}[T_n - E_{I\!F_n}(T_n)]^3 / \sigma_n^3(I\!F_n)$.

D′. $H_n(x, I\!F_n) \equiv P_{I\!F_n}(\sqrt{n}(T_n - T(I\!F_n)) \le x)$.

E′. $K_n(x, I\!F_n) \equiv P_{I\!F_n}(\sqrt{n}||I\!F_n^* - I\!F_n||_\infty \le x)$.

F′. $L_n(x, I\!P_n) \equiv Pr_{I\!P_n}(\sqrt{n}||I\!P_n^* - I\!P_n||_{\mathbf{F}} \le x)$.

Because we usually lack closed - form expressions for the ideal bootstrap estimators in A - F, evaluation of A′ - F′ is usually indirect. Since the empirical df $I\!F_n$ is discrete, we could, in principle, enumerate all possible samples of size n from $I\!F_n$ (or $I\!P_n$) with replacement. If n is large, this is a large number, however: n^n. [Problem: Show that the number of *distinct* bootstrap samples is $\binom{2n-1}{n}$.]

On the other hand, Monte - Carlo approximations to $A' - F'$ are easy: Let

$$(X^*_{j1}, \cdots, X^*_{jn}) \qquad j = 1, \cdots, B$$

be B independent samples of size n drawn *with replacement* from $I\!F_n$ (or $I\!P_n$); let

$$I\!F^*_{j,n}(x) \equiv n^{-1} \sum_{i=1}^{n} 1_{[X^*_{ji} \le x]}$$

be the empirical df of the $j-$th sample, and let

$$T^*_{j,n} \equiv T(I\!F^*_{j,n}) , \quad j = 1, \cdots, B .$$

Then, approximation of $A' - F'$ are given by:

A″. $b^*_{n,B} \equiv \{ \frac{1}{B} \sum_{j=1}^{B} T^*_{j,n} - T_n \}$.

B″. $n\sigma^{*2}_{n,B} \equiv n \frac{1}{B} \sum_{j=1}^{B} \{ T^*_{j,n} - \overline{T^*_n} \}^2$.

C″. $\kappa^*_{3,n,B} \equiv \frac{1}{B} \sum_{j=1}^{B} \{ T^*_{j,n} - \overline{T^*_n} \}^3 / \sigma^{*3}_{n,B}$.

D″. $H^*_{n,B}(x) \equiv \frac{1}{B} \sum_{j=1}^{B} 1_{[\sqrt{n}(T^*_{j,n} - T_n) \le x]}$.

E″. $K^*_{n,B}(x) \equiv \frac{1}{B} \sum_{j=1}^{B} 1_{[\sqrt{n}||I\!F^*_{j,n} - I\!F_n||_\infty \le x]}$.

F″. $L^*_{n,B}(x) \equiv \frac{1}{B} \sum_{j=1}^{B} 1_{[\sqrt{n}||I\!P^*_{j,n} - I\!P_n||_{\mathbf{F}} \le x]}$.

For fixed sample size n and data $I\!F_n$, it follows from the Glivenko - Cantelli theorem that

$$\sup_x |H^*_{n,B}(x) - H_n(x, I\!F_n)| \to_{a.s.} 0 \quad \text{as} \quad B \to \infty ,$$

and by Donsker's theorem,

$$\sqrt{B}(H^*_{n,B}(x) - H_n(x, I\!F_n)) \Longrightarrow I\!U^{**}(H_n(x, I\!F_n)) \quad \text{as} \quad B \to \infty .$$

Moreover, by the Dvoretzky, Kiefer, Wolfowitz (1956) inequality ($P(||I\!U_n|| \ge \lambda) \le C\exp(-2\lambda^2)$ for all n and $\lambda > 0$), it follows that

$$P(\sup_x |H^*_{n,B}(x) - H_n(x, I\!F_n)| \ge \epsilon) \le C\exp(-2B\epsilon^2) ;$$

it is now known that the constant C can be taken to be 2 ; see Massart (1990).

For a given $\epsilon > 0$ we can make this probability as small as we please by choosing B, over which we have complete control given sufficient computing power, sufficiently large. Since the deviations of $H^*_{n,B}$ from $H_n(x, IF_n)$ are so well - controlled and understood, much of our discussion below will focus on the differences between $H_n(x, IF_n)$ and $H_n(x, F)$.

Sometimes it is possible to compute the distribution of the bootstrap estimator explicitly without resort to Monte-Carlo; here is an example of this kind.

Example 1.1. (The distribution of the bootstrap estimator of the median). Suppose that $T(F) = F^{-1}(1/2)$. Then

$$T(IF_n) = IF_n^{-1}(1/2) = X_{([n+1]/2)}$$

and

$$T(IF_n^*) = IF_n^{*-1}(1/2) = X^*_{([n+1]/2)} .$$

Let $m \equiv [n+1]/2$, and let $M_j \equiv \#\{X_i^* = X_j(\omega) : i = 1, \cdots, n\}$, $j = 1, \cdots, n$ so that

$$\underline{M} \equiv (M_1, \cdots, M_n) \sim Mult_n(n, (\frac{1}{n}, \cdots, \frac{1}{n})) .$$

Now $[X^*_{(m)} \geq X_{(k)}(\omega)] = [\, nIF_n^*(X_{(k)}(\omega)) \leq m-1\,]$, and hence

$$\begin{aligned} P(T(IF_n^*) = X^*_{(m)} > X_{(k)}(\omega) \mid IF_n) &= P(\, nIF_n^*(X_{(k)}(\omega)) \leq m-1 \mid IF_n) , \\ &= P(Binomial(n, \frac{k}{n}) \leq m-1) \\ &= \sum_{j=0}^{m-1} \binom{n}{j} (\frac{k}{n})^j (1-\frac{k}{n})^{n-j} . \end{aligned}$$

This implies that

$$P(T(IF_n^*) = X_{(k)}(\omega) \mid IF_n) = \sum_{j=0}^{m-1} \{\binom{n}{j} (\frac{k-1}{n})^j (1-\frac{k-1}{n})^{n-j} - \binom{n}{j} (\frac{k}{n})^j (1-\frac{k}{n})^{n-j}\} ,$$

for $k = 1, \cdots, n$.

Example 1.2. (Standard deviation of a correlation coefficient estimator). Let $T(F) = \rho(F)$ where F is the bivariate distribution of a pair of random variables (X, Y) with finite fourth moments. We know from Chapter 2 that the sample correlation coefficient $\hat\rho_n \equiv T(IF_n)$ satisfies

$$\sqrt{n}(\hat\rho_n - \rho) \equiv \sqrt{n}(\rho(IF_n) - \rho(F)) \to_d N(0, V^2)$$

where $V^2 = Var[Z_1 - (\rho/2)[Z_2 + Z_3]]$ where $\underline{Z} \equiv (Z_1, Z_2, Z_3) \sim N_3(0, \Sigma)$ and Σ is given by

$$\Sigma = E(X_s Y_s - \rho, X_s^2 - 1, Y_s^2 - 1)^{\otimes 2} ;$$

here $X_s \equiv (X - \mu_X)/\sigma_X$ and $Y_s \equiv (Y - \mu_Y)/\sigma_Y$ are the standardized variables. If F is bivariate normal, then $V^2 = (1 - \rho^2)^2$.

Consider estimation of the standard deviation of $\hat{\rho}_n$:

$$\sigma_n(F) \equiv \{Var_F(\hat{\rho}_n)\}^{1/2} .$$

The normal theory estimator of $\sigma_n(F)$ is

$$(1 - \hat{\rho}_n^2)/\sqrt{n-3} .$$

The delta - method estimate of $\sigma_n(F)$ is

$$\frac{\hat{V}_n}{\sqrt{n}} = \{\hat{Var}[Z_1 - (\rho/2)[Z_2 + Z_3]]\}^{1/2}/\sqrt{n} .$$

The (Monte-Carlo approximation to) the bootstrap estimate of $\sigma_n(F)$ is

$$\sqrt{B^{-1}\sum_{j=1}^{B}[\hat{\rho}_j^* - \overline{\rho^*}]^2} .$$

Finally the *jackknife estimate* of $\sigma_n(F)$ is

$$\sqrt{\frac{n-1}{n}\sum_{j=1}^{n}[\hat{\rho}_{(i)} - \overline{\hat{\rho}_{(\cdot)}}]^2} ;$$

see the beginning of section 2 for the notation used here. We will discuss the jackknife further in sections 2 and 4.

Parametric Bootstrap Methods

Once the idea of nonparametric bootstrapping (sampling from the empirical measure $I\!P_n$) becomes clear, it seems natural to consider sampling from other estimators of the unknown P . For example, if we are quite confident that some parametric model holds, then it seems that we should consider bootstrapping by sampling from an estimator of P based on the parametric model. Here is a formal description of this type of model - based bootstrap procedure.

Let $(\mathbf{X}, \mathbf{A})$ be a measurable space, and let $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ be a model, parametric, semiparametric, or nonparametric. We do not insist that Θ be finite - dimensional. For example, in a parametric extreme case, $\mathbf{P}$ could be the family of all normal (Gaussian) distributions on $(\mathbf{X}, I\!A) = (R^d, \mathbf{B}^d)$. Or, to give a nonparametric example with only a smoothness restriction, $\mathbf{P}$ could be the family of all distributions on $(\mathbf{X}, I\!A) \equiv (R^d, \mathbf{B}^d)$ with a density with respect to Lebesgue measure which is uniformly continuous.

Let $X_1, X_2, \cdots, X_n, \cdots$ be iid with distribution $P_\theta \in \mathbf{P}$. We assume that there exists an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \cdots, X_n)$ of θ. Then Efron's parametric (or model - based) bootstrap proceeds by sampling from the estimated or fitted model $P_{\hat{\theta}_n(\omega)} \equiv \hat{P}_n^{\omega}$: suppose that $X_{n1}^*, \cdots, X_{nn}^*$ are independent and identically distributed with distribution $\hat{P}_n^{\omega}$ on $(\mathbf{X}, \mathit{IA})$, and let

$$(3.1) \qquad \mathit{IP}_n^* \equiv n^{-1} \sum_{i=1}^{n} \delta_{X_{ni}^*} \equiv \text{the } \textit{parametric bootstrap empirical measure} .$$

The key difference between this parametric bootstrap procedure and the nonparametric bootstrap discussed earlier in this section is that we are now sampling from the model - based estimator $\hat{P}_n \equiv P_{\hat{\theta}_n}$ of P rather than from the nonparametric estimator IP_n.

Example 1.3. Suppose that $X_1, \cdots, X_n$ are i.i.d. $P_\theta = N(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$. Let $\hat{\theta}_n = (\hat{\mu}_n, \hat{\sigma}_n^2) = (\overline{X}_n, S_n^2)$ where S_n^2 is the usual unbiased estimator of σ^2, so that

$$\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\hat{\sigma}_n} \sim t_{n-1}, \qquad \frac{(n-1)\hat{\sigma}_n^2}{\sigma^2} \sim \chi_{n-1}^2 .$$

Now $P_{\hat{\theta}_n} = N(\hat{\mu}, \hat{\sigma}_n^2)$, and if $X_1^*, \cdots, X_n^*$ are i.i.d. $P_{\hat{\theta}_n}$, then the bootstrap estimators $\hat{\theta}_n^* = (\hat{\mu}_n^*, \hat{\sigma}_n^{*2})$ satisfy, conditionally on IF_n,

$$\frac{\sqrt{n}(\hat{\mu}_n^* - \hat{\mu}_n)}{\hat{\sigma}_n^*} \sim t_{n-1}, \qquad \frac{(n-1)\hat{\sigma}_n^{*2}}{\hat{\sigma}_n^2} \sim \chi_{n-1}^2 .$$

Thus the bootstrap estimators have exactly the same distributions as the original estimators in this case.

Example 1.4. Suppose that $X_1, \cdots, X_n$ are i.i.d. $P_\theta = \text{exponential}(1/\theta)$: $P_\theta(X_1 > t) = \exp(-t/\theta)$ for $t \geq 0$. Then $\hat{\theta}_n = \overline{X}_n$ and $(n\hat{\theta}_n/\theta \,|\, \mathit{IF}_n) \sim \text{Gamma}(n, 1)$. Now $P_{\hat{\theta}_n} = \text{exponential}(1/\hat{\theta}_n)$, and if $X_1^*, \cdots, X_n^*$ are i.i.d. $P_{\hat{\theta}_n}$, then $\hat{\theta}_n^* = \overline{X}_n^*$ has $(n\hat{\theta}_n^*/\hat{\theta}_n \,|\, \mathit{IF}_n) \sim \text{Gamma}(n, 1)$, so the bootstrap distribution replicates the original estimator exactly.

Example 1.5. (Bootstrapping from a "smoothed empirical measure"; or, the "smoothed bootstrap"). Suppose that

$$\mathbf{P} = \{ P \text{ on } (R^d, \mathbf{B}^d) : \; p \equiv \frac{dP}{d\lambda} \text{ exists and is uniformly continuous} \} .$$

Then one way to estimate P so that our estimator $\hat{\mathit{IP}}_n \in \mathbf{P}$ is via a kernel estimator of the density p :

$$\hat{p}_n(x) = \frac{1}{b_n^d} \int k\left(\frac{y - x}{b_n}\right) d\mathit{IP}_n(y)$$

where $k : R^d \rightarrow R$ is a uniformly continuous density function. Then $\hat{\mathit{IP}}_n$ is defined for $C \in \mathit{IA}$ by

$$I\hat{P}_n(C) = \int 1_C(x)\hat{p}_n(x)dx ,$$

and the model - based bootstrap proceeds by sampling from $I\hat{P}_n$.

There are many other examples of this type involving nonparametric or semiparametric models $\mathbf{P}$. For some work on "smoothed bootstrap" methods see e.g. Silverman and Young (1987) and Hall, DiCiccio, and Romano (1989).

"Bayesian" and other "Exchangeably - weighted" Bootstrap Methods

In the course of example 1.1 we introduced the vector $\underline{M}$ of counts of how many times the bootstrap variables X_i^* equal the observations $X_j(\omega)$ in the underlying sample. Thinking about the process of sampling at random (with replacement) from the population described by the empirical measure IP_n , it becomes clear that we can think of the bootstrap empirical measure IP_n^* as the empirical measure with multinomial random weights:

$$IP_n^* = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^*} = \frac{1}{n}\sum_{i=1}^n M_i\,\delta_{X_i(\omega)}$$

This view of Efron's nonparametric bootstrap as the empirical measure with random weights suggests that we could obtain other random measures which would behave much the same way as Efron's nonparametric bootstrap -- but without the same random sampling interpretation -- by replacing the vector of multinomial weights by some other random vector $\underline{W}$. One of the possible defiencies of the nonparametric bootstrap involves its “discreteness” via missing observations in the original sample: Note that the number of points of the original sample which are missed (or not given any bootstrap weight) is $N_n \equiv \#\{j \le n : M_j = 0\} = \sum_{j=1}^n 1_{[M_j = 0]}$. Hence the proportion of observations missed by the bootstrap is $n^{-1}N_n$, and the expected proportion of missed observations is

$$E(n^{-1}N_n) = P(M_1 = 0) = (1 - \frac{1}{n})^n \rightarrow e^{-1} \doteq .36787\ldots$$

[Moreover, from occupancy theory for urn models

$$\sqrt{n}(n^{-1}N_n - (1 - \frac{1}{n})^n) \rightarrow_d N(0, e^{-1}(1 - 2e^{-1})) = N(0, .09720887\cdots) ;$$

see e.g. Johnson and Kotz (1977), page 317, 3. with $r = 0$.] By using some other vector of exchangeable weights $\underline{W}$ rather than $\underline{M} \sim Mult_n(n, (1/n\,, \cdots, 1/n))$ we might be able to avoid some of this discreteness caused by multinomial weights.

Since the resulting measure should be a probability measure, it seems reasonable to require that the components of $\underline{W}$ should sum to n . Since the multinomial random vector with cell probabilities all equal to $1/n$ is exchangeable, it seems reasonable to require that the vector $\underline{W}$ have an *exchangeable distribution*: i.e. $\pi\underline{W} \equiv (W_{\pi(1)}, \cdots, W_{\pi(n)}) \stackrel{d}{=} \underline{W}$ for all permutations π of $\{1, \cdots, n\}$. Then

$$I\!P_n^W \equiv \frac{1}{n}\sum_{i=1}^n W_{ni}\delta_{X_i(\omega)}$$

is called the *exchangeably weighted bootstrap empirical measure* corresponding to the weight vector $\underline{W}$. Here are several examples:

Example 1.6. (Dirichlet weights). Suppose that $Y_1, Y_2, \cdots$ are i.i.d. exponential(1) random variables, and set

$$W_{ni} \equiv \frac{nY_i}{Y_1 + \cdots + Y_n}, \qquad i = 1, \cdots, n\,.$$

The resulting random vector $\underline{W}/n$ has a Dirichlet$(1, \cdots, 1)$ distribution; i.e. $n^{-1}\underline{W} \stackrel{d}{=} \underline{D}$ where the D_i 's are the *spacings* of a random sample of $n-1$ Uniform$(0, 1)$ random variables.

Example 1.7 (More general continuous weights). Other weights $\underline{W}$ of the same form as in example 1.6 are obtained by replacing the exponential distribution of the Y_i 's by some other distribution on R^+ . It will turn out that the limit theory can be easily established for any of these weights as long as the Y_i 's satisfy $Y_i \in L_{2,1}$: $\int_0^\infty \sqrt{P(|Y| \geq t)}\,dt < \infty$.

Example 1.8 (Jackknife weights). Suppose that $\underline{w} = (w_{n1}, \cdots, w_{nn})$ is a vector of constants which sum to n : $\sum_1^n w_{ni} = n$. Set $\underline{W}$ be a random permutation of the coordinate of $\underline{w}$: if $\underline{R}$ is uniformly distributed over $\Pi \equiv$ {all permutations of $\{1, \cdots, n\}$}, then $\underline{W} \equiv \underline{R}\underline{w} \equiv (w_{nR_1}, \cdots, w_{nR_n})$. If we take $\underline{w} = (n/(n-d))\underline{1}_{n-d} = (n/n-d)(1, \cdots, 1, 0, \cdots, 0)$ where $\underline{1}_{n-d}$ is the vector with all 1 's in the first $n-d$ coordinates and 0 's in the remaining d coordinates, then these weights W_{ni} correspond to the *delete - d jackknife*. It turns out that these weights yield behavior like that of Efron's nonparametric bootstrap (with multinomial weights) only if $d = d_n$ satisfies $n^{-1}d_n \to \alpha > 0$.

Other weights $\underline{W}$ based on various urn schemes are possible: see Praestgaard and Wellner (1993) for some of these.

3. The Jackknife

The jackknife preceded the bootstrap -- mostly due to its simplicity and relative ease of computation. The original work on the "delete - one" jackknife is due to Quenouille (1949) and Tukey (1958). Here is how it works.

Suppose that $T(I\!F_n)$ estimates $T(F)$. Let

$$T_{n,i} \equiv T(I\!F_{n-1,i}) \qquad \text{where} \qquad I\!F_{n-1,i}(x) \equiv \frac{1}{n-1}\sum_{j \neq i} 1_{[X_j \leq x]} \ ;$$

thus $T_{n,i}$ is the estimator based on the data with X_i *deleted* or left out. Let

$$T_{n,\cdot} \equiv \frac{1}{n}\sum_{i=1}^{n} T_{n,i} \ .$$

We also set

$$T^*_{n,i} \equiv nT_n - (n-1)T_{n,i} \equiv \text{i th pseudo - value}$$

and $\overline{T}^*_n \equiv n^{-1}\sum_1^n T^*_{n,i} = n\,T_n - (n-1)T_{n,\cdot}$.

The Jackknife estimator of bias, and the Jackknife estimator of $T(F)$

Now let $E_n \equiv E_F T_n = E_F T(I\!F_n)$, and suppose that we can expand E_n in powers of n^{-1} as follows:

$$E_n \equiv E_F T_n = T(F) + \frac{a_1(F)}{n} + \frac{a_2(F)}{n^2} + \cdots .$$

Then the bias of the estimator $T_n = T(I\!F_n)$ is

$$bias_n(F) \equiv E_F T_n - T(F) = \frac{a_1(F)}{n} + \frac{a_2(F)}{n^2} + \cdots .$$

We can also write

$$T(F) = E_F(T_n) - bias_n(F) .$$

Note that

$$E_F T_{n,\cdot} = E_{n-1} = T(F) + \frac{a_1(F)}{n-1} + \frac{a_2(F)}{(n-1)^2} + \cdots .$$

Hence it follows that

$$\begin{aligned} E_F(\overline{T}^*_n) &= n\,E_n - (n-1)\,E_{n-1} \\ &= T(F) + a_2(F)\{\frac{1}{n} - \frac{1}{n-1}\} + a_3(F)\{\frac{1}{n^2} - \frac{1}{(n-1)^2}\} + \cdots \\ &= T(F) - \frac{a_2(F)}{n(n-1)} + \cdots . \end{aligned}$$

Thus $\overline{T}_n^*$ has bias $O(n^{-2})$ whereas T_n has bias of the order $O(n^{-1})$ if $a_1(F) \neq 0$. We will call $\overline{T}_n^*$ the *jackknife estimator of* $T(F)$; similarly, by writing

$$\overline{T}_n^* = T_n - \hat{bias}_n$$

we find that

$$\hat{bias}_n = T_n - \overline{T}_n^* = (n-1)\{T_{n,\cdot} - T_n\} .$$

Example 2.1. If $T(F) = E_F(X) = \int x\,dF(x)$ so that $T_n = \bar{X}_n$, then $T_{n,i}^* = nT_n - (n-1)T_{n,i} = X_i$, so $\overline{T}_n^* = \bar{X} = T_n$ and $\hat{bias}_n = 0$.

Example 2.2. If $T(F) = Var_F(X) = \int (x - \int y\,dF(y))^2\,dF(x)$ so that $T_n = T(IF_n) = n^{-1}\sum_1^n (X_i - \bar{X})^2$, the empirical (biased!) estimator of $T(F)$, then $E_n = [(n-1)/n]T(F) = T(F) - T(F)/n$, and algebra shows that the jackknife estimator of $T(F)$ is $\overline{T}_n^* = \sum_1^n (X_i - \bar{X})^2/(n-1)$, the usual unbiased estimator of $T(F)$. The bias estimator is just

$$\hat{bias}_n = -\frac{1}{n(n-1)}\sum_{i=1}^n (X_i - \bar{X})^2 .$$

The Jackknife estimator of variance

Now consider estimation of

$$Var_n \equiv Var_F(T_n) = Var_F(T(IF_n)) .$$

Tukey's jackknife estimator of Var_n is

$$\hat{Var}_n = \frac{n-1}{n}\sum_{i=1}^n [T_{n,i} - T_{n,\cdot}]^2$$

$$= \frac{1}{n(n-1)}\sum_{i=1}^n [T_{n,i}^* - \overline{T}_n^*]^2 \equiv \frac{n-1}{n}\tilde{Var}_{n-1} . \tag{2.1}$$

Since

$$Var(\bar{X}_n) = \frac{\sigma^2}{n} = \frac{n-1}{n}\frac{\sigma^2}{n-1} = \frac{n-1}{n}Var(\bar{X}_{n-1}) ,$$

we can regard the factor of $(n-1)/n$ as an adjustment from sample size $n-1$ to sample size n and $\tilde{Var}_{n-1}$ as an estimator of $Var_{n-1} \equiv Var_F(T_n)$. The following result of Efron and Stein (1981) shows that the jackknife estimate $\tilde{Var}_{n-1}$ of Var_{n-1} is always biased upwards:

Theorem 2.3. (Efron and Stein, 1981). $E(\tilde{Var}_{n-1}) \geq Var_{n-1}$.

Proof. See Efron (1982), chapter 4, or Efron and Stein (1981). The proof proceeds by way of the (Hoeffding) U-statistic decomposition of an arbitrary symmetric statistic. □

For further discussion of the relationship between the jackknife and the bootstrap, see Efron and Tibshirani, pages 145 - 148 and 287. They show that the jackknife can be view as an approximation to the bootstrap (via linearization - i.e. the delta method).

Unfortunately, the jackknife estimate of variance *fails* for many functionals $T(F)$ which are not sufficiently smooth. In fact, it fails for the median functional $T(F) = F^{-1}(1/2)$: For this $T(F)$ and $n = 2m$, if $g \equiv F^{-1}$ has a continuous derivative in a neighborhood of $1/2$,

$$(2.2) \qquad n\hat{Var}_n = n(n-1)\{\frac{X_{(m+1)} - X_{(m)}}{2}\}^2 \rightarrow_d \frac{1}{4f^2(F^{-1}(1/2))}(\frac{\chi_2^2}{2})^2 .$$

Now $Y \equiv (\chi_2^2/2)^2$ has $E(Y) = 2$, $Var(Y) = 20$, *and is random*! On the other hand,

$$\sqrt{n}(T(I\!F_n) - T(F)) \rightarrow_d N(0, \frac{1}{4f^2(F^{-1}(1/2))})$$

and if $E_F|X|^r < \infty$ for some $r > 0$, then

$$n\, Var_F(T(I\!F_n)) \rightarrow \frac{1}{4f^2(F^{-1}(1/2))}$$

by uniform integrability arguments. Thus the jackknife estimator of variance is not consistent for the median functional.

Proof of (2.2): Now $X_{(i)} \stackrel{d}{=} F^{-1}(\xi_{(i)})$, $1 \leq i \leq n$ where $0 \leq \xi_{(1)} \leq \cdots \leq \xi_{(n)} \leq 1$ are the order statistics of a sample of n Uniform$(0,1)$ random variables. Moreover, for any $i = 1, \cdots, n$,

$$n(\xi_{(i)} - \xi_{(i-1)}) \stackrel{d}{=} n\,\xi_{(1)} \rightarrow_d \text{exponential}(1) .$$

Thus if $g \equiv F^{-1}$ has a continuous derivative in a neighborhood of $1/2$, by the mean value theorem

$$\begin{aligned} n[X_{(m+1)} - X_{(m)}] &\stackrel{d}{=} \frac{g(\xi_{(m+1)}) - g(\xi_{(m)})}{\xi_{(m+1)} - \xi_{(m)}}\, n\,(\xi_{(m+1)} - \xi_{(m)}) \\ &= g'(\xi_{(m)} + \theta(\xi_{(m+1)} - \xi_{(m)}))\, n\,(\xi_{(m+1)} - \xi_{(m)}) \\ &\qquad \text{where } |\theta| \leq 1 \\ &= g'(I\!G_n^{-1}(1/2) + \theta n(\xi_{(m+1)} - \xi_{(m)})/n)\, n\,(\xi_{(m+1)} - \xi_{(m)}) \end{aligned}$$

$$\to_d \; g'(1/2+0)\exp(1)$$

by Slutsky's theorem, continuity of g', and $||\mathbb{G}_n^{-1} - I||_\infty \to_{a.s.} 0$. Hence we have

$$\begin{aligned} n\hat{Var}_n &= \frac{n-1}{4n}\{n[X_{(m+1)} - X_{(m)}]\}^2 \\ &\to_d \frac{1}{4}\, g'(1/2)^2\, exp(1)^2 \qquad \text{by the Mann - Wald theorem} \\ &\stackrel{d}{=} \frac{1}{4f^2(F^{-1}(1/2))}\,(\frac{\chi_2^2}{2})^2 \end{aligned}$$

since $g' = 1/f(F^{-1})$ and $2\,exp(1) \stackrel{d}{=} \chi_2^2$. □

The delete - d Jackknife

See Shao and Wu (1989), Shao (1993), and Praestgaard and Wellner (1993) for more on delete - d jackknife methods.

4. Some limit theory for bootstrap methods

We begin again with Efron's nonparametric bootstrap. Our goal will be to show that the asymptotic behavior of the distribution of the nonparametric bootstrap estimator "mimics" the behavior of the original estimator in probability or almost surely: if we are estimating $T(P)$ by $T(I\!P_n)$ and we know (perhaps by a delta method argument) that

$$\sqrt{n}(T(I\!P_n) - T(P)) \rightarrow_d N(0, V^2(P)) ,$$

then our goal will be to show that the bootstrap estimator satisfies

$$\sqrt{n}(T(I\!P_n^*) - T(I\!P_n)) \rightarrow_d N(0, V^2(P)) \qquad \text{in probability or a.s.} .$$

For concreteness, first consider the sample mean of a distribution P on R: if $X \sim P$ and $E X^2 < \infty$, then for $T(P) = \int x\, dP(x) = \mu(P)$ we know that

$$\sqrt{n}(T(I\!P_n) - T(P)) = \sqrt{n}(\overline{X}_n - \mu(P)) \rightarrow_d N(0, Var(X)) .$$

The corresponding statement for the bootstrap is:

Theorem 3.1. If $E X^2 < \infty$, then for a.e. sequence $X_1, X_2, \cdots$,

$$\sqrt{n}(T(I\!P_n^*) - T(I\!P_n)) = \sqrt{n}(\overline{X}_n^* - \overline{X}_n) \rightarrow_d N(0, Var(X)) .$$

Proof. Now $E_* X_{ni}^* = n^{-1}\sum_{i=1}^n X_i^\omega = \overline{X}^\omega$, and

$$Var_*(X_{ni}^*) = \frac{1}{n}\sum_{i=1}^n (X_i^\omega - \overline{X}^\omega)^2 \equiv S_n^2 .$$

It follows that

$$\sqrt{n}(\overline{X}_n^* - \overline{X}^\omega) = \sum_{i=1}^n Z_{ni}$$

where $Z_{ni} \equiv n^{-1/2}(X_{ni}^* - \overline{X}^\omega)$, $i = 1, \cdots, n$ have $E_* Z_{ni} = 0$, $\sigma_{ni}^2 = Var_*(Z_{ni}) = \frac{1}{n} S_n^2$, and $\sigma_n^2 = \sum_{i=1}^n \sigma_{ni}^2 = S_n^2 \rightarrow_{a.s.} \sigma^2$. Finally, for $\epsilon > 0$, the Lindeberg condition is

$$\frac{1}{\sigma_n^2}\sum_{i=1}^n E_*|Z_{ni}|^2 1_{[|Z_{ni}| > \epsilon\sigma_n]}$$

$$= \frac{1}{S_n^2} n E_* |n^{-1/2}(X_{ni}^* - \overline{X}_n^\omega)|^2 1_{[|X_{ni}^* - \overline{X}^\omega| > \sqrt{n}\epsilon S_n]}$$

$$= \frac{1}{S_n^2}\frac{1}{n}\sum_{i=1}^n |X_i^\omega - \overline{X}_n^\omega|^2 1_{[|X_i^\omega - \overline{X}^\omega| > \epsilon\sqrt{n} S_n]}$$

$$\leq 1_{[\max_{1 \leq i \leq n} |X_i^\omega - \bar{X}^\omega| > \epsilon\sqrt{n}\, S_n]}$$

$$\to 0 \quad \text{a.s.}$$

since $E|X - \mu|^2 < \infty$ implies that

$$\frac{1}{\sqrt{n}} \max_{1 \leq i \leq n} |X_i^\omega - \bar{X}^\omega| \leq \frac{1}{\sqrt{n}} \max_{1 \leq i \leq n} |X_i^\omega - \mu| + \frac{1}{\sqrt{n}} |\mu - \bar{X}^\omega| \to_{a.s.} 0 .$$

hence the theorem follows from the Lindeberg - Feller CLT. □

The above proof is basically from Bickel and Freedman (1981). The more refined statements of the following theorem are due to K. Singh (1981).

Theorem.

A. If $E(X^2) < \infty$, then

$$D_n \equiv D_n(\underline{X}) \equiv ||P^*(\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x) - P(\sqrt{n}(\bar{X}_n - E_F(X)) \leq x)||_\infty \to_{a.s.} 0 .$$

B. If $E(X^4) < \infty$, then

$$\limsup_{n \to \infty} \frac{\sqrt{n}}{(\log\log n)^{1/2}} D_n = \frac{\sqrt{Var[(X - \mu)^2]}}{2\sigma^2\sqrt{2\pi e}} \quad \text{a.s.}$$

C. If $E|X|^3 < \infty$, and

$$D_n^s \equiv ||P^*(\sqrt{n}(\bar{X}_n^* - \bar{X}_n)/S_n \leq x) - P(\sqrt{n}(\bar{X}_n - E_F(X))/\sigma \leq x)||_\infty$$

where $S_n^2 \equiv n^{-1}\sum_1^n (X_i - \bar{X})^2$, then

$$\limsup_{n \to \infty} \sqrt{n}\, D_n^s \leq K\rho/\sigma^3 \qquad \text{a.s.}$$

where $\rho \equiv E|X - \mu|^3 < \infty$ and K is the universal constant of the Berry - Esseen bound.

D. If $E|X|^3 < \infty$ and F is nonlattice, then

$$P^*(\sqrt{n}(\bar{X}_n^* - \bar{X}_n)/S_n \leq x) = \Phi(x) + \{\mu_3(1 - x^2)/(6\sigma^3 n^{1/2})\}\phi(x) + o(n^{-1/2})$$

uniformly in x a.s. where Φ and ϕ are the standard normal d.f. and standard normal density function respectively; hence in this case

$$\lim_{n \to \infty} \sqrt{n}\, D_n^s = 0 \qquad \text{a.s. .}$$

Proof. See Singh (1981). □

Now we turn to the corresponding behavior of the bootstrap empirical distribution function $\mathbb{F}_n^*$ (or bootstrap empirical measure $\mathbb{P}_n^*$). We know that for $\mathbf{X} = R$ we have, by the inverse transformation,

$$\sqrt{n}(\mathbb{F}_n - F) \stackrel{d}{=} \mathbb{U}_n(F) \Longrightarrow \mathbb{U}(F)$$

where $\mathbb{U}_n$ is the empirical process of n i.i.d. Uniform$(0,1)$ random variables and $\mathbb{U}$ is a Brownian bridge process on $[0,1]$. The following theorem says that the bootstrap mimics this behavior for almost every sequence $X_1, X_2, \cdots$.

Theorem 3.2. If $m \wedge n \to \infty$, then for almost every sequence $X_1, X_2, \cdots$,

$$\sqrt{m}(\mathbb{F}_m^* - \mathbb{F}_n) \Longrightarrow \mathbb{U}^*(F)$$

where $\mathbb{U}^*$ is a Brownian bridge process on $[0,1]$.

Proof. The following proof is due to Shorack (1982). Let $\xi_1^*, \xi_2^*, \cdots$ be i.i.d. Uniform$(0,1)$, let $\mathbb{G}_m^*$ be the empirical d.f. of the first m ξ_i^*'s, and let $\mathbb{U}_m^* \equiv \sqrt{m}(\mathbb{G}_m - I)$ be the corresponding empirical process. By the Skorokhod construction we can construct the sequence $\{\mathbb{U}_m^*\}$ on a common probability space with a Brownian bridge process $\mathbb{U}^*$ so that $||\mathbb{U}_m^* - \mathbb{U}^*|| \to_{a.s.} 0$. [In fact by the Hungarian construction, this can be carried out with a sequence of Brownian bridge processes $\mathbb{B}_m^0$ so that $||\mathbb{U}_m^* - \mathbb{B}_m^0|| \le M(\log m)/\sqrt{m}$ almost surely; at the moment we only need the less precise result.]

Now we construct the bootstrap sample in terms of the uniform random variables ξ_i^*: by the inverse transformation the random variables

$$X_i^* \equiv \mathbb{F}_n^{-1}(\xi_i^*), \quad i = 1, \cdots, m$$

are, conditional on $\mathbb{F}_n$, i.i.d. with d.f. $\mathbb{F}_n$, and furthermore the empirical d.f. $\mathbb{F}_m^*$ thereof satisfies $\mathbb{F}_m^* = \mathbb{G}_m^*(\mathbb{F}_n)$. Hence we have

$$\sqrt{m}(\mathbb{F}_m^* - \mathbb{F}_n) = \sqrt{m}(\mathbb{G}_m^*(\mathbb{F}_n) - \mathbb{F}_n) = \mathbb{U}_m^*(\mathbb{F}_n).$$

But

$$\begin{aligned} ||\mathbb{U}_m^*(\mathbb{F}_n) - \mathbb{U}^*(F)|| &\le ||\mathbb{U}_m^*(\mathbb{F}_n) - \mathbb{U}^*(\mathbb{F}_n)|| \\ &\quad + ||\mathbb{U}^*(\mathbb{F}_n) - \mathbb{U}^*(F)|| \\ &\le ||\mathbb{U}_m^* - \mathbb{U}^*|| + ||\mathbb{U}^*(\mathbb{F}_n) - \mathbb{U}^*(F)|| \\ &\to_{a.s.} 0 + 0 = 0 \end{aligned}$$

since $\mathbb{U}^*$ is uniformly continuous and $||\mathbb{F}_n - F|| \to_{a.s.} 0$ by Glivenko - Cantelli. □

Example 3.3. (Bootstrap confidence bands for an arbitrary distribution function). Consider the distribution of the Kolmogorov statistic $D_n \equiv \sqrt{n} \sup_x |\mathbb{F}_n(x) - F(x)|$ as in example 1.E:

$$K_n(x, F) = P_F(D_n \le x).$$

If F is continuous this distribution does not depend on F and is tabled for small n; the asymptotic distribution is then also independent of F and is just the distribution of $||\mathbb{U}||_\infty$ where $\mathbb{U}$ is a Brownian bridge process on $[0,1]$. If F is discontinuous, however, then both K_n and the asymptotic distribution K_∞ depend on F. The bootstrap offers a way around this difficulty: the bootstrap estimator of $K_n(\cdot, F)$ is just

$$K_n(x, I\!F_n) \;=\; P_{I\!F_n}(\sqrt{n}||I\!F_n^* - I\!F_n|| \;\leq\; x\,)$$

and a Monte - Carlo approximation of it is

$$K^*_{n,B}(x) \;\equiv\; \frac{1}{B}\sum_{j=1}^{B} 1_{[\sqrt{n}||I\!F^*_{n,j} \,-\, I\!F_n|| \,\leq\, x]}$$

where $X^*_{j,1}, \cdots, X^*_{j,n}$ is a random sample from $I\!F_n$ for each $j = 1, \cdots, B$.

If we could find approximate upper α percentage points of the distribution of D_n; i.e. numbers $c_n(\alpha, F)$ so that

$$\lim_{n\to\infty} K_n(c_n(\alpha, F), F) \;=\; \lim_{n\to\infty} P_F(\sqrt{n}||I\!F_n - F|| \leq c_n(\alpha, F)) \;=\; 1-\alpha\,,$$

then we could construct an asymptotic $1-\alpha$ confidence band for F:

$$\lim_{n\to\infty} P_F\{I\!F_n(x) - n^{-1/2}c_n(\alpha, F) \leq F(x) \leq I\!F_n(x) + n^{-1/2}c_n(\alpha, F) \text{ for all } x \in R\,\} \;=\; 1-\alpha\,.$$

But our natural bootstrap estimator of $c_n(\alpha, F)$ is just $c_n(\alpha, I\!F_n) \;=\; K_n^{-1}(1-\alpha\,, I\!F_n)$; and a Monte-Carlo approximation of this is just $K^{-1}_{n,B}(1-\alpha) \;\equiv\; c^*_{n,B}(\alpha)$. Thus we obtain an asymptotically valid family of confidence bands for an arbitrary distribution function F:

Corollary 3.4. The bootstrap confidence bands $\{I\!F_n \;\pm\; n^{-1/2}c_n(\alpha\,, I\!F_n)\}$ satisfy

$$\lim_{n\to\infty} P_F\{I\!F_n(x) - n^{-1/2}c_n(\alpha, I\!F_n) \leq F(x) \leq I\!F_n(x) + n^{-1/2}c_n(\alpha, I\!F_n) \text{ for all } x \in R\,\} \;=\; 1-\alpha\,.$$

The behavior of these bands -- and the savings over the (conservative) asymptotic or finite - sample Kolmogorov bands -- has been investigated by Bickel and Krieger (1989).

Bootstrapping Empirical Measures

Does theorem 3.2 carry over to Efron's bootstrap for empirical measures? The answer is yes as shown by Giné and Zinn (1990). For a class of functions $\mathbf{F} \subset L_2(P)$, we let the envelope function F be defined by $F(x) \;\equiv\; \sup_{f\in\mathbf{F}} |f(x)|$. Here are the two bootstrap limit theorems of Giné and Zinn (1990):

Theorem 3.5. (Giné and Zinn, 1990). (Almost sure bootstrap limit theorem). Suppose that $\mathbf{F} \in NLDM(P)$. Then the following are equivalent:

A. $\mathbf{F} \in CLT(P)$ and $P(F^2) < \infty$; i.e. $\sqrt{n}(I\!P_n - P) \Longrightarrow I\!G_P$ where $I\!G_P$ is a ρ_P − uniformly continuous P − Brownian bridge process on $\mathbf{F}$.

B. $\sqrt{n}(I\!P_n^* - I\!P_n^\omega) \Longrightarrow I\!G_P$ in $l^\infty(\mathbf{F})$ almost surely.

Theorem 3.6. (Giné and Zinn, 1990). (In probablity bootstrap limit theorem). Suppose that $\mathbf{F} \in NLDM(P)$. Then the following are equivalent:

A. $\mathbf{F} \in CLT(P)$: i.e. $\sqrt{n}(I\!P_n - P) \Longrightarrow I\!G_P$ where $I\!G_P$ is a ρ_P − uniformly continuous P − Brownian bridge process on $\mathbf{F}$.

B. $\sqrt{n}(I\!P_n^* - I\!P_n^\omega) \Longrightarrow I\!G_P$ in $l^\infty(\mathbf{F})$ in probability.

Proofs: Multiplier inequalities together with lots of tricks and tools from empirical process theory. See Giné and Zinn (1990) or Klaassen and Wellner (1992). □

The spirit of the Giné and Zinn theorems carry over to the exchangeably - weighted bootstrap methods as shown by Praestgaard and Wellner (1993). Here are the hypotheses needed on the weights:

W1. The vectors $\underline{W} \equiv \underline{W}_n$ in R^n are exchangeable for each n.

W2. $W_{nj} \geq 0$ for all $j = 1, \cdots, n$ and $\sum_1^n W_{nj} = n$.

W3. $\sup_n ||W_{n1}||_{2,1} \equiv M(\underline{W}) < \infty$ where
$$||W_{n1}||_{2,1} \equiv \int_0^\infty \sqrt{P(W_{n,1} \geq t)}\, dt\,.$$

W4. $\lim_{\lambda \to \infty} \limsup_{n \to \infty} \sup_{t \geq \lambda} t^2 P(W_{n1} \geq t) = 0$.

W5. $n^{-1} \sum_{j=1}^n (W_{nj} - 1)^2 \to_p c^2 > 0$.

Theorem 3.7. (Praestgaard and Wellner, 1993). (Exchangeably weighted bootstrap limit theorems). Suppose that $\mathbf{F} \in NLDM(P)$ and that the random weight vectors $\{\underline{W}_n\}$ satisfy W1 - W5. Then:

A. $\mathbf{F} \in CLT(P)$ and $P(F^2) < \infty$ implies that

$$\sqrt{n}(I\!P_n^W - I\!P_n) \Longrightarrow I\!G_P \tag{3.1}$$

in $l^\infty(\mathbf{F})$ almost surely.

B. $\mathbf{F} \in CLT(P)$ implies that the convergence in (3.1) holds in probability.

Proof. See Praestgaard and Wellner (1993), or van der Vaart and Wellner (1994). □

The methods developed in Praestgaard and Wellner (1993) also lead to the following limit theorem for Efron's (multinomial) bootstrap with a bootstrap sample size $m = m_n$ possibly different than n .

Corollary 3.8. Suppose that $\mathbf{F} \in NLDM(P)$. Then:

A. $\mathbf{F} \in CLT(P)$ and $P(F^2) < \infty$ implies that

$$\sqrt{m}(I\!P^*_m - I\!P_n) \Longrightarrow I\!\!G_P \tag{3.2}$$

in $l^\infty(\mathbf{F})$ almost surely if $m \wedge n \to \infty$.

B. $\mathbf{F} \in CLT(P)$ implies that the convergence in (3.2) holds in probability.

Failures of Efron's Nonparametric Bootstrap

Just as the Jackknife fails for functions $T(F)$ which are not sufficiently smooth -- as we saw in section 2 -- the nonparametric bootstrap fails in a variety of situations involving "tail behavior". The following example is typical of these situations in which the empirical distribution is not a sufficiently accurate estimator of the population (true) distribution for the bootstrap to succeed.

Example 3.9. (Bootstrapping the estimator of θ for the Uniform$(0, \theta)$ distribution). Suppose that $X_1, \cdots, X_n$ are i.i.d. Uniform$(0, \theta)$. Then $\hat{\theta}_n = X_{(n)} \equiv \max_{1 \le i \le n} X_i$ and

$$n(\theta - \hat{\theta}_n) = n\theta(1 - \max_{1 \le i \le n} X_i/\theta) \to_d \theta Y$$

where $Y \sim \exp(1)$. Thus the limiting distribution is exponential$(1/\theta)$. Now let $X_1^*, \cdots, X_n^*$ be i.i.d. $I\!F_n$, and let $\hat{\theta}^*_n \equiv \max_{1 \le i \le n} X_i^* = X^*_{(n)}$. Then

$$\begin{aligned} P(\hat{\theta}^*_n = X_{(n)} = \hat{\theta}_n \,|\, I\!F_n) &= 1 - P(X^*_{(n)} < X_{(n)} \,|\, I\!F_n) \\ &= 1 - P(\text{all } X_i^* < X_{(n)} \,|\, I\!F_n) \\ &= 1 - (\frac{n-1}{n})^n \\ &= 1 - (1 - \frac{1}{n})^n \to 1 - e^{-1} \doteq .62\cdots , \end{aligned}$$

and, more generally,

$$\begin{aligned} P(n(X_{(n)} - \hat{\theta}^*_n) > n(X_{(n)} - X_{(n-k+1)}) \,|\, I\!F_n) &= P(X^*_{(n)} < X_{(n-k+1)} \,|\, I\!F_n) \\ &= P(\text{all } X_i^* < X_{(n-k+1)} \,|\, I\!F_n) \\ &= (1 - \frac{k}{n})^n \to e^{-k} . \end{aligned}$$

[In fact, this can be pushed further to show that the limiting distribution of the bootstrap is a *random* distribution.] Thus the bootstrap distribution differs dramatically from the actual distribution for large sample sizes, and this is also reflected in the finite sample distributions; see Efron and Tibshirani (1993), pages

81 and Figure 7.11 on page 83.

Example 3.10. (Bootstrapping a $V-$ statistic). Suppose that $X_1, \cdots, X_n$ are i.i.d. with common distribution function F, and let $h : R\,2 \rightarrow R$ be a symmetric function; i.e. $h(x,y) = h(y,x)$. Then

$$V_n \equiv \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h(X_i, X_j) = \int\int h(x,y)dIF_n(x)dIF_n(y)$$

is the $V-$ statistic based on the function h while
is the $U-$ statistic based on the function h. Note that U_n and V_n are closely related since

$$\begin{aligned} U_n &= \frac{2}{n(n-1)}\sum_{1 \le i < j = n} h(X_i, X_j) \\ &= \frac{1}{n(n-1)}\sum_{i \ne j} h(X_i, X_j) \\ &= \frac{1}{n(n-1)}\left\{ n\,2V_n - \sum_{i=1}^{n} h(X_i, X_i)\right\} \\ &= \frac{n}{n-1}V_n - \frac{1}{n-1}\int h(x,x)\,dIF_n(x)\,. \end{aligned}$$

Now suppose that $E_F X = 0$ and $h(x,y) = x\,y$. Then

$$V_n = \int\int xy\,dIF_n(x)dIF_n(y) = \left\{\int x\,dIF_n(x)\right\}^2 = \overline{X}_n^2\,,$$

and hence if $E_F X^2 = Var_F(X) < \infty$, then

$$n\,V_n = (\sqrt{n}\overline{X}_n)^2 \rightarrow_d (\sigma_F Z)^2 = \sigma_F^2 Z^2 \stackrel{d}{=} \sigma_F^2 \chi_1^2\,.$$

Does the nonparametric bootstrap mimic this? Unfortunately, the answer is "no": with

$$V_n^* \equiv \int\int x\,y\,dIF_n^*(x)dIF_n(y) = (\overline{X}_n^*)^2$$

we have

$$\begin{aligned} n\,V_n^* &= (\sqrt{n}\overline{X}_n^*)^2 \\ &= (\sqrt{n}(\overline{X}_n^* - \overline{X}_n) + \sqrt{n}\overline{X}_n)^2 \end{aligned}$$

where the first term converges in distribution a.s. conditionally on $X_1, X_2, \cdots$ to $\sigma_F Z^*$ where $Z^* \sim N(0,1)$, but the second term does not converge to zero, but instead converges instead to something random and nondegenerate (namely $\sigma_F^2\chi_1^2$), and marginally (over both the bootstrap and orginal data randomness) we see that $nV_n^* \rightarrow_d \sigma_F^2 \chi_2^2$. Hence the nonparametric bootstrap fails. This example is due to Bretagnolle (1983), and a solution to the failure is due to Arcones and Gine (1992).

General Bootstrap Without Replacement Limit Theory

One way around the difficulty in the preceding example is to take a bootstrap sample size $m = m_n$ much smaller than n and to try to estimate the distribution (or standard deviation or bias or ...) of $\hat{\theta}_{m_n}$ rather than $\hat{\theta}_n$. It turns out that a better way to do this is to draw a sample of size m_n *without replacement.* The following quite general result in this spirit is due to Politis and Romano (1993).

Suppose that $T_n \equiv T(IF_n)$ is an estimator of $\theta \equiv T(F)$ and that

(3.3) $$\tau_n(T_n - \theta) \rightarrow_d Z ,$$

(3.4) $$\tau_n \rightarrow \infty ,$$

(3.5) $$m_n = o(n) , \quad \text{and} \quad \tau_{m_n}/\tau_n \rightarrow 0 .$$

Let $\hat{T}_{m_n} \equiv T_{m_n}(\hat{X}_1, \cdots, \hat{X}_m)$ where $\hat{X}_1, \cdots, \hat{X}_{m_n}$ is a sample *without replacement* from $\{X_1, \cdots, X_n\}$.

Theorem 3.10. (Politis and Romano without replacement bootstrap limit theorem). If (3.3), (3.4), and (3.5) hold, then

$$\tau_{m_n}(\hat{T}_{m_n} - T_n) \rightarrow_d Z$$

in probability as $m_n \bigwedge n \rightarrow \infty$.

Proof. We will drop the subscript n on the sample size m_n in the proof. First note that

$$\tau_m(T_n - \theta) = \frac{\tau_m}{\tau_n} \tau_n(T_n - \theta) = o(1)O_p(1) = o_p(1)$$

by (3.5), so by writing

$$\tau_m(\hat{T}_m - T_n) = \tau_m(\hat{T}_m - \theta) - \tau_m(T_n - \theta) ,$$

it suffices, by Slutsky's theorem, to show that

$$P(\tau_m(\hat{T}_m - \theta) \leq z \,|\, IF_n) \rightarrow_p P(Z \leq z)$$

for $z \in C(\mathbf{L}(Z))$, the continuity set of the distribution function of Z. But

$$P(\tau_m(\hat{T}_m - \theta) \leq z \,|\, IF_n)$$
$$= \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \cdots < i_m \leq n} 1\{\tau_m(T_m(X_{i_1}, \cdots, X_{i_m}) - \theta) \leq z\} \equiv U_{n,m}$$

is an $m-$th order U - statistic with

$$E\,U_{n,m} \;=\; P(\tau_m(T_m(X_{i_1},\cdots,X_{i_m})-\theta)\;\le\;z)\;\to\;P(Z\le z)$$

for $z \in C(\mathbf{L}(Z))$ since $m = m_n \to \infty$. Hence it suffices to show that

$$U_{n,m} \;-\; EU_{n,m} \;\to_p\; 0\,.$$

This follows from Hoeffding's inequality for $U-$ statistics: since the kernel of the $U-$ statistic in question is bounded above by 1 and below by 0 ,

$$P(|U_{n,m} \;-\; EU_{n,m}| > t) \;\le\; 2\exp(\,-\,2\,[n/m]t^2/(1-0)^2) \;\to\; 0$$

since $n/m = n/m_n \to \infty$. □

Here is a proof of Hoeffding's inequality in several steps.

Proposition 3.11. (Hoeffding, 1963). If $X_1,\cdots,X_n$ are independent and $a_i \le X_i \le b_i$, $i=1,\cdots,n$, then with $\bar\mu_n \equiv E(\bar X_n)$,

(3.6) $$P(\bar X_n - \bar\mu_n \;\ge\; t\,) \;\le\; \exp(\,-\,2n^2t^2/\sum\nolimits_1^n (b_i - a_i)^2) \qquad \text{for all } t>0\,;$$

equivalently,

$$P(\sqrt{n}(\bar X_n - \bar\mu_n) \;\ge\; t\,) \;\le\; \exp(\,-\,2t^2/\{n^{-1}\sum\nolimits_1^n (b_i - a_i)^2\}\,) \qquad \text{for all } t>0\,.$$

Proof. By Markov's inequality and independence of the X_i 's it follows that, for $r>0$

$$\begin{aligned} P(\bar X_n - \bar\mu_n \;\ge\; t\,) &= P(\,\exp(r\,n\,(\bar X_n - \bar\mu_n) \;\ge\; \exp(r\,n\,t)\,) \\ &\le \frac{E\,e^{rn(\bar X_n - \bar\mu_n)}}{e^{r\,n\,t}} \\ &= \frac{\prod_{i=1}^n E\,e^{r(X_i-\mu_i)}}{e^{rnt}} \end{aligned}$$

where $\mu_i = E\,X_i$, $i=1,\cdots,n$. But since e^{rx} is convex on $[a,b]$, on $[a,b]$ it lies below the line passing through (a,e^{ra}) and (b,e^{rb}) :

$$e^{rx} \;\le\; \frac{b-x}{b-a}\,e^{ra} \;+\; \frac{x-b}{b-a}\,e^{rb}\,, \qquad a\le x\le b\,.$$

Hence

$$\begin{aligned} Ee^{r(X_i-\mu_i)} &\le e^{-r\mu_i}\left\{\frac{b_i-\mu_i}{b_i-a_i}\,e^{ra_i} \;+\; \frac{\mu_i-a_i}{b_i-a_i}\,e^{rb_i}\right\} \\ &= (1-p_i)e^{-r(\mu_i-a_i)} \;+\; p_i\,e^{r(b_i-\mu_i)} \\ &= \exp(L(r_i)) \end{aligned}$$

where $p_i \equiv (\mu_i - a_i)/(b_i-a_i)$, $r_i \equiv r(b_i-a_i)$, and

$$L(r_i) \;=\; -\,r_ip_i \;+\; \log(1-p_i+p_ie^{r_i})\,.$$

Now

$$L'(r_i) \;=\; -\,p_i \;+\; \frac{p_i}{(1-p_i)e^{-r_i} \,+\, p_i}$$

and

$$\begin{aligned} L''(r_i) \;&=\; \frac{p_i(1-p_i)e^{-r_i}}{[(1-p_i)e^{-r_i} \,+\, p_i]^2} \\ &=\; \frac{p_i}{[(1-p_i)e^{-r_i} \,+\, p_i]} \cdot \frac{(1-p_i)e^{-r_i}}{[(1-p_i)e^{-r_i} \,+\, p_i]} \\ &\equiv\; u_i(1-u_i) \;\le\; 1/4 \,. \end{aligned}$$

Thus by Taylor's theorem, with $0 \le s_i \le r_i$,

$$\begin{aligned} L(r_i) \;&=\; L(0) \;+\; L'(0)r_i \;+\; \frac{1}{2}\,L''(s_i)r_i^2 \\ &\le\; 0 \;+\; 0 \;+\; \frac{1}{2}\,\frac{1}{4}\,r_i^2 \;=\; \frac{1}{8}\,r^2(b_i-a_i)^2 \,. \end{aligned}$$

Hence

$$E\,e^{r(X_i-\mu_i)} \;\le\; \exp\!\big(\frac{1}{8}\,r^2(b_i-a_i)^2\big)$$

and

$$P(\overline{X}_n - \overline{\mu}_n \;\ge\; t) \;\le\; \exp\!\big(-\,nrt \;+\; \frac{1}{8}\,r^2 \sum_1^n (b_i-a_i)^2\big) \;\equiv\; e^{-g(r)}$$

for all $r > 0$. Since $g(r)$ is maximized (and the resulting bounds is minimized) if

$$r \;=\; r_0 \;\equiv\; 4nt/\sum_1^n (b_i-a_i)^2 \,,$$

with

$$g(r_0) \;=\; 2\,n^2\,t^2/\sum_1^n (b_i-a_i)^2 \,,$$

the inequality (3.6) follows. □

Corollary 3.12. (Hoeffding, 1963). If $X_1, \cdots, X_n$ are i.i.d. and

$$U_{n,m} \;=\; \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < \cdots < i_m \le n} h(X_{i_1}, \cdots, X_{i_m})$$

with $a \le h(x_1, \cdots, x_m) \le b$ for all $x_1, \cdots, x_m$ and h symmetric in its arguments, then

(3.7) $$P(U_{n,m} - EU_{n,m} \ge t) \;\le\; \exp(\,-\,2\,[n/m]\,t^2/(b-a)^2) \quad \text{for all } t > 0\,.$$

Proof. Suppose that

$$T \;=\; p_1 T_1 + \cdots + p_N T_N$$

where T_i an an average of independent random variables and $p_1 + \cdots + p_N = 1$. (The random variables $T_1, \cdots, T_N$ need *not* be independent.) For $r > 0$

$$\begin{aligned}
P(\,T - ET \ge t\,) \;&\le\; e^{-rt}\,E\,e^{r(T-ET)} \\
&=\; e^{-rt}\,E\,e^{r\sum_1^N p_i(T_i - \mu_i)} \\
&\le\; e^{-rt}\,E\sum_1^N p_i\,e^{r(T_i - \mu_i)} \qquad \text{since } e^x \text{ is convex} \\
\text{(a)} \qquad &=\; \sum_{i=1}^N p_i\,E\,e^{r(T_i - \mu_i - t)}\,.
\end{aligned}$$

If we can bound $E\exp(r(T_i - \mu_i - t))$ by something not depending on i (for example if $T_1, \cdots, T_N$ are identically distributed), then this bound will be a bound for $P(T \ge t)$ since $\sum_1^N p_i = 1$.

Now let $k \;\equiv\; [n/m]$ and define

$$\begin{aligned}
V(X_1, \cdots, X_n) \;&\equiv\; \frac{1}{k}\,\{h(X_1, \cdots, X_m) \;+\; h(X_{m+1}, \cdots, X_{2m}) \;+\; \cdots \\
&\qquad +\; h(X_{k(m-1)+1}, \cdots, X_{km})\} \\
&=\; \overline{Y}_k
\end{aligned}$$

where $Y_i \;\equiv\; h(X_{i(m-1)+1}, \cdots, X_{im})$, $i = 1, \cdots, k$ are independent. Note that

$$U_{n,m} \;=\; \frac{1}{n!}\sum_{\pi\,\in\,\Pi} V(X_{\pi(1)}, \cdots, X_{\pi(n)}) \;=\; \sum_{j=1}^N p_j T_j$$

where $p_j \;\equiv\; 1/n!$, $j = 1, \cdots, n! \;\equiv\; N$, and $T_j \;\equiv\; V(X_{\pi_j(1)}, \cdots, X_{\pi_j(n)})$ is an average of k independent random variables for $j = 1, \cdots, N$. Now

$$\begin{aligned}
P(T_j - ET_j \;\ge\; t) \;&=\; P(\overline{Y}_k - E\overline{Y}_k \;\ge\; t) \\
&\le\; \inf_{r\,>\,0} \frac{E\,e^{rk(\overline{Y}_k - E\overline{Y}_k)}}{e^{rkt}} \\
&\le\; \exp(\,-\,2\,k\,t^2/(b-a)^2)
\end{aligned}$$

by the proof of Hoeffding's inequality (3.6). Hence it follows from (a) that (3.7) holds. □

Corollary 3.13. If $\epsilon_1, \cdots, \epsilon_n$ are i.i.d. as $2\,Bernoulli(1/2) - 1$ (i.e. $P(\epsilon_i = \pm 1) = 1/2$), and $c_1, \cdots, c_n$ are constants, then

$$P(n^{-1/2}|\sum_{i=1}^n c_i\epsilon_i| \geq t) \leq 2\exp(-\frac{t^2}{2(n^{-1}\sum_{i=1}^n c_i^2)}) \quad \text{for all } t > 0 .$$

Some Limit Theory for Parametric Bootstrapping

It often holds that

$$(3.8) \qquad \sqrt{n}(\hat{\theta}_n - \theta) \Longrightarrow Y \qquad \text{as } n \to \infty ,$$

and, if $\theta \to P_\theta$ is differentiable in an appropriate sense,

$$(3.9) \qquad \sqrt{n}(P_{\hat{\theta}_n} - P_\theta) \Longrightarrow \dot{P}_\theta Y \qquad \text{as } n \to \infty$$

where $\dot{P}_\theta$ is a derivative map. The parametric bootstrap would proceed by forming $\hat{\theta}_n^* \equiv \hat{\theta}_n(X_{n1}^*, \cdots, X_{nn}^*)$. We then want to show that: for almost all sample sequences $X_1, X_2, \cdots$

$$(3.10) \qquad \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \Longrightarrow Y^* \sim Y$$

and

$$(3.11) \qquad \sqrt{n}(P_{\hat{\theta}_n^*} - P_{\hat{\theta}_n}) \Longrightarrow \dot{P}_\theta Y^* \sim \dot{P}_\theta Y \qquad \text{as } n \to \infty .$$

The following result is a useful first step toward proving (3.10) or (3.11), especially if $\hat{\theta}_n = \theta(I\!P_n)$ so that $\hat{\theta}_n^* = \theta(I\!P_n^*)$. This type of theorem for "model - based" or "parametric" bootstrap empirical processes was also suggested by Giné and Zinn (1991).

Theorem 3.14. (Convergence of the "parametric bootstrap" empirical process). Suppose that $\mathbf{F}$ is $\mathbf{P}$ – measurable with envelope function F and that:

(i) $\mathbf{F} \in CLT_u(\mathbf{P})$.

(ii) $||P_{\hat{\theta}_n} - P_\theta||_{\mathbf{G}}^* \equiv ||\hat{P}_n - P_\theta||_{\mathbf{G}}^* \to_{a.s.} 0$ where $\mathbf{G} = \mathbf{F} \cup \mathbf{F}^2 \cup (\mathbf{F}')^2$ and $\mathbf{F}' = \{f - g : f, g \in \mathbf{F}\}$.

(iii) F is $\mathbf{P}$ – uniformly square integrable.

Then, for P^∞ almost all sample sequences $X_1, X_2, \cdots$,

$$(3.12) \qquad I\!G_{n,n}^* \equiv \sqrt{n}(I\!P_n^* - P_{\hat{\theta}_n}) \Longrightarrow I\!G_0^* \sim I\!G_{P_\theta} \qquad \text{in } l^\infty(\mathbf{F})$$

as $n \to \infty$.

Proof. First note that (i) and (iii) imply that $\mathbf{F} \in AEC_u(\mathbf{P}, \rho)$ and $(\mathbf{F}, \rho_P)$ is totally bounded uniformly in $P \in \mathbf{P}$ by Sheehy and Wellner (1991), theorem 2.2. Hence, in particular, $\mathbf{F} \in AEC_u(\{\hat{P}_n\}, \rho)$ and $(\mathbf{F}, \rho_{P_\theta})$ is totally bounded. Furthermore, (iii) implies that for P^∞ a.e. ω the envelope F is $\{\hat{P}_n^\omega\}$ – uniformly square integrable. Thus, for P^∞ – a.e. ω, the hypotheses of Sheehy and Wellner (1991) theorem 3.1 are satisfied by $\mathbf{F}$ for the sequence $\{\hat{P}_n^\omega\} \equiv \{P_{\hat{\theta}_n(\omega)}\}$. Then the conclusion follows from theorem 3.1 with $P_0 \equiv P_\theta$. □

To give an example where this result is immediately useful, consider the non-parametric example mentioned briefly above:

Example 3.15. (Bootstrapping from a "smoothed empirical measure"; or, the "smoothed bootstrap"). Suppose that

$$\mathbf{P} = \{ P \text{ on } (R^d, \mathbf{B}^d) : \ p \equiv \frac{dP}{d\lambda} \text{ exists and is uniformly continuous} \} .$$

Suppose $\mathbf{C}$ is a measurable Vapnik - Chervonenkis class of subsets of R^d. Then $\mathbf{F} = \{1_C : C \in \mathbf{C}\} \in CLT_u \equiv CLT_u(\mathbf{M}) \subset CLT_u(\mathbf{P})$, so (i) holds. Suppose that $\hat{I\!P}_n$ is defined for $C \in I\!A$ by

$$\hat{I\!P}_n(C) = \int 1_C(x) \hat{p}_n(x) dx$$

where

$$\hat{p}_n(x) = \frac{1}{b_n^d} \int k(\frac{y-x}{b_n})\, dI\!P_n(y)$$

where $k : R^d \to R$ is a uniformly continuous density function. It follows that $\hat{I\!P}_n \in \mathbf{P}$ and, if $b_n \to 0$ and $nb_n^d \to \infty$, then

$$\int |\hat{p}_n(x) - p(x)|\, dx \to_{a.s.} 0 ;$$

see Devroye (1983), theorem 1. When $\mathbf{F}$ is all indicators of a subclass of Borel sets $\mathbf{C}$ the supremum in (ii) is bounded by the total variation distance between $\hat{I\!P}_n$ and P, which, in turn, is well-known to equal half the L_1 – distance between the respective densities (see e.g. the statement of Scheffé's theorem in Billingsley (1968), page 224)). Hence

$$||\hat{I\!P}_n - P||_{\mathbf{G}}^* \le ||\hat{I\!P}_n - P||_{\mathbf{B}^d} = \frac{1}{2} \int |\hat{p}_n(x) - p(x)|\, dx \to_{a.s.} 0 ,$$

so (ii) holds. Since (iii) holds trivially (with $F \equiv 1$), theorem 3.14 shows that "the bootstrap from $\hat{I\!P}_n$ works:" i.e. for P^∞ almost all sample sequences $X_1, X_2, \cdots$, $I\!G_{n,n}^* \Longrightarrow I\!G_0^* \sim I\!G_P$ in $l^\infty(\mathbf{F})$.

For more general classes $\mathbf{F}$, the results of Yukich (1989) could be used to verify hypothesis (ii) of theorem 4.5.

Silverman and Young (1987) have studied several smoothed bootstrap methods, and give criteria for determining when $\alpha(P_{\hat{\theta}_n})$ will give a better estimator of $\alpha(P)$ than $\alpha(I\!P_n)$ for functionals $\alpha : \mathbf{P} \rightarrow R$; see also Hall, DiCiccio, and Romano (1989) for further work in this direction.

5. The Bootstrap and the Delta Method

Now we combine the results established for the bootstrap empirical process with differentiability hyptheses on functionals $T(F)$ or $T(P)$ to establish asymptotic validity of the bootstrap for nonlinear functionals $T(F)$ or $T(P)$. The following theorem is due to Gill (1989).

Theorem 4.1. Suppose that $T : \mathbf{F} \to R$ is Hadamard - differentiable at F wrt $||\cdot||_\infty$ tangentially to the subspace $C_u(R\,, \rho_F)$ of uniformly continuous functions (with respect to the pseudo-metric ρ_F defined by $\rho_F^2(s,t) = Var_F(1_{(-\infty,s]}(X) - 1_{(-\infty,t]}(X))$). Let ψ_F denote the influence function of T at $F \in \mathbf{F}$. Then

$$\sqrt{n}(T(I\!F_n) - T(F)) \to_d N(0\,, E\psi_F^2(X))$$

and furthermore

$$\sqrt{n}(T(I\!F_n^*) - T(I\!F_n)) \to_d N(0\,, E\psi_F^2(X)) \qquad \text{in probability} .$$

Proof. The first part has been proved already in theorem 7.4.11. The second part proceeds by a "double - differencing" argument as follows. Suppose that we have constructed versions of both the original empirical process and the bootstrap empirical process in terms of uniform empirical processes $\{I\!U_n\}$ and $\{I\!U_n^*\}$ satisfying

$$||I\!U_n - I\!U||_\infty \to_{a.s.} 0 \qquad \text{and} \qquad ||I\!U_n^* - I\!U^*||_\infty \to_{a.s.} 0 .$$

Thus with $\tilde{I\!F}_n \equiv I\!G_n(F)$

$$\sqrt{n}(I\!F_n - F) \stackrel{d}{=} \sqrt{n}(\tilde{I\!F}_n - F) = I\!U_n(F) \to_{a.s.} I\!U(F) \quad \text{in } (D(\bar{R}), ||\cdot||_\infty)$$

and, with $\tilde{I\!F}_n^* \equiv I\!G_n^*(\tilde{I\!F}_n)$,

$$\sqrt{n}(I\!F_n^* - I\!F_n) \stackrel{d}{=} \sqrt{n}(\tilde{I\!F}_n^* - \tilde{I\!F}_n) = I\!U_n^*(\tilde{I\!F}_n) \to_{a.s.} I\!U^*(F) \quad \text{in } (D(\bar{R}), ||\cdot||_\infty) .$$

Now write

$$\begin{aligned} I\!F_n^* &= F + n^{-1/2}n^{1/2}(I\!F_n^* - I\!F_n) + n^{-1/2}n^{1/2}(I\!F_n - F) \\ &\stackrel{d}{=} F + n^{-1/2}\{I\!U_n^*(\tilde{I\!F}_n) + I\!U_n(F)\} \end{aligned}$$

and

$$I\!F_n = F + n^{-1/2}n^{1/2}(I\!F_n - F) \stackrel{d}{=} F + n^{-1/2} I\!U_n(F) .$$

Thus we may write

$$\sqrt{n}(T(I\!F_n^*) - T(I\!F_n)) \stackrel{d}{=} \sqrt{n}(T(F + n^{-1/2}\{I\!U_n^*(\tilde{I\!F}_n) + I\!U_n(F)\}) - T(F))$$

$$- \sqrt{n}(T(F + n^{-1/2} \mathbb{U}_n(F)) - T(F))$$
$$\rightarrow_{a.s.} \dot{T}(F ; \mathbb{U}^*(F) + \mathbb{U}(F)) - \dot{T}(F ; \mathbb{U}(F))$$
$$= \dot{T}(F ; \mathbb{U}^*(F)) \sim N(0 ; E_F \psi_F^2(X))$$

using linearity of $\dot{T}(F ; \cdot)$ in the last step. Thus for the constructed empirical distributions, for a.e. $\tilde{X}_1, \tilde{X}_2, \cdots$

$$\sqrt{n}(T(\tilde{IF}_n^*) - T(\tilde{IF}_n)) \rightarrow_{a.s.} \dot{T}(F ; \mathbb{U}^*(F)) \sim N(0, E\psi_F^2(X)) .$$

Since $\rightarrow_{a.s.}$ implies $\rightarrow_d$, this implies that for a.e. sequence $\tilde{X}_1, \tilde{X}_2, \cdots$

$$\sqrt{n}(T(\tilde{IF}_n^*) - T(\tilde{IF}_n)) \rightarrow_d N(0, E\psi_F^2(X)) ,$$

and this implies that with

$$H_n(x ; \tilde{IF}) \equiv P_{\tilde{IF}_n}(\sqrt{n}(T(\tilde{IF}_n^*) - T(\tilde{IF}_n)) \le x) ,$$

for a.e. sequence $\tilde{X}_1, \tilde{X}_2, \cdots$ we have

$$d_{BL^*}(H_n(\cdot , \tilde{IF}_n), N(0, E\psi_F^2(X)) \rightarrow 0 .$$

But this just means that

$$d_{BL^*}(H_n(\cdot , \tilde{IF}_n), N(0, E\psi_F^2(X)) \rightarrow_{a.s.} 0$$

for the constructed sequence $\tilde{IF}_n$. But $H_n(\cdot ; \tilde{IF}_n) \stackrel{d}{=} H_n(\cdot ; IF_n)$ and $\rightarrow_{a.s.}$ implies $\rightarrow_p$, so we conclude that

$$d_{BL^*}(H_n(\cdot , IF_n), N(0, E\psi_F^2(X)) \rightarrow_p 0 .$$

□

For further results of this type, see Gill (1989), Arcones and Gine (1992), and van der Vaart and Wellner (1994).

Example 4.2. Suppose F is a df which is differentiable at its median $T(F) \equiv F^{-1}(1/2) \equiv m(F)$, and that $f(m(F)) \equiv F'(m(F)) > 0$. If $X_1, \cdots, X_n$ are iid rv's with df F, let $M_n \equiv IF_n^{-1}(1/2)$ be the sample median, and let

(1) $$H_n(x, F) \equiv Pr_F(\sqrt{n} (M_n - m(F)) \le x) .$$

Of course it is well - known that

$$H_n(x, F) \rightarrow Pr(N(0, \frac{1/4}{f^2(m(F))}) \le x) \quad \text{as } n \rightarrow \infty$$

for every $x \in R$.

The natural "bootstrap estimate" of the df $H_n(x, F)$ is simply $H_n(x, IF_n)$ where IF_n is the empirical df of the X_i 's. It follows from theorem 4.1 and Hadamard differentiability of the median functional $T(F) = F^{-1}(1/2)$ as proved in Gill (1989) (or see van der Vaart and Wellner (1994), section 3.x), that

$$(2) \qquad H_n(x\,, I\!F_n) \;\to\; Pr(\, N(0\,, \frac{1/4}{f^2(m(F))}\,) \;\le\; x\,) \qquad \text{for all } x \in R \text{ in probability}$$

as $n \to \infty$. This type of result was first established by Bickel and Freedman (1981). They showed that under the hypothesis of *continuous* differentiability at $m(F)$ the bootstrap works almost surely:

Theorem 4.3. (Bickel and Freedman, 1981). Suppose that F is continuously differentiable in a neighborhood of $m(F)$ with $f(m(F)) > 0$. Then for almost every sample sequence $X_1\,, X_2\,, \cdots$

$$(2) \qquad H_n(x\,, I\!F_n) \;\to\; Pr(\, N(0\,, \frac{1/4}{f^2(m(F))}\,) \;\le\; x\,) \qquad \text{as } \; n \;\to\; \infty$$

for all $x \in R$.

In view of Polya's lemma, (2) can be re-expressed as

$$(3) \qquad \sup_x |\, H_n(x\,, I\!F_n) \;-\; \Phi(\, x\, 2\, f(m(F))\,)\,| \;\to_{a.s.}\; 0$$

as $n \to \infty$ where Φ denotes the standard $N(0,1)$ df.

Proof. Represent the iid X_i 's as $F^{-1}(\xi_i)$ where $\xi_1\,, \xi_2\,, \cdots$ are iid $U(0,1)$ rv's. Let $I\!F_n \;=\; I\!G_n(F)$ denote the empirical df of the X_i 's, so that the empirical process of the X_i 's is

$$(a) \qquad \sqrt{n}(I\!F_n \;-\; F) \;=\; I\!U_n(F)\,.$$

Then represent the bootstrap sample $X^*_{n1}\,, \cdots, X^*_{nn}$ as $X^*_{ni} \;\equiv\; I\!F_n^{-1}(\xi^*_i)$ where $\xi^*_1\,, \xi^*_2\,, \cdots$ is another sequence of independent $U(0,1)$ rv's independent of the ξ_i 's. Thus the empirical df of the X^*_{ni} 's is $I\!F^*_n \;=\; I\!G^*_n(I\!F_n)$, and the bootstrap empirical process is

$$(b) \qquad \sqrt{n}(I\!F^*_n \;-\; I\!F_n) \;=\; I\!U^*_n(I\!F_n)\,.$$

We give the proof of (2) for $x \ge 0$; the argument for $x < 0$ is similar. Now

$$
\begin{aligned}
(c) \qquad H_n(x\,, I\!F_n) \;&=\; Pr_{I\!F_n}\{\, \sqrt{n}(M^*_n \;-\; m(I\!F_n)) \;\le\; x\,\} \\
&=\; Pr_{I\!F_n}\{\, I\!F_n^{*\,-1}(1/2) \;\le\; m(I\!F_n) \;+\; n^{-1/2}\,x\,\} \\
&=\; Pr_{I\!F_n}\{\, I\!F^*_n(m(I\!F_n) + n^{-1/2}\,x) \;\ge\; 1/2\,\} \\
&=\; Pr_{I\!F_n}\{\, I\!U^*_n(I\!F_n(m(I\!F_n) + n^{-1/2}\,x\,) \;\ge\; -\,D_n\,\}
\end{aligned}
$$

where

$$D_n \equiv \sqrt{n}(I\!F_n(m(I\!F_n) + n^{-1/2}x) - \frac{1}{2})$$

$$= \sqrt{n}(I\!F_n(m(I\!F_n) + n^{-1/2}x) - F(m(I\!F_n) + n^{-1/2}x))$$

$$+ \sqrt{n}(F(m(I\!F_n) + n^{-1/2}x) - F(m(I\!F_n)))$$

$$+ \sqrt{n}(F(m(I\!F_n)) - I\!F_n(m(I\!F_n))) + o(n^{-1/2})$$

(d)
$$= I\!U_n(F(m(I\!F_n) + n^{-1/2}x) - I\!U_n(F(m(I\!F_n)))$$

$$+ \sqrt{n}(F(m(I\!F_n) + n^{-1/2}x) - F(m(I\!F_n))]$$

(e)
$$\equiv I\!U_n(b_n) - I\!U_n(a_n)$$

$$+ \sqrt{n}(b_n - a_n).$$

Now since F is continuously differentiable in a neighborhood of $m(F)$,

(f)
$$\sqrt{n}(b_n - a_n)$$

$$= \sqrt{n}(F(m(I\!F_n) + n^{-1/2}x) - F(m(I\!F_n))]$$

$$\rightarrow_{a.s.} x\, f(m(F)) \equiv c > 0 \quad \text{as} \quad n \rightarrow \infty.$$

Hence, for n sufficiently large, the first term on the right side in (e) is bounded by

$$\omega_n((c+1)n^{-1/2}) \rightarrow_{a.s.} 0 \quad \text{as} \quad n \rightarrow \infty,$$

by well known properties of the oscillation modulus of $I\!U_n$; see e.g. Shorack and Wellner (1986), page 542. Hence

(g)
$$D_n \rightarrow_{a.s.} x\, f(m(F)).$$

But $I\!U_n^*$ converges weakly to a Brownian bridge process $I\!U^*$, and without loss of generality we can assume that the ξ_i^* 's have been constructed on a common probability space with the $I\!U^*$ process so that

(h)
$$||I\!U_n^* - I\!U^*|| \rightarrow 0.$$

Using (g) and (h) with the last line of (c) yields (or, instead of (h), use a similar argument as above involving the oscillation of $I\!U_n^*$!)

$$H_n(x, I\!F_n) \rightarrow_{a.s.} Pr(I\!U^*(1/2) > -x\, f(m(F))) \quad \text{as} \quad n \rightarrow \infty$$

$$= Pr(N(0, \frac{1/4}{f^2(m(F))} \leq x).$$

□

I conjecture that the asymptotic validity of the bootstrap in the almost sure **8.6** sense fails to hold if F is not continuously differentiable at $m(F)$.

6. Bootstrap Tests and Confidence Intervals

8.7

7. M - Estimates and the Bootstrap

One of the most common type of estimators used in statistics are those derived from *estimating equations* or *M - estimates.* The system of estimating equations to be solved for estimators $\hat{\theta}_n$ of some (vector of) parameters θ is often derived from likelihood considerations by considering a model

$$\mathbf{P} \;=\; \{P_\theta : \theta \in \Theta\}$$

in which the distributions P_θ have densities $p_\theta \;\equiv\; p(\,\cdot\,;\theta)$ with respect to some dominating measure μ . Then we define estimators $\hat{\theta}_n \;\equiv\; T(I\!P_n)$ by solving the likelihood equations based on the scores $\dot{\mathbf{l}}_\theta(\,\cdot\,;\theta)$: $\hat{\theta}_n$ satisfies

$$(1) \qquad 0 \;=\; I\!P_n \dot{\mathbf{l}}_\theta(\,\cdot\,;\hat{\theta}_n) \;=\; \int \dot{\mathbf{l}}_\theta(x\,;\hat{\theta}_n)\,dI\!P_n(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} \dot{\mathbf{l}}_\theta(X_i\,;\hat{\theta}_n)\,.$$

However, we frequently want to study the properties of the estimators $\hat{\theta}_n \;=\; T(I\!P_n)$ even when the true distribution P generating the data is not in the collection $\mathbf{P}$. In this case it is traditional to relabel the score functions $\dot{\mathbf{l}} \;\equiv\; \nabla \log p_\theta(\,\cdot\,)$ as $\psi(\,\cdot\,,\theta)$; then the estimator $\hat{\theta}_n$ satisfies

$$(2) \qquad 0 \;=\; I\!P_n \psi(\,\cdot\,,\hat{\theta}_n) \;=\; \int \psi(x\,,T(I\!P_n))\,dI\!P_n(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} \psi(X_i\,,T(I\!P_n))\,.$$

Our goal in this section is to treat the large sample theory of such estimators and to investigate whether the bootstrap "works". The first part of this is very much as in section 4.7, but now we focus on estimators defined in terms of zeros of the estimating equations, and we also seek to drop the convexity hypotheses used there.

One - Dimensional M - Estimates with Monotone ψ

We begin with a simple case: $\theta \in \Theta \;\subset\; R$ and $\psi(\,x\,,t)$ monotone decreasing in t . The following development is from Huber (1981).

Suppose that $\psi(x\,,t) \downarrow$ in t for each $x \in \mathbf{X}$, and $\psi(x\,,-\infty) \;>\; 0$, $\psi(x\,,\infty) \;<\; 0$, and $\hat{\theta}_n \;\equiv\; T_n \;\equiv\; T(I\!P_n)$ is defined by (2). Set

$$T_n^* \;\equiv\; \sup\{t : \sum_{i=1}^{n} \psi(X_i, t) \;>\; 0\}\,,$$

$$T_n^{**} \;\equiv\; \inf\{t : \sum_{i=1}^{n} \psi(X_i, t) \;<\; 0\}\,.$$

Now

$$(3) \qquad [T_n^* < t] \ \subset \ [\sum_{i=1}^n \psi(X_i, t) \ \le \ 0] \ \subset \ [T_n^* \le t]$$

and

$$(4) \qquad [T_n^{**} < t] \ \subset \ [\sum_{i=1}^n \psi(X_i, t) \ < \ 0] \ \subset \ [T_n^* \le t] \, .$$

Hence

$$(5) \qquad P(T_n^* \le t) \ = \ P(\sum_{i=1}^n \psi(X_i, t) \ \le \ 0)$$

and

$$(6) \qquad P(T_n^{**} \le t) \ = \ P(\sum_{i=1}^n \psi(X_i, t) \ < \ 0)$$

at continuity points of the left side.

Theorem 8.6.1. (Consistency). Suppose that there is a number $\theta = \theta(P)$ so that $\lambda(t) \equiv P\,\psi(\,\cdot\,, t) = \int \psi(x\,, t)\, dP(x)$ satisfies $\lambda(t) > 0$ for $t < \theta$, $\lambda(t) < 0$ for $t > \theta$. Then both T_n^* and T_n^{**} converge in probability to θ.

Proof. For $\epsilon > 0$ so that $\theta + \epsilon$ is a continuity point of $F_{T_n^*}$,

$$P(T_n^* \ \le \ \theta + \epsilon) \ = \ P(\frac{1}{n} \sum_{i=1}^n \psi(X_i\,, \theta + \epsilon) \ \le \ 0) \ \to \ 1$$

since

$$\frac{1}{n} \sum_{i=1}^n \psi(X_i, \theta + \epsilon) \ \to_{a.s.} \ \lambda(\theta + \epsilon) \ < \ 0 \, ;$$

and for $\epsilon > 0$ so that $\theta - \epsilon$ is a continuity point of $F_{T_n^*}$

$$P(T_n^* \ \le \ \theta - \epsilon) \ = \ P(\frac{1}{n} \sum_{i=1}^n \psi(X_i\,, \theta - \epsilon) \ \le \ 0) \ \to \ 0$$

since

$$\frac{1}{n} \sum_{i=1}^n \psi(X_i, \theta - \epsilon) \ \to_{a.s.} \ \lambda(\theta - \epsilon) \ > \ 0 \, .$$

Hence $T_n^* \to_p \theta$. The proof for T_n^{**} is similar. □

Remark. Note that $\lambda(t) \equiv \lambda(t, P)$ is continuous at t_0 if $\psi(x, \cdot)$ is continuous for P - a.e. x and $\{\psi(X, t)\}_{t \in N(t_0;\delta)}$ is uniformly integrable for some $\delta > 0$.

Now we are ready to establish a central limit theorem for $T_n \equiv T(IP_n)$. First, note that since $\psi(x, t)$ is monotone decreasing in t, it follows that the function λ is also monotone decreasing. Hence we have

$$[-\lambda(T_n) < -\lambda(t)] \subset [T_n < t] \subset [T_n \le t] \subset [-\lambda(T_n) \le -\lambda(t)] .$$

Here are the rest of the assumptions we will need:

A1. $\psi(x, t)$ is measurable in x, $\downarrow$ in t.

A2. $\Gamma_0 \equiv \{t : \lambda(t, P) = 0\} \neq \emptyset$; let $\theta \in \Gamma_0$.

A3. λ is continuous in a neighborhood of Γ_0.

A4. $\sigma^2(t) \equiv E_P \psi^2(X, t) - \lambda(t, P)^2$ is finite, $\neq 0$, and continuous in a neighborhood of Γ_0. Set $\sigma_0^2 \equiv \sigma^2(\theta)$.

A5. The derivative $\lambda'(\theta)$ exists and $\lambda'(\theta) < 0$.

Theorem 8.6.2. If A1 - A4 hold, then

$$-\sqrt{n}\,\lambda(T_n) \to_d N(0, \sigma_0^2) . \tag{7}$$

Furthermore, if A5 holds, then

$$\sqrt{n}(T_n - \theta) \to_d N(0, \frac{\sigma_0^2}{\lambda'(\theta)^2}) . \tag{8}$$

Proof. Let $T_n = T_n^*$ and fix $y \in R$. Define t_n by $y = \sqrt{n}\lambda(t_n)$ using A3. Then, if y is a continuity point of the distribution of $\lambda(T_n)$,

$$\begin{aligned} P(-\sqrt{n}\,\lambda(T_n) < y) &= P(T_n < t_n) = P(\sqrt{n}\, IP_n \psi(\cdot, t_n) \le 0) \\ &= P(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\psi(X_i, t_n) - \lambda(t_n)}{\sigma(t_n)} \le -\frac{\sqrt{n}\lambda(t_n)}{\sigma(t_n)} = \frac{y}{\sigma(t_n)}) \\ &\equiv P(\frac{1}{\sqrt{n}} \sum_{i=1}^n Y_{ni} \le \frac{y}{\sigma(t_n)}) \end{aligned} \tag{9}$$

where the $Y_{ni} \equiv \{\psi(X_i, t_n) - \lambda(t_n)\}/\sigma(t_n)$ are i.i.d with $E\,Y_{ni} = 0$ and $Var(Y_{ni}) = 1$.

Now

$$\sigma(t_n) \to \begin{cases} \sigma(\theta_*) & \text{if } y < 0 \\ \sigma(\theta_{**}) & \text{if } y > 0 \end{cases} \tag{10}$$

by considering the natural picture. Furthermore, note that if $\lambda(s, P) = \lambda(t, P) = 0$ for $s \le t$ then we have

$$\lambda(s,P) - \lambda(t,P) = \int \{\psi(x,s) - \psi(x,t)\}\, dP(x) = 0\,,$$

where the integrand is ≥ 0. Hence $\psi(x,s) = \psi(x,t)$ a.e. P, and it follows that $\sigma^2(t) = \sigma^2(\theta_*) = \sigma^2(\theta_{**})$ for all $t \in \Gamma_0$. It follows that to prove (i) we need only to verify the Lindeberg condition. Since the Y_{ni} 's are identically distributed, this becomes

$$EY_{n1}^2 1_{[|Y_{n1}| > \sqrt{n}\epsilon]} \to 0 \quad \text{as} \quad n \to \infty\,. \tag{11}$$

But since λ and σ are continuous by A3 and A4, this is equivalent to

$$E\,\psi^2(X\,,t_n)1_{[|\psi(X,t_n)| > \sqrt{n}\epsilon]} \to 0\,. \tag{12}$$

Since $\psi(x,t)$ is monotone in t,

$$\psi^2(X,s) \leq \psi^2(X,s_0) + \psi^2(X,s_1) \equiv \text{a fixed integrable function} \tag{13}$$

for all $s_0 \leq s \leq s_1$, and hence (9) holds.

To prove (ii), note that when $\lambda'(\theta) < 0$, then $\Gamma_0 = \{\theta\}$ and $T_n \to_p \theta$, so by (i), existence of $\lambda'(\theta) < 0$, and Slutsky's theorem,

$$\begin{aligned}\sqrt{n}(T_n - \theta) &= \frac{T_n - \theta}{\lambda(T_n) - \lambda(\theta)}\,\sqrt{n}\lambda(T_n)\\ &\to_d \frac{1}{\lambda'(\theta)}\,N(0\,,\sigma_0^2) = N(0\,,\sigma_0^2/\lambda'(\theta)^2)\,.\end{aligned}$$

□

Remark 1. Note that the asymptotic variance of the M - estimator T_n established above agrees with the heuristic influence curve calculation whenever

$$\lambda'(t,P) = \frac{\partial}{\partial t}\int \psi(x,t)\,dP(x) = \int \frac{\partial}{\partial t}\psi(x,t)\,dP(x) \qquad \text{at} \quad t = \theta(P)\,.$$

Remark 2. The case of $\psi(x,t) = \psi(x-t)$ with $\psi : R \to R$ a fixed monotone $\uparrow$ function is of particular interest. Then $T(P)$ is a *location functional.* If ψ is bounded, continuous, and strictly increasing with $\psi(+\infty) > 0$ and $\psi(-\infty) < 0$, with a bounded and continuous derivative ψ', then A1, A3, A4, and A5 hold automatically.

Generalized M - Estimates and the Master Theorem

To treat a wider variety of estimators, it is useful to consider estimators defined by more general equations than the M - estimates defined above. Note that for M - estimates, the function $\mathbb{W}_n$ defined by

$$\mathbb{W}_n(t) \equiv \mathbb{P}_n\psi(\,\cdot\,,t) = \int \psi(x\,,t)\,d\mathbb{P}_n(x) \tag{14}$$

is linear in $\mathbb{P}_n$, and its expectation $\lambda(t) \equiv W(t,P)$ is linear in P:

$$\lambda(t) \equiv W(t,P) = P\psi(\cdot\,,t) = \int \psi(x\,,t)\,dP(x)\,.$$

In many situations of interest this linearity of $I\!W_n$ in $I\!P_n$ (and of W in P) fails. Here are two examples.

Example 8.6.3. Suppose that $X_1\,, X_2 \sim P$ on R are independent. Consider the functional $T(P)$ defined by $W(T(P),P) = 0$ where

$$\begin{aligned} W(t,P) &\equiv P((X_1 + X_2)/2 > t) - 1/2 \\ &= \int F(2t - x)\,dF(x) - 1/2\,; \end{aligned}$$

here $X_1\,, X_2$ are independent with distribution P on R with corresponding distribution function F. The corresponding estimator $T(I\!P_n)$ satisfies (approximately) $I\!W_n(T(I\!P_n), I\!P_n) = 0$ where

$$I\!W_n(t) = W(t, I\!P_n) = \int I\!F_n(2t - x)\,dI\!F_n(x) - 1/2$$

and $I\!F_n(x) \equiv I\!P_n 1_{(-\infty\,,x]}(\cdot)$. $T(I\!P_n)$ is called the *Hodges - Lehmann estimator* of location.

Example 8.6.4. (The Cox partial likelihood estimator). Suppose that $X = (Z_{m\times 1}, Y, \Delta)$ where $Y = T \bigwedge C$, $\Delta = 1_{[T \geq C]}$, and, given Z, T and C are independent. Assume also that the conditional distribution of T given Z is $1 - \bar{G}^r$ where $r = r(\nu^T z) = \exp(\nu^T z)$, $\bar{G} = 1 - G$. Then $X \sim P$ belongs to the generalization $\mathbf{P}$ of the Cox model, example 3.4.2, in which right censoring of T by C is allowed and the distribution of the censoring variable C can depend on the covariate. For simplicity in the sequel we take $m = 1$. The argument for $m > 1$ is only notationally more complicated.

Our goal is obtain asymptotic properties of the Cox partial likelihood estimator *even when the model fails*. Suppose that P is *any* distribution of $(Z, Y\,, \Delta)$ on $R \times R \times \{0,1\}$ satisfying

(i) Z is not degenerate a.s. P ; i.e. $P(Z = z_0) \neq 1$ for all z_0 .

(ii) Z is bounded a.s. P ; i.e. $P(|Z| \leq C) = 1$.

Let $\mathbf{Q}$ denote the collection of all such distributions P. We suppose that $X_1\,, \cdots, X_n$ are iid $P \in \mathbf{Q}$; note that P is not necessarily in the Cox model $\mathbf{P} \subset \mathbf{Q}$.

We define the Cox estimate of θ using the notation of Tsiatis (1981). For $Q \in \mathbf{Q}$, let

$$S_j(t,\theta,Q) = \int z^j\, r(\theta z)\, 1_{[s \geq t]}\, dQ^{(12)}(z,s)\,, \qquad \text{for} \quad j = 0, 1, 2\,,$$

where $Q^{(12)}$ is the marginal distribution of (Z, Y). Let

(15) $$D(\theta, Q) \;\equiv\; -\,\theta \int z\, dQ^{(1)}(z) \;+\; \int \log S_0(t, \theta, Q)\, dQ^{(2)}(t)$$

for $\theta \in \Theta \equiv R$ where $Q^{(1)}(z) \equiv Q(Z \le z\,, \Delta = 1)$ and $Q^{(2)}(t) \equiv Q(Y \le t\,, \Delta = 1)$.

In fact this is a case of a *convex* function D. Note that

$$D(\theta, I\!P_n) \;=\; -\,\frac{1}{n}\,\theta \sum_{i=1}^{n} \Delta_i\, Z_i \;+\; \frac{1}{n} \sum_{i=1}^{n} \Delta_i \log \left\{ \frac{1}{n} \sum_{j \,\in\, R_i} e^{\theta Z_j} \right\}$$

$$=\; -\,\frac{1}{n} \sum_{i=1}^{n} \Delta_i \log \left\{ \frac{e^{\theta Z_i}}{\frac{1}{n} \sum_{j \,\in\, \mathbf{R}_i} e^{\theta Z_j}} \right\},$$

where the "risk set" $\mathbf{R}_i \equiv \{ j : T_j \ge T_i \}$. Hence the resulting generalized M - estimator $\theta(I\!P_n)$ is the Cox partial likelihood estimator of θ. It satisfies

$$I\!\!W_n(\theta(I\!P_n)) \;=\; W(\theta, I\!P_n) \;=\; 0$$

for

$$W(\theta, Q) \;=\; -\int z\, dQ^{(1)}(z) \;+\; \int \frac{S_1}{S_0}\,(t, \theta, Q)\, dQ^{(2)}(t)\,.$$

To handle examples such as these, we will need to allow for rather general functions $W(t, P)$ of P and hence for estimators $\hat\theta_n \equiv \theta(I\!P_n) \equiv T(I\!P_n)$ defined by $I\!\!W_n(\theta(I\!P_n)) = W(\theta(I\!P_n)\,, I\!P_n) = 0$.

Suppose that $\mathbf{P} \subset \mathbf{Q}$ where $\mathbf{Q}$ is a collection of distributions containing all distributions with finite support (i.e. all the empirical measures). Suppose that $W : R^m \times \mathbf{M}_0 \to R^m$ and that

(16) $$W(\theta(P)\,, P) \;=\; 0 \quad \text{for all} \quad P \in \mathbf{Q}$$

Let $I\!\!W_n(\theta) \equiv W(\theta, I\!P_n)$. $I\!\!W_n(\hat\theta_n) = 0$, then $\hat\theta_n$ is a *generalized M - estimate* (or *GM - estimate* for short) of $\theta(P)$. If

(17) $$I\!\!W_n(\hat\theta_n) \;=\; o_P(n^{-1/2}) \qquad \text{for all} \quad P \in \mathbf{P}\,,$$

then $\hat\theta_n$ is an *asymptotic generalized M - estimate* (or *AGM - estimate* for short) of $\theta(P)$. Define

$$I\!\!V_n(\theta) \;\equiv\; \sqrt{n}\,\{ I\!\!W_n(\theta) \;-\; W(\theta\,, P) \}$$

for $\theta \in \Theta$ *open* $\subset R^m$. Here are the key assumptions:

(GM0) There exists $\theta : \mathbf{Q} \to R^m$ such that $\theta(P)$ satisfies $W(\theta(P), P) = 0$ for all $P \in \mathbf{Q}$.

(GM1) For any $\epsilon_n \downarrow 0$ we have

$$\sup\{ \frac{|\mathbb{W}_n(\theta) - \mathbb{W}_n(\theta(P))|}{1 + \sqrt{n}|\theta - \theta(P)|} : |\theta - \theta(P)| \le \epsilon_n \} = o_p(1) .$$

(GM2) There is a function $\psi : \mathbf{X} \times \mathbf{Q} \rightarrow R^m$ with $\int \psi(x, P) dP(x) = 0$ and $|\psi(\cdot, P)| \in L_2(P)$ such that

$$\mathbb{W}_n(\theta(P)) = n^{-1} \sum_{i=1}^{n} \psi(X_i, P) + o_p(n^{-1/2}) .$$

(GM3) $W(\cdot, P) = (W_1(\cdot, P), \cdots, W_m(\cdot, P))^T$ is differentiable with derivative $\dot{W}(\theta, P) \equiv [\frac{\partial}{\partial \theta_j} W_i(\theta, P)]_{m \times m}$ and $\dot{W}(P) \equiv \dot{W}(\theta(P), P)$ is non-singular.

We introduce the model $\mathbf{Q}$ to emphasize that even though W may be motivated by features of $\mathbf{P}$, the AGM-estimates corresponding to W can be thought of as estimates of parameters on a larger model.

Theorem 8.6.5. (AGM - estimates). Suppose $P \in \mathbf{Q}$. Let $\hat{\theta}_n$ be an AGM - estimate of $\theta(P)$ on $\mathbf{Q}$. If $\hat{\theta}_n$ is consistent and if (GM0) - (GM3) hold, then

$$\sqrt{n}(\hat{\theta}_n - \theta(P)) = - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{W}^{-1}(P) \psi(X_i, P) + o_p(1)$$
$$\rightarrow_d N_m(0, \Sigma(P)) \qquad \text{as } n \rightarrow \infty$$

where

$$\Sigma(P) = \dot{W}^{-1}(P) E[\psi(X, P) \psi^T(X, P)] [\dot{W}^{-1}(P)]^T .$$

If (GM1) holds for $\epsilon_n = O(n^{-1/2})$ only, and if $\hat{\theta}_n$ is $\sqrt{n}$ – consistent, this asymptotic linearity remains valid.

Proof. By (GM0) and (1),

(a)
$$\mathbb{W}_n(\theta(P)) + \sqrt{n}\{W(\hat{\theta}_n, P) - W(\theta(P), P)\}$$
$$= \mathbb{W}_n(\theta(P)) - \mathbb{W}_n(\hat{\theta}_n) + o_p(1) .$$

Dividing both sides of (a) by $1 + \sqrt{n}|\hat{\theta}_n - \theta(P)|$, and applying (GM1) yields, since $\hat{\theta}_n$ is consistent,

$$\frac{\mathbb{W}_n(\theta(P)) + \sqrt{n}\{W(\hat{\theta}_n, P) - W(\theta(P), P)\}}{1 + \sqrt{n}|\hat{\theta}_n - \theta(P)|} = o_p(1)$$

and hence, by (GM2),

(b) $$\frac{n^{-1/2}\sum_{i=1}^{n}\psi(X_i,P) + \sqrt{n}\{W(\hat{\theta}_n,P) - W(\theta(P),P)\}}{1 + \sqrt{n}|\hat{\theta}_n - \theta(P)|} = o_p(1)\ .$$

We now complete the proof by arguing as in Huber (1967). Fix $\epsilon > 0$ small, and let $M^2 \equiv 2\int |\psi(x,P)|^2\, dP(x)/\epsilon < \infty$ by (GM2). Then it follows from (GM2) and (b) that, with probability at least $1 - \epsilon$ for n sufficiently large, the following two inequalities hold:

(c) $$|n^{-1/2}\sum_{i=1}^{n}\psi(X_i,P)| < M$$

and

(d) $$|n^{-1/2}\sum_{i=1}^{n}\psi(X_i,P) + \sqrt{n}\,\{W(\hat{\theta}_n,P) - W(\theta(P),P)\}\,| \leq \epsilon(1 + \sqrt{n}|\hat{\theta}_n - \theta(P)|)\ .$$

Since $\hat{\theta}_n$ is consistent, (GM3) implies that with probability converging to 1 we have

(e) $$|W(\hat{\theta}_n,P) - W(\theta(P),P)| \geq \alpha\,|\hat{\theta}_n - \theta(P)|$$

for some $\alpha = \alpha(P) > 0$. Combining (c) - (e) yields

$$\begin{aligned}\epsilon(1 + \sqrt{n}|\hat{\theta}_n - \theta(P)|) &\geq |\sqrt{n}\{W(\hat{\theta}_n,P) - W(\theta(P),P)\}| - M\\ &\geq \alpha\sqrt{n}|\hat{\theta}_n - \theta(P)| - M\end{aligned}$$

and hence, for $\epsilon < \alpha$ and with probability of at least $1 - \epsilon$,

(f) $$\sqrt{n}|\hat{\theta}_n - \theta(P)| \leq \frac{M + \epsilon}{\alpha - \epsilon}\ ,$$

so that $\hat{\theta}_n$ is $\sqrt{n}$ - consistent. Finally, since $\epsilon M = O(\epsilon^{1/2})$, (d) and (f) imply that

(g) $$\sqrt{n}\{W(\hat{\theta}_n,P) - W(\theta(P),P)\} = -\,n^{-1/2}\sum_{i=1}^{n}\psi(X_i,P) + o_p(1)\ ,$$

and the first part of the theorem follows from (g) and (GM3). The modifications of this proof needed for the second part are simple. □

We now specialize theorem 8.6.5 to the case of M - estimates. Suppose that $\psi : \mathbf{X} \times R^m \to R^m$ is such that $\int |\psi(x,\theta)|^2\, dP(x) < \infty$ and $\int \psi(x,\theta(P))\, dP(x) = 0$, and define

$$\boldsymbol{W}_n(\theta) \;\equiv\; \frac{1}{n}\sum_{i=1}^{n} \psi(X_i,\theta) = \int \psi(x,\theta)\, d\boldsymbol{P}_n(x)\ ,$$

and

$$W(\theta,P) \;\equiv\; \int \psi(x,\theta)\, dP(x) \;=\; E_P\psi(X,\theta)\ .$$

Note that (GM0) is satisfied by the assumptions on ψ:

(M0) $\qquad W(\theta(P),P) \;=\; 0\ .$

The key hypothesis (GM1) of theorem 8.6.5 becomes:

(M1) For any $\epsilon_n \downarrow 0$ we have

$$\sup_{|\theta - \theta(P)| \le \epsilon_n} \frac{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\psi(X_i,\theta) - \psi(X_i,\theta(P)) - E_P[\psi(X,\theta) - \psi(X,\theta(P))]\}\right|}{1 + \sqrt{n}\,|\theta - \theta(P)|}$$

$$= \; o_p(1)\ .$$

Note that

(M2) $$\sqrt{n}\,\boldsymbol{W}_n(\theta(P)) \;=\; n^{-1/2}\sum_{i=1}^{n}\psi(X_i,\theta(P))\ ,$$

so that (GM2) holds trivially. Finally, (GM3) becomes

(M3) $W(\theta,P) \;=\; \int \psi(x,\theta)\, dP(x)$ is differentiable and the derivative at $\theta(P)$, $\dot{W}(P)$, is nonsingular.

These identifications prove the following corollary of theorem 1:

Corollary 8.6.6. (M - estimates). Suppose that (M1) and (M3) hold, $\hat{\theta}_n$ is consistent, and is an asymptotic M - estimate: i.e. $\sqrt{n}\,\boldsymbol{W}_n(\hat{\theta}_n) \;=\; o_p(1)$. Then $\hat{\theta}_n$ is asymptotically linear with influence function $-\dot{W}^{-1}(P)\,\psi(\,\cdot\,,P)$.

Suppose that $\dot{\boldsymbol{W}}_n \equiv [\,\frac{\partial}{\partial\theta_j}\boldsymbol{W}_{ni}(\theta)\,]$ exists where $\boldsymbol{W}_n \equiv (\boldsymbol{W}_{n1}, \cdots, \boldsymbol{W}_{nm})^T$ and that:

(U) For some sequence $\{\epsilon_n\}$ with $\epsilon_n \downarrow 0$, $\epsilon_n\, n^{1/2} \to \infty$,

$$\sup\{|\,\dot{\boldsymbol{W}}_n(\theta) \;-\; \dot{W}(P)\,| : |\,\theta - \theta(P)\,| \;<\; \epsilon_n\} \;=\; o_P(1)\,.$$

Let $T_n^{(0)}$ be a preliminary estimate and define

(18) $$T_n^{(j+1)} \;=\; T_n^{(j)} \;-\; \dot{\boldsymbol{W}}_n^{-}(T_n^{(j)})\boldsymbol{W}_n(T_n^{(j)})\ , \quad j \;=\; 0, 1, \cdots\ .$$

Theorem 8.6.7. (Iteration). Suppose (GM0), (GM2), (GM3) and (U) hold.

A. With probability converging to 1, $\mathbb{W}_n(\theta)$ has a unique root $\theta_n^{(\infty)}$ in $\{\theta : |\theta - \theta(P)| \le \epsilon_n\}$. $\theta_n^{(\infty)}$ is $\sqrt{n}$ – consistent and asymptotically linear with influence function $-\dot{W}^{-1}(P)\psi(\,\cdot\,, \theta(P))$.

B. If there exists an estimator $T_n^{(0)}$ satisfying $P(\,|T_n^{(0)} - \theta(P)\,| < \epsilon_n) \to 1$ for $\{\epsilon_n\}$ as in (U) and if the Newton Raphson iteration (18) starts at $T_n^{(0)}$, then

$$P(\,T_n^{(\infty)} \text{ exists and equals } \theta_n^{(\infty)}) \;\to\; 1\,, \text{ as } n \to \infty\,.$$

Proof. We begin by establishing two preliminary results. First

$$\text{(a)} \qquad \begin{aligned} a_n &\equiv \sup\{||I - \dot{W}^{-1}(P)\dot{\mathbb{W}}_n(\theta)|| : |\theta - \theta(P)| \le \epsilon_n\} \\ &\le \sup\{||\dot{W}^{-1}(P)||\,||\dot{W}(P) - \dot{\mathbb{W}}_n(\theta)|| : |\theta - \theta(P)| \le \epsilon_n\} \\ &= o_p(1) \qquad \text{by (U) and (GM3)}\,. \end{aligned}$$

Note that (U) and (GM3) imply also that

$$P\left\{\begin{matrix} \dot{\mathbb{W}}_n^{-1}(\theta) \quad \text{exists and} \quad ||\dot{\mathbb{W}}_n^{-1}(\theta)|| \le 2\,||\dot{W}^{-1}(P)|| \\ \text{for all} \quad \theta \quad \text{with} \quad |\theta - \theta(P)| \le \epsilon_n \end{matrix}\right\} \;\to\; 1\,.$$

Hence, just as in (a),

$$\text{(b)} \qquad \begin{aligned} b_n &\equiv \sup\{||I - \dot{\mathbb{W}}_n^{-1}(\theta)\,\dot{\mathbb{W}}_n(\theta')|| : |\theta - \theta(P)| \le \epsilon_n\,, |\theta' - \theta(P)| \le \epsilon_n\} \\ &= o_p(1)\,. \end{aligned}$$

Proof of A: Let $h_n(\theta) = \theta - \dot{W}^{-1}(P)\,\mathbb{W}_n(\theta)$, and let $\theta_n^{(0)} = \theta(P)$, $\theta_n^{(j+1)} = h_n(\theta_n^{(j)})$, $j \ge 1$.

Suppose that

$$|\theta_n^{(j)} - \theta(P)| \le (1 - a_n)^{-1}\,|\dot{W}^{-1}(P)\,\mathbb{W}_n(\theta(P))| \equiv (1 - a_n)^{-1}A_n\,,$$

where $A_n = O_p(n^{-1/2})$ by (GM2). Since

$$\text{(c)} \qquad \begin{aligned} \theta_n^{(j+1)} - \theta(P) &= \theta_n^{(j)} - \theta(P) - \dot{W}^{-1}(P)\,\mathbb{W}_n(\theta_n^{(j)}) \\ &= \theta_n^{(j)} - \theta(P) - \dot{W}^{-1}(P)(\mathbb{W}_n(\theta_n^{(j)}) - \mathbb{W}_n(\theta(P))) \\ &\qquad - \dot{W}^{-1}(P)\,\mathbb{W}_n(\theta(P)) \\ &= (I - \dot{W}^{-1}(P)\,\dot{\mathbb{W}}_n(\tilde{\theta}_n^{(j)}))\,(\theta_n^{(j)} - \theta(P)) \\ &\qquad - \dot{W}^{-1}(P)\,\mathbb{W}_n(\theta(P))\,, \end{aligned}$$

for some intermediate point $\tilde{\theta}_n^{(j)}$, it follows from (a) and (U) that

$$|\theta_n^{(j+1)} - \theta(P)| \leq a_n \frac{A_n}{1 - a_n} + A_n = \frac{A_n}{1 - a_n} .$$

Since $\theta_n^{(0)} - \theta(P) = 0$, induction yields

(d) $$\sup_{j \geq 1} |\theta_n^{(j)} - \theta(P)| \leq (1 - a_n)^{-1} A_n = O_p(n^{-1/2}) .$$

Now consider

(e) $$\begin{aligned} &|h_n(\theta_n^{(j+1)}) - h_n(\theta_n^{(j)})| \\ &\quad = |\theta_n^{(j+1)} - \theta_n^{(j)} - \dot{W}^{-1}(P)(W_n(\theta_n^{(j+1)}) - W_n(\theta_n^{(j)}))| \\ &\quad = |(I - \dot{W}^{-1}(P)\dot{W}_n(\tilde{\theta}_n^{(j)}))(\theta_n^{(j+1)} - \theta_n^{(j)})| \\ &\quad \leq a_n |\theta_n^{(j+1)} - \theta_n^{(j)}| \end{aligned}$$

where $\tilde{\theta}_n^{(j)}$ is another intermediate point, which, by (d), satisfies $|\tilde{\theta}_n^{(j)} - \theta(P)| = O_p(n^{-1/2})$. It follows from (a) and (e) that h_n is a contraction on the set $\{\theta : |\theta - \theta(P)| \leq (1 - a_n)^{-1} A_n\}$, and hence that $\theta_n^{(j)} \to \theta_n^{(\infty)}$ as $j \to \infty$ on a set with probability arbitrarily close to 1 for large n . Further $\theta_n^{(\infty)}$ satisfies

(f) $$\theta_n^{(\infty)} - \theta(P) = O_p(n^{-1/2}) = o_p(\epsilon_n) .$$

Moreover, $\theta_n^{(\infty)}$ is a fixed point of h_n and satisfies $W_n(\theta_n^{(\infty)}) = 0$. The linearity of $\theta_n^{(\infty)}$ follows from theorem 8.6.5.

Proof of B: Consider the sequence (2). Assume that $|T_n^{(j)} - \theta(P)| \leq \epsilon_n$. Then

(g) $$\begin{aligned} |T_n^{(j+1)} - \theta_n^{(\infty)}| &= |T_n^{(j)} - \theta_n^{(\infty)} - \dot{W}_n^{-}(T_n^{(j)})(W_n(T_n^{(j)}) - W_n(\theta_n^{(\infty)}))| \\ &= |(I - \dot{W}_n^{-}(T_n^{(j)})\,\dot{W}_n(\tilde{T}_n^{(j)}))(T_n^{(j)} - \theta_n^{(\infty)})| \\ &\leq b_n |T_n^{(j)} - \theta_n^{(\infty)}| \end{aligned}$$

where $\tilde{T}_n^{(j)}$ is another intermediate point. Hence

$$\begin{aligned} |T_n^{(j+1)} - \theta(P)| &\leq b_n |T_n^{(j)} - \theta(P)| + |\theta_n^{(\infty)} - \theta(P)| \\ &\leq b_n \epsilon_n + o_p(\epsilon_n) \\ &\leq \epsilon_n \end{aligned}$$

on the event $\{|\theta_n^{(\infty)} - \theta(P)| \leq \epsilon_n/2\} \cap \{b_n < 1/2\}$. This event has probability converging to 1 by (b) and (f). We obtain from (g) that, for n sufficiently large,

$$|T_n^{(j)} - \theta_n^{(\infty)}| \leq b_n^j |\tilde{\theta}_n - \theta_n^{(\infty)}| \to 0 \quad \text{as } j \to \infty$$

with probability converging to 1.

□

M - Estimates with Monotone ψ for $\theta \in \Theta \subset R^m$

Note that the asymptotic equicontinuity type condition GM1 was avoided in our treatment of the one-dimensional situation by imposing a monotonicity hypothesis. Similarly, in the case of $\Theta \subset R^m$, convexity hypotheses on ρ, which lead to a type of monotonicity of the gradient functions ψ and hence of W, can allow GM1 to be efficiently bypassed. Briefly, convexity assumptions can be coupled with much weaker conditions than (GM0) - (GM3) to yield the conclusion of theorem 8.6.5. The approach taken here uses a theorem of Brown (1985) which is reproved by Ritov (1987); see Bickel, Klaassen, Ritov, and Wellner (1993), theorem 7.4.2, page 328, theorem A.10.3, page 519, and the key uniformity result contained in theorem A.7.8, page 473.

Let Θ be an open convex subset of R^m. Let $\mathbf{W}$ be the class of all functions $W : \Theta \to R^m$ such that for all $u \in R^m$, $t \in \Theta$, the maps $\lambda \to u^T W(t + \lambda u)$ from $\{\lambda \in R : t + \lambda u \in \Theta\}$ to R are monotone nondecreasing. Let $\mathbf{W}_0 \subset \mathbf{W}$ be the subclass of functions which have a unique root in Θ.

Assume that (GM0) and (GM3) hold, and let $\theta_0 \equiv \theta(P)$. Suppose that $\mathbf{Q}$ contain all realizations of the the empirical measures $I\!P_n$, $n \geq 1$. Here $\mathbf{Q} \supset \mathbf{P}$. Fix $P \in \mathbf{Q}$ and $W : \Theta \times \mathbf{Q} \to R^m$, and consider the assumptions:

(C1) $P(W(\,\cdot\,, I\!P_n) \in \mathbf{W}_0) \to 1$ as $n \to \infty$.

(C2) For each fixed $\tau \in R^m$,
$$\sqrt{n}(W(\theta_0 + n^{-1/2}\tau, I\!P_n) - W(\theta_0, I\!P_n)) = \dot{W}(P)\tau + o_p(1)\,.$$

(C3) $W(\theta_0, I\!P_n) = \int \psi(x, P)\, dI\!P_n + o_p(n^{-1/2})$ where $|\psi| \in L_2(P)$,
$$\int \psi(x, P)\, dP(x) = 0\,.$$

We will also use a strengthening of (C1):

(C1′) $W(\,\cdot\,, Q) \in \mathbf{W}_0$ for all $Q \in \mathbf{Q}$.

The following useful result is due to Brown (1985) and Ritov (1987).

Theorem 8.6.8. (Convexity). Suppose that $X_1, \cdots, X_n$ are iid $P \in \mathbf{Q}$. Suppose that (GM0), (GM3), (C1′), (C2), and (C3) hold. Then $\hat{\theta}_n$ corresponding to $W_n(\theta) \equiv W(\theta, I\!P_n) = 0$ is uniquely defined and asymptotically linear with influence function $-\dot{W}^{-1}(P)\psi(\,\cdot\,, P)$. If the hypothesis (C1′) is replaced by (C1), then the AGM - estimate $\hat{\theta}_n$ exists and asymptotic linearity continues to hold.

Theorem 8.6.8 is a result similar to theorem 8.6.1, but for $\Theta \subset R^m$ with $m > 1$. It can also be viewed as an alternative to theorem 8.6.5 with weak smoothness hypotheses counter-balanced by the strong assumptions (C1) or (C1′). The proof hinges on Bickel, Klaassen, Ritov, and Wellner (1993), theorem A.7.8, which is a close relative of the uniform convergence lemma 4.7.1 for convex functions.

Proof. Let $U_n(\tau) = \sqrt{n} W(\theta_0 + n^{-1/2}\tau, IP_n)$ Since $U_n(\cdot) - U_n(0) \in \mathbf{W}$ and $\tau \to \dot{W}(P)\tau$ is continuous, (C2) and theorem A.7.8 imply that

(a) $$\sup_{|\tau| \le M} |U_n(\tau) - U_n(0) - \dot{W}(P)\tau| = o_p(1) .$$

We proceed to check the conditions of the second part of theorem 8.6.5 for $W_n(\theta, IP_n) = W(\theta, IP_n)$. Note that

(b) $$V_n(\theta) - V_n(\theta_0) = U_n(n^{1/2}(\theta - \theta_0)) - U_n(0) - n^{1/2}[W(\theta, P) - W(\theta_0, P)] .$$

So, by (a) and differentiability of W,

$$\begin{aligned} &\sup\{|V_n(\theta) - V_n(\theta_0)| : n^{1/2}|\theta - \theta_0| \le M\} \\ &= o_p(1) \\ &\quad + \sup\{|n^{1/2}(W(\theta_0 + n^{-1/2}\tau, P) - W(\theta_0, P)) - \dot{W}(P)\tau| : |\tau| \le M\} \\ &= o_p(1) \quad \text{by (GM3)} . \end{aligned}$$

Hence condition (GM1) of theorem 8.6.5 follows, at least for $\epsilon_n = O(n^{-1/2})$. Suppose (C1′) holds. We now need only to verify the $\sqrt{n}$ – consistency of $\hat{\theta}_n$ which is uniquely defined by $W(\hat{\theta}_n, IP_n) = 0$. It is enough to show that for all $\epsilon > 0$ there exists an $M = M(\epsilon)$, so that, for n sufficiently large

(c) $$P(A_n) > 1 - \epsilon$$

where

$$A_n = [\inf\{\tau^T U_n(\tau) : |\tau| = M\} > 0] .$$

This is enough since $U_n(\tau) \in \mathbf{W}$ implies $\tau^T U_n(\lambda\tau)$ is increasing in λ and hence that, on A_n, $\tau^T U_n(\tau) > 0$ for all $|\tau| \ge M$, and hence $\sqrt{n}\,|\hat{\theta}_n - \theta_0| < M$. (The argument is spelled out in the proof of theorem 7.5.3.)

We obtain from (a) that

(d) $$\sup\{|\tau^T U_n(\tau) - \tau^T U_n(0) - \tau^T \dot{W}(P)\tau| : |\tau| = M\} = o_p(1) .$$

Let $A(P) = \frac{1}{2}(\dot{W}(P) + \dot{W}^T(P))$. Then

(e) $$\tau^T A(P)\tau = \tau^T \dot{W}(P)\tau \;, \qquad \text{for all} \quad \tau \;.$$

$W(\cdot\,, P) \in \mathbf{W}_0$ implies that $A(P)$ is symmetric, positive definite. In particular its minimum eigenvalue $\lambda_1(P)$ is strictly positive.

Fix $\epsilon > 0$. By (C3), for M sufficiently large and all n :

(f) $$P(\,|U_n(0)| \;<\; \frac{1}{2}\,\lambda_1(P)M\;) \;\geq\; 1 \;-\; \frac{1}{2}\,\epsilon\,.$$

We obtain from (d) - (f) that with probability at least $1-\epsilon$ for all n and M sufficiently large

$$\inf\{\tau^T U_n(\tau) \colon |\tau| \;=\; M\} \;>\; \frac{1}{3}\,\lambda_1(P)M^2\,.$$

Hence (c) holds, and $\hat{\theta}_n$ is $\sqrt{n}$ - consistent. If (C1) holds then (c) still holds and implies that the minimum of $|\,W(\,\cdot\,, IP_n)\,|$ is assumed on $\{\theta\colon \sqrt{n}\,|\,\theta - \theta_0\,| < M\}$. To complete the proof we need only to show therefore, that the minimizer is an AGM - estimate. That is, it suffices to show that $\inf_\tau\,|\,U_n(\tau)\,| \;=\; o_p(1)$. Let $\tau_n \;=\; -\,\dot{W}^{-1}(P)\,U_n(0)$. We obtain from (a) that

$$\inf_\tau\,|\,U_n(\tau)\,| \;\leq\; |\,U_n(\tau_n)\,| \;=\; o_p(1)\;,$$

and the theorem follows. □

The Bootstrap for M - estimates

With several good theorems for M - estimators and generalized M - estimators in hand, the question now is: does the bootstrap "work" in probability or almost surely for such estimators? While generalized M - estimators are a bit too general to answer this question easily, good results for M - estimators are now available. The following treatment has been adapted from the work of Arcones and Giné (1992) -- but also see Lele (1991), ???, and ??? .

We adopt the notation used in corollary 8.6.6.

Theorem 8.6.?. Suppose that $P \in \mathbf{Q}$, and that (M0), (M1), and (M3) hold. Suppose that $\hat{\theta}_n$ is a consistent estimator of $\theta(P)$ and that the bootstrap estimator $\hat{\theta}^*_n \equiv \theta(IP^*_n)$ is consistent in probability; i.e. for every $\epsilon > 0$

$$P^*(|\hat{\theta}^*_n - \hat{\theta}_n| > \epsilon|\, IP_n\,) \;\to_p\; 0\,.$$

Then

$$\sqrt{n}(\hat{\theta}_n \;-\; \theta(P)) \;\to_d\; N(0\,, \Sigma(P))$$

and

$$\sqrt{n}(\hat{\theta}^*_n \;-\; \hat{\theta}_n) \;\to_d\; N(0\,, \Sigma(P)) \qquad \text{in probability}$$

where, as in theorem 8.6.6 and corollary 8.6.7,

$$\Sigma(P) \;=\; \dot{W}^{-1}(P)\,E[\psi(X\,,P)\,\psi^{T}(X\,,P)][\dot{W}^{-1}(P)]^{T}\,.$$