Chapter 8
Bootstrap and Jackknife Estimation of Sampling Distributions

1. **A General View of the Bootstrap**
2. **Bootstrap Methods**
3. **The Jackknife**
4. **Some limit theory for bootstrap methods**
5. **The bootstrap and the delta method**
6. **Bootstrap Tests and Bootstrap Confidence Intervals**
7. **M - Estimators and the Bootstrap**

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

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

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 ideal bootstrap estimators are often difficult to compute exactly, we can often obtain Monte-Carlo estimates thereof by sampling fromm $P_{\hat{\theta}}$: let $\underline{X}_1^*, \ldots, \underline{X}_B^*$ be i.i.d. with common distribution $P_{\hat{\theta}}$, and calculate $\hat{\theta}(\underline{X}_j^*)$ for $j = 1, \ldots, B$. 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″. $B^{-1} \sum_{j=1}^B (\underline{a}^T \hat{\theta}(\underline{X}_j^*) - B^{-1} \sum_{j=1}^B \underline{a}^T \hat{\theta}(\underline{X}_j^*))^2$.

If $\mathcal{P}$ is a parametric model, the above approach yields a *parametric bootstrap*. If $\mathcal{P}$ 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 $\underline{X} = (X_1, \ldots, X_n)$ i.i.d. F or P with $\mathcal{P}$ nonparametric so that $P_\theta = F \times \cdots \times F$ and $P_{\hat{\theta}} = \mathbb{F}_n \times \cdots \times \mathbb{F}_n$. Or, if the basic underlying sample space for each X_i is not $\mathbb{R}$, $P_\theta = P \times \cdots \times P$ and $P_{\hat{\theta}} = \mathbb{P}_n \times \cdots \times \mathbb{P}_n$.

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 $X_1, \ldots, X_n$ are i.i.d. with distribution function F, then we estimate $T(F)$ by $T(\mathbb{F}_n) \equiv T_n$ where $\mathbb{F}_n$ is the empirical d.f. $\mathbb{F}_n \equiv n^{-1}\sum_{i=1}^n 1\{X_i \le x\}$. More generally, if $T(P)$ is some functional of P and $X_1, \ldots, X_n$ are i.i.d. P, then a natural estimator of $T(P)$ is just $T(\mathbb{P}_n)$ where $\mathbb{P}_n$ is the empirical measure $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$.

Consider estimation of:

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

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

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. $K_n(x, F) \equiv P_F(\sqrt{n}\|\mathbb{F}_n - F\|_\infty \le x)$.

F. $L_n(x, P) \equiv Pr_P(\sqrt{n}\|\mathbb{P}_n - P\|_{\mathcal{F}} \le x)$ where $\mathcal{F}$ is a class of functions for which the central limit theorem holds uniformly over $\mathcal{F}$ (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 $\mathbb{F}_n$ (or the empirical measure $\mathbb{P}_n$). This yields the following nonparametric bootstrap estimates in examples A - F:

A′. $b_n(\mathbb{F}_n) \equiv n\{E_{\mathbb{F}_n}(T_n) - T(\mathbb{F}_n)\}$.

B′. $n\sigma_n^2(\mathbb{F}_n) \equiv nVar_{\mathbb{F}_n}(T_n)$.

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

D′. $H_n(x, \mathbb{F}_n) \equiv P_{\mathbb{F}_n}(\sqrt{n}(T_n - T(\mathbb{F}_n)) \le x)$.

E′. $K_n(x, \mathbb{F}_n) \equiv P_{\mathbb{F}_n}(\sqrt{n}\|\mathbb{F}_n^* - \mathbb{F}_n\|_\infty \le x)$.

F′. $L_n(x, \mathbb{P}_n) \equiv Pr_{\mathbb{P}_n}(\sqrt{n}\|\mathbb{P}_n^* - \mathbb{P}_n\|_{\mathcal{F}} \le x)$ where $\mathcal{F}$ is a class of functions for which the central limit theorem holds uniformly over $\mathcal{F}$ (i.e. a *Donsker class*).

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 d.f. $\mathbb{F}_n$ is discrete (with all its mass at the data), we could, in principle enumerate all possible samples of size n from $\mathbb{F}_n$ (or $\mathbb{P}_n$) 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

$$(X_{j1}^*, \ldots, X_{jn}^*) \qquad j = 1, \ldots, B$$

be B independent samples of size n drawn *with replacement* from $\mathbb{F}_n$ (or $\mathbb{P}_n$); let

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

be the empirical d.f. of the $j-$th sample, and let

$$T^*_{j,n} \equiv T(\mathbb{F}^*_{j,n}), \qquad j = 1, \ldots, B.$$

Then approximations of $A' - F'$ are given by:

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

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″. $K^*_{n,B}(x) \equiv \frac{1}{B} \sum_{j=1}^{B} 1\{\sqrt{n}\|\mathbb{F}^*_{j,n} - \mathbb{F}_n\|_\infty \le x\}$.

F″. $L^*_{n,B}(x) \equiv \frac{1}{B} \sum_{j=1}^{B} 1\{\sqrt{n}\|\mathbb{P}^*_{j,n} - \mathbb{P}_n\|_{\mathcal{F}} \le x\}$.

For fixed sample size n and data $\mathbb{F}_n$, it follows from the Glivenko - Cantelli theorem (applied to the bootstrap sampling) that

$$\sup_x |H^*_{n,B}(x) - H_n(x, \mathbb{F}_n)| \to_{a.s.} 0 \qquad \text{as} \;\; B \to \infty,$$

and, by Donsker's theorem,

$$\sqrt{B}(H^*_{n,B}(x) - H_n(x, \mathbb{F}_n)) \Rightarrow \mathbb{U}^{**}(H_n(x, \mathbb{F}_n)) \qquad \text{as} \;\; B \to \infty.$$

Moreover, by the Dvoretzky, Kiefer, Wolfowitz (1956) inequality ($P(\|\mathbb{U}_n\| \ge \lambda) \le 2\exp(-2\lambda^2)$ for all n and $\lambda > 0$ where the constant 2 before the exponential comes via Massart (1990)),

$$P(\sup_x |H^*_{n,B}(x) - H_n(x, \mathbb{F}_n)| \ge \epsilon) \le 2\exp(-2B\epsilon^2).$$

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 $H_n(x, \mathbb{F}_n)$ are so well -understood and controlled, much of our discussion below will focus on the differences between $H_n(x, \mathbb{F}_n)$ and $H_n(x, F)$.

Sometimes it is possible to compute the distribution of the bootstrap estimator explicitly with out resort to Monte-Carlo; here is an example of this kind.

Example 2.1 (The distribution of the bootstrap estimator of the median). Suppose that $T(F) = F^{-1}(1/2)$. Then

$$T(\mathbb{F}_n) = \mathbb{F}_n^{-1}(1/2) = X^*_{([n+1]/2)}$$

and

$$T(\mathbb{F}^*_n) = \mathbb{F}_n^{*-1}(1/2) = X_{([n+1]/2)}.$$

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

$$\underline{M} \equiv (M_1, \ldots, M_n) \sim \text{Mult}_n(n, (1/n, \ldots, 1/n)).$$

Now $[X_{(m)}^* > X_{(k)}(\omega)] = [n\mathbb{F}_n^*(X_{(k)}(\omega)) \le m-1]$, and hence

$$\begin{aligned} P(T(\mathbb{F}_n^*) = X_{(m)}^* > X_{(k)}(\omega)|\mathbb{F}_n) &= P(n\mathbb{F}_n^*(X_{(k)}(\omega)) \le m-1|\mathbb{F}_n) \\ &= P(\text{Binomial}(n, k/n) \le m-1) \\ &= \sum_{j=0}^{m-1} \binom{n}{j} (k/n)^j (1-k/n)^{n-j}. \end{aligned}$$

This implies that

$$\begin{aligned} &P(T(\mathbb{F}_n^*) = X_{(k)}(\omega)|\mathbb{F}_n) \\ &= \sum_{j=0}^{m-1} \left\{ \binom{n}{j} \left(\frac{k-1}{n}\right)^j \left(1 - \frac{k-1}{n}\right)^{n-j} - \binom{n}{j} \left(\frac{k}{n}\right)^j \left(1 - \frac{k}{n}\right)^{n-j} \right\} \end{aligned}$$

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

Example 2.2 (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 $\hat{\rho}_n \equiv T(\mathbb{F}_n)$ satisfies

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

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

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

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

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

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

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

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

Let $X_1, \ldots, X_n, \ldots$ be i.i.d. with distribution $P_\theta \in \mathcal{P}$. We assume that there exists an estimator $\hat\theta_n = \hat\theta_n(X_1, \ldots, X_n)$ of θ. Then Efron's parametric (or model - based) bootstrap proceeds by sampling from the estimated or fitted model $P_{\hat\theta(\omega)} \equiv \hat P_n^\omega$: suppose that $X^*_{n,1}, \ldots, X^*_{n,n}$ are independent and identically distributed with distribution $\hat P_n^\omega$ on $(\mathcal{X}, \mathcal{A})$, and let

$$(1) \qquad \mathbb{P}^*_n \equiv n^{-1} \sum_{i=1}^n \delta_{X^*_{n,i}} \equiv \text{the } \textit{parametric bootstrap empirical measure}\ .$$

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 $\hat P_n = p_{\hat\theta_n}$ of P rather than from the nonparametric estimator $\mathbb{P}_n$.

Example 2.3 Suppose that $X_1, \ldots, X_n$ 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, and hence

$$\frac{\sqrt{n}(\hat\mu_n - \mu)}{\hat\sigma_n} \sim t_{n-1}, \qquad \frac{(n-1)\hat\sigma_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Now $P_{\hat\theta_n} = N(\hat\mu_n, \hat\sigma_n^2)$, and if $X^*_1, \ldots, X^*_n$ are i.i.d. $P_{\hat\theta_n}$, then the bootstrap estimators $\hat\theta^*_n = (\hat\mu^*_n, \hat\sigma^{*2}_n)$ satisfy, conditionally on $\mathbb{F}_n$,

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

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

Example 2.4 Suppose that $X_1, \ldots, X_n$ are i.i.d. $P_\theta = \text{exponential}(1/\theta)$: $P_\theta(X_1 > t) = \exp(-t/\theta)$ for $t \ge 0$. Then $\hat\theta_n = \overline{X}_n$ and $n\hat\theta_n/\theta \sim \text{Gamma}(n, 1)$. Now $P_{\hat\theta_n} = \text{exponential}(1/\hat\theta_n)$, and if $X^*_1, \ldots, X^*_n$ are i.i.d. $P_{\hat\theta_n}$, then $\hat\theta^*_n = \overline{X}^*_n$ has $(n\hat\theta^*_n/\hat\theta_n | \mathbb{F}_n) \sim \text{Gamma}(n, 1)$, so the bootstrap distribution replicates the original estimator exactly.

Example 2.5 (Bootstrapping from a "smoothed empirical measure"; or the "smoothed bootstrap"). Suppose that

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

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

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

where $k : \mathbb{R}^d \to \mathbb{R}$ is a uniformly continuous density function. Then $\hat{\mathbb{P}}_n$ is defined for $C \in \mathcal{A}$ by

$$\hat{\mathbb{P}}_n(C) = \int_C \hat{p}_n(x)dx,$$

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

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

Exchangeably - weighted and "Bayesian" bootstrap methods

In the course of example 5.1 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 $\mathbb{P}_n$, it becomes clear that we can think of the bootstrap empirical measure $\mathbb{P}_n^*$ as the empirical measure with multinomial random weights:

$$\mathbb{P}_n^* = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^*} = \frac{1}{n}\sum_{i=1}^n M_i \delta_{X_i(\omega)}.$$

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 deficiencies 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 number proportion of missed observations is

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

[Moreover, from occupancy theory for urn models

$$\sqrt{n}(n^{-1}N_n - (1-1/n)^n) \to_d N(0, e^{-1}(1-2e^{-1})) = N(0, .09720887\ldots);$$

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 $M_n \sim \text{Mult}_n(n, (1/n, \ldots, 1/n))$, 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. $\pi \underline{W} \equiv (W_{\pi(1)}, \ldots, W_{\pi(n)}) \stackrel{d}{=} \underline{W}$ for all permutations π of $\{1, \ldots, n\}$. Then

$$\mathbb{P}_n^W \equiv \frac{1}{n} \sum_{i=1}^n W_{ni} \delta_{X_i(\omega)}$$

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

Example 2.6 (Dirichlet weights). Suppose that $Y_1, Y_2, \ldots$ are i.i.d. exponential(1) random variables, and set

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

The resulting random vector $\underline{W}/n$ has a Dirichlet$(1, \ldots, 1)$ 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 2.7 (More general continuous weights). Other weights $\underline{W}$ of the same for as in example 1.6 are obtained by replacing the exponential distribution of the Y's by some other distribution on $\mathbb{R}^+$. It will turn out that the limit theory can be established for any of these weights as long as the Y_i's satisfy $Y_i \in L_{2,1}$; i.e. $\int_0^\infty \sqrt{P(|Y| > t)}dt < \infty$.

Example 2.8 (Jackknife weights). Suppose that $\underline{w} = (w_{n,1}, \ldots, w_{n,n})$ is a vector of constants which sum to n: $\sum_{i=1}^n w_{n,i} = n$. Let $\underline{W}$ be a random permutation of the coordinates of $\underline{w}$: if $\underline{R}$ is uniformly distributed over $\Pi \equiv \{$all permutations of $\{1, \ldots, n\}\}$, then $\underline{W} \equiv \underline{R}\,\underline{w} \equiv (w_{n,R_1}, \ldots, w_{n,R_n})$. If we take $\underline{w} = (n/(n-d))\underline{1}_{n-d} = (n/(n-d)(1, \ldots 1, 0, \ldots 0)$ 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 $W_{n,i}$ 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 also 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 $T(\mathbb{F}_n)$ estimates $T(F)$. Let

$$T_{n:i} \equiv T(\mathbb{F}_{n-1,i}) \qquad \text{where} \qquad \mathbb{F}_{n-1,i}(x) \equiv \frac{1}{n-1}\sum_{j\neq i} 1_{(-\infty,x]}(X_j);$$

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

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

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

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

Now let $E_n \equiv E_F T_n = E_F T(\mathbb{F}_n)$, 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 $T_n = T(\mathbb{F}_n)$ is

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

We can also write

$$T(F) = E_F(T_n) - \text{bias}_n(F).$$

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

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

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 call $\overline{T}_n^*$ the *jackknife estimator of* $T(F)$; similarly, by writing

$$\overline{T}_n^* = T_n - \widehat{\text{bias}}_n,$$

we find that

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

Example 3.1 If $T(F) = E_F(X) = \int x dF(x)$ so that $T_n = \overline{X}_n$, then $T^*_{n,i} = nT_n - (n-1)T_{n,i} = X_i$, so $\overline{T}^*_n = \overline{X}_n = T_n$, and $\widehat{\text{bias}}_n = 0$.

Example 3.2 If $T(F) = Var_F(X) = \int(x - \int y dF(y))^2 dF(x)$ so that $T_n = T(\mathbb{F}_n) = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$, 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 $\overline{T}^*_n = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1)$, the usual unbiased estimator of $T(F)$. The bias estimtor is just

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

The Jackknife estimator of variance

Now consider estimation of

$$Var_n \equiv Var_F(T_n) = Var_F(T(\mathbb{F}_n)).$$

Tukey's jackknife estimator of Var_n is

$$\begin{aligned}\widehat{Var}_n &= \frac{n-1}{n}\sum_{i=1}^n (T_{n,i} - T_{n,\cdot})^2 \\ &= \frac{1}{n(n-1)}\sum_{i=1}^n [T^*_{n,i} - \overline{T}^*_n]^2 \equiv \frac{n-1}{n}\widetilde{Var}_{n-1},\end{aligned}$$

and hence

$$\widetilde{Var}_{n-1} = \frac{1}{(n-1)^2}\sum_{i=1}^n (T^*_{n,i} - \overline{T}^*_n)^2.$$

Since

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

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

Theorem 3.1 (Efron and Stein, 1981). $E(\widetilde{Var}_{n-1}) \geq Var_{n-1}$.

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 viewed 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 $g = F^{-1}$ has a continuous derivative in a neighborhood of $1/2$,

$$n\widehat{Var}_n = n(n-1)\left\{\frac{X_{(m+1)} - X_{(m)}}{2}\right\}^2 \to_d \frac{1}{4f^2(F^{-1}(1/2))}\left(\frac{\chi_2^2}{2}\right)^2 . \tag{1}$$

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

$$\sqrt{n}(T(\mathbb{F}_n) - T(F)) \to_d N(0, \frac{1/4}{f^2(F^{-1}(1/2))}),$$

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

$$nVar_F(T(\mathbb{F}_n)) \to \frac{1/4}{f^2(F^{-1}(1/2))}$$

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

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

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

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

$$\begin{aligned}
n(X_{(m+1)} - X_{(m)}) &\stackrel{d}{=} \frac{g(\xi_{(m+1)}) - g(\xi_{(m)})}{\xi_{(m+1)} - \xi_{(m)}} n(\xi_{(m+1)} - \xi_{(m)}) \\
&= g'(\xi_{(m)} + \theta(\xi_{(m+1)} - \xi_{(m)}))n(\xi_{(m+1)} - \xi_{(m)}), \quad |\theta| \le 1, \\
&= g'(\mathbb{G}_n^{-1}(1/2) + \theta n(\xi_{(m+1)} - \xi_{(m)})/n)n(\xi_{(m+1)} - \xi_{(m)}) \\
&\to_d g'(1/2 + 0)\exp(1)
\end{aligned}$$

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

$$\begin{aligned}
n\widehat{Var}_n &= \frac{n-1}{4n}\{n(X_{(m+1)} - X_{(m)})\}^2 \\
&\to_d \frac{1}{4}g'(1/2)^2 \exp(1) \qquad \text{by continuous mapping} \\
&\stackrel{d}{=} \frac{1/4}{f^2(F^{-1}(1/2))}(\chi_1^2/2)^2
\end{aligned}$$

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

The delete - d Jackknife

See Shao and Wu (1989), Shao (1993), and Praestgaard (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 $T(\mathbb{P}_n)$ and we know (perhaps by a delta method argument) that

$$\sqrt{n}(T(\mathbb{P}_n) - T(P)) \to_d N(0, V^2(P)),$$

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

$$\sqrt{n}(T(\mathbb{P}_n^*) - T(\mathbb{P}_n)) \to_d N(0, V^2(P)) \qquad \text{in probability or a.s..}$$

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

$$\sqrt{n}(T(\mathbb{P}_n) - T(P)) = \sqrt{n}(\overline{X}_n - \mu(P)) \to_d N(0, Var_P(X)).$$

The corresponding statement for the bootstrap is:

Theorem 4.1 If $EX^2 < \infty$, then for a.e. sequence $X_1, X_2, \ldots,$

$$\sqrt{n}(T(\mathbb{P}_n^*) - T(\mathbb{P}_n)) = \sqrt{n}(\overline{X}_n^* - \overline{X}_n) \ \to_d N(0, Var(X)).$$

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

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

It follows that

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

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

$$\begin{aligned}
&\frac{1}{\sigma_n^2}\sum_{i=1}^n E_*|Z_{ni}|^2 1\{|Z_{ni}| > \epsilon\sigma_n\} \\
&\quad = \frac{1}{S_n^2} n E_* |n^{-1/2} E_* |n^{-1/2}(X_{n1}^* - \overline{X}_n(\omega)|^2 1\{|X_{ni}^* - \overline{X}_n| > \sqrt{n}\epsilon S_n\} \\
&\quad = \frac{1}{S_n^2}\frac{1}{n}\sum_{i=1}^n |X_i(\omega) - \overline{X}_n(\omega)|^2 1\{|X_i - \overline{X}_n| > \sqrt{n}\epsilon S_n\} \\
&\quad \le 1\{\max_{1\le i\le n}|X_i - \overline{X}_n| > \epsilon\sqrt{n}S_n\} \\
&\quad \to_{a.s.} 0
\end{aligned}$$

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

$$\frac{1}{\sqrt{n}}\max_{1\le i\le n}|X_i - \overline{X}_n| \le \frac{1}{\sqrt{n}}\max_{1\le i\le n}|X_i - \mu| + \frac{1}{\sqrt{n}}|\mu - \overline{X}_n| \to_{a.s.} 0,$$

and hence the theorem follows from the Lindeberg-Feller Central Limit Theorem. □

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

Theorem 4.2 (Singh, 1981).
A. If $E(X^2) < \infty$, then

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

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

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

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

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

where $S_n^2 = n^{-1}\sum_1^n (X_i - \overline{X}_n)^2$, then

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

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 non-lattice, then

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

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

$$\sqrt{n} D_n^s \rightarrow_{a.s.} 0.$$

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 $\mathcal{X} = \mathbb{R}$ we have, by the inverse transformation,

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

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 $X_1, X_2, \ldots$.

Theorem 4.3 If $m \wedge n \to \infty$, then for almost every sequence $X_1, X_2, \ldots$,

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

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

Proof. The following proof is due to Shorack (1982). Let $\xi_1^*, \xi_2^*, \ldots$ be i.i.d. Uniform$(0,1)$, let $\mathbb{G}_m^*$ be the empirical d.f. of the first m of the 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 $\|\mathbb{U}_m^* - \mathbb{U}^*\|_\infty \rightarrow_{a.s.} 0$. [In fact by the Hungarian construction, this can be carried out with a sequence of Brownian bridge processes $\mathbb{B}_m^0$ so that $\|\mathbb{U}_m^* - \mathbb{B}_m^0\|_\infty \le M(\log m)/\sqrt{m}$ 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^*: by the inverse transformation the random variables

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

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

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

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

Example 4.1 (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:

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

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

$$K_n(x, \mathbb{F}_n) = P_{\mathbb{F}_n}(\sqrt{n}\|\mathbb{F}_n^* - \mathbb{F}_n\| \leq x),$$

and a Monte-Carlo approximation of it is

$$K_{n,B}^*(x) \equiv \frac{1}{B}\sum_{j=1}^{B} 1\{\sqrt{n}\|\mathbb{F}_{n,j}^* - \mathbb{F}_n\|_\infty \leq x\}$$

where $X_{j,1}^*, \ldots, X_{j,n}^*$ is a random sample from $\mathbb{F}_n$ for each $j = 1, \ldots, B$.

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

$$\lim_{n\to\infty} K_n(c_n(\alpha, F), F) = \lim_{n\to\infty} P_F(\sqrt{n}\|\mathbb{F}_n - F\|_\infty \leq c_n(\alpha, F)) = 1 - \alpha,$$

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

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

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

Corollary 1 The bootstrap confidence bands $\{\mathbb{F}_n \pm n^{-1/2} c_n(\alpha, \mathbb{F}_n)\}$ satisfy

$$\lim_{n\to\infty} P_F\{\mathbb{F}_n(x) - n^{-1/2} c_n(\alpha, \mathbb{F}_n) \le F(x) \le \mathbb{F}_n(x) + n^{1/3} c_n(\alpha, \mathbb{F}_n) \ \text{ for all } \ x \in \mathbb{R}\} = 1 - \alpha.$$

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 4.3 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 $\mathcal{F} \subset L_2(P)$, we let the envelope function F be defined by $F(x) \equiv \sup_{f\in\mathcal{F}} |f(x)|$. Here are the two bootstrap limit theorems of Giné and Zinn (1990):

Theorem 4.4 (Giné and Zinn, 1990). (Almost sure bootstrap limit theorem). Suppose that $\mathcal{F}$ is $P-$measurable. Then the following are equivalent:

A. $\mathcal{F} \in CLT(P)$ and $P(F^2) < \infty$; i.e. $\sqrt{n}(\mathbb{P}_n - P) \Rightarrow \mathbb{G}_P$ where $\mathbb{G}_P$ is a ρ_P- uniformly continuous $P-$ Brownian bridge process on $\mathcal{F}$.

B. $\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n^\omega) \Rightarrow \mathbb{G}_P$ in $\ell^\infty(\mathcal{F})$ almost surely.

Theorem 4.5 (Giné and Zinn, 1990). (In probability bootstrap limit theorem). Suppose that $\mathcal{F}$ is $P-$measurable. Then the following are equivalent:

A. $\mathcal{F} \in CLT(P)$; i.e. $\sqrt{n}(\mathbb{P}_n - P) \Rightarrow \mathbb{G}_P$ where $\mathbb{G}_P$ is a ρ_P- uniformly continuous $P-$ Brownian bridge process on $\mathcal{F}$.

B. $\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n) \Rightarrow \mathbb{G}_P$ in $\ell^\infty(\mathcal{F})$ in probability.

The proofs of these two theorem rely on "multiplier inequalities" closedly related to the "multiplier central limit theorem", Poissonization inequalities, and other tools from empirical process theory. See Giné and Zinn (1990), Klaassen and Wellner (1992), and van der Vaart and Wellner (1996), chapters 3.6 and 2.9. In particular, van der Vaart and Wellner (1996), theorems 3.6.1 and 3.6.2, give a version of theorems 4.4 and 4.5 with a somewhat improved treatment of the measurability issues.

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 $\underline{W} = \underline{W}_n$ in $\mathbb{R}^n$ are exchangeable for each n.

W2. $W_{nj} \ge 0$ for all $j = 1, \ldots, n$ and $\sum_{j=1}^n W_{nj} = n$.

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

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

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

Theorem 4.6 (Exchangeably weighted bootstrap limit theorem). Suppose that $\mathcal{F}$ is $P-$ measurable and that the random weight vectors $\{\underline{W}_n\}$ satisfy W1 - W5. Then:

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

$$(1) \qquad \sqrt{n}(\mathbb{P}_n^W - \mathbb{P}_n^\omega) \Rightarrow \mathbb{G}_P$$

in $\ell^\infty(\mathcal{F})$ almost surely.

B. $\mathcal{F} \in CLT(P)$ implies that the convergence in (1) holds in probability.

For the proofs, see Praestgaard and Wellner (1993), or van der Vaart and Wellner (1996), chapter 3.6.

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 1 Suppose that $\mathcal{F}$ is $P-$measurable. Then:

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

$$(2) \qquad \sqrt{m}(\mathbb{P}_m^* - \mathbb{P}_n) \Rightarrow \mathbb{G}_P$$

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

B. $\mathcal{F} \in CLT(P)$ implies that the convergence in (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 4.2 (Bootstrapping the estimator of θ for the Uniform$(0,\theta)$ distribution.) Suppose that $X_1, \ldots, X_n$ are i.i.d Uniform$(0,\theta)$. Then $\hat{\theta}_n = X_{(n)} \equiv \max_{1\le i\le n} X_i$ and

$$n(\theta - \hat{\theta}_n) = n\theta(1 - X_{(n)}/\theta) \stackrel{d}{=} \theta n(1 - \xi_{(n)}) \to_d \theta Y$$

where $Y \sim \exp(1)$. Thus the limiting distribution is exponential$(1/\theta)$. Now let $X_1^*, \ldots, X_n^*$ be i.i.d. $\mathbb{F}_n$, and let $\hat{\theta}_n^* \equiv X_{(n)}^*$. Then

$$\begin{aligned} P(\hat{\theta}_n^* = X_{(n)}|\mathbb{F}_n) &= 1 - P(X_{(n)}^* < X_{(n)}|\mathbb{F}_n) \\ &= 1 - P(\text{all } X_i^* < X_{(n)}|\mathbb{F}_n) = 1 - \left(\frac{n-1}{n}\right)^n \\ &= 1 - \left(1 - \frac{1}{n}\right)^n \to 1 - e^{-1} \doteq .62\cdots, \end{aligned}$$

and, more generally,

$$\begin{aligned} P(n(X_{(n)} - \hat{\theta}_n^*) > n(X_{(n)} - X_{(n-k+1)})|\mathbb{F}_n) &= P(X_{(n)}^* < X_{(n-k+1)}|\mathbb{F}_n) \\ &= (1 - k/n)^n \to e^{-k}. \end{aligned}$$

[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 4.3 (Bootstrapping a $V-$statistic). Suppose that $X_1, \ldots, X_n$ are i.i.d. with common distribution function F, and let $h : \mathbb{R}^2 \to \mathbb{R}$ be a symmetric function; i.e. $h(x, y) = h(y, x)$. Then

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

is the V-statistic based on the function h while $U_n = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} h(X_i, X_j) = n^{-1}(n-1)^{-1} \sum_{i \ne j} h(X_i, X_j)$ is the U-statistic based on the function h. Note that U_n and V_n are closely related since

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

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

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

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

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

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

$$V_n^* \equiv \int\int xy d\mathbb{F}_n^*(x) d\mathbb{F}_n^*(y) = (\overline{X}_n^*)^2$$

we have

$$nV_n^* = (\sqrt{n} \overline{X}_n^*)^2 = \left(\sqrt{n}(\overline{X}_n^* - \overline{X}_n) + \sqrt{n} \overline{X}_n \right)^2$$

where the first term converges in distribution a.s. conditionally on $X_1, X_2, \ldots$ to $\sigma_F Z^*$ where $Z^* \sim N(0, 1)$, but the second term does not converge to zero, but instead converges to something random and non-degenerate (namely $\sigma_F Z$), and marginally (over both the bootstrap and original data randomness) we $nV_n^* \to_d \sigma_F^2 \chi_2^2$. Hence the nonparametric bootstrap fails. This example is due to Bretagnolle (1983), and a solution to the failure is due to Arcones and Giné (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 other functional) 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 $T_n \equiv T(\mathbb{F}_n)$ is an estimator of $\theta \equiv T(F)$ and that

$$\tau_n(T_n - \theta) \to_d Z, \tag{3}$$
$$\tau_n \to \infty, \tag{4}$$
$$m_n = o(n), \qquad \text{and} \quad \tau_{m_n}/\tau_n \to 0. \tag{5}$$

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

Theorem 4.7 (Politis and Romano without replacement bootstrap limit theorem). If (3), (4), and (5) hold, then

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

in probability as $m_n \wedge n \to \infty$.

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 (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

$$P(\tau_m(\hat{T}_m - \theta) \le z|\mathbb{F}_n) \to_p P(Z \le z)$$

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

$$P(\tau_m(\hat{T}_m - \theta) \le z|\mathbb{F}_n) \quad = \quad \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < \cdots < i_m \le n} 1\{\tau_m(T_m(X_{i_1}, \ldots, X_{i_m}) - \theta) \le z\} \equiv U_{n,m}$$

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

$$EU_{n,m} = P(\tau_m(T_m(X_{i_1}, \ldots, X_{i_m}) - \theta) \le z) \to P(Z \le z)$$

for all $z \in C(\mathcal{L}(Z))$ since $m \to \infty$. 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,

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

since $n/m = n/m_n \to \infty$. $\square$

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

Proposition 4.1 (Hoeffding, 1963). If $X_1, \ldots, X_n$ are independent and $a_i \le X_i \le b_i$, $i = 1, \ldots, n$, then with $\overline{\mu}_n \equiv E(\overline{X}_n)$,

$$P(\sqrt{n}(\overline{X}_n - \overline{\mu}_n) \ge t) \le \exp(-2t^2/\{n^{-1}\sum_1^n (b_i - a_i)^2\}) \qquad \text{for all} \quad t > 0;$$

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

$$\begin{aligned} P(\overline{X}_n - \overline{\mu}_n \ge t) &= P\left(\exp(rn(\overline{X}_n - \overline{\mu}_n)) \ge \exp(rnt)\right) \\ &\le \frac{E\exp(rn(\overline{X}_n - \overline{\mu}_n))}{e^{rnt}} \\ &= \frac{\prod_{i=1}^n Ee^{r(X_i - \mu_i)}}{e^{rnt}} \end{aligned}$$

where $\mu_i = EX_i$, $i = 1, \ldots, n$. But since e^{rx} is convex on $[a, b]$, on $[a, b]$ it lies below the line passing through (ae^{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

$$L(r_i) = -r_i p_i + \log(1 - p_i + p_i e^{r_i}).$$

Now

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

and

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

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

$$Ee^{r(X_i - \mu_i)} \le \exp(r^2(b_i - a_i)^2/8)$$

and

$$P(\overline{X}_n - \overline{\mu}_n \ge t) \le \exp(-nrt + r^2 \sum_1^n (b_i - a_i)^2/8) \equiv \exp(-g(r))$$

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

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

with

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

and the inequality (6) follows. □

Corollary 1 (Hoeffding, 1963). If $X_1, \ldots, X_n$ are i.i.d. and

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

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

$$(6) \qquad P(U_{n,m} - EU_{n,m} \ge t) \ \le \exp(-2[n/m]t^2/(b-a)^2) \qquad \text{for all} \quad t > 0.$$

Proof. Suppose that

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

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

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

If we can bound $E\exp(r(T_i - \mu_i - t))$ by something not depending on i (for example if $T_1, \ldots, T_N$ 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

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

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

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

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

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

by the proof of Hoeffding's inequality (6). Hence it follows from (a) that (6) holds. □

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

$$P\left(n^{-1/2}\Big|\sum_{i=1}^n c_i\epsilon_i\Big| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2(n^{-1}\sum_{i=1}^n c_i^2)}\right)$$

for all $t > 0$.

Some Limit Theory for Parametric Bootstrapping

It often holds that

$$(7) \qquad \sqrt{n}(\hat{\theta}_n - \theta) \to_d Y \qquad \text{as } n \to \infty,$$

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

$$(8) \qquad \sqrt{n}(P_{\hat{\theta}} - P_\theta) \to_d \dot{P}_\theta Y \qquad \text{as } n \to \infty$$

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

$$(9) \qquad \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \to_d Y^* \stackrel{d}{=} Y$$

and

$$(10) \qquad \sqrt{n}(P_{\hat{\theta}^*} - P_{\hat{\theta}}) \to_d \dot{P}_\theta Y^* \stackrel{d}{=} Y \qquad \text{as } n \to \infty.$$

The following result is a useful first step toward proving (9) and (10), especially if $\hat{\theta} = \theta(\mathbb{P}_n)$ so that $\hat{\theta}_n^* = \theta(\mathbb{P}_n^*)$. This type of theorem for a "model-based" or "parametric" bootstrap empirical process was also suggested by Giné and Zinn (1991).

Theorem 4.8 (Convergence of the "parametric bootstrap" empirical process). Suppose that $\mathcal{F}$ is $\mathcal{P}$ - measurable with envelope function F and that:

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

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

(iii) F is $\mathcal{P}-$uniformly square integrable.

Then, for P^∞ almost all sample sequence $X_1, X_2, \ldots,$

$$\mathbb{G}^*_{n,n} \equiv \sqrt{n}(\mathbb{P}_n - P_{\hat\theta_n}) \Rightarrow \mathbb{G}^*_0 \stackrel{d}{=} \mathbb{G}_{P_\theta} \qquad \text{in} \quad \ell^\infty(\mathcal{F})$$

as $n \to \infty$.

Proof. First note that (i) and (iii) imply that $\mathcal{F} \in \mathrm{AEC}_u(\mathcal{P}, \rho_P)$ and $(\mathcal{F}, \rho_P)$ is totally bounded uniformly in $P \in \mathcal{P}$ by Sheehy and Wellner (1991), theorem 2.2. Hence, in particular, $\mathcal{F} \in \mathrm{AEC}_u(\{\hat P_n\}, \rho)$, and $(\mathcal{F}, \rho_{P_\theta})$ is totally bounded. Furthermore, (iii) implies that for P^∞ a.e. ω the envelope F is $\{\hat P^\omega_n\}-$ uniformly square integrable. Thus, for $P^\infty-$ a.e. ω, the hypotheses of Sheehy and Wellner (1991), theorem 3.1, are satisfied by $\mathcal{F}$ for the sequence $\{\hat P^\omega_n\} \equiv \{P_{\hat\theta_n(\omega)}\}$. 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 nonparametric example mentioned briefly above:

Example 4.4 (Bootstrapping from a "smoothed empirical measure"; or, the "smoothed bootstrap"). Suppose that

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

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

$$\hat{\mathbb{P}}_n(C) = \int 1_C(x)\hat p_n(x)dx$$

where

$$\hat p_n(x) = \frac{1}{b_n^d}\int k\left(\frac{y-x}{b_n}\right) d\mathbb{P}_n(y)$$

where $k : \mathbb{R}^d \to \mathbb{R}$ is a uniformly continuous density function. It follows that $\hat{\mathbb{P}}_n \in \mathcal{P}$, 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 $\mathcal{F}$ is all indicators of a subclass of Borel set $\mathcal{C}$, the supremum in (ii) is bounded by the total variation distance between $\hat{\mathbb{P}}_n$ and P, which, in turn, is well-known to equal half the L_1-distanced between the respective densities (see e.g. xxx ???). Hence

$$\|\hat{\mathbb{P}}_n - P\|^*_{\mathcal{G}} \le \|\hat{\mathbb{P}}_n - P\|_{\mathcal{B}^b} = \frac{1}{2}\int |\hat p_n(x) - p(x)|dx \to_{a.s.} 0,$$

so (ii) holds. Since (iii) holds trivially (with $F \equiv 1$), theorem 3.14 shows that "the bootstrap from $\hat{\mathbb{P}}_n$ works:" i.e. for P^∞ almost all sample sequences $X_1, X_2, \ldots, \mathbb{G}^*_{n,n} \Rightarrow \mathbb{G}^*_0 \sim \mathbb{G}_P$ in $\ell^\infty(\mathcal{F})$.

For more genera classes $\mathcal{F}$, the results of Yukich (1989) or van der Vaart (1994) 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 $\alpha(\mathbb{P}_n)$ for functionals $\alpha : \mathcal{P} \to \mathbb{R}$; 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 hypotheses 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 5.1 Suppose that $T : \mathcal{F} \rightarrow \mathbb{R}$ is Hadamard - differentiable at F with respect to $\|\cdot\|_\infty$ tangentially to the subspace $C_u(\mathbb{R}, \rho_P)$ of uniformly continuous functions (with respect to the pseudo-metric ρ_F defined by $\rho^2(s,t) = Var_F(1_{(-\infty,s]}(X) - 1_{(-\infty,t]}(X))$. Let ψ_F denote the influence function of T at $F \in \mathcal{F}$. Then

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

and furthermore

$$\sqrt{n}(T(\mathbb{F}_n^*) - T(\mathbb{F}_n)) \rightarrow_d N(0, E\psi_F^2(X)) \qquad \text{in probability.}$$

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 $\{\mathbb{U}_n\}$ and $\{\mathbb{U}_n^*\}$ satisfying

$$\|\mathbb{U}_n - \mathbb{U}\|_\infty \rightarrow_{a.s.} 0, \qquad \text{and} \qquad \|\mathbb{U}_n^* - \mathbb{U}^*\|_\infty \rightarrow_{a.s.} 0.$$

Thus with $\tilde{\mathbb{F}}_n \equiv \mathbb{G}_n(F)$,

$$\sqrt{n}(\mathbb{F}_n - F) \stackrel{d}{=} \sqrt{n}(\tilde{\mathbb{F}}_n - F) = \mathbb{U}_n(F) \rightarrow_{a.s.} \mathbb{U}(F) \quad \text{in} \quad (D(\overline{\mathbb{R}}, \|\cdot\|_\infty),$$

and, with $\tilde{\mathbb{F}}_n^* \equiv \mathbb{G}_n^*(\tilde{\mathbb{F}}_n)$,

$$\sqrt{n}(\mathbb{F}_n^* - \mathbb{F}_n) \stackrel{d}{=} \sqrt{n}(\tilde{\mathbb{F}}_n^* - \tilde{\mathbb{F}}_n) = \mathbb{U}_n^*(\tilde{\mathbb{F}}_n) \rightarrow_{a.s.} \mathbb{U}^*(F) \quad \text{in} \quad (D(\overline{\mathbb{R}}, \|\cdot\|_\infty).$$

Now write

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

and

$$\mathbb{F}_n = F + n^{-1/2} n^{1/2}(\mathbb{F}_n - F) \stackrel{d}{=} F + n^{-1/2}\mathbb{U}_n(F).$$

Thus we may write

$$\begin{aligned} \sqrt{n}(T(\mathbb{F}_n^*) - T(\mathbb{F}_n)) \quad &\stackrel{d}{=} \quad \sqrt{n}(T(F + n^{-1/2}\{\mathbb{U}_n^*(\tilde{\mathbb{F}}_n) + \mathbb{U}_n(F)\}) - T(F)) \\ &\qquad - \sqrt{n}(T(F + n^{-1/2}\mathbb{U}_n(F)) - T(F)) \\ &\rightarrow_{a.s.} \quad \dot{T}(F; \mathbb{U}^*(F) + \mathbb{U}(F)) - \dot{T}(F; \mathbb{U}(F)) \\ &= \quad \dot{T}(F; \mathbb{U}^*(F)) \sim N(0, E_F\psi_F^2(X)) \end{aligned}$$

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

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

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

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

and this implies that with

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

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

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

But this just means that

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

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

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

□

For further results of the type, see Gill (1989), Arcones and Giné (1992), and van der Vaart and Wellner (1996), chapter 3.9.

Example 5.1 Suppose F is a d.f. 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 $X_1, \ldots, X_n$ are i.i.d. random variables with d.f. F, let $M_n \equiv \mathbb{F}_n^{-1}(1/2)$ 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

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

for every $x \in \mathbb{R}$.

The natural "bootstrap estimate" of the d.f. $H_n(x, F)$ is simply $H_n(x, \mathbb{F}_n)$ where $\mathbb{F}_n$ is the empirical d.f. 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 (1996)), section 3.9), that

$$H_n(x, \mathbb{F}_n) \rightarrow_p Pr\left(N(0, \frac{1/4}{f^2(m(F))}) \le x\right) \qquad \text{for all} \quad x \in \mathbb{R}$$

as $n \rightarrow \infty$. 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 5.2 (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 $X_1, X_2, \ldots$

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

as $n \to \infty$. In view of Polya's lemma, (2) can be re-expressed as

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

Proof. Represent the i.i.d. X_i's as $F^{-1}(\xi_i)$ where $\xi_1, \xi_2, \ldots$ are i.i.d. $U(0,1)$ random variables. Let $\mathbb{F}_n = \mathbb{G}_n(F)$ denote the empirical d.f. of the X_i's, so that the empirical process of the X_i's is

$$(a) \qquad \sqrt{n}(\mathbb{F}_n - F) = \mathbb{U}_n(F).$$

Then represent the bootstrap sample $X_{n1}^*, \ldots, X_{nn}^*$ as $X_{ni}^* \equiv \mathbb{F}_n^{-1}(\xi_i^*)$ where $\xi_1^*, \xi_2^*, \ldots$ is another sequence of independent $U(0,1)$ random variables independent of the ξ_i's. Thus the empirical d.f. of the X_{ni}^*'s is $\mathbb{F}_n^* = \mathbb{G}_n^*(\mathbb{F}_n)$, and the bootstrap empirical process is

$$(b) \qquad \sqrt{n}(\mathbb{F}_n^* - \mathbb{F}_n) = \mathbb{U}_n^*(\mathbb{F}_n).$$

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

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

where

$$\begin{aligned}
D_n &\equiv \sqrt{n}(\mathbb{F}_n(m(\mathbb{F}_n) + n^{-1/2}x) - 1/2) \\
&= \sqrt{n}(\mathbb{F}_n(m(\mathbb{F}_n) + n^{-1/2}x) - F(m(\mathbb{F}_n) + n^{-1/2}x)) \\
&\quad + \sqrt{n}(F(m(\mathbb{F}_n) + n^{-1/2}x) - F(m(\mathbb{F}_n))) \\
&\quad + \sqrt{n}(F(m(\mathbb{F}_n)) - \mathbb{F}_n(m(\mathbb{F}_n))) + o(n^{-1/2}) \\
(d) \qquad &= \mathbb{U}_n(F(m(\mathbb{F}_n) + n^{-1/2}x)) - \mathbb{U}_n(F(m(\mathbb{F}_n))) \\
&\quad + \sqrt{n}[F(m(\mathbb{F}_n) + n^{-1/2}x) - F(m(\mathbb{F}_n))] \\
(e) \qquad &\equiv \mathbb{U}_n(b_n) - \mathbb{U}_n(a_n) + \sqrt{n}(b_n) - a_n).
\end{aligned}$$

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

$$\begin{aligned}
(f) \qquad \sqrt{n}(b_n - a_n) &= \sqrt{n}[F(m(\mathbb{F}_n) + n^{-1/2}x) - F(m(\mathbb{F}_n))] \\
(g) \qquad &\to_{a.s.} x f(m(F)) \equiv c > 0 \qquad \text{as} \quad n \to \infty.
\end{aligned}$$

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

$$\omega_n((c+1)n^{-1/2}) \equiv \sup_{s,t:\ |t-s| \le (c+1)n^{-1/2}} |\mathbb{U}_n(t) - \mathbb{U}_n(s)| \to_{a.s.} 0 \qquad \text{as} \quad n \to \infty$$

by well known properties of the oscillation modulus $\omega_n(a)$ of the empirical process $\mathbb{U}_n$; see e.g. Shorack and Wellner (1986), page 542. Hence

$$(h) \qquad D_n \to_{a.s.} x f(m(F)).$$

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

(i) $\quad \|\mathbb{U}_n^* - \mathbb{U}^*\|_\infty \to_{a.s.} 0.$

Using (h) and (i) with the last line of (c) yields (or, instead of (i), use a similar argument as above involving the oscillation of $\mathbb{U}_n^*$!)

$$\begin{aligned} H_n(s, \mathbb{F}_n) &\to_{a.s.} Pr(\mathbb{U}^*(1/2) > -xf(m(F))) \qquad \text{as} \ \ n \to \infty \\ &= Pr(N(0, \frac{1/4}{f^2(m(F))}) \le x). \end{aligned}$$

□

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