Bootstrap Methods for dependent data

Jedrzej Sarna

November 7, 2023

1 Introduction

In this paper, we describe commonly used bootstrap methods that have been proposed in the past. Firstly we begin with a brief description of Efron's (1979) bootstrap method based on simple random sampling of the data, which forms the basis for almost all other bootstrap methods. Later we explore the example of Singh (1981), which points out the limitation of this resampling scheme for dependent variables. In the end we will get to know a bootstrap method for time-series models driven by iid variables, such as the autoregression model.

2 IID Bootstrap

We begin with the formulation of the IID bootstrap method of Efron (1979). For the discussion in this section, we assume that $X_1, X_2, \ldots$ is a sequence of iid random variables with common distribution F. Suppose, $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ generate the data at hand and let $T_n = t_n(\mathcal{X}_n; F)$, $n \geq 1$ be a random variable of interest. Note that T_n depends on the data as well as on the underlying unknown distribution F. Typical examples of T_n include:

- the normalized sample mean $T_n = \sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma}$
- the studentized sample mean $T_n = \sqrt{n}\frac{(\bar{X}_n - \mu)}{s_n}$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \quad \mu = E(X_1), \quad \sigma^2 = Var(X_1).$

Let G_n denote the sampling distribution of T_n. The goal is to find an accurate approximation to the unknown distribution of T_n or to some population characteristics, e.g., the standard error of T_n. The bootstrap method of Efron (1979) provides an effective way of addressing these problems without any model assumptions on F.

Given $\mathcal{X}_n$, we draw a simple random sample $\mathcal{X}_m^* = \{X_1^*, \ldots, X_m^*\}$ of size m with replacement from $\mathcal{X}_n$. Thus, conditional on $\mathcal{X}_n$, $\{X_1^*, \ldots, X_m^*\}$ are iid random variables with

$$P_*(X_1^* = X_i) = \frac{1}{n}, \;\; 1 \leq i \leq n$$

where P_* denotes the conditional probability given $\mathcal{X}_n$. Hence, the common distribution of X_i^*'s is given by the empirical distribution

$$F_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i},$$

where δ_{X_i} denotes the probability measure putting unit mass at X_i. Usually, one chooses the resample size $m = n$. However, there are several known examples where a different choice of m is desirable.

Next define the bootstrap version $T^*_{m,n}$ of T_n by replacing $\mathcal{X}_n$ with $\mathcal{X}^*_m$ and F with F_n as

$$T^*_{m,n} = t_m(\mathcal{X}^*_m; F_n)$$

Also, let $\hat{G}_{m,n}$ denote the conditional distribution of $T^*_{m,n}$ given $\mathcal{X}_n$. Then the bootstrap principle advocates $\hat{G}_{m,n}$ as an estimator of the unknown sampling distribution G_n of T_n. If, instead of G_n, one is interested in estimating only a certain functional $\varphi(G_n)$ of the sampling distribution of T_n, then the corresponding bootstrap estimator is given by *plugging-in* $\hat{G}_{m,n}$ for G_n , i.e., the bootstrap estimator of $\varphi(G_n)$ is given by $\varphi(\hat{G}_{m,n})$. For example, if

$$\varphi(G_n) = Var(T_n) = \int x^2\, dG_n(x) - (\int x\, dG_n(x))^2,$$

the bootstrap estimator of $Var(T_n)$ is given by

$$\varphi(\hat{G}_{m,n}) = Var(T^*_{m,n}|\mathcal{X}_n) = \int x^2\, d\hat{G}_{m,n}(x) - (\int x\, d\hat{G}_{m,n}(x))^2.$$

Once the variables X_n have been observed, the common distribution F_n of X^*_i's becomes known, and, hence, it is possible (at least theoretically) to find the conditional distribution $\hat{G}_{m,n}$ and the bootstrap estimator $\varphi(\hat{G}_{m,n})$ from the knowledge of the data. In practice, however, finding $\hat{G}_{m,n}$ exactly may be a daunting task. This is because the number of possible distinct values of $\mathcal{X}^*_m$ grows very rapidly, at the rate $O(n^m)$ as $n \to \infty$, $m \to \infty$ under the IID bootstrap. Consequently, the conditional distribution of $T^*_{m,n}$ is further approximated by Monte-Carlo simulations.

To illustrate the main ideas, let's consider an example where T_n is the centered and scaled sample mean $T_n = \sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma}$. Here $\mu = EX_1$ is the parameter we want to infer about. Following the description given above, the bootstrap version $T^*_{m,n}$ of T_n based on a bootstrap sample of size m is given by

$$T^*_{m,n} = \sqrt{m}\frac{(\bar{X}^*_m - E_*X^*_1)}{\sqrt{Var_*(X^*_1)}},$$

where $\bar{X}^*_m = \frac{1}{m}\sum_{i=1}^m X^*_i$ denotes the bootstrap sample mean based on $X^*_1, \ldots, X^*_m$, and E_* and Var_* respectively denote the conditional expectation and conditional variance, given $\mathcal{X}_n$. For any $k \geq 1$

$$E_*(X^*_1)^k = \int x^k\, dF_n(x) = \frac{1}{n}\sum_{i=1}^n X^k_i.$$

In particular, this implies $E_*(X^*_1) = \bar{X}_n$ and $Var_*(X^*_1) = s^2_n = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Hence we define $T^*_{m,n}$ by replacing $\bar{X}_n$ with $\bar{X}^*_m$ and μ and σ^2 by $E_*(X^*_1)$ and $Var_*(X^*_1)$ respectively.

Thus, the bootstrap version of T_n is given by

$$T^*_{m,n} = \sqrt{m}\frac{(\bar{X}^*_m - \bar{X}_n)}{s_n}.$$

If, for example, we are interested in estimating $\varphi_\alpha(G_n) =$ the αth quantile of T_n for some $\alpha \in (0, 1)$, then the bootstrap estimator of $\varphi_\alpha(G_n)$ is $\varphi_\alpha(\hat{G}_{m,n})$, the αth quantile of the conditional distribution of $T^*_{m,n}$.

As mentioned above, determining $\hat{G}_{m,n}$ exactly is not very easy even in this simple case. However, when $EX^2_1 < \infty$, and $m = n$, we have the following result.

Theorem 1. *If $X_1, X_2, \ldots$ are iid with $\sigma^2 = Var(X_1) \in (0, \infty)$, then*

$$\sup_x |P_*(T^*_{n,n} \leq x) - \Phi(x/\sigma)| = o(1) \ \ as \ \ n \to \infty \ \ a.s.$$

Proof: Since $X^*_1, \ldots, X^*_n$ are iid, by the Berry-Esseen Theorem

$$\sup_x |P_*(T^*_{n,n} \leq x) - \Phi(x)| \leq 2.75\hat{\Delta}_n$$

where $\hat{\Delta}_n = \frac{E_*|X^*_1 - \bar{X}_n|^3}{s^3_n \sqrt{n}}$ and $s^2_n = E_*(X^*_1 - \bar{X}_n)^2$.

Theorem (Berry-Esseen Theorem)**.** *Let $X_1, \ldots, X_n$ be a collection of $n \in \mathbb{N}$ independent but not necessarily identically distributed random variables with $EX_j = 0$ and $E|X_j|^3 < \infty$ for $1 \leq j \leq n$. If $\sigma^2 = \frac{1}{n}\sum_{j=1}^n EX^2_j > 0$, then*

$$\sup_x |P(\frac{1}{\sqrt{n}\sigma_n}\sum_{j=1}^n X_j \leq x) - \Phi(x)| \leq 2.75\frac{1}{n^{3/2}}\sum_{j=1}^n (\frac{E|X_j|^3}{\sigma^3_n}),$$

where $\Phi(x)$ denotes the distribution function of the standard normal distribution on $\mathbb{R}$.

Coming back to our proof, by the Strong Law of Large Numbers (SLLN)

$$s^2_n = \frac{1}{n}\sum_{i=1}^n X^2_i - (\bar{X}_n)^2 \to \sigma^2 \ \ a.s.$$

Theorem (SLLN)**.** *Let $\{X_n\}_{n\geq 1}$ be a sequence of iid random variables with $E|X_1| < \infty$. Then,*

$$\frac{1}{n}\sum_{i=1}^n X_i \to EX_1 \ \ as \ \ n \to \infty \ \ a.s.$$

and by the Marcinkiewicz-Zygmund SLLN

$$\frac{1}{n^{3/2}}\sum_{i=1}^n |X_i|^3 \to 0 \ \ a.s.$$

Theorem (Marcinkiewicz-Zygmund SLLN)**.** *Let $\{X_n\}_{n\geq 1}$ be a sequence of iid random variables and $p \in (0, \infty)$. If $E|X_1|^p < \infty$, then*

$$\frac{1}{n^{1/p}}\sum_{i=1}^n (X_i - c) \to 0 \ \ as \ \ n \to \infty \ \ a.s.$$

for any $c \in \mathbb{R}$ if $p \in (0, \infty)$ and for $c = EX_1$ if $p \in [1, \infty)$.

Hence, $\hat{\Delta}_n \to 0$ a.s. as $n \to \infty$, and Theorem 1 follows.

□

Note that by the Central Limit Theorem (CLT), T_n also converges in distribution to the $N(0,1)$ distribution. Hence, it follows that

$$\tilde{\Delta}_n \equiv \sup_x |\hat{G}_{n,n}(x) - G_n(x)| = \sup_x |P_*(T^*_{n,n} \leq x) - P(T_n \leq x)| = o(1) \;\; as \;\; n \to \infty \;\; a.s.$$

i.e., the conditional distribution $\hat{G}_{n,n}$ of $T^*_{n,n}$ generated by the IID bootstrap method provides a valid approximation for the sampling distribution G_n of T_n. Under some additional conditions

$$\tilde{\Delta}_n = O(n^{-1}(loglogn)^{1/2}) \;\; as \;\; n \to \infty \;\; a.s.$$

Therefore, the bootstrap approximation for $P(T_n \leq \cdot)$ is far more accurate than the classical normal approximation, which has an error of order $O(n^{-\frac{1}{2}})$.

Here we have described Efron's (1979) bootstrap for iid data mainly as a prelude to the bootstrap methods for dependent data, as the basic principles in both cases are the same. Furthermore, it provides a historical account of the developments that culminated in formulation of the bootstrap methods for dependent data.

3 Inadequacy of IID Bootstrap for Dependent Data

The lID bootstrap method of Efron (1979), being very simple and general, has found application to a lot of statistical problems. However, the general perception that the bootstrap is method giving accurate results in all problems automatically is misleading. A prime example of this appears in the seminal paper by Singh (1981), which pointed out its inadequacy for dependent data, which we are going to consider now.

Suppose $X_1, X_2, \ldots$ is a sequence of m-dependent random variables with $EX_1 = \mu$ and $EX_1^2 < \infty$. Sequence $\{X_n\}_{n\geq 1}$ is called m-dependent for some integer $m \geq 0$ if $\{X_1, \ldots, X_k\}$ and $\{X_{k+m+1}, \ldots\}$ are independent for all $k \geq 1$. Thus, an iid sequence of random variables $\{\epsilon_n\}_{n\geq 1}$ is 0-dependent and if we define using this iid sequence $\{\epsilon_n\}_{n\geq 1}$:

$$X_n = \epsilon_n + 0.5\epsilon_{n+1}, \; n \geq 1,$$

then $\{X_n\}_{n\geq 1}$ is 1-dependent.

Next, let $\sigma^2_m = Var(X_1) + \sum_{i=1}^{m-1} Cov(X_1, X_{1+i})$ and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. If $\sigma^2_m \in (0, \infty)$, then by the CLT for m-dependent variables

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2_m)$$

Theorem (CLT for m-dependent Sequences). *Let $\{X_i\}_{i\in\mathbb{Z}}$ be a sequence of stationary m-dependent random variables for some integer $m \geq 0$. If $EX_1^2 < \infty$ and*
$\sigma^2_m = Var(X_1) + \sum_{k=1}^{m-1} Cov(X_1, X_{1+k}) > 0$ then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - EX_1) \xrightarrow{d} N(0, \sigma^2_m)$$

Now, suppose that we want to estimate the sampling distribution of the random variable

$$T_n = \sqrt{n}(\bar{X}_n - \mu)$$

using the IID bootstrap. For simplicity, assume that the resample size equals the sample size, i.e., from $\mathcal{X}_n = (X_1, \ldots, X_n)$, an equal number of bootstrap variables $(X_1^*, \ldots, X_n^*)$ are generated. Then, the bootstrap version $T_{n,n}^*$ of T_n is given by

$$T_{n,n}^* = \sqrt{n}(\bar{X}_n^* - \bar{X}_n),$$

where $\bar{X}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^*$. The conditional distribution of $T_{n,n}^*$ under the IID bootstrap method still converges to a normal distribution, but with a "wrong" variance, as shown below.

Theorem 2. *Suppose $\{X_n\}_{n\geq 1}$ is a sequence of stationary m-dependent random variables with $EX_1 = \mu$, and $\sigma^2 = Var(X_1) \in (0, \infty)$. Then*

$$\sup_x |P_*(T_{n,n}^* \leq x) - \Phi(x/\sigma)| = o(1) \;\; as \;\; n \to \infty \;\; a.s.$$

Proof: Note that conditional on $\mathcal{X}_n$, $X_1^*, \ldots, X_n^*$ are iid random variables. As in the proof of Theorem 1, by the Berry-Esseen Theorem, it is enough to show that

$$s_n^2 \to \sigma^2 \;\; as \;\; n \to \infty \;\; a.s.$$

and

$$\frac{1}{n^{3/2}} \sum_{i=1}^n |X_i|^3 \to 0 \;\; as \;\; n \to \infty \;\; a.s.$$

These follow easily from the following lemma:

Lemma: Let $\{X_n\}_{n\geq 1}$ be a sequence of stationary m-dependent random variables. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a Borel measurable function with $E|f(X_1)|^p < \infty$ for some $p \in (0, \infty)$, and that $Ef(X_1) = 0$ if $p \geq 1$. Then

$$\frac{1}{n^{1/p}} \sum_{i=1}^n f(X_i) \to 0 \;\; as \;\; n \to \infty \;\; a.s.$$

Hence Theorem 2 is proved.

□

Corollary: Under the conditions of Theorem 2, if $\sum_{i=1}^m Cov(X_1, X_{1+i}) \neq 0$ and $\sigma_\infty^2 \neq 0$, then for any $x \neq 0$,

$$\lim_{n\to\infty} [P_*(T_{n,n}^* \leq x) - P(T_n \leq x)] = [\Phi(x/\sigma) - \Phi(x/\sigma_\infty)] \neq 0 \;\; a.s.$$

Proof: Follows from Theorem 2 and CLT for m-dependent variables.

□

Thus, for all $x \neq 0$ the IID bootstrap estimator $P_*(T^*_{n,n} \leq x)$ of parameter $P(T_n \leq x)$ has a mean squared error that tends to a nonzero number in the limit and the bootstrap estimator of $P(T_n \leq x)$ is not consistent. Therefore, the IID bootstrap method fails drastically for dependent data. Resampling individual X_i's from the data $\mathcal{X}_n$ ignores the dependence structure of the sequence $\{X_n\}_{n\geq 1}$ completely, and thus, fails to account for the lag-covariance terms $Cov(X_1, X_{1+i}), 1 \leq i \leq m)$ in the asymptotic variance.

Following this result, there have been several attempts to extend the IID bootstrap method to the dependent case. Now we will look at one of extensions of this method to certain dependent models generated by iid random variables.

4 Bootstrap Based on IID Innovations

Suppose $\{X_n\}_{n\geq 1}$ is a sequence of random variables satisfying the equation

$$X_n = h(X_{n-1}, \ldots, X_{n-p}; \beta) + \epsilon_n, \quad n > p, \tag{1}$$

where β is a $q \times 1$ vector of parameters, $h : \mathbb{R}^{p+q} \to \mathbb{R}$ is a known Borel measurable function, and $\{\epsilon_n\}_{n>p}$ is a sequence of iid random variables with common distribution F that are independent of the random variables $X_1, \ldots, X_p$. For identifiability of the model (1), assume that $E\epsilon_1 = 0$. A commonly used example of this model is the autoregressive process of order p (AR(p)) which we will discuss later. Noting that the process $\{X_n\}_{n\geq 1}$ is driven by the innovations ϵ_i's that are iid, the IID bootstrap method can be easily extended to the dependent model (1).

As before, suppose that $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ denotes the sample and that we want to approximate the sampling distribution of a random variable $T_n = t_n(\mathcal{X}_n; F, \beta)$. Let $\hat{\beta}_n$ be an estimator, e.g., the least squares estimator, of β based on $\mathcal{X}_n$. Define the residuals

$$\hat{\epsilon}_i = X_i - h(X_{i-1}, \ldots, X_{i-p}; \hat{\beta}_n), \quad p < i \leq n.$$

Note that, in general

$$\bar{\epsilon}_n \equiv \frac{1}{n-p} \sum_{i=1}^{n-p} \hat{\epsilon}_{i+p} \neq 0.$$

Hence, we center the residuals $\hat{\epsilon}_i$'s and define the "centered" residuals

$$\tilde{\epsilon}_i = \hat{\epsilon}_i - \bar{\epsilon}_n, \quad p < i \leq n.$$

Without such a centering, the resulting bootstrap approximation often has a random bias that does not vanish in the limit and makes the approximation useless.

Next draw a simple random sample $\epsilon^*_{p+1}, \ldots, \epsilon^*_m$ of size $(m-p)$ from $\{\tilde{\epsilon}_i : p < i \leq n\}$ with replacement and define the bootstrap pseudo-observations, using the model structure (1), as:

$$X^*_i = X_i \quad for \;\; i = 1, \ldots, p \;\; and$$

$$X^*_i = h(X^*_{i-1}, \ldots, X^*_{i-p}; \hat{\beta}_n) + \epsilon^*_i, \quad p < i \leq m.$$

Note that by construction $\epsilon^*_i, \;\; p < i \leq m$ are iid and $E_*\epsilon^*_1 = 0$. The bootstrap version of the random variable $T_n = t_n(\mathcal{X}_n; F, \beta)$ is defined as

$$T^*_{m,n} = t_m(\mathcal{X}^*_m; F_n, \hat{\beta}_n),$$

where $\mathcal{X}^*_m = \{X^*_1, \ldots, X^*_m\}$ and F_n denotes the empirical distribution of the centered residuals $\tilde{\epsilon}_i, p < i \leq n$. The sampling distribution of T_n is approximated by the conditional distribution of $T^*_{m,n}$ given

$\mathcal{X}_n$. For certain time-series models satisfying (1), different versions of this resampling scheme have been proposed. A special case of model (1) is the autoregression model of order p (AR(p)), given by

$$X_n = \beta_1 X_{n-1} + \ldots + \beta_p X_{n-p} + \epsilon_n, \quad n > p, \tag{2}$$

where $\beta = (\beta_1, \ldots, \beta_p)$ is the vector of autoregressive parameters, and $\{\epsilon_n\}_{n>p}$ is an iid sequence satisfying the requirements of model (1).

For AR(p)-models, validity and the rate of approximation of the IID-Innovation bootstrap have been well-studied in the literature.

- When the sequence $\{X_n\}_{n\geq 1}$ is stationary, Bose (1988) shows that under suitable regularity conditions, a version of the IID-innovation bootstrap approximation to the sampling distribution of the standardized least square estimator is more accurate than the normal approximation.
- For nonstationary cases, performance of this method has been studied and it has been shown that the IID-innovation bootstrap method is very sensitive to the values of the autoregression parameter vector β. Indeed, if the value of β is such that the roots of the characteristic equation $z^p + \beta_1 z^{p-1} + \ldots + \beta_p = 0$ lie on the unit circle, then the IID-innovation bootstrap fails. Because of its dependence on the validity of the model, and drastic change in the performance with a small change in the parameter value, one needs to be particularly careful when applying the IID-innovation bootstrap method.