

Lecture Note 7: Bootstrap and Subsampling

Bootstrap of Sample Mean - take two

Here is an alternative, hopefully somewhat more useful explanation of the nonparametric bootstrap of the sample mean.

As before, let $X_n = (x_1, \dots, x_n)$ where $x_i \stackrel{\text{iid}}{\sim} F$, and let

$$\theta(F) = E[x_i] = \int x dF(x).$$

Assume

$$0 < \sigma^2(F) = V[x_i] < \infty.$$

Root:

$$R_n(X_n, \theta(F)) = \sqrt{n}(\bar{x}_n - \theta(F)),$$

where $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$.

Let \hat{F}_n denote the empirical CDF. Now,

$$\theta(\hat{F}_n) = \int x d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}_n,$$

$$\sigma^2(\hat{F}_n) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2.$$

Reminder: definition of almost sure convergence: let $y_1(\omega), y_2(\omega), \dots$, and $y(\omega)$ be random variables defined on a probability space (Ω, P) . Then

$$y_n \xrightarrow{as} y$$

means that

$$P(\{\omega \in \Omega : y_n(\omega) \rightarrow y(\omega)\}) = 1.$$

Now, by the Strong Law of Large Numbers,

$$\theta(\hat{F}_n) \xrightarrow{as} \theta(F).$$

Formally, we define $x_1(\omega), x_2(\omega), \dots$ as random variables on a common probability space (Ω, F) . So X_n , the vector of the first n observations, is also a function $X_n(\omega)$. In turn, \hat{F}_n is a function of X_n , so it is also a function of ω . To make this more clear, we can write

$$\hat{F}_n = \hat{F}_n(\cdot | X_n(\omega)).$$

So $\theta(\hat{F}_n) \xrightarrow{as} \theta(F)$ means that

$$P(\{\omega \in \Omega : \theta(\hat{F}_n(\cdot|X_n(\omega))) \rightarrow \theta(F)\}) = 1.$$

Similarly, by the SLLN,

$$\sigma^2(\hat{F}_n) \xrightarrow{as} \sigma^2(F).$$

Also, a result called the Glivenko-Cantelli Theorem says that

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{as} 0.$$

Now, let $\omega \in \Omega$ satisfy

$$\begin{aligned} \theta(\hat{F}_n(\cdot|X_n(\omega))) &\rightarrow \theta(F), \\ \sigma^2(\hat{F}_n(\cdot|X_n(\omega))) &\rightarrow \sigma^2(F), \\ \sup_x |\hat{F}_n(x|X_n(\omega)) - F(x)| &\rightarrow 0 \end{aligned}$$

(This holds for almost all ω , by the a.s. convergence results.) We will continue to hold ω fixed, so that $\hat{F}_n(\cdot|X_n(\omega))$ is just a deterministic sequence.

Now, consider the bootstrap. We draw x_1^b, \dots, x_n^b IID from \hat{F}_n , and form

$$\sqrt{n}(\theta(\hat{F}_n^b) - \theta(\hat{F}_n)).$$

By doing this for many bootstrap samples, we essentially can determine the distribution of this root under \hat{F}_n . Let $J_n(\hat{F}_n)$ denote this distribution. We want this distribution to be a good approximation to the limiting distribution of $\sqrt{n}(\theta(\hat{F}_n) - \theta(F))$, which we know is $N(0, \sigma^2(F))$.

Array CLT idea: suppose

$$\begin{aligned} x_{11} &\sim F_1, \\ x_{21}, x_{22} &\stackrel{\text{iid}}{\sim} F_2, \\ x_{31}, x_{32}, x_{33} &\stackrel{\text{iid}}{\sim} F_3, \end{aligned}$$

and so on. Also, suppose that $F_n \rightsquigarrow F$, $\theta(F_n) \rightarrow \theta(F)$, $\sigma^2(F_n) \rightarrow \sigma^2(F)$. Then, under some conditions,

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n x_{ni} - \theta(F_n) \right) \xrightarrow{d} N(0, \sigma^2(F)).$$

So this implies that the distribution of

$$\sqrt{n}(\theta(\hat{F}_n^b) - \theta(\hat{F}_n))$$

converges weakly to $N(0, \sigma^2(F))$, which is what we wanted to show. As mentioned above, this holds

for almost all ω . So the bootstrap distribution converges to the right distribution with probability one.

Having shown this, it can then be shown that the quantiles of the bootstrap distribution converges to the appropriate quantiles of the normal distribution with probability one, and that the bootstrap confidence interval has asymptotically correct coverage with probability one.

Bootstrap of Mean-like Statistics

Now suppose the parameter of interest has the form

$$\theta(F) = \int \psi(x) dF(x).$$

Examples:

$$\theta(F) = E[x_i^2] = \int x^2 dF(x),$$

$$\theta(F) = E[\log(x_i)] = \int \log(x) dF(x).$$

Then

$$\theta(\hat{F}_n) = \frac{1}{n} \sum_{i=1}^n \psi(x_i).$$

Let

$$\tau(F) = V[\psi(x_i)] = \int \psi^2(x) dF(x) - \theta^2(F).$$

Assume $0 < \tau(F) < \infty$. We can define $y_i = \psi(x_i)$. So if $x_1, x_2, \dots \stackrel{\text{iid}}{\sim} F$, then $y_1, y_2, \dots \stackrel{\text{iid}}{\sim} F_Y$, where F_Y is the distribution $\psi(x_i)$. Then the exact same argument we used above applies here, and we conclude that the bootstrap is asymptotically valid.

Nonlinear Functionals

A lot of interesting estimators can be written as functionals¹ of \hat{F}_n , but not necessarily mean-like in the previous sense.

Example: suppose that $x_i \stackrel{\text{iid}}{\sim} \text{Exp}(1/\theta)$, the exponential distribution with mean $1/\theta$. Then, letting $F = \text{Exp}(1/\theta)$,

$$\theta(F) = \frac{1}{E_F[x_i]} = \frac{1}{\int x dF(x)}.$$

¹A functional is a function that takes as its input another function. Here, $\theta(F)$ is a function of a distribution function.

The MLE can be easily derived as

$$\hat{\theta} = \frac{1}{\bar{x}_n} = \frac{1}{\int x d\hat{F}_n(x)}.$$

This is a functional of \hat{F}_n , but not of the form $\theta(\hat{F}_n) = \int \psi(x) d\hat{F}_n(x)$.

More generally, the MLE is a nonlinear functional of the empirical CDF:

$$\begin{aligned} \theta(F) &= \arg \max_{\tilde{\theta} \in \Theta} E[\log f(x_i | \tilde{\theta})] \\ &= \arg \max_{\tilde{\theta} \in \Theta} \int \log f(x | \tilde{\theta}) dF(x); \end{aligned}$$

and

$$\begin{aligned} \hat{\theta}_{ML} &= \arg \max_{\tilde{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^n \log f(x_i | \tilde{\theta}) \\ &= \arg \max_{\tilde{\theta} \in \Theta} \int \log f(x | \tilde{\theta}) d\hat{F}_n(x) \\ &= \theta(\hat{F}_n). \end{aligned}$$

So, for the general nonlinear case, we want to see if the bootstrap provides a valid approximation to the distribution of the root

$$\sqrt{n}(\theta(\hat{F}_n) - \theta(F)).$$

As we have seen, the argument we have used requires a version of the central limit theorem for arrays. So if we can show something similar here, we will be fine.

First, consider the case where x_i has finite support $\{\xi_1, \xi_2, \dots, \xi_K\}$. Then, F can be summarized by K numbers

$$\begin{aligned} F_1 &= P(x_i \leq \xi_1) \\ F_2 &= P(x_i \leq \xi_2) \\ &\vdots \\ F_K &= P(x_i \leq \xi_K) \end{aligned}$$

Likewise, the empirical CDF can be summarized by

$$\begin{aligned} \hat{F}_{n1} &= \frac{1}{n} \sum_{i=1}^n 1(x_i \leq \xi_1) \\ &\vdots \\ \hat{F}_{nK} &= \frac{1}{n} \sum_{i=1}^n 1(x_i \leq \xi_K) \end{aligned}$$

By the multivariate CLT,

$$\sqrt{n} \left(\begin{pmatrix} \hat{F}_{n1} \\ \vdots \\ \hat{F}_{nK} \end{pmatrix} - \begin{pmatrix} F_1 \\ \vdots \\ F_K \end{pmatrix} \right) \xrightarrow{d} N(0, V),$$

where

$$V = \begin{pmatrix} (1 - F_1)F_1 & (1 - F_2)F_1 & (1 - F_3)F_1 & \cdots \\ & (1 - F_2)F_2 & (1 - F_3)F_2 & \cdots \\ & & \ddots & \\ & & & \ddots \end{pmatrix}.$$

So, if $\theta(F)$ is a function of F (now considered as a K -vector), with derivative matrix D , then by the delta method,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, DVD').$$

Then it turns out that our previous argument can be used with slight modification.

Now, what about the case where x_i has continuous support? Now $F(\cdot)$ and $\hat{F}_n(\cdot)$ are functions over the real line, and $\theta(F)$ is a functional.

Basically, we need a functional version of differentiability of $\theta(F)$. There are different notions of functional differentiability. One that works for us is:

Fréchet differentiability at F : $\theta(F)$ is Fréchet differentiable if there exists a function $g_F(x)$ with $\int g_F(x)dF(x) = 0$ and

$$\theta(G) = \theta(F) + \int g_F(x)dG(x) + o(\sup_x |G(x) - F(x)|).$$

This says that for G close to F , $\theta(G) - \theta(F)$ is well approximated by a mean-like function. In particular, for the empirical CDF,

$$\theta(\hat{F}_n) - \theta(F) \approx \int g_F(x)d\hat{F}_n(x).$$

Then, if the variance of $\theta(\hat{F}_n)$ is finite, it turns out that asymptotic normality holds, and under similar assumptions to the previous cases, the bootstrap is generally valid.

When does the bootstrap fail?

Primarily, when the smoothness or bounded variance assumptions do not hold. One example is for the sample mean, when the variance of x_i is infinite.

Another example: Parametric Model of a Procurement Auction

Suppose that there is a public work project being auctioned off by a government agency. There are M bidders, each of which has a private cost c_m drawn IID from an exponential distribution with mean γ . So the density of cost c_m is

$$f_c(c_m|\gamma) = \frac{1}{\gamma} \exp(-c/\gamma).$$

We want to estimate γ , because knowing γ would help the agency design future auctions to maximize expected revenue.

Suppose that the auction is a first-price sealed bid auction (so the lowest bid wins the contract). There is a symmetric Nash equilibrium, in which bidders bid

$$b = c + \frac{\gamma}{M-1}.$$

So the observed bids have a shifted exponential density

$$f_b(b_m|\gamma) = \frac{1}{\gamma} \exp\left(-\frac{1}{\gamma} \left(b - \frac{\gamma}{M-1}\right)\right) \cdot 1\left(b \geq \frac{\gamma}{M-1}\right).$$

In practice, we might have observations from a series of such auctions, and pool them together to estimate γ .

The problem here is that the density has support which depends on the unknown parameter γ . This introduces a nonsmoothness in how F relates to γ . It turns out that the MLE is not asymptotically normal, and that the bootstrap is not valid here.

Subsampling

Now, we will let the estimator $\hat{\theta}_n$ be a general function of the data (not necessarily a function only of the empirical CDF). We will assume that the root has the form

$$\tau_n(\hat{\theta}_n - \theta(F)),$$

where τ_n is a sequence increasing with n . Usually, we have

$$\tau_n = \sqrt{n},$$

although in some cases (for example the procurement auction example) it turns out we need a different choice

$$\tau_n = n.$$

Assume that the root converges in distribution to some nondegenerate distribution with CDF $J(t)$.

Let

$$J_n(t, F) = P\left(\tau_n(\hat{\theta}_n - \theta(F)) \leq t\right)$$

be the CDF of the root. So $J_n(t, F) \rightsquigarrow J(t)$.

Subsampling Algorithm:

Let $b < n$. There are $S(n) = \binom{n}{b}$ subsets of size b from the original sample.

Let $Y_1, Y_2, \dots, Y_{S(n)}$ denote the different possible subsamples.

For each $j = 1, \dots, S(n)$, let $\hat{\theta}_{n,j}$ be the estimate based on subsample Y_j .

Then, let the subsampling approximation to the distribution of the root be

$$\hat{J}_n(t) = \frac{1}{S(n)} \sum_{j=1}^{S(n)} 1(\tau_b(\hat{\theta}_{n,j} - \hat{\theta}_n) \leq t).$$

Theorem: Suppose that $n \rightarrow \infty$, $b \rightarrow \infty$, $b/n \rightarrow 0$, and $\tau_b/\tau_n \rightarrow 0$. Then

$$\hat{J}_n(t) \xrightarrow{p} J(t)$$

at each continuity point of J . Further, confidence intervals formed using $\hat{J}_n(t)$ will have asymptotically correct coverage.

(See Politis, Romano, and Wolf for the proof.)

What's remarkable is that the conditions for subsampling are so weak. Essentially, all that is required is that the root have some limiting distribution, and that the subsample size b is not too large (but still going to infinity).

So the subsampling algorithm can be used in many cases where the bootstrap fails. However, if the bootstrap is valid, then the bootstrap approximation is often better.