8 Bootstrap

8.1 Basic Idea

The main principle of the bootstrap is that our data can be used to approximate the population.

1. We can observe a sample $X = (X_1, ..., X_n)$ with distribution function F and the statistical model $\mathcal{P} = \{\mathrm{P}_\theta;\ \theta \in \Theta\}$.

2. We are interested in the parameter $\theta = \theta(F)$, the distribution properties of a given estimator $\hat{\theta} = \hat{\theta}(X)$, and the distribution properties of a given function $T(X, F)$.

3. Estimate F or P_θ by some estimator $\hat{F}_n$ or P^*.

4. Draw independently B random samples $x^*_{(j)}$ of size n from P^*, $j = 1, ..., B$, that is $X^*_{(j)} \sim P^*$. The samples $x^*_{(j)}$ are called the bootstrap samples.

5. For each j, estimate θ using the bootstrap sample $x^*_{(j)}$ as $\hat{\theta}^*_j = \hat{\theta}\left(x^*_{(j)}\right)$. The bootstrap replications of $\hat{\theta}$ are $\hat{\theta}^*_1$, ..., $\hat{\theta}^*_B$. The bootstrap replications of the function $T(X, F)$ are $T^*_{(j)} = T\left(X^*_{(j)}, \hat{F}_n\right)$, $j = 1, ..., B$.

6. Pretend that the bootstrap replications are an iid sample from the distribution of $T(X, F)$.

Example 1. Suppose that we have observed iid data x_1, ..., x_n from a statistical model $X \sim P$ with distribution function F. Our focus is the median of this distribution, denoted by $m(F)$. We use the sample median $m_{(n)}$ to estimate the population median $m(F)$. The "bias" can be computed by $T(\boldsymbol{X}, \boldsymbol{F}) = m_{(n)} - m(F)$. In bootstrap, the first steps are the same as classical inference:

1. Our data is $\boldsymbol{x} = (x_1, ..., x_n)$ from $\boldsymbol{X} = (X_1, ..., X_n)$.

2. The parameter is $m(F)$. We are interested in $T = m_{(n)} - m(F)$.

3. If we choose the nonparametric approach, we estimate $\boldsymbol{F}$ by some estimator $\hat{\boldsymbol{F}}_n$ using $\boldsymbol{x}$ as

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

   The estimated distribution P^* is $\mathrm{P}(X = x_i) = n^{-1}$ for any i. The estimator of $m(F)$ is $m_{(n)} = m\left(\hat{F}_n\right)$.

4. Daw independently B bootstrap samples $\boldsymbol{x}^*_{(j)} = \left(x^*_{1j}, ..., x^*_{nj}\right)$ of size n from P^*, $j = 1, ..., B$.

5. For each j, estimate $m\left(\hat{F}_n\right)$ using the bootstrap sample $\boldsymbol{x}^*_{(j)}$ as $m_{(n),j}$, where $m_{(n),j}$ is the sample median of $x^*_{1j}, ..., x^*_{nj}$. Compute the difference

$$T_j = m_{(n),j} - m_{(n)},$$

   since $m_{(n)}$ is the median of the distribution P^*.

6. View $m_{(n),j} - m_{(n)}$, $j = 1, ..., B$, as an iid sample from the distribution of $m_{(n)} - m(F)$. For example, the bias of the estimator $m_{(n)}$ can be approximated by

$$\frac{1}{B}\sum_{j=1}^{B} m_{(n),j} - m_{(n)}.$$

Parametric and nonparametric approaches can make some difference.

1. The parametric bootstrap assumes a parametric form of the statistical model P index by some parameter θ.
   - For example, we assume $X \sim N\left(\mu, \sigma^2\right)$. The bootstrap samples are drawn from $N\left(\hat{\mu}, \hat{\sigma}^2\right)$.
2. The nonparametric bootstrap estimates F by the empirical distribution function.
   - The estimated distribution is $\mathrm{P}\left(X = x_i\right) = n^{-1}$ for any i.

Definition 1 (ecdf)**.** Let x_1, ..., x_n be a sequence of iid measurements, then the fraction

$$\hat{F}_n\left(x\right) = \frac{1}{n}\sum_{i=1}^{n} 1\left(x_i \leq x\right)$$

is called the empirical cumulative distribution function (ecdf) or the empirical distribution function.

8.2 Why Does Bootstrap Work?

By Law of large numbers, $\hat{F}_n\left(x\right) \xrightarrow{P} F\left(x\right)$, where $F\left(x\right)$ is the true cdf, for every fixed x.

Theorem 1 (Glivenko-Cantelli Theorem for Unidimensional Case)**.** *Let Z_1, ..., Z_n be iid real valued random variables with distribution function $F\left(z\right) = P\left(Z \leq z\right)$. Denote the empirical distribution function by*

$$F_n\left(z\right) = \frac{1}{n}\sum_{i=1}^{n} 1\left(Z_i \leq z\right).$$

Then,

$$P\left\{\sup_{z\in\mathbb{R}} \left|F_n\left(z\right) - F\left(z\right)\right| > \epsilon\right\} \leq 8\left(n+1\right)\exp\left\{-\frac{1}{32}n\epsilon^2\right\}.$$

In particular, by the Borel-Cantelli lemma,

$$\lim_{n\to\infty}\sup_{z\in\mathbb{R}}\left|F_n\left(z\right) - F\left(z\right)\right| = 0, \text{ with probability 1.}$$

In fact, we can improve the result to

$$\mathrm{P}\left\{\sup_{z\in\mathbb{R}}\left|F\left(z\right) - F_n\left(z\right)\right| > \epsilon\right\} \leq 2\exp\left\{-2n\epsilon^2\right\}.$$

Recall that bootstrap aims to approximate the exact distribution of $T\left(X, F\right)$ using the distribution of $T\left(X^*, \hat{F}_n\right)$, if we take nonparametric bootstrap as an example.

Definition 2. The bootstrap is consistent if

$$\sup_t\left|\mathrm{P}\left(T\left(X, F\right) \leq t\right) - \mathrm{P}\left(T\left(X^*, \hat{F}_n\right) \leq t \mid \hat{F}_n\right)\right| \xrightarrow{P} 0,$$

where the second probability $\mathrm{P}\left(T\left(X^*, \hat{F}_n\right) \leq t \mid \hat{F}_n\right)$ is still a random variable.

By the Glivenko-Cantelli theorem, we know that

$$\mathrm{P}\left(\sup_x\left|\hat{F}_n\left(x\right) - F\left(x\right)\right| > \epsilon\right) \to 0.$$

Hence for sufficiently large n, sampling from F is similar to sampling from $\hat{F}_n$. More assumptions are needed in order for bootstrap to work.

Proposition 1. *Suppose that*

$$\begin{aligned} P\left(T\left(X,F\right)\le x\right) &\to F_A\left(x\right), \\ P\left(T\left(X^*,\hat{F}_n\right)\le x \mid \hat{F}_n\right) &\xrightarrow{P} F_A\left(x\right), \end{aligned}$$

where $F_A(x)$ is a continuous function. Then bootstrap is consistent.

Example 2. Suppose that X_1, ..., X_n is an iid $(\mu, 1)$, where $\mathrm{E}[X] = \mu$ and $\mathrm{Var}[X] = 1$, but not necessarily normal. We are interested in $T\left(\bar{X}, \mu\right) = \sqrt{n}\left(\bar{X} - \mu\right)$.

1. By CLT, $T\left(\bar{X}, \mu\right) = \sqrt{n}\left(\bar{X} - \mu\right) \xrightarrow{d} N(0,1)$. Hence,

$$\mathrm{P}\left(\sqrt{n}\left(\bar{X}-\mu\right)\le x\right) \to \Phi\left(x\right),$$

where $\Phi(x)$ is a continuous function in x.

2. In the bootstrap world, we simulate $(X_1^*, ..., X_n^*)$ from $\hat{F}_n$. The conditional mean and variance is

$$\begin{aligned} \mathrm{E}\left[X_i^* \mid \hat{F}_n\right] &= \bar{X}, \\ \mathrm{Var}\left[X_i^* \mid \hat{F}_n\right] &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2. \end{aligned}$$

By Strong Law of Large Numbers, $\bar{X} \to \mu$ and $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \to 1$ for almost every sequence X_1, X_2, ... (almost surely convergence). To obtain the asymptotic distribution of $\bar{X}^*$, we need CLT for a triangular array, since we don't have identical distribution anymore when we change n. The Lindeberg-Fuller CLT says that the sequence $\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n}\left(Y_{n,i} - \mathrm{E}\left[Y_{n,i}\right]\right) \xrightarrow{d} N\left(0, \sigma^2\right)$ under some assumptions. Using this CLT, conditional on $\hat{F}_n$, we have

$$\sqrt{n}\left(\bar{X}^* - \bar{X}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(X_i^* - \bar{X}\right) \xrightarrow{d} N\left(0,1\right).$$

Equivalently,

$$\mathrm{P}\left(\sqrt{n}\left(\bar{X}^* - \bar{X}\right)\le x \mid \hat{F}_n\right) \to \Phi\left(x\right),$$

conditionally on X_1, X_2, ... The convergence holds for almost every sequence X_1, X_2, ... , since $\bar{X} \to \mu$ and $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \to 1$ for almost every sequence X_1, X_2, ... (almost surely convergence). This suggests that

$$\mathrm{P}\left(\sqrt{n}\left(\bar{X}^* - \bar{X}\right)\le x \mid \hat{F}_n\right) \xrightarrow{P} \Phi\left(x\right).$$

The extra convergence in probability is needed because the LHS probability depends on $\hat{F}_n$.

Example 3 (Bootstrap can fail)**.** Suppose that X_1, ..., X_n is an iid uniform $[0, \theta]$, we can show that $\sqrt{n}\left(X_{[n]} - \theta\right)$ does not converge to a normal distribution. In fact,

$$T = \frac{n\left(\theta - X_{[n]}\right)}{\theta} \xrightarrow{d} \mathrm{Exp}\left(1\right).$$

If nonparametric bootstrap works, the distribution of $T^* = n\left(x_{[n]} - X_{[n]}^*\right)/x_{[n]}$ should converge to distribution of T. But

$$\mathrm{P}\left(T^* = 0\right) \to 1 - e^{-1},$$

because

$$\begin{aligned}
\mathrm{P}\left(\max_i X_i^* = x_{[n]} \mid x_1, ..., x_n\right) &= 1 - \underbrace{\left[P\left(X_i^* < x_{[n]} \mid x_1, ..., x_n\right)\right]^n}_{\text{all bootstrap sample } < x_{[n]}} \\
&= 1 - \left(1 - \frac{1}{n}\right)^n \to 1 - e^{-1} > 0.
\end{aligned}$$