The plug-in principle of the bootstrap

François Portier

October 12, 2016

The main goal of the bootstrap is to measure the accuracy of an estimate $\hat{\theta}$ by reproducing/approximating the distribution of $n^{1/2}(\hat{\theta} - \theta_0)$.

1 The framework

The original bootstrap applies to estimators based on identically distributed and independent sequences of random variables. Let P be a probability distribution on $\mathbb{R}^d$ and define

$$\theta_0 = \theta(P),$$

where θ is a map valued in $\mathbb{R}$. The parameter θ_0 stands here for the parameter of interest. It could be for instance the theoretical mean of an unknown distribution or the regression vector in a regression model. Examples of such map θ are given below. Suppose that $X_1, X_2 \ldots$ are identically distributed and independent random variables with common distribution P. Define the empirical measure

$$\hat{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}.$$

In this context, a natural estimator $\hat{\theta}$ of θ_0 is given by

$$\hat{\theta} = \theta(\hat{P}_n).$$

This kind of estimators are usually called "plug-in" estimators as the estimated distribution $\hat{P}_n$ has been "plugged" in the expression in place of the theoretical distribution P.

Example 1. *The expectation of g with respect to P is given by $\theta(P) = \int g(x)\,\mathrm{d}P(x)$. The mean corresponds to $g(x) = x$.*

Example 2. *The variance corresponds to*

$$\theta(P) = \int x^2\,\mathrm{d}P(x) - \left(\int x\,\mathrm{d}P(x)\right)^2.$$

Example 3. *The covariance between X and Y is given by*

$$\theta(P) = \int xy\,\mathrm{d}P(x,y) - \left(\int x\,\mathrm{d}P_1(x)\right)\left(\int y\,\mathrm{d}P_2(y)\right),$$

where P_1 and P_2 stands for the marginals of X and Y.

Example 4. *The correlation coefficient between X and Y is given by*

$$\theta(P) = \frac{\int xy\,\mathrm{d}P(x,y) - \left(\int x\,\mathrm{d}P_1(x)\right)\left(\int y\,\mathrm{d}P_2(y)\right)}{\sqrt{\left(\int x^2\,\mathrm{d}P_1(x) - \left(\int x\,\mathrm{d}P_1(x)\right)^2\right)\left(\int y^2\,\mathrm{d}P_2(y) - \left(\int y\,\mathrm{d}P_2(y)\right)^2\right)}}.$$

Example 5. *The regression coefficient β defined through the model*

$$Y = \beta^T X + \epsilon,$$

where $\mathbb{E}[\epsilon \mid X] = 0$ and $X \in \mathbb{R}^p$, $p \geq 1$, is given by

$$\theta(P) = \left(\int x x^T \mathrm{d}P_1(x) \right)^{-1} \int xy \mathrm{d}P(x, y)$$

Example 6. *The distribution function F evaluated at y is given by*

$$\theta(P) = \int \mathbb{1}_{\{x \leq y\}} \mathrm{d}P(x).$$

Example 7. *The median is given by*

$$\theta(P) = F^{-}(1/2).$$

where F^{-} is the inverse of F.

The goal of the following exercise is to show that most of the common statistical estimators are in fact plug-in estimator.

Exercise 1. *For every examples, express the associated plug-in estimator.*

In the following we shall use the notion of root defined as follows. This notion is crucial because it provides a central quantity of which the behaviour should be reproducible (by bootstrap, see bellow) and that contains all the information needed to assess the accuracy of $\hat{\theta}$ estimating θ_0.

Definition 1. *A statistical root is a function of $X_1, X_2, \ldots X_n$ that converges in distribution to a tight distribution.*

In regular cases (on which we shall focus here), it can be shown, based on classical asymptotic theory (central limit theorem, Slutsky's lemma, Delta-method,...) that

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma^2),$$

where Σ is a positive definite matrix. Consequently, the rescaled sequence $n^{1/2}(\hat{\theta} - \theta_0)$ is a statistical root. Note that indeed the rescaled sequence $n^{1/2}(\hat{\theta} - \theta_0)$ is appropriate to describe the accuracy of $\hat{\theta}$ estimating θ_0.

Exercise 2. *Show the root property in examples 3 and 5.*

2 The plug-in principle

The bootstrap approximation is based on the following "plug-in" principle. As described before, we consider the estimator $\hat{\theta} = \theta(\hat{P}_n)$ of $\theta_0 = \theta(P)$, where $\hat{P}_n$ is based on $X_1, \ldots X_n$ with common law P. To reproduce the same context one could simply replace the distribution P by $\hat{P}_n$. Doing this we define

$$\hat{\theta}^* = \theta(\hat{P}_n^*),$$

where $\hat{P}_n^*$ is based on $X_{n,1}^*, \ldots X_{n,1}^*$, identically distributed and independent random variables with common law $\hat{P}_n$, i.e.,

$$\hat{P}_n^* = n^{-1} \sum_{i=1}^{n} \delta_{X_{n,1}^*}.$$

Real world		Bootstrap world
$X_1, \dots X_n \sim P$ (**unknown**)		$X_1^*, \dots X_n^* \sim \hat{P}_n$ (**known**)
$\downarrow$		$\downarrow$
$\hat{\theta}$		$\hat{\theta}^*$
$\downarrow$		$\downarrow$
$n^{1/2}(\hat{\theta} - \theta_0)$	"$\simeq$"	$n^{1/2}(\hat{\theta}^* - \hat{\theta})$

Table 1: The bootstrap "plug-in" principle.

From this "plug-in" approach we expect that $n^{1/2}(\hat{\theta}^* - \hat{\theta})$ would "behave similarly" as $n^{1/2}(\hat{\theta} - \theta_0)$.

Moreover since $\hat{P}_n$ is known, computing versions $n^{1/2}(\hat{\theta}^* - \hat{\theta})$ can be done as many times as we wish. Consequently, the "behaviour" of $n^{1/2}(\hat{\theta}^* - \hat{\theta})$ can be completely determined. This permits to state the two main steps of the bootstrap:

(i) (**definition step**)* Verify that the bootstrap root $R^* = n^{1/2}(\hat{\theta}^* - \hat{\theta})$ that "mimics" the behaviour of the root of interest $n^{1/2}(\hat{\theta} - \theta_0)$.

(ii) (**simulation step**)** For some B, compute $\hat{R}_1^*, \dots, \hat{R}_B^*$ to approximate the law of $\hat{R}$.

We purposefully left unspecified what is behind the words "behaves similarly" and "mimics" as the definition of those concepts might bring confusion. The reader should understand this words as "has a very similar distribution". For instance, we have the following result for the empirical mean which actually can be extended to all the examples raised before and to many other estimators $\hat{\theta}$.

Theorem 1. *Let $X_1, X_2, \dots$ be independent and identically distributed random variables such that $\operatorname{var}(X_1) = \Sigma$. If $\theta(P) = \int x \, \mathrm{d}P(x)$, then the bootstrap root $n^{1/2}(\hat{\theta}^* - \hat{\theta})$ has the same law (asymptotically) as the root $n^{1/2}(\hat{\theta}^* - \hat{\theta})$.*

2.1 The bootstrap algorithm

The following algorithm is the original bootstrap algorithm as introduced by Efron. To make clear that it corresponds to the plug-in principle described previously, one should answer the following exercise.

Exercise 3. *Remark that X^* generated from a uniform draw among $\{X_1, \dots X_n\}$ has law $\hat{P}_n$.*

Algorithm.

Input : $X_1, \dots X_n$, number of bootstrap iterations B
Output: Bootstrap estimators $(\hat{\theta}_1^*, \dots, \hat{\theta}_B^*)$
for $b = 1, \dots, B$ **do**

Draw (uniformly) with replacement among $X_1, \dots, X_n$, and provide the new **random** sample:

$$\text{Bootstrap sample : } X_{n,1}^*, \dots, X_{n,n}^*$$

Apply the estimation over the bootstrap sample:

$$\hat{\theta}_b^* = \theta(\hat{P}_n^*)$$

end

The bootstrap estimators $\hat{\theta}_b^*$, $b = 1, \dots B$ are identically distributed and independent random variables.