Bootstrap resampling

Applied Statistics
Université Paris-Saclay

september 29th 2022

What is the bootstrap?

The bootstrap is a method for estimating standard errors and computing confidence intervals.

Bootstrap (B. Efron 1979)

"*To pull oneself up by one's own bootstraps*"

To estimate the variability of an estimator $\widehat{\theta}$, you need to observe many possible values and/or use probability calculations to approximate its distribution. But

- we only see ONE value of $\widehat{\theta}$ from ONE observed sample;
- in complex models, probability calculations are intractable.

$\hookrightarrow$ Idea: Replace theoretical calculations by Monte Carlo simulations

How do we sample from an *unknown* population?

we cannot do this exactly!

Observation: an i.i.d. sample $X_1, \ldots, X_n$
$\hookrightarrow$ we simulate *sampling from the population* by sampling from the sample
= resampling

- The resamples are drawn with replacement from $X_1, \ldots, X_n$,
- each resample has the same sample size n as the original sample.

(model-free resampling=basic bootstrap)

bootstrap samples

θ parameter, $\widehat{\theta} = g(X_1, \ldots, X_n)$: estimator from the sample.

- draw n observations with replacement from $X_1, \ldots, X_n$
  $\hookrightarrow X_1^*, \ldots, X_n^*$ bootstrap sample
- equivalently: $X_1^*, \ldots, X_n^* \sim F_n$, where F_n is the ECDF of $X_1, \ldots, X_n$

F_n is an estimator of the unknown CDF of $X_1, \ldots, X_n$: $X_1^*, \ldots, X_n^*$ is as similar as possible to the original sample.

Bootstrap sampling distribution

n^n different bootstrap samples $\hookrightarrow$ in practice, simulate B resamples

- In the real world, compute $\widehat{\theta} = g(X_1, \dots, X_n)$;
- In the bootstrap world, compute $\hat{\theta}^* = g(X_1^*, \dots, X_n^*)$;
- Repeat B times

$$\begin{array}{ccc}
(X_1^{*(1)}, \dots, X_n^{*(1)}) & \to & \hat{\theta}_n^{*(1)} \\
\dots & \dots & \dots \\
(X_1^{*(b)}, \dots, X_n^{*(b)}) & \to & \hat{\theta}_n^{*(b)} \\
\dots & \dots & \dots \\
(X_1^{*(B)}, \dots, X_n^{*(B)}) & \to & \hat{\theta}_n^{*(B)}
\end{array}$$

to get
$(\hat{\theta}_n^{*(1)}, \hat{\theta}_n^{*(2)}, \dots, \hat{\theta}_n^{*(B)})$, the bootstrap sampling distribution of $\widehat{\theta}$.

Bootstrap variance estimation

$$\widehat{V}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{b=B}\left(\hat{\theta}^{*(b)} - \frac{1}{B}\sum_{b=1}^{b=B}\hat{\theta}^{*(b)}\right)^2$$

- What are the bootstrap estimates of Bias, Standard error and MSE?

Validity of the bootstrap

We are using two approximations:

$$\begin{aligned}\text{Law } (\widehat{\theta}-\theta) &\approx \text{bootstrap law}(\widehat{\theta}^{*}-\widehat{\theta}) \\ &\approx \text{bootstrap sampling law } (\widehat{\theta}^{*1}-\widehat{\theta},\widehat{\theta}^{*2}-\widehat{\theta},\ldots,\widehat{\theta}^{*B}-\widehat{\theta})\end{aligned}$$

- The first approximation is from the bootstrap approximation of the population's law by the empirical distribution: it becomes small as n becomes large
- The second approximation is due to Monte Carlo error and can be made small by choosing B large (B=200 or B =1000)

Basic bootstrap intervals

The law of $\widehat{\theta} - \theta$ is approximated by the law of $\widehat{\theta}^* - \widehat{\theta}$.

bootstrap quantiles: let $b^*_{\alpha/2}$ and $b^*_{1-\alpha/2}$ be the quantiles of the sampling bootstrap law

$$(\widehat{\theta}^{*1} - \widehat{\theta}, \widehat{\theta}^{*2} - \widehat{\theta}, \ldots, \widehat{\theta}^{*B} - \widehat{\theta})$$

i.e. the fraction of bootstrap estimates that satisfy $b^*_{\alpha/2} \leq \widehat{\theta}^{*b} - \widehat{\theta} \leq b^*_{1-\alpha/2}$ is $1 - \alpha$.

$\hookrightarrow IC(\theta) = \left[\widehat{\theta} - b^*_{1-\alpha/2};\ \widehat{\theta} - b^*_{\alpha/2}\right]$ is a confidence interval for θ. Its approximate confidence level is $1 - \alpha$.

Why?

Bootstrap-t intervals

If a standard error for $\widehat{\theta}$ is available, let $\widehat{s} = \widehat{s}(X_1, \ldots, X_n))$ be the estimate of the s.e. of $\widehat{\theta}$ (from the Fisher information for example), then we compute the *t-statistic*

$$t = \frac{\widehat{\theta} - \theta}{\widehat{s}}$$

Then the bth bootstrap t-statistic is

$$t^{*b} = \frac{\widehat{\theta}^{*b} - \widehat{\theta}}{\widehat{s}^{*b}}$$

where $\widehat{s}^{*b} = \widehat{s}(X_1^{*(b)}, \ldots, X_n^{*(b)})$.
Let $qt^*_{\alpha/2}$ and $qt^*_{1-\alpha/2}$ be the $\alpha/2$ lower and upper quantiles of the sampling distribution of these bootstrap t-statistics, then the confidence interval for θ is

$$[\widehat{\theta} - qt^*_{1-\alpha/2}\widehat{s},\ \widehat{\theta} - qt^*_{\alpha/2}\widehat{s}]$$