Chapter 5 - Beyond Monte Carlo

Introduction to the Bootstrap Method.

Prof. Alex Alvarez, Ali Raisolsadat

School of Mathematical and Computational Sciences
University of Prince Edward Island

Introduction to Bootstrap Method

In Chapter 3 we studied how to use Monte Carlo methods to solve some statistical problems (bias estimation, standard error estimation, validation of confidence intervals).

By generating many Monte Carlo samples we can obtain important information about the probability distribution of some estimators that cannot be obtained using analytical methods.

In all of these studied cases we generate the Monte Carlo samples from some statistical model (that is, generating values from specific probability distributions).

What should we do to solve similar problems in the absence of a probability model?

Introduction to Bootstrap Method

In Chapter 5 (specifically in Section 5.2) we will study the use of **resampling methods** in Statistics.

The main point of resampling methods is being able to generate random samples in the absence of a probabilistic model. In particular we will study a resampling method called **Bootstrap**.

The Bootstrap method was introduced by Bradley Efron in 1979. In 2019 Efron was awarded the **Prize in Statistics** for his initial work on the Bootstrap method. Since 1980 the terms Bootstrap/Bootstrapping have been mentioned in more than 200,000 scientific documents.

Bootstrap Method Framework (Vector Notation)

Sample data: Let

$$\mathbf{X}_n = (X_1, X_2, \ldots, X_n)^\top \in \mathbb{R}^n \qquad X_i \overset{\text{i.i.d.}}{\sim} F$$

where the distribution F is unknown. We observe a realization

$$\mathbf{x}_n = (x_1, x_2, \ldots, x_n)^\top$$

General Problem: Estimation of some statistic $\theta = \theta(F)$ which is a functional of the unknown distribution F.

Examples:

$$\theta(F) = \int y \, dF(y) \quad \text{(mean)} \qquad \theta(F) = F^{-1}(1/2) \quad \text{(median)}$$

Estimator: Let $\widehat{\theta}_n = T(\mathbf{X}_n)$ where $T : \mathbb{R}^n \to \mathbb{R}$ is a statistic computed from the data vector. Is this estimator unbiased? What is the standard deviation? What is its distribution?

All these problems are difficult to solve if we do not have a statistical model to generate random samples of X.

Bootstrap Method Rationale (Vector Notation)

Main Idea: Generate new random samples by sampling (with replacement) from the sample data.

As we will be sampling from the sample data, the Bootstrap method is part of a larger set of **resampling methods**.

Bootstrap sample: For $b = 1, \ldots, B$,

$$\mathbf{X}_n^{*(b)} = (X_1^{*(b)}, \ldots, X_n^{*(b)})^\top$$

Example: Observed data

$$\mathbf{x}_n = (5.2,\ 3.1,\ 3.4,\ 4.7,\ 2.2)^\top$$

Three bootstrap samples:

$$\mathbf{x}_1^* = (3.1, 4.7, 2.2, 4.7, 5.2)^\top$$
$$\mathbf{x}_2^* = (2.2, 3.1, 5.2, 2.2, 2.2)^\top$$
$$\mathbf{x}_3^* = (2.2, 4.7, 3.4, 2.2, 3.4)^\top$$

R code (single bootstrap sample):

```
SampleData <- c(5.2, 3.1, 3.4, 4.7, 2.2)
sample(SampleData, size=5, replace=TRUE)
```

Bootstrap Method Rationale

These Bootstrap samples can be used in the solution of many problems. For instance, let us compute $\widehat{\theta}_n(x_j^*)$ for $j = 1, 2, ..., N$ (that is for each Bootstrap sample). Doing this will allow us to create many replications $\widehat{\theta}_n(x_j^*)$ of estimates of θ.

By looking at the whole set of values $\widehat{\theta}_n(x_j^*)$ we will have more information to assess the uncertainty in $\widehat{\theta}_n(x)$.

The term "Bootstrap" comes from the metaphor of *"pulling yourself up by your bootstraps"*, roughly meaning to succeed in doing something with own efforts (without outside help).

Theoretical Justification of the Method

From a more conceptual point of view, in the Bootstrap method the original sample has the same role that populations play in standard statistical analysis.

The empirical cumulative distribution function F_n defined as

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, t]}(x_i)$$

converges to the actual unknown c.d.f. distribution F as n approaches infinity. Moreover, the approximation $F_n \approx F$ may be valid even for relatively small values of n.

Example

Example: Compare the empirical cumulative distribution function obtained from a sample of 50 standard normally distributed random numbers, and the theoretical cumulative distribution function.

R Code

```
xseq = seq(-3,3,length=100)
x = rnorm(50)
plot(ecdf(x))
lines(xseq,pnorm(xseq))
```

Resampling from the observations $x = (x_1, x_2, \ldots, x_n)$ is equivalent as sampling from F_n. If F_n and F are close then sampling from F_n is roughly equivalent as sampling from F.

Theoretical Justification of the Method

Having established that the distributions F_n and F are close, we would hope that

$$\theta(F_n) \approx \theta(F)$$

This is not true in general (for all statistics θ) but theoretical results can be obtained on a case by case basis.

This means that if we want to study $\theta(F)$ we can do so by studying $\theta(F_n)$ using Bootstrap samples that follow the **known** distribution F_n. This part of the analysis is not much different than what we did in Chapter 3.

During the next few lectures we will apply these ideas to specific statistical problems.