Bootstrap

High-Dimensional Data Analysis and Machine Learning

Camille Mondon

Introduction

Let $X_1, \ldots, X_n \sim P_\theta$ i.i.d. Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be an estimator for θ.

One often wants to evaluate the **variance** $\operatorname{Var}[\hat{\theta}]$ to quantify the uncertainty of $\hat{\theta}$.

The bootstrap is a powerful, broadly applicable method:

- to estimate $\operatorname{Var}[\hat{\theta}]$
- to estimate $\mathbb{E}[\hat{\theta}] - \theta$ (**bias**)
- to construct confidence intervals for θ
- …

The method is nonparametric and can deal with small n.

A motivating example

Optimal portfolio

- Let Y and Z be the values of two random assets and consider the portfolio:

$$W_\lambda = \lambda Y + (1-\lambda) Z, \qquad \lambda \in [0,1]$$

  allocating a proportion λ of your wealth to Y and a proportion $1-\lambda$ to Z.
- A common, risk-averse, strategy is to minimize the risk $\operatorname{Var}[W_\lambda]$.
- It can be shown that this risk is minimized at

$$\lambda_{\text{opt}} = \frac{\operatorname{Var}[Z] - \operatorname{Cov}[Y,Z]}{\operatorname{Var}[Y] + \operatorname{Var}[Z] - 2\operatorname{Cov}[Y,Z]}$$

- But in practice, $\operatorname{Var}[Y]$, $\operatorname{Var}[Z]$ and $\operatorname{Cov}[Y,Z]$ are **unknown**.

Sample case

Now, if historical data $X_1 = (Y_1, Z_1), \ldots, X_n = (Y_n, Z_n)$ are available, then we can estimate λ_{opt} by

$$\hat{\lambda}_{\text{opt}} = \frac{\widehat{\operatorname{Var}[Y]} - \widehat{\operatorname{Cov}[Y,Z]}}{\widehat{\operatorname{Var}[Y]} + \widehat{\operatorname{Var}[Z]} - 2\widehat{\operatorname{Cov}[Y,Z]}}$$

where

- $\widehat{\operatorname{Var}[Y]}$ is the sample variance of the Y_i's
- $\widehat{\operatorname{Var}[Z]}$ is the sample variance of the Z_i's
- $\widehat{\operatorname{Cov}[Y,Z]}$ is the sample covariance of the Y_i's and Z_i's.

How to estimate the accuracy of $\hat{\lambda}_{\text{opt}}$?

- … i.e., its standard deviation $\text{Std}[\hat{\lambda}_{\text{opt}}]$)?
- On the basis of the available sample, we observe $\hat{\lambda}_{\text{opt}}$ **only once**.
- We need further samples leading to further observations of $\hat{\lambda}_{\text{opt}}$.

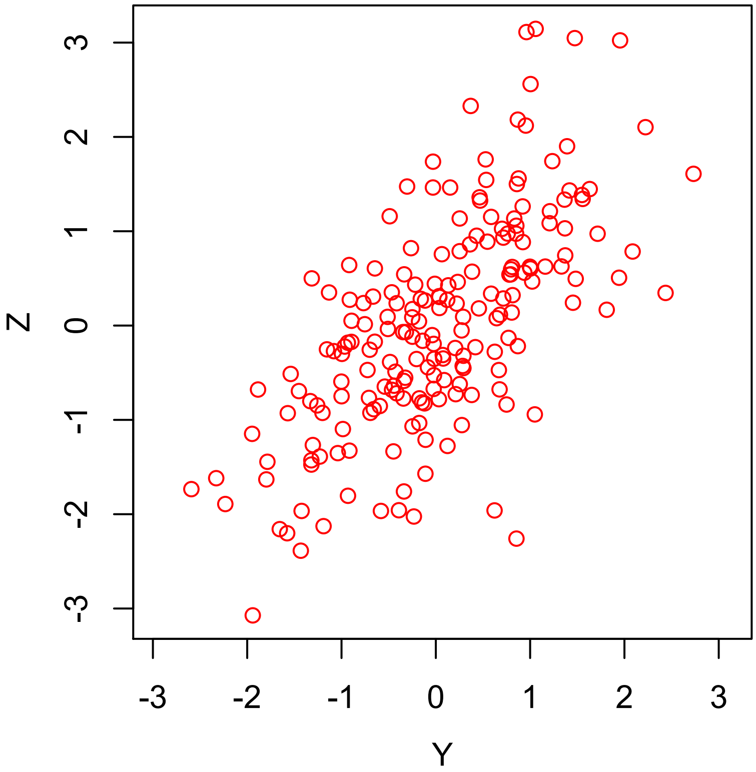

Figure 1: Portfolio data. For this sample, $\hat{\lambda}_{\text{opt}} = 0.283$.

Sampling from the population

We generated 1000 samples from the population. The first three are

- This allows us to compute: $\bar{\lambda}_{\text{opt}} = \frac{1}{1000}\sum_{i=1}^{1000} \hat{\lambda}_{\text{opt}}^{(i)}$
- Then: $\widehat{\text{Std}[\hat{\lambda}_{\text{opt}}]} = \sqrt{\frac{1}{999}\sum_{i=1}^{1000}(\hat{\lambda}_{\text{opt}}^{(i)} - \bar{\lambda}_{\text{opt}})^2}$.

Here:

$$\widehat{\text{Std}[\hat{\lambda}_{\text{opt}}]} \approx .077, \qquad \bar{\lambda}_{\text{opt}} \approx .331 \ \left(\approx \lambda_{\text{opt}} = \frac{1}{3} = .333\right)$$

and the distribution of $\hat{\lambda}_{\text{opt}}$ is described by

(This could also be used to estimate quantiles of $\hat{\lambda}_{\text{opt}}$.)

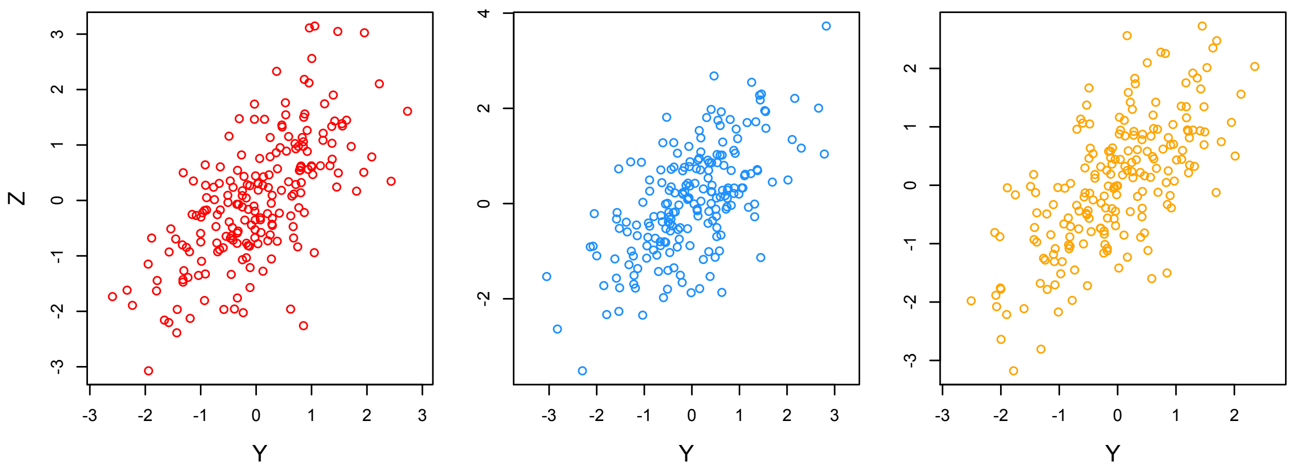

Figure 2: $\hat{\lambda}^{(1)}_{\text{opt}} = 0.283$, $\hat{\lambda}^{(2)}_{\text{opt}} = 0.357$, $\hat{\lambda}^{(3)}_{\text{opt}} = 0.299$.

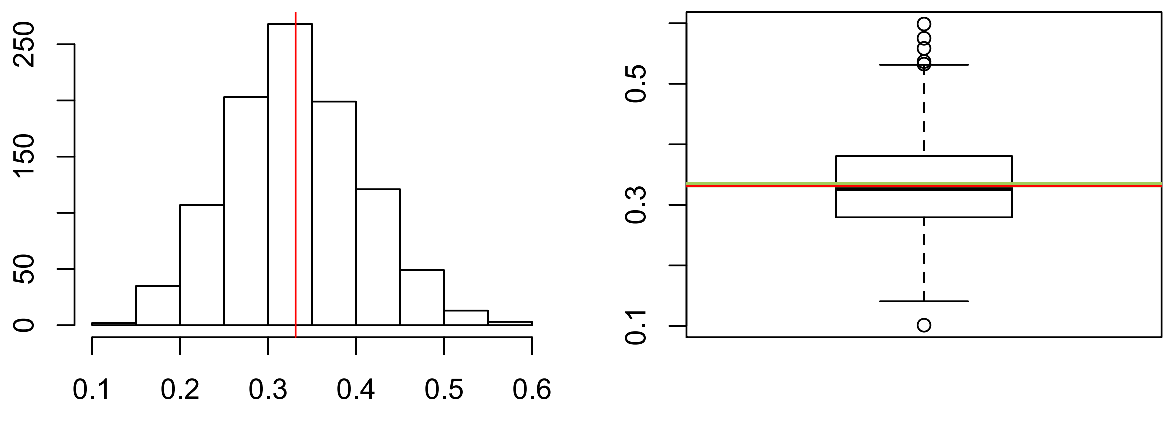

Figure 3: Histogram and boxplot of the empirical distribution of the $\hat{\lambda}^{(i)}_{\text{opt}}$.

Sampling from the sample: the bootstrap

- It is important to realize that **this cannot be done in practice**. One cannot sample from the population P_θ since it is **unknown**.
- However, one may sample instead from the empirical distribution P_n (i.e., the uniform distribution over $\{X_1,\ldots,X_n\}$, that is close to P_θ for large n.
- This means that we sample with replacement from $\{X_1,\ldots,X_n\}$, providing a first **bootstrap sample** $(X_1^{*1},\ldots,X_n^{*1})$ which allows us to evaluate $\hat\lambda^{*(1)}_{\mathrm{opt}}$.
- Further generating bootstrap samples $(X_1^{*b},\ldots,X_n^{*b})$, $b=2,\ldots,B=1000$, one can compute

$$\widehat{\mathrm{Std}[\hat\lambda_{\mathrm{opt}}]}^* = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat\lambda^{*(b)}_{\mathrm{opt}} - \bar\lambda^*_{\mathrm{opt}})^2}$$

with

$$\bar\lambda^*_{\mathrm{opt}} = \frac{1}{1000}\sum_{b=1}^{B}\hat\lambda^{*(b)}_{\mathrm{opt}}$$

Sampling from the sample: the bootstrap

This provides

$$\widehat{\mathrm{Std}[\hat\lambda_{\mathrm{opt}}]}^* \approx .079$$

and the distribution of $\hat\lambda_{\mathrm{opt}}$ is described by

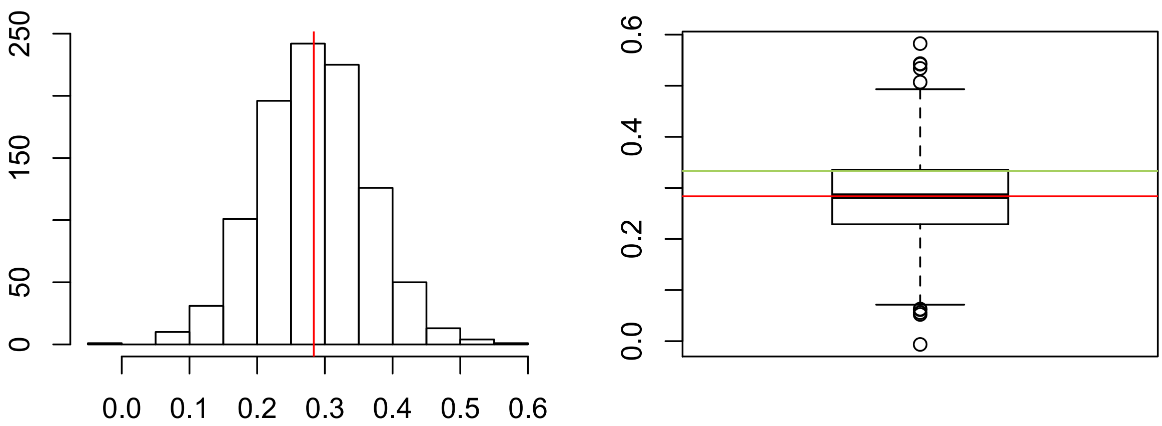

Figure 4: Histogram and boxplot of the bootstrap distribution of $\hat\lambda_{\mathrm{opt}}$.

(This could again be used to estimate quantiles of $\hat\lambda_{\mathrm{opt}}$.)

A comparison between both samplings

Results are close: $\widehat{\mathrm{Std}[\hat\lambda_{\mathrm{opt}}]} \approx 0.077$ and $\widehat{\mathrm{Std}[\hat\lambda_{\mathrm{opt}}]}^* \approx 0.079$.

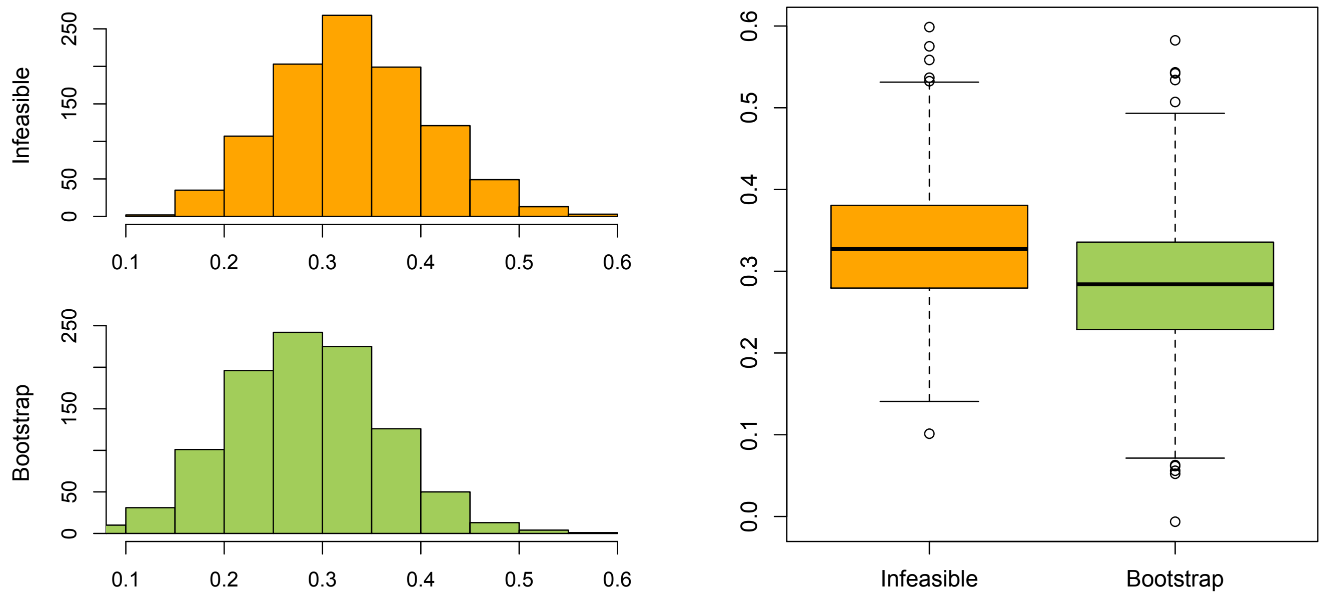

Figure 5: Bootstrap distributions from portfolio data.

The general procedure

The bootstrap

- Let $X_1, \ldots, X_n$ be i.i.d P_θ.
- Let $T = T(X_1, \ldots, X_n)$ be a statistic of interest.
- The bootstrap allows us to say something about the distribution of T:

$$\begin{array}{ccc} (X_1^{*1}, \ldots, X_n^{*1}) & \rightsquigarrow & T^{*1} = T(X_1^{*1}, \ldots, X_n^{*1}) \\ & \vdots & \\ (X_1^{*b}, \ldots, X_n^{*b}) & \rightsquigarrow & T^{*b} = T(X_1^{*b}, \ldots, X_n^{*b}) \\ & \vdots & \\ (X_1^{*B}, \ldots, X_n^{*B}) & \rightsquigarrow & T^{*B} = T(X_1^{*B}, \ldots, X_n^{*B}) \end{array}$$

- Under mild conditions, the empirical distribution of $T^{*1}, \ldots, T^{*B}$ provides a good approximation of the sampling distribution of T under P_θ.

The bootstrap

Above, each bootstrap sample $(X_1^{*b}, \ldots, X_n^{*b})$ is obtained by sampling (uniformly) with replacement among the original sample $(X_1, \ldots, X_n)$.

Possible uses:

- $\frac{1}{B-1}\sum_{b=1}^{B}(T^{*b} - \bar{T}^*)^2$, with $\bar{T}^* = \frac{1}{B}\sum_{b=1}^{B} T^{*b}$, estimates **Var[T]**
- The sample α-quantile q_α^* of $T^{*1}, \ldots, T^{*B}$ estimates **T's α-quantile**

Possible uses when T is an estimator of θ:

- $(\frac{1}{B}\sum_{b=1}^{B} T^{*b}) - T$ estimates **the bias** $\mathbb{E}[T] - \theta$ **of** T

- $[q^*_{\alpha/2}, q^*_{1-(\alpha/2)}]$ is an approximate $(1-\alpha)$**-confidence interval for** θ.
- …

About the implementation in R

A toy illustration

- Let $X_1, \dots, X_n$ $(n = 4)$ be i.i.d t-distributed with 6 degrees of freedom.
- Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean.
- How to estimate the variance of $\bar{X}$ through the bootstrap?

```
n <- 4
(X <- rt(n,df=6))
```

```
[1] -0.08058779  0.28044078  1.19011050 -1.25212790
```

```
Xbar <- mean(X)
Xbar
```

```
[1] 0.0344589
```

Obtaining a bootstrap sample

```
X
```

```
[1] -0.08058779  0.28044078  1.19011050 -1.25212790
```

```
d <- sample(1:n,n,replace=TRUE)
d
```

```
[1] 2 4 4 4
```

```
Xstar <- X[d]
Xstar
```

```
[1]  0.2804408 -1.2521279 -1.2521279 -1.2521279
```

Generating $B = 1000$ bootstrap means

```
B <- 1000
Bootmeans <- vector(length = B)
for (b in (1:B)) {
  d <- sample(1:n, n, replace = TRUE)
  Bootmeans[b] <- mean(X[d])
}
Bootmeans[1:4]
```

```
[1]  0.2370868 -0.3486833  0.3521335  0.2370868
```

Bootstrap estimates

Bootstrap estimates of $\mathbb{E}[\bar{X}]$ and $\text{Var}[\bar{X}]$ are then given by

```
mean(Bootmeans)
```

```
[1] 0.03679914
```

```
var(Bootmeans)
```

```
[1] 0.1789107
```

The practical sessions will explore how well such estimates behave.

The `boot` function

A better strategy is to use the `boot` function from

```
library(boot)
```

The `boot` function takes typically 3 arguments:

- `data`: the original sample
- `statistic`: a **user-defined function** with the statistic to bootstrap
  - **1st argument:** a generic sample
  - **2nd argument:** a vector of indices pointing to a subsample on which the statistic is to be evaluated…
- `R`: the number B of bootstrap samples to consider

If the statistic is the mean, then a suitable **user-defined function** is

```
boot.mean <- function(x,d) {
  mean(x[d])
}
```

The bootstrap estimate of $\text{Var}[\bar{X}]$ is then

```
res.boot <- boot(X,boot.mean,R=1000)
var(res.boot$t)
```

```
          [,1]
[1,] 0.1844024
```