High-Dimensional Data Analysis and Machine Learning

From a course by Davy Paindaveine and Nathalie Vialaneix

Camille Mondon

Last updated on October 1, 2024

Toulouse School of Economics

1. Bootstrap

Efron (1979)

1. Bootstrap

1.1 Introduction

1.1 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** $\mathrm{Var}[\hat{\theta}]$ to quantify the uncertainty of $\hat{\theta}$.

The bootstrap is a powerful, broadly applicable method:

- to estimate the **variance** $\mathrm{Var}[\hat{\theta}]$
- to estimate the **bias** $\mathbb{E}[\hat{\theta}] - \theta$
- to construct **confidence intervals** for θ
- more generally, to estimate the distribution of $\hat{\theta}$.

The method is **nonparametric** and can deal with small n.

Toulouse School of Economics

1. Bootstrap

1.2 A motivating example

James et al. (2021)

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** $\mathrm{Var}[W_\lambda]$.
- It can be shown that this risk is minimized at

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

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

Toulouse School of Economics

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

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

where

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

Toulouse School of Economics

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

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

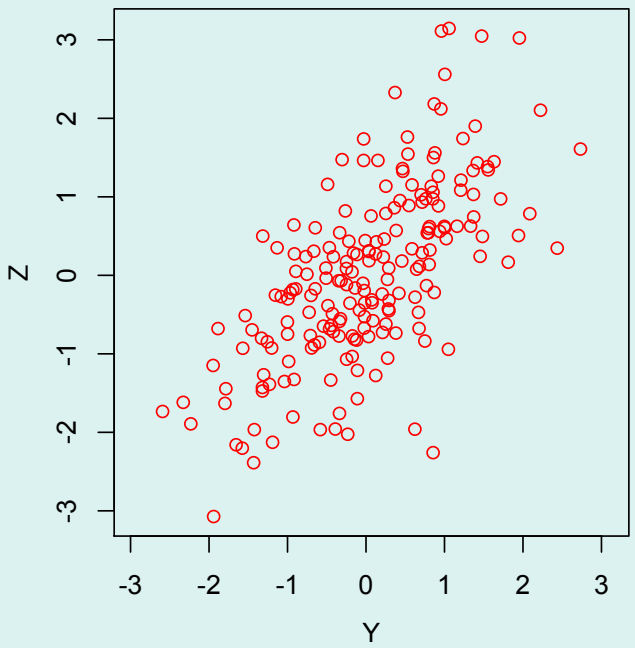


Figure 1: Portfolio data. For this sample, $\hat{\lambda}_{\text{opt}} = 0.283$ (James et al. 2021).

Sampling from the population: infeasible

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

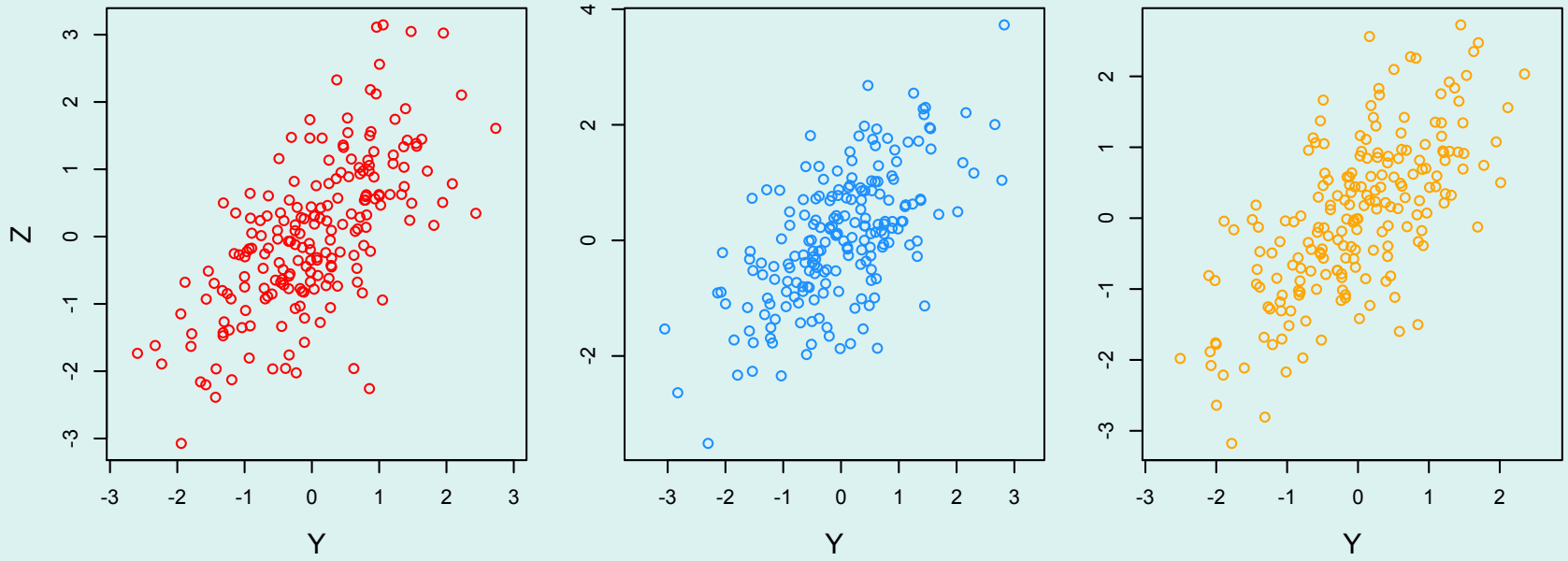


Figure 2: $\hat{\lambda}^{(1)}_{\text{opt}} = 0.283$, $\hat{\lambda}^{(2)}_{\text{opt}} = 0.357$, $\hat{\lambda}^{(3)}_{\text{opt}} = 0.299$ (James et al. 2021).

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

Toulouse School of Economics

Here:

$$\widehat{\mathrm{Std}[\hat{\lambda}_{\mathrm{opt}}]} \approx 0.077, \qquad \bar{\lambda}_{\mathrm{opt}} \approx 0.331 \quad (\approx \lambda_{\mathrm{opt}} = \frac{1}{3} = 0.333)$$

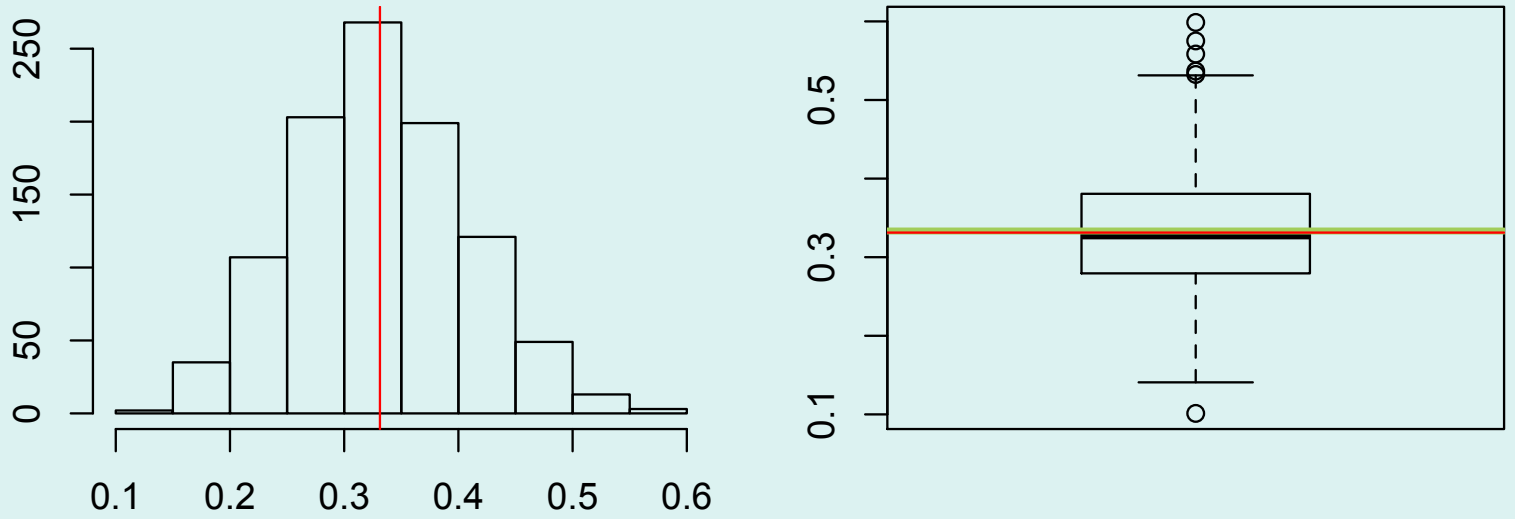


Figure 3: Histogram and boxplot of the empirical distribution of the $\hat{\lambda}_{\mathrm{opt}}^{(i)}$ (James et al. 2021).

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

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 $\hat{\lambda}_{\text{opt}}^{*(1)}$.
- Further generating bootstrap samples $(X_1^{*b}, \ldots, X_n^{*b})$, $b = 2, \ldots, B = 1000$, one can compute

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

with

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

Toulouse School of Economics

This provides

$$\widehat{\mathrm{Std}[\hat{\lambda}_{\mathrm{opt}}]}^{*} \approx 0.079$$

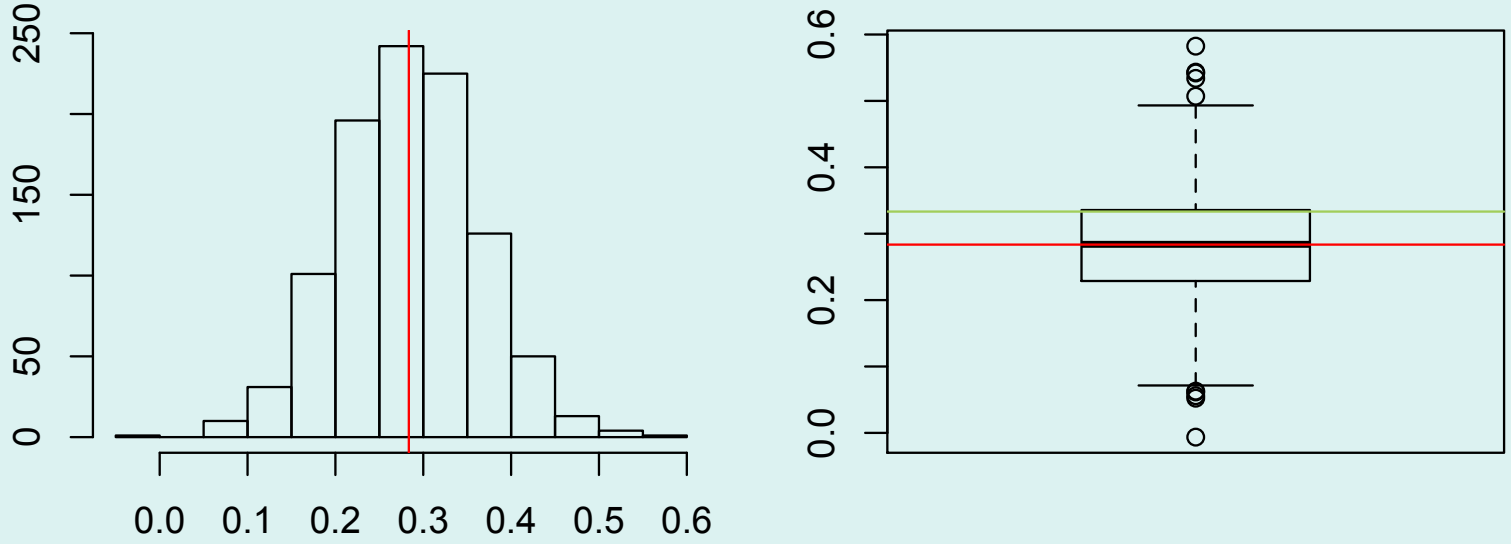


Figure 4: Histogram and boxplot of the bootstrap distribution of $\hat{\lambda}_{\mathrm{opt}}$ (James et al. 2021).

(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 $\widehat{\mathrm{Std}[\hat{\lambda}_{\mathrm{opt}}]}^{*} \approx 0.079$.

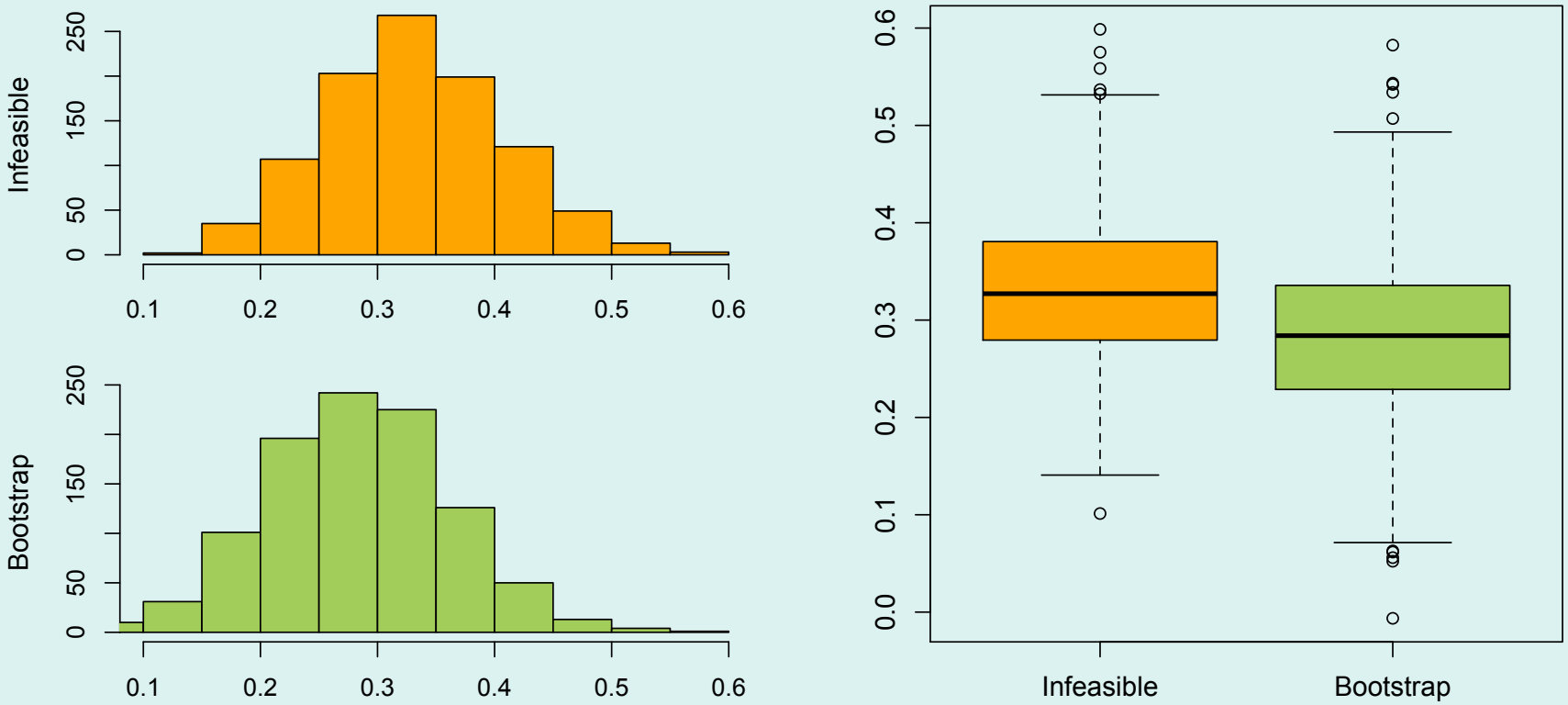


Figure 5: Bootstrap distributions from portfolio data (James et al. 2021).

1. Bootstrap

1.3 The general procedure

The bootstrap

- Let $X_1, \ldots, X_n$ be i.i.d $\sim P_\theta$.
- 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:

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

- 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.

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.
- ...

1. Bootstrap

1.4 About the implementation in R

A toy illustration

- Let $X_1, \ldots, X_n$ ($n = 4$) 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
```

Toulouse School of Economics

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
```

Toulouse School of Economics

Generating $B = 1000$ bootstrap means

```r
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
```

Toulouse School of Economics

Bootstrap estimates

Bootstrap estimates of $\mathbb{E}[\bar{X}]$ and $\mathrm{Var}[\bar{X}]$ 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

Toulouse School of Economics

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 $\mathrm{Var}[\bar{X}]$ is then

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

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

2. Bagging

Breiman (1996)

2. Bagging

2.1 Introduction

2.1 Introduction

- The bootstrap has other uses than those described above.
- In particular, it allows us to design **ensemble methods** in **statistical learning**.
- **Bagging** (**B**ootstrap **Agg**regat**ing**), which is the most famous approach in this direction, can be applied to both **regression** and **classification**.
- Below, we mainly focus on **bagging of classification trees**, but it should be clear that bagging of regression trees can be performed similarly.

Toulouse School of Economics

2. Bagging

2.2 Classification trees

Breiman et al. (1984)

The classification problem

- In classification, one observes (X_i, Y_i), $i = 1, \ldots, n$, where
  - X_i collects the values of p predictors on individual i, and
  - $Y_i \in \{1, 2, \ldots, K\}$ is the class to which individual i belongs.
- The problem is to classify a new observation for which we only see x, that is, to bet on the corresponding value $y \in \{1, 2, \ldots, K\}$.
- A classifier is a mapping

$$\begin{aligned} \phi_{\mathcal{S}} : \mathcal{X} &\rightarrow \{1, 2, \ldots, K\} \\ x &\mapsto \phi_{\mathcal{S}}(x), \end{aligned}$$

that is designed using the sample $\mathcal{S} = \{(X_i, Y_i),\ i = 1, \ldots, n\}$.

Toulouse School of Economics

```r
library(boot)
data(channing)
channing <- channing[,c("sex","entry","time","cens")]
channing[1:4,]
```

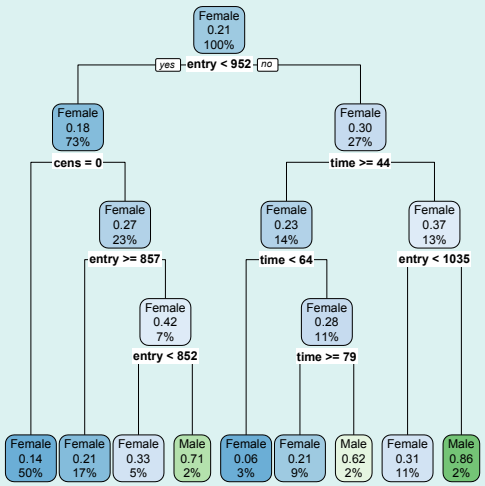
```
   sex entry time cens
1 Male   782  127    1
2 Male  1020  108    1
3 Male   856  113    1
4 Male   915   42    1
```

Predict `sex` $\in \{\text{Male}, \text{Female}\}$ on the basis of two numerical predictors (`entry`, `time`) and a binary one (`cens`).

Classification trees

In Part 1 of this course, we learned about a special type of classifiers $\phi_{\mathcal{S}}$, namely classification trees.

```
library(rpart)
library(rpart.plot)
fitted.tree <- rpart(sex~., data=channing, method="class")
rpart.plot(fitted.tree)
```



(+) Interpretability
(+) Flexibility
(−) Stability
(−) Performance

The process of **averaging** will reduce variability, hence, **improve stability**. Recall indeed that, if $U_1, \ldots, U_n$ are uncorrelated with variance σ^2, then

$$\mathrm{Var}[\bar{U}] = \frac{\sigma^2}{n}.$$

Since unpruned trees have low bias (but high variance), this reduced variance will lead to a low value of

$$\mathrm{MSE} = \mathrm{Var} + (\mathrm{Bias})^2$$

which will ensure a **good performance**.

How to perform this **averaging**?

2. Bagging

2.3 Bagging of classification trees

Bagging

Denote as $\phi_{\mathcal{S}}(x)$ the predicted class for predictor value x returned by the classification tree associated with sample $\mathcal{S} = \{(X_i, Y_i),\ i = 1, \ldots, n\}$.

Bagging of this tree considers predictions from B bootstrap samples

$$\begin{array}{rcl} \mathcal{S}^{*1} & = & ((X_1^{*1}, Y_1^{*1}), \ldots, (X_n^{*1}, Y_n^{*1})) \quad \rightsquigarrow \quad \phi_{\mathcal{S}^{*1}}(x) \\ \vdots & & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ \mathcal{S}^{*b} & = & ((X_1^{*b}, Y_1^{*b}), \ldots, (X_n^{*b}, Y_n^{*b})) \quad \rightsquigarrow \quad \phi_{\mathcal{S}^{*b}}(x) \\ \vdots & & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ \mathcal{S}^{*B} & = & ((X_1^{*B}, Y_1^{*B}), \ldots, (X_n^{*B}, Y_n^{*B})) \quad \rightsquigarrow \quad \phi_{\mathcal{S}^{*B}}(x) \end{array}$$

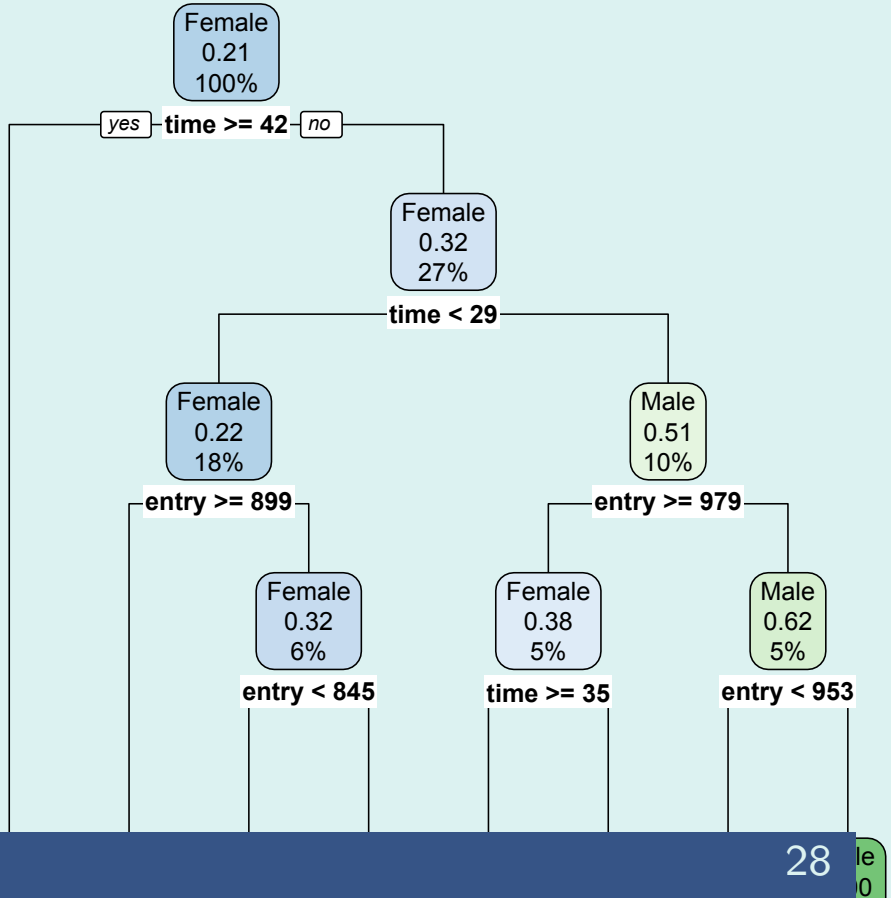
then proceeds by **majority voting** (i.e., the most frequently predicted class wins):

$$\phi_{\mathcal{S}}^{\text{Bagging}}(x) = \underset{k \in \{1,\ldots,K\}}{\text{argmax}}\ \#\{b : \phi_{\mathcal{S}^{*b}}(x) = k\}$$

Toulouse School of Economics

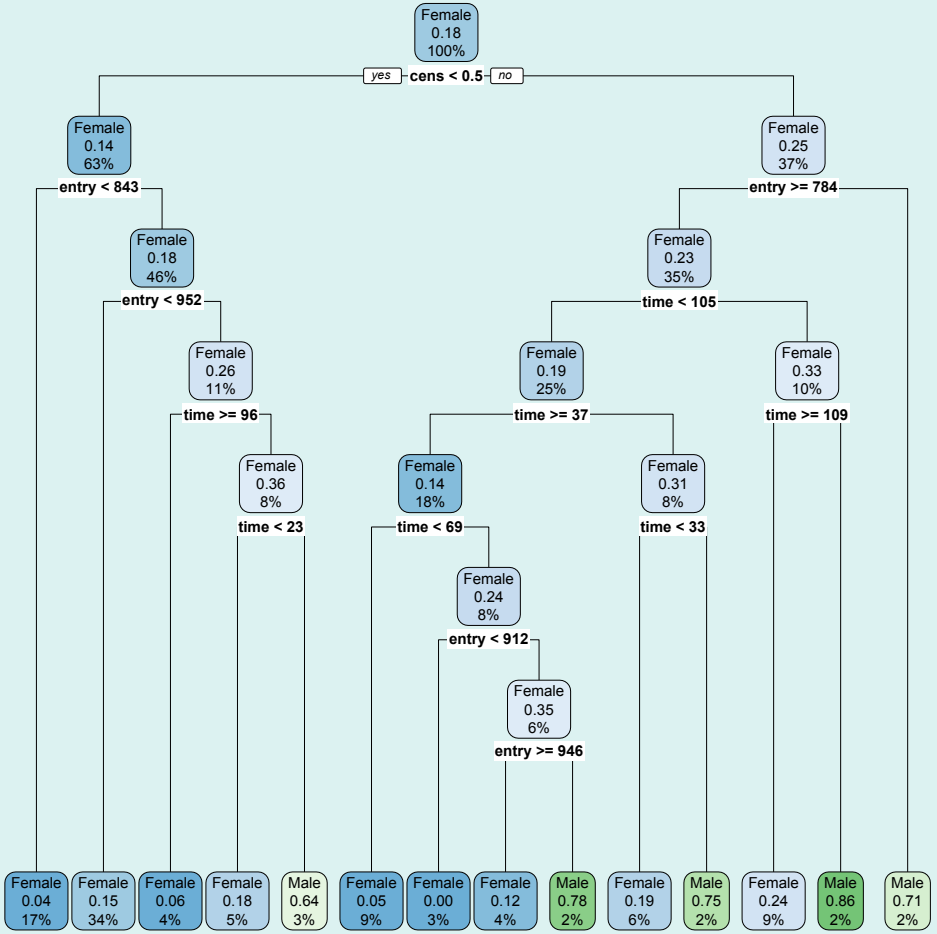
Toy illustration: bagging with $B = 3$ trees

```
d=sample(1:n,n,replace=TRUE)
fitted.tree <- rpart(sex~.,data=channing[d,],method="class")
rpart.plot(fitted.tree)
predict(fitted.tree, channing[1,], type="class")
```



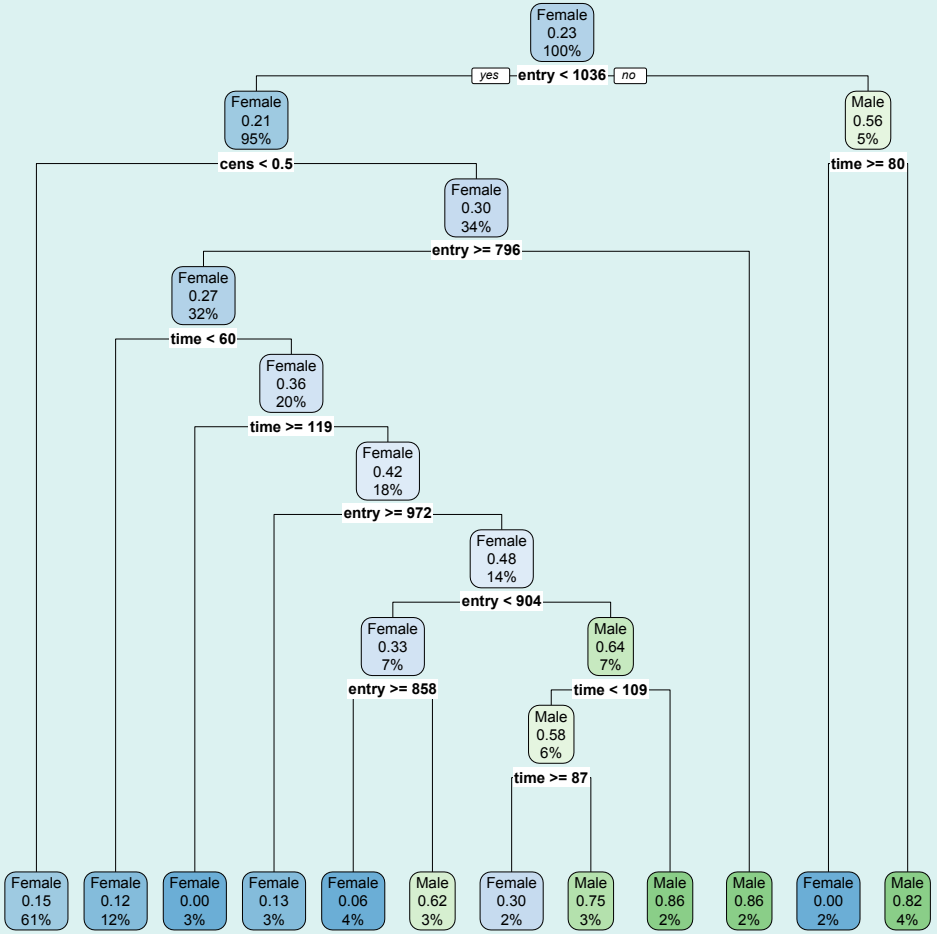
entry=782
time=127
cens=1
$\Downarrow$
Female

```r
d=sample(1:n,n,replace=TRUE)
fitted.tree <- rpart(sex~.,data=channing[d,],method="class")
rpart.plot(fitted.tree)
predict(fitted.tree, channing[1,], type="class")
```



entry=782
time=127
cens=1
$\Downarrow$
Male

```r
d=sample(1:n,n,replace=TRUE)
fitted.tree <- rpart(sex~.,data=channing[d,],method="class")
rpart.plot(fitted.tree)
predict(fitted.tree, channing[1,], type="class")
```



entry=782
time=127
cens=1
$\Downarrow$
Male

Toulouse School of Economics

For $x = (\text{entry},\text{time},\text{cens}) = (782,127,1)$,

- **two** (out of the $B = 3$ trees) **voted for Male**
- **one** (out of the $B = 3$ trees) **voted for Female**, the bagging classifier will thus classify x into **Male**.

Of course, B is usually much larger ($B = 500$? $B = 1000$?), which requires automating the process (through, e.g., the `boot` function).

2. Bagging

2.4 How much do you gain?

A simulation

We repeat $M = 1000$ times the following experiment:

(1) Split the data set into a training set (of size 300) and a test set (of size 162);

(2) (a) **train** a classification tree on the training set and evaluate its **test** error (i.e., misclassification rate) on the test set;

(b) do the same with a bagging classifier using $B = 500$ trees.

This provides $M = 1000$ test errors for the **direct** (single-tree) approach, and $M = 1000$ test errors for the **bagging** approach.

Toulouse School of Economics

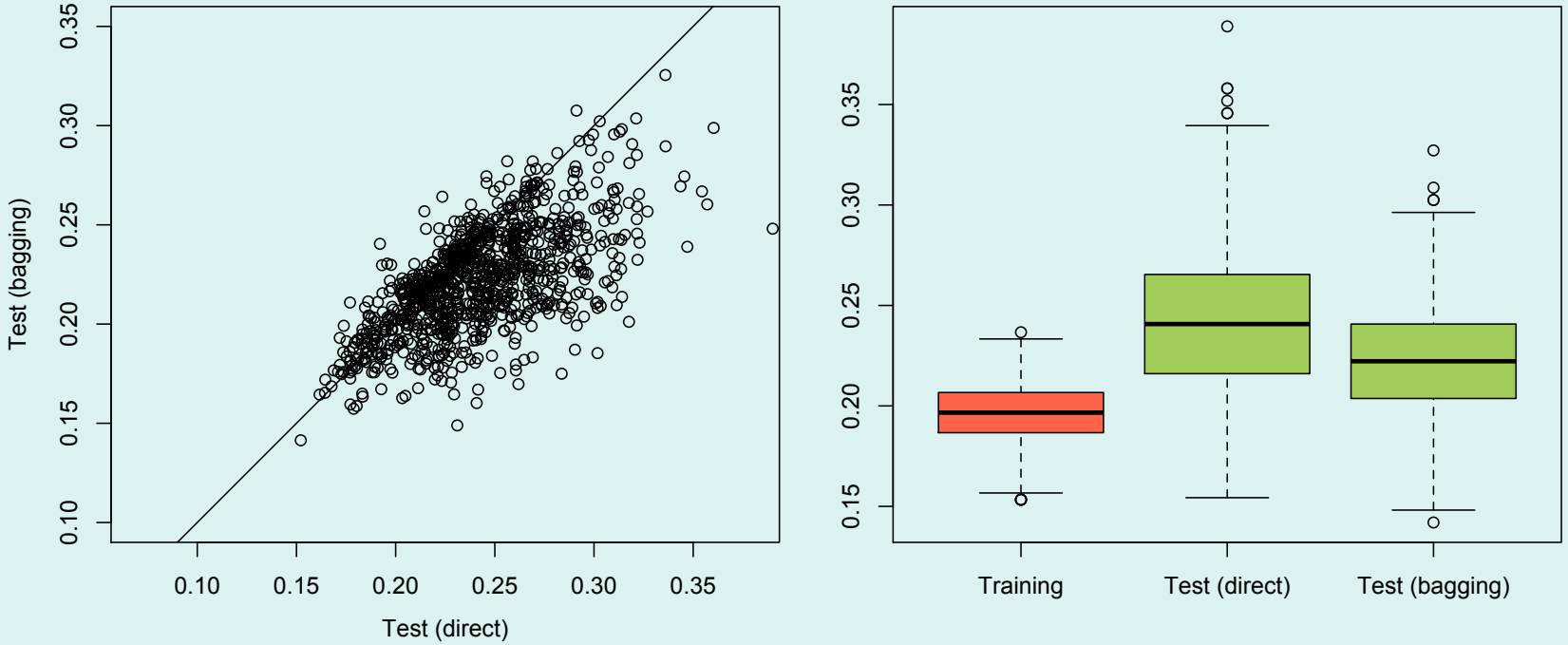


Figure 6: Results of the simulation (Q-Q plot and boxplot).

2. Bagging

2.5 Estimating the prediction accuracy

Estimating the prediction (lack of) accuracy

Several strategies to estimate prediction accuracy of a classifier:

(1) Compute a test error (as above): Partition the data set $\mathcal{S}$ into a training set $\mathcal{S}_{\text{train}}$ (to train the classifier) and a test set $\mathcal{S}_{\text{test}}$ (on which to evaluate the misclassification rate e_{test}).

(2) Compute an L-fold cross-validation error:

Partition the data set $\mathcal{S}$ into L folds $\mathcal{S}_\ell$, $\ell = 1, \ldots, L$. For each ℓ, evaluate the test error $e_{\text{test},\ell}$ associated with training set $\mathcal{S} \setminus \mathcal{S}_\ell$ and test set $\mathcal{S}_\ell$.

Run ℓ=1	Test	Train	Train	Train	Train
Run ℓ=2	Train	Test	Train	Train	Train
Run ℓ=3	Train	Train	Test	Train	Train
Run ℓ=4	Train	Train	Train	Test	Train
Run ℓ=5	Train	Train	Train	Train	Test

Figure 7

The quantity

$$e_{\text{CV}} = \frac{1}{L} \sum_{\ell=1}^{L} e_{\text{test},\ell}$$

is then the (L-fold) 'cross-validation error'.

Toulouse School of Economics

(3) Compute the Out-Of-Bag (OOB) error[1]:

For each observation X_i from $\mathcal{S}$, define the OOB prediction as

$$\phi_{\mathcal{S}}^{\text{OOB}}(X_i) = \underset{k \in \{1,\ldots,K\}}{\text{argmax}} \ \#\{b : \phi_{\mathcal{S}^{*b}}(X_i) = k \text{ and } (X_i, Y_i) \notin \mathcal{S}^{*b}\}$$

This is a **majority voting** discarding, quite naturally, bootstrap samples that use (X_i, Y_i) to train the classification tree. The OOB error is then the corresponding misclassification rate

$$e_{\text{OOB}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[\phi_{\mathcal{S}}^{\text{OOB}}(X_i) \neq Y_i]$$

[1] This is for bagging procedures only.

Final remarks

- **Bagging of trees can also be used for regression**. The only difference is that majority voting is then replaced with an averaging of individual predicted responses.
- **Bagging is a general device that applies to other types of classifiers**. In particular, it can be applied to kNN classifiers (we will illustrate this in the practical sessions).
- **Bagging affects interpretability of classification trees**. There are, however, solutions that intend to measure importance of the various predictors (see the next section).

Toulouse School of Economics

3. Random forests

Ho (1995)

3. Random forests

3.1 What can be improved?

Averaging reduces variability...

- We argued that bagging of trees will work because **averaging** reduces variability: if $U_1, \ldots, U_n$ are uncorrelated with variance σ^2, then
$$\mathrm{Var}[\bar{U}] = \frac{\sigma^2}{n}.$$
- However, the B trees that are 'averaged' in bagging are **not uncorrelated**! This will result into a smaller reduction of the variability.
- Random forests have the flavour of bagging-of-trees, but they incorporate a modification that aims at **decorrelating the trees**.

Toulouse School of Economics

3. Random forests

3.2 The procedure

- Like bagging-of-trees, random forests predict via majority voting from classification trees trained on B bootstrap samples.
- **However, whenever a split is designed in each tree, the split is only allowed among $m(\ll p)$ predictors randomly selected out of the p predictors**.
- **The rationale**: in a situation where there would be one strong predictor only, most bagged trees would use this predictor in the top split, which would result in highly correlated trees. The **tweak** above will prevent this, hence will lead to **less correlated trees**.

Toulouse School of Economics

Choosing the number of features

Using $m = p$ would simply provide bagging-of-trees. Using m small is appropriate when there are many correlated predictors. Common practice:

- For **classification** (where majority voting is used), $m \approx \sqrt{p}$.
- For **regression** (where tree predictions are averaged), $m \approx \frac{p}{3}$.

In both cases, results are actually not very sensitive to m.

Random forests

- ☒ are **nonparametric** and **efficient**
- ☒ can deal with **a large number of predictors** (high dimension)
- ☒ can cope with **both small and large sample sizes** (Big Data)

but they

- ☐ rely on a rather **black box** model, and
- ☐ are **not supported by strong theoretical results**

Toulouse School of Economics

3. Random forests

3.3 A simulation

Let us look at efficiency...

We repeated $M = 1000$ times the following experiment:

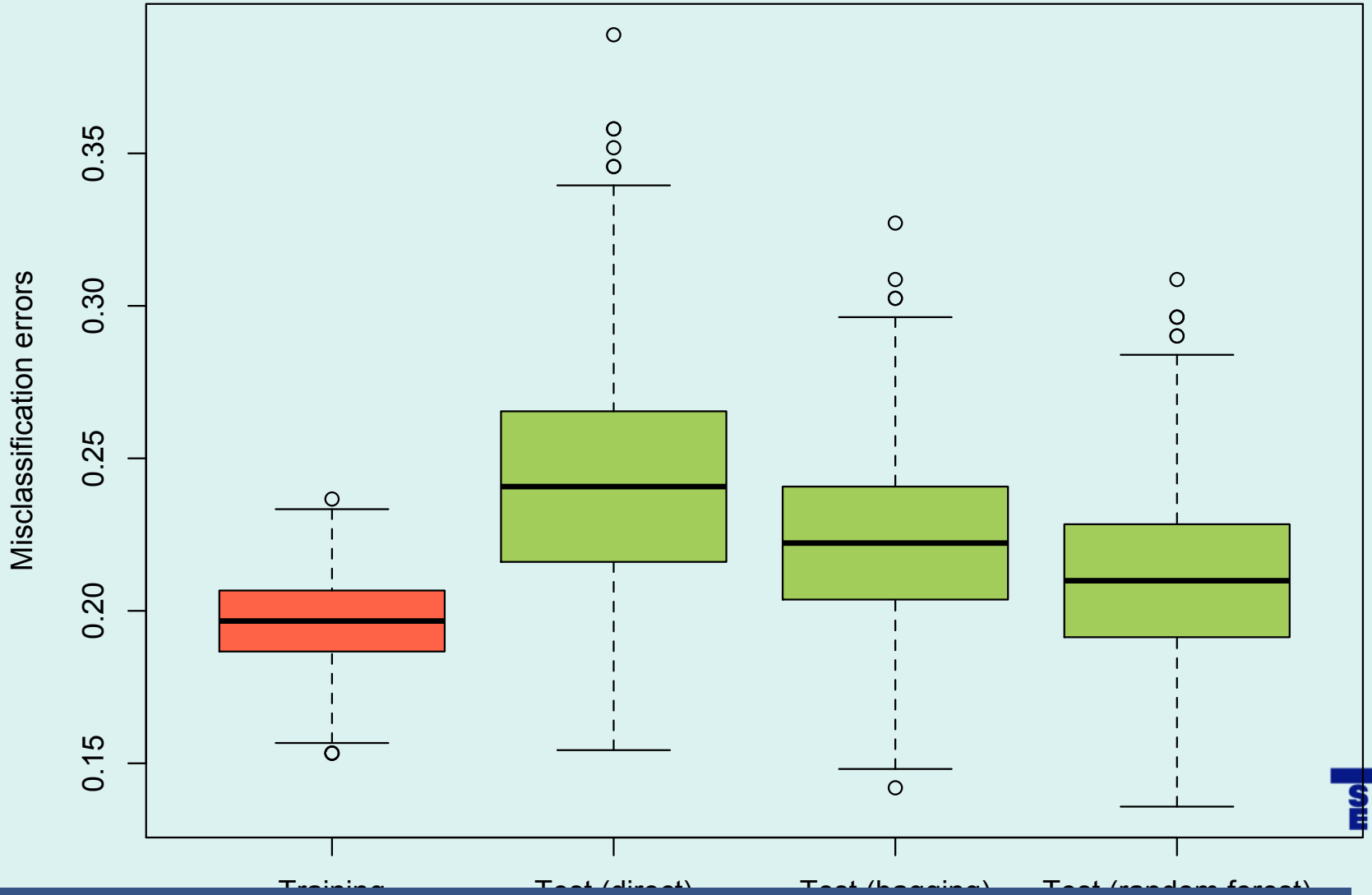
(1) Split the `channing` data set into a training set (of size 300) and a test set (of size 162);

(2) (a) train a classification tree on the training set and evaluate its test error (i.e., misclassification rate) on the test set;
    (b) do the same with a bagging classifier using $B = 500$ trees;
    (c) **do the same with a random forest classifier using the `randomForest` function in `R` with default parameters ($B = 500$ trees, $m \approx \sqrt{p}$).**

This provided $M = 1000$ test errors for the direct (single-tree) approach, $M = 1000$ test errors for the bagging approach, and $M = 1000$ test errors for the random forest approach.

Toulouse School of Economics

The results



3. Random forests

3.4 Importance of each predictor

Measuring importance of each predictor

Because bagging-of-trees and random forests are poorly interpretable compared to classification trees, the following is useful.

The **importance v_j of the jth predictor** is measured as follows.

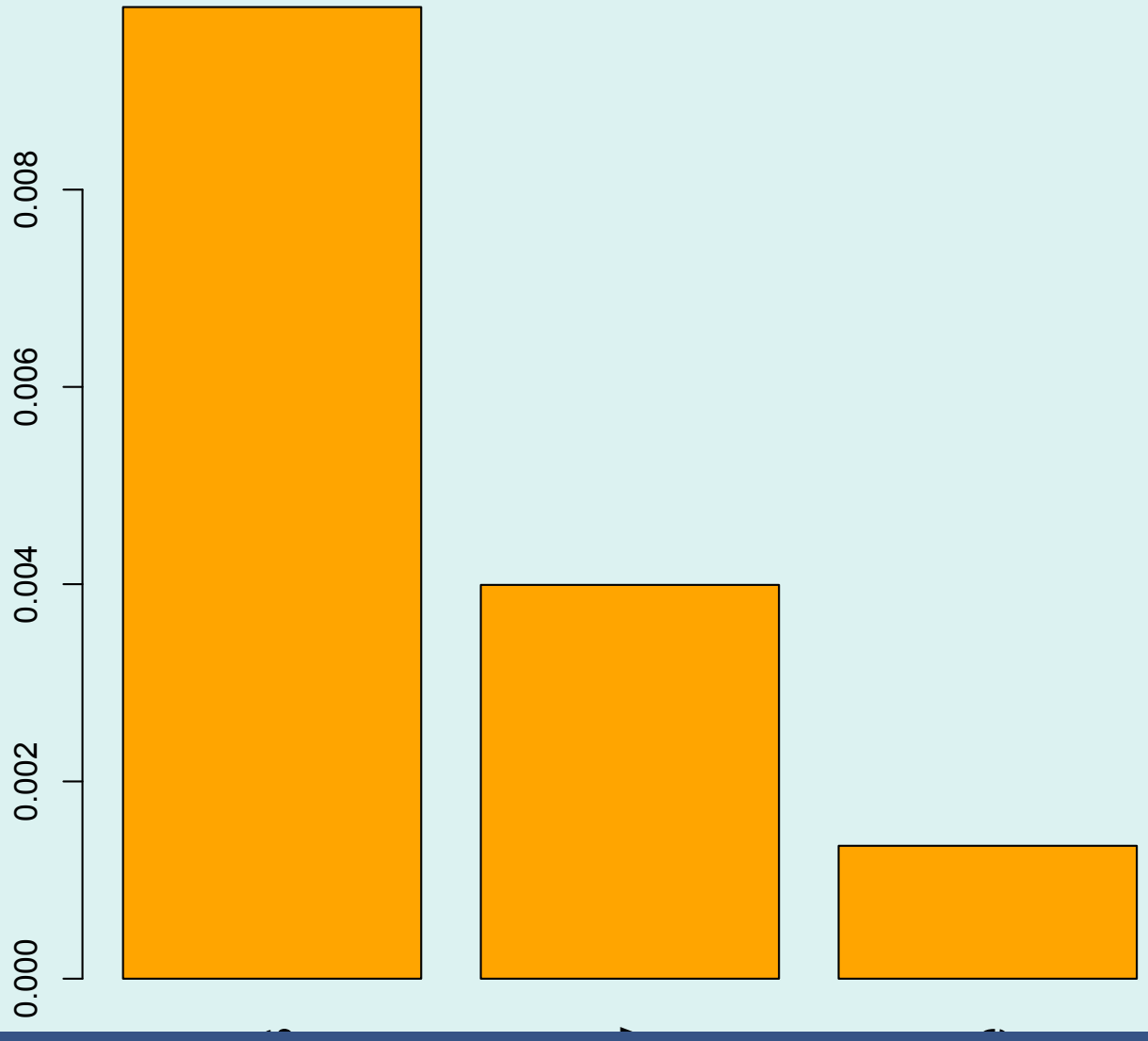
For each tree (i.e., for any $b = 1, \ldots, B$),

- the **prediction error on OOB observations** is recorded, and
- the same is done after **permuting randomly** all values of the jth predictor (which essentially turns this predictor into noise), the **difference between both errors** is then averaged over $b = 1, \ldots, B$ (and normalized by the standard error—if it is positive), yielding v_j.

(A similar measure is used for regression, based on MSEs).

Toulouse School of Economics

Measuring importance of each predictor



References I

- Breiman, Leo (Aug. 1996). **"Bagging predictors".** en. In: *Machine Learning* 24.2, pp. 123–140. ISSN: 1573-0565. DOI: 10.1007/BF00058655 (cit. on p. 27).
- Breiman, Leo et al. (1984). ***Classification And Regression Trees.*** en. 1st ed. Routledge. ISBN: 978-1-315-13947-0. DOI: 10.1201/9781315139470 (cit. on p. 31).
- Efron, B. (Jan. 1979). **"Bootstrap Methods: Another Look at the Jackknife".** In: *The Annals of Statistics* 7.1. ISSN: 0090-5364. DOI: 10.1214/aos/1176344552 (cit. on p. 3).
- Ho, Tin Kam (Aug. 1995). **"Random Decision Forests".** In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1, 278–282 vol.1. DOI: 10.1109/ICDAR.1995.598994 (cit. on p. 51).

Toulouse School of Economics

References II

James, Gareth et al. (2021). ***An Introduction to Statistical Learning: with Applications in R.*** en. Springer Texts in Statistics. New York, NY: Springer US. isbn: 978-1-07-161417-4. doi: 10.1007/978-1-0716-1418-1 (cit. on pp. 7, 10–12, 14, 15).

Toulouse School of Economics