# Bootstrap – Bagging – Random Forests

Olivier Roustant

Mines Saint-Étienne

2017/11

# Outline

1. **Bootstrap**
2. **Aggregation, bagging and random forests**

## Warning

- This is only a very short introduction to bootstrap, aggregation and random forests, aiming at giving some insights to the future case study
- This has to be completed by your own reading on these topics, in particular Chapters 9 and 15 of [ESL]

# Bootstrap

## Purpose

- The idea of bootstrap is to resample in the data

  $\rightarrow$ Allows creating variability without extra information.

  Etymology : To go up by pulling on the bootstraps (without extra force !)

  $\rightarrow$ Allows simulating from an unknown distribution.

- Application to the case study 2015
  Compute forecast intervals without assuming the normality of the residuals $\varepsilon_t$ in the linear model with AR(2) errors

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_p x_{p,t} + u_t$$

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \varepsilon_t$$

## Principle

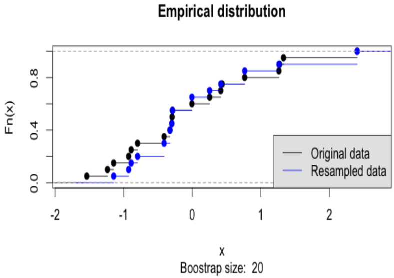
Denote $\hat{F}_n$ the empirical distribution, i.e. the discrete distribution supported by the data $\{x_1, \dots, x_n\}$, with uniform weights :

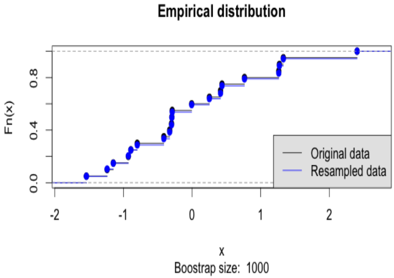
$$d\hat{F}_n(x) = \frac{1}{n}\delta_{x_1}(x) + \cdots + \frac{1}{n}\delta_{x_n}(x)$$

Assume that $x_1, \dots, x_n$ is a sample of $F$ (an unknown distribution).
Then If $n$ is large enough, simulating from $\hat{F}_n$ or $F$ with be very similar.

## Principle

Ex. Explain why if $U \sim \mathcal{U}(\{1, \dots, n\})$ then $x_U \sim \hat{F}_n$. Thus, simulating from $\hat{F}_n$ is achieved by resampling the data (with replacement).

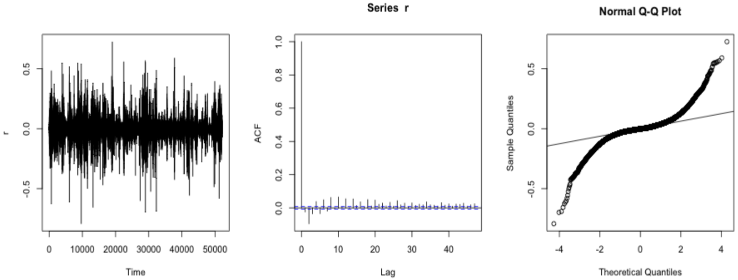



$\rightarrow$ R code : `sample(data, size = nboot, replace = TRUE)`

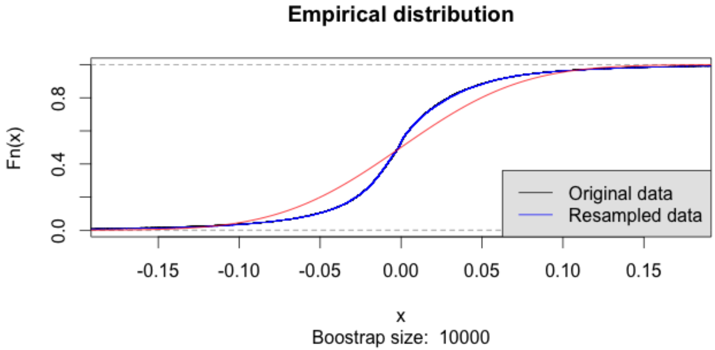
## Application to the case study 2015

Possible residuals $\varepsilon_t$ are represented below.
$\rightarrow$ They look approx. independent (ignoring variance variations...)
$\rightarrow$ They have fatter tails than the normal distribution ('leptokurticity')



## Application to the case study 2015

Compare the cdf of bootstrapped residuals (drawn from $\hat{F}_n$) to the cdf of the Gaussian distribution (in red), here different from $F$.



## Correlation of a bootstrapped sample

Since boostrapped data are drawn from the same data, they are correlated.

Ex. Let $X_1, \ldots, X_n$ i.i.d. $(0, \sigma^2)$. Define :

$$X_1^* = X_{U_1}, \ldots, X_B^* = X_{U_B}$$

boostrapped data, where $U_1, \ldots, U_B$ are i.i.d. $\sim \mathcal{U}(\{1, \ldots, n\})$ and independent from $X_1, \ldots, X_n$.

Prove that $X_1^*, \ldots, X_B^*$ are i.d. $(0, \sigma^2)$ but with $cor(X_i^*, X_j^*) = \frac{1}{n}$.

## Correlation of bootstrapped sample means

Ex. Let $X_1, \dots, X_n$ i.i.d. $(0, \sigma^2)$ and let $\overline{X}_1^*, \overline{X}_2^*$ two sample means computed (independently) by bootstrap. Prove that

$$cor(\overline{X}_1^*, \overline{X}_2^*) = \frac{n}{2n-1} \approx 50\%$$

# Aggregation, bagging and random forests

## Bagging : Bootstrap + Aggregating

Principle. Consider a set of data $z_1, \dots, z_N$.

- Obtain new data by bootstrapping the original data
  $\rightarrow$ each bootstrap sample $Z_1^{\star b}, \dots, Z_N^{\star b}$ gives a new learner
- Aggregate (here : average) the learners

## Idea #1 : Bagging is most useful for instable models

Notations

- $Z = \{(Y_n, X_n), n = 1, \ldots, N\}$ : i.i.d. r.v. representing the data
- $\phi(x, Z)$ : Prediction of $y$ for a new $x$
- $\phi_A(x) = E_Z(\phi(x, Z))$ : Aggregated prediction
  *In Bagging,* $\phi_A(x) \approx \frac{1}{B} \sum_{b=1}^{B} \phi(x, Z^{\star b})$

## Idea #1 : Bagging is most useful for instable models

Notations

- $Z = \{(Y_n, X_n), n = 1, \dots, N\}$ : i.i.d. r.v. representing the data
- $\phi(x, Z)$ : Prediction of $y$ for a new $x$
- $\phi_A(x) = E_Z(\phi(x, Z))$ : Aggregated prediction
  *In Bagging,* $\phi_A(x) \approx \frac{1}{B} \sum_{b=1}^{B} \phi(x, Z^{\star b})$

Define, for given $x, y$ :

- $e(x, y) = E_Z\left[(y - \phi(x, Z))^2\right]$ : The mean square error
- $e_A(x, y) = (y - \phi_A(x))^2$ : The aggregate error

## Idea #1 : Bagging is most useful for instable models

Notations

- $Z = \{(Y_n, X_n), n = 1, \dots, N\}$ : i.i.d. r.v. representing the data
- $\phi(x, Z)$ : Prediction of $y$ for a new $x$
- $\phi_A(x) = E_Z(\phi(x, Z))$ : Aggregated prediction
  *In Bagging,* $\phi_A(x) \approx \frac{1}{B}\sum_{b=1}^{B} \phi(x, Z^{\star b})$

Define, for given $x, y$ :

- $e(x, y) = E_Z\left[(y - \phi(x, Z))^2\right]$ : The mean square error
- $e_A(x, y) = (y - \phi_A(x))^2$ : The aggregate error

Exercise. By interpreting $e$ and $e_A$ with risk and bias, show that

$$e_A(x, y) - e(x, y) = -\mathrm{var}_Z(\phi(x, Z)) \le 0$$

## Idea #2 : Bagging is improved by reducing correlation

Fact. The 'weak' learners $\phi(x, Z^{\star b})$ are independent conditionaly to initial data $(X_1, Y_1), \ldots, (X_n, Y_n)$, but not independent.

## Idea #2 : Bagging is improved by reducing correlation

Fact. The 'weak' learners $\phi(x, Z^{\star b})$ are independent conditionaly to initial data $(X_1, Y_1), \ldots, (X_n, Y_n)$, but not independent.

Ex. #1. The $\phi(x, Z^{\star b})$ have common variance and correlation.

## Idea #2 : Bagging is improved by reducing correlation

Fact. The 'weak' learners $\phi(x, Z^{\star b})$ are independent conditionaly to initial data $(X_1, Y_1), \dots, (X_n, Y_n)$, but not independent.

Ex. #1. The $\phi(x, Z^{\star b})$ have common variance and correlation.

Ex. #2. Let $B$ r.v. $W_1, \dots, W_B$ with common variance $\sigma^2$ and correlation $\rho \geq 0$. Then the variance of $\frac{1}{B}\sum_{b=1}^{B} W_b$ is :

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

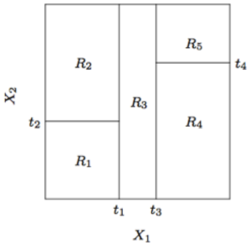
$\rightarrow$ All the more efficient as $\rho$ is small.

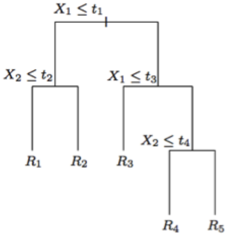
## Principles of random forest

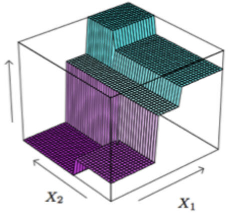
- Use non-linear and unstable weak learners
  $\rightarrow$ Averaging of linear learners result in a linear learner !
  $\rightarrow$ Unstable : see above 'bagging and unstability'
  $\rightarrow$ Trees are good candidates
- Resample the observations as in bagging
- Resample the variables in order to decrease $\rho$ ("feature sampling")

## Trees in 1 slide (from [ESL, chapter 9])

Example with CART (Classification and Regression Trees).



# Algorithm (from [ESL, Chapter 15])

**Algorithm 15.1** *Random Forest for Regression or Classification.*

1. For $b = 1$ to $B$:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size $N$ from the training data.
   (b) Grow a random-forest tree $T_b$ to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{min}$ is reached.
      i. Select $m$ variables at random from the $p$ variables.
      ii. Pick the best variable/split-point among the $m$.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point $x$:

*Regression:* $\hat{f}_{\mathrm{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$.

*Classification:* Let $\hat{C}_b(x)$ be the class prediction of the $b$th random-forest tree. Then $\hat{C}_{\mathrm{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

# References

**ESL** T. Hastie, R. Tibshirani and J. Friedman (2009), The Elements of Statistical Learning, Springer, 2nd edition, print 10.

**BRE** L. Breiman (1994), Bagging Predictors, Technical Report 421, University of California at Berkeley.