Forschungszentrum Jülich
Bergische Universität Wuppertal

TRAINING COURSE

Introduction to descriptive and parametric statistic with R

Antoine Tordeux — | tordeux@uni-wuppertal.de
vzu.uni-wuppertal.de

Content

Introduction to descriptive and parametric statistic with R

The objectives are both to propose useful statistical methods allowing to analyse univariate and multivariate data or to develop and calibrate models, as well as to learn how to use R.

The course is organized in three sessions :

- Session 1 : **Statistics for uni- and bivariate dataset**
- Session 2 : **Statistics for multivariate dataset**
- Session 3 : **Parametric statistic and statistical inference**

Git :	`gitlab.version.fz-juelich.de`
Homepage :	`www.vzu.uni-wuppertal.de/lehre`
Download R :	`cran.r-project.org`

Statistic

Origin : 'Statistic' initially refers to the collection of information by states

- Etymology from the New Latin *statisticum* and the German words *Statistik* and *Staatskunde* (18th century)
- Counting of demographic and economic data

Modern sense : Collection, visualization, analysis, modelling, interpretation, prediction of information of all types

- Physics, social science, biology, ... — Models for understanding
- Engineering, neuroscience, ... — Models for prediction
- Applied mathematics, physics, ... — Statistical inference

Context

Data : n observations of characteristics (of individuals, systems, ...) or results of experiments

⚠ Sample is not a time series (order of the observations has no importance)

→ Stochastic processes for dynamical systems

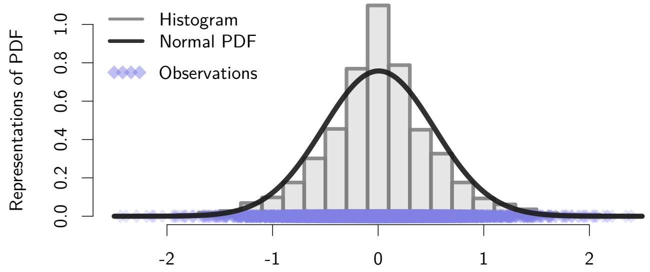
Statistic : Mathematical tools allowing to present, resume, explain or predict some data, and to develop and calibrate models

- Loose of information (data too big to individually analyze each observation)
- Focus on phenomena of interest, tendencies, global performances

Descriptive statistic : Tools describing data with no probabilist assumptions

Parametric statistic : Probabilist assumptions on the distributions of the data

Illustrative example



Representations of PDF by	Histogram :	**Descriptive estimation**
	Normal PDF :	**Parametric estimation**

Statistical packages

Product	Description	Creation Date	Open Source	Written in Scripting	Support
MatLab `mathworks.com`	Platform for numerical computing	1970's		C++, java MatLab	Windows, Mac OS, Linux
SAS `sas.com`	Statistical analysis system	1974		C SAS language	Windows, Linux
SPSS `ibm.com`	Software package for statistical analysis	1968		java R, Python	Windows, Mac OS, Linux
Stata `stata.com`	General-purpose statistical software	1985		C ado, Mata	—
Statistica `dell.com`	Advanced analytics software package	1991		C++ R, SVB	Windows
R `r-project.org`	Software environment for statistical computing	1993	×	C, Fortran R language	Windows, Mac OS, Linux
SciLab `scilab.org`	Open-source alternative to MatLab	1990	×	C, C++, java SciLab	—
PSPP `gnu.org`	Open-source alternative to SPSS	1998	×	C Pearl	—
SciPy `scipy.org`	Python library for scientific computing	1992	×	C, Fortran Python	—

And many others ... (see, e.g., `Wikipedia : Statistical packages`)

R software environment[1]

R is a open source programming language and environment for statistical computing and graphics

Windows : **The terminal** — **The script** (eventual) — **The plots** (eventual)

Help with R : ?*name_of_a_function* or help(*name_of_a_function*)

Implementation of S language — Functional programming

Computation in R consists of sequentially evaluating statements separated by semi-colon or new line, and that can be grouped using braces

 # Variable, vector, operations

```
pi*sqrt(10)+exp(4)
2:7
seq(0,1,0.1)
x=c(1,2,3);y=c(4,5)
z=c(x,y)
z∧2;log(z)
```

 # Main control structures

```
x=7
if(x>0) y=0
for(i in 1:7)
  x=x+i
while(y>1)
  y=y/2
```

 # Functions

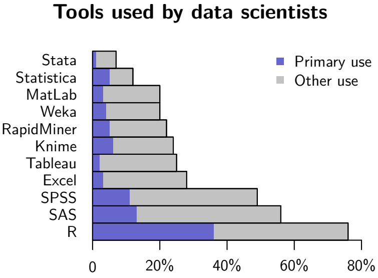
```
exp(2)
?exp
exp_app=function(x,n)
  sum(x∧n/factorial(n))
exp_app(2,1:5)
```

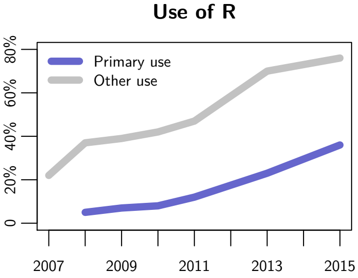
Integrated development environments for R : RStudio, Jupyter (online), Rattle, Red-R, R Commander, ...

[1] 1993, GNU General Public License, r-project.org

Use of R

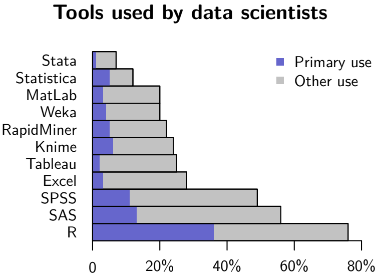
Source : Rexer Analytics, 2016

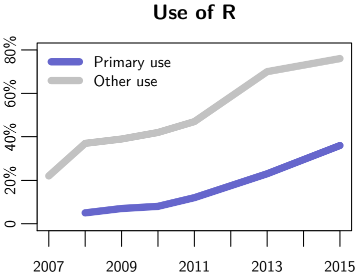




Use of R

Source : Rexer Analytics, 2016





- R is the most used tool of data scientists and analysts (with tendency to increase)
- R is solely dedicated to statistical computing and graphics
- More general languages such as Python (see, e.g., package `scipy`) can compute statistical methods as well, but the implementation in R is generally easier

  → See Python & R codes for common machine learning algorithms at `analyticsvidhya.com` or R vs Python at `blog.dominodatalab.com`

Overview

Part 1 | ## Descriptive statistics for univariate and bivariate data

Repartition of the data (histogram, kernel density, empirical cumulative distribution function), order statistic and quantile, statistics for location and variability, boxplot, scatter plot, covariance and correlation, QQplot

Part 2 | ## Descriptive statistics for multivariate data

Least squares and linear and non-linear regression models, principal component analysis, principal component regression, clustering methods (K-means, hierarchical, density-based), linear discriminant analysis, bootstrap technique, artificial neural networks

Part 3 | ## Parametric statistic

Likelihood, estimator definition and main properties (bias, convergence), punctual estimate (maximum likelihood estimation, Bayesian estimation), confidence and credible intervals, information criteria, test of hypothesis, parametric clustering

Appendix LaTeX plots with R and Tikz

Overview

Part 1 — Descriptive statistics for univariate and bivariate data

Repartition of the data (histogram, kernel density, empirical cumulative distribution function), order statistic and quantile, statistics for location and variability, boxplot, scatter plot, covariance and correlation, QQplot

Part 2 — Descriptive statistics for multivariate data

Least squares and linear and non-linear regression models, principal component analysis, principal component regression, clustering methods (K-means, hierarchical, density-based), linear discriminant analysis, bootstrap technique, artificial neural networks

Part 3 — Parametric statistic

Likelihood, estimator definition and main properties (bias, convergence), punctual estimate (maximum likelihood estimation, Bayesian estimation), confidence and credible intervals, information criteria, test of hypothesis, parametric clustering

Appendix LaTeX plots with R and Tikz

Data used

Experiments with pedestrians on a ring

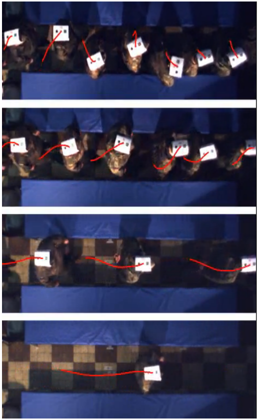
→ 11 experiments done for different density levels

Measurement of :

Spacing
(position difference with predecessor)

Speed
(position time-difference)

Acceleration rate
(speed time-difference)



Descriptive statistics for univariate data

$$(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$$

Histogram — R : `hist(x)`

Histogram : Counting of the observations on a regular partition $(I_j)_j$ with window δ

$$\forall j, x \in I_j, \quad \tilde{h}(x) = \sum_{i=1}^{n} \mathbb{1}_{I_j}(x_i), \qquad \text{with} \quad \mathbb{1}_I(x) = \begin{cases} 1 & \text{if } x \in I \\ 0 & \text{otherwise} \end{cases}$$

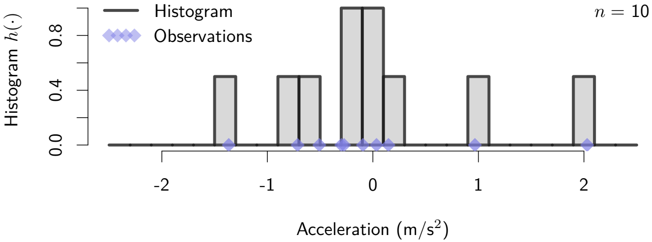
→ **Normalized histogram** $h(x) = \frac{1}{\delta n} \tilde{h}(x)$ for estimation of PDF

Histogram — R : hist(x)

Histogram : Counting of the observations on a regular partition $(I_j)_j$ with window δ

$$\forall j, x \in I_j, \quad \tilde{h}(x) = \sum_{i=1}^{n} \mathbb{1}_{I_j}(x_i), \qquad \text{with} \quad \mathbb{1}_I(x) = \begin{cases} 1 & \text{if } x \in I \\ 0 & \text{otherwise} \end{cases}$$

$\rightarrow$ **Normalized histogram** $h(x) = \frac{1}{\delta n} \tilde{h}(x)$ for estimation of PDF

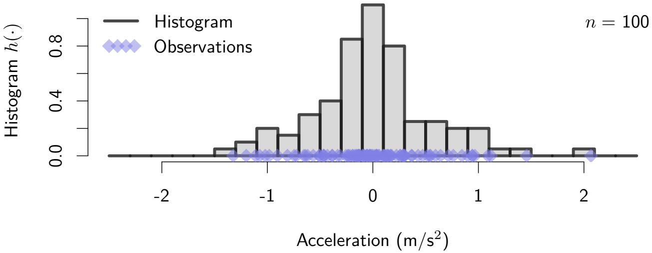


Histogram — R : hist(x)

Histogram : Counting of the observations on a regular partition $(I_j)_j$ with window δ

$$\forall j, x \in I_j, \quad \tilde{h}(x) = \sum_{i=1}^{n} \mathbb{1}_{I_j}(x_i), \qquad \text{with} \quad \mathbb{1}_I(x) = \begin{cases} 1 & \text{if } x \in I \\ 0 & \text{otherwise} \end{cases}$$

→ **Normalized histogram** $h(x) = \frac{1}{\delta n}\,\tilde{h}(x)$ for estimation of PDF

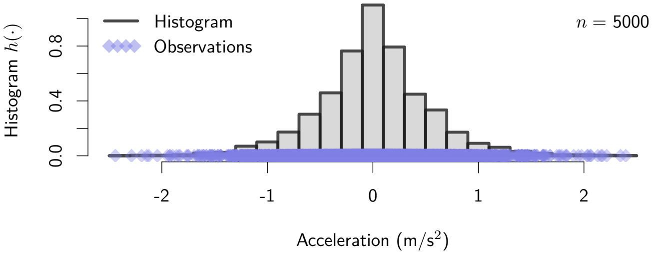


Histogram — R : hist(x)

Histogram : Counting of the observations on a regular partition $(I_j)_j$ with window δ

$$\forall j, x \in I_j, \quad \tilde{h}(x) = \sum_{i=1}^{n} \mathbb{1}_{I_j}(x_i), \qquad \text{with} \quad \mathbb{1}_I(x) = \begin{cases} 1 & \text{if } x \in I \\ 0 & \text{otherwise} \end{cases}$$

$\rightarrow$ **Normalized histogram** $h(x) = \frac{1}{\delta n}\tilde{h}(x)$ for estimation of PDF



Kernel density — `R : density(x)`

Kernel continuous estimation of the PDF

$$d(x) = \frac{1}{nb} \sum_{i=1}^{n} k((x - x_i)/b), \qquad \text{with } b > 0 \text{ the bandwidth}$$

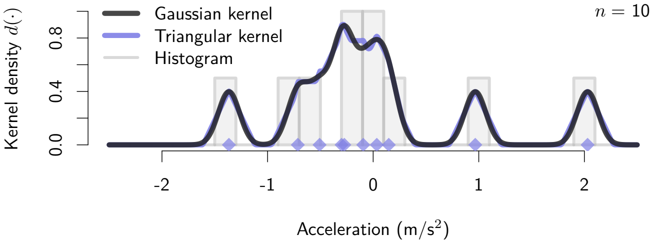
→ **Kernel** $k(.)$ such that $\int k(x)\,\mathrm{d}x = 1$ and $k(x) = k(-x)$

Kernel density — R : density(x)

Kernel continuous estimation of the PDF

$$d(x) = \frac{1}{nb} \sum_{i=1}^{n} k((x - x_i)/b), \qquad \text{with } b > 0 \text{ the bandwidth}$$

→ **Kernel** $k(.)$ such that $\int k(x)\,\mathrm{d}x = 1$ and $k(x) = k(-x)$

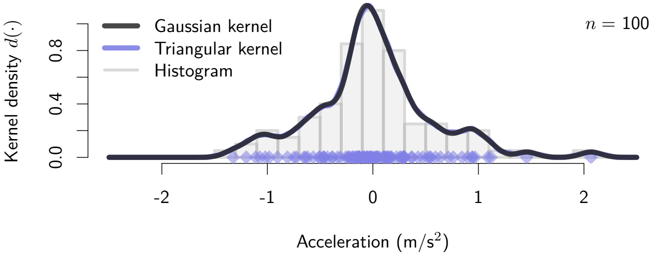


Kernel density — R : density(x)

Kernel continuous estimation of the PDF

$$d(x) = \frac{1}{nb}\sum_{i=1}^{n} k((x - x_i)/b), \qquad \text{with } b > 0 \text{ the bandwidth}$$

→ **Kernel** $k(.)$ such that $\int k(x)\,\mathrm{d}x = 1$ and $k(x) = k(-x)$

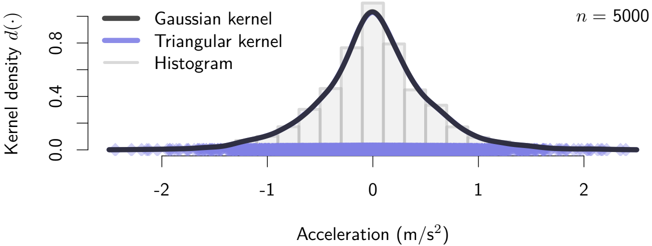


Kernel density — `R : density(x)`

Kernel continuous estimation of the PDF

$$d(x) = \frac{1}{nb}\sum_{i=1}^{n} k((x - x_i)/b), \qquad \text{with } b > 0 \text{ the bandwidth}$$

→ **Kernel** $k(.)$ such that $\int k(x)\,\mathrm{d}x = 1$ and $k(x) = k(-x)$



Cumulative distribution function — R : `ecdf(x)`

Empirical cumulative distribution function (ECDF)

$$D(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \leq x}, \qquad \text{with} \quad \mathbb{1}_R = \begin{cases} 1 & \text{if } R \\ 0 & \text{otherwise} \end{cases}$$

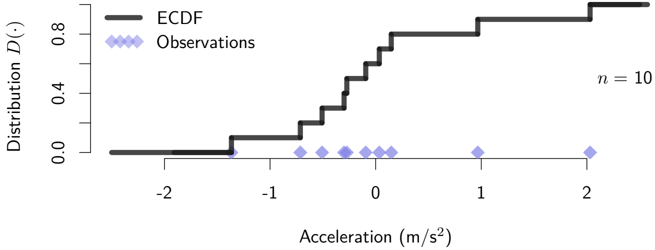
→ Does not depend on a width to calibrate

Cumulative distribution function — R : ecdf(x)

Empirical cumulative distribution function (ECDF)

$$D(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \leq x}, \qquad \text{with} \quad \mathbb{1}_R = \begin{cases} 1 & \text{if } R \\ 0 & \text{otherwise} \end{cases}$$

→ Does not depend on a width to calibrate

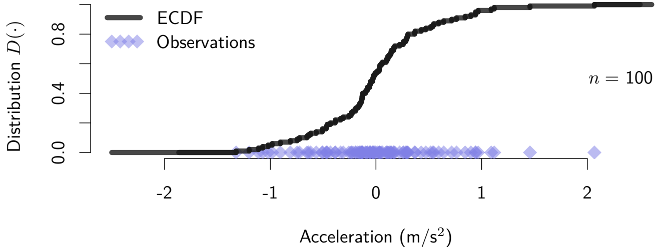


Cumulative distribution function — R : ecdf(x)

Empirical cumulative distribution function (ECDF)

$$D(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \leq x}, \qquad \text{with} \quad \mathbb{1}_R = \begin{cases} 1 & \text{if } R \\ 0 & \text{otherwise} \end{cases}$$

→ Does not depend on a width to calibrate

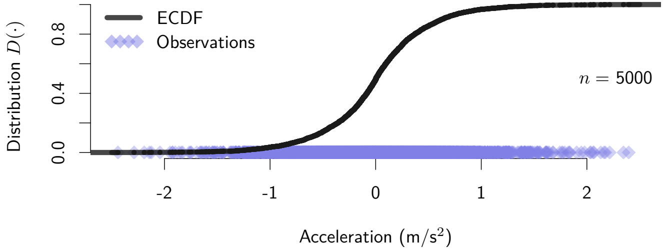


Cumulative distribution function — R : ecdf(x)

Empirical cumulative distribution function (ECDF)

$$D(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \leq x}, \qquad \text{with} \quad \mathbb{1}_R = \begin{cases} 1 & \text{if } R \\ 0 & \text{otherwise} \end{cases}$$

→ Does not depend on a width to calibrate

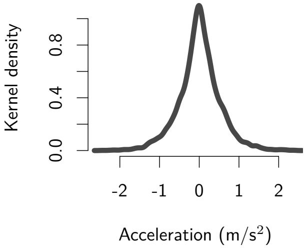


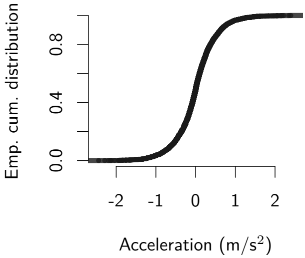
Cumulative distribution function — R : `ecdf(x)`

Empirical cumulative distribution function (ECDF)

$$D(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \le x}, \qquad \text{with} \quad \mathbb{1}_R = \begin{cases} 1 & \text{if } R \\ 0 & \text{otherwise} \end{cases}$$

→ Does not depend on a width to calibrate



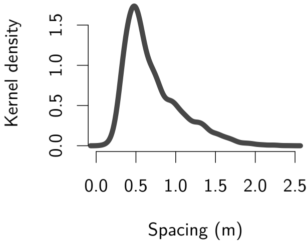


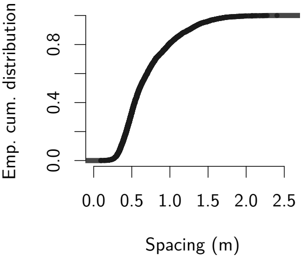
Cumulative distribution function — R : `ecdf(x)`

Empirical cumulative distribution function (ECDF)

$$D(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \leq x}, \qquad \text{with} \quad \mathbb{1}_R = \begin{cases} 1 & \text{if } R \\ 0 & \text{otherwise} \end{cases}$$

→ Does not depend on a width to calibrate



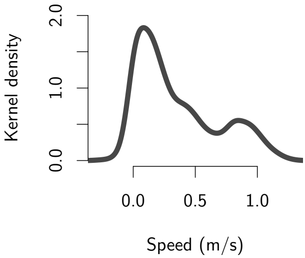


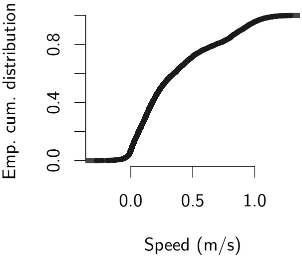
Cumulative distribution function — R : ecdf(x)

Empirical cumulative distribution function (ECDF)

$$D(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \le x}, \qquad \text{with} \quad \mathbb{1}_R = \begin{cases} 1 & \text{if } R \\ 0 & \text{otherwise} \end{cases}$$

→ Does not depend on a width to calibrate





Order statistic and quantile — R : `sort(x)`, `quantile(x,·)`

Univariate data : $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$

$(i_1, \ldots, i_n)$ is a permutation of the ID $(1, \ldots, n)$ such that $x_{i_1} \leq x_{i_2} \leq \ldots \leq x_{i_n}$

- k-**th order statistic** : $x^{(k)} = x_{i_k}, \qquad k = 1, \ldots, n$
  → k is the rank variable : $k - 1$ observations smaller, $n - k + 1$ bigger
- α-**quantile** : $q_x(\alpha) = x^{([\alpha n])}, \qquad \alpha \in [0, 1]$
  → α % of the data smaller, $1 - \alpha$ % bigger

Order statistic and quantile — R : sort(x), quantile(x,·)

Univariate data : $x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$

$(i_1, \dots, i_n)$ is a permutation of the ID $(1, \dots, n)$ such that $x_{i_1} \leq x_{i_2} \leq \dots \leq x_{i_n}$

- k**-th order statistic** : $x^{(k)} = x_{i_k}, \quad k = 1, \dots, n$
  - → k is the rank variable : $k-1$ observations smaller, $n-k+1$ bigger
- α**-quantile** : $q_x(\alpha) = x^{([\alpha n])}, \quad \alpha \in [0, 1]$
  - → α % of the data smaller, $1-\alpha$ % bigger

* Unique values if $x_{i_1} < x_{i_2} < \dots < x_{i_n}$

* Minimum and maximum values are : $\min_i x_i = q_x(0) = x^{(1)}$, $\max_i x_i = q_x(1) = x^{(n)}$

* Statistics stable by monotone transformation f :

$$(f(x))^{(k)} = \begin{cases} f\left(x^{(k)}\right) \\ f\left(x^{(n-1-k)}\right) \end{cases} \quad \text{and} \quad q_{f(x)}(\alpha) = \begin{cases} f\left(q_x(\alpha)\right) & \text{if } f \nearrow \\ f\left(q_{fx}(1-\alpha)\right) & \text{if } f \searrow \end{cases}$$

Statistic for the location — R : mean(x), median(x)

Three main statistics for the central position of univariate data $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$

- **Arithmetic mean** value (or mean value) $\bar{x} = \frac{1}{n} \sum_i x_i$ — `R : mean(x)`
- **Median** (central observation) $med_x = x^{([n/2])} = q_x(0.5)$ — `median(x)`
- **Mode** (most probable value) $mod_x = sup_z \, \mathrm{PDF}_x(z)$ — `x[pdf(x)==max(pdf(x))]`

Statistic for the location — R : `mean(x)`, `median(x)`

Three main statistics for the central position of univariate data $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$

- **Arithmetic mean** value (or mean value) $\bar{x} = \frac{1}{n} \sum_i x_i$ — R : `mean(x)`
- **Median** (central observation) $med_x = x^{([n/2])} = q_x(0.5)$ — `median(x)`
- **Mode** (most probable value) $mod_x = sup_z \, \mathrm{PDF}_x(z)$ — `x[pdf(x)==max(pdf(x))]`

* $\bar{x} = med_x = mod_x$ for uni-modal symmetric repartition of the data
* Mean and median solution of : $\bar{x} = \arg\min_a \sum_i (x_i - a)^2$ and $med_x = \arg\min_a \sum_i |x_i - a|$
* Mean sensible to extreme values, median or mode not : If $x_i \to \infty$ then $\bar{x} \to \infty$ but $med_x, mod_x \not\to \infty$
* Median and mode stable by monotone transform $med_{f(x)} = f(med_x), \; mod_{f(x)} = f(mod_x)$

But the mean is not :

$$\frac{1}{n} \sum_i f(x_i) \quad \begin{matrix} \leq \\ = \\ \geq \end{matrix} \quad f(\bar{x}) \quad \begin{matrix} \text{if } f \text{ is concave} \\ \text{if } f \text{ is affine} \\ \text{if } f \text{ is convex} \end{matrix}$$

(Jensen inequality)

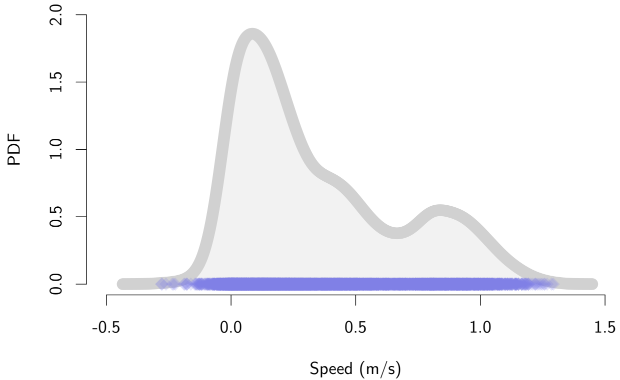
Other statistics for the location

Average		Example (1, 2, 3)	R
Harmonic	$\bar{x}_H = \left(\frac{1}{n}\sum_i 1/x_i\right)^{-1}$	1.65	`1/mean(1/x)`
Geometric	$\bar{x}_G = \sqrt[n^{-1}]{\prod_i x_i}$	1.82	`prod(x)^{1/length(x)}`
Arithmetic	$\bar{x}_A = \frac{1}{n}\sum_i x_i$	2	`mean(x)`
Quadratic	$\bar{x}_Q = \sqrt{\frac{1}{n}\sum_i x_i^2}$	2.16	`sqrt(mean(x^2))`
Contraharmonic	$\bar{x}_T = \sum_i x_i^2 / \sum_i x_i$	2.33	`mean(x^2)/mean(x)`

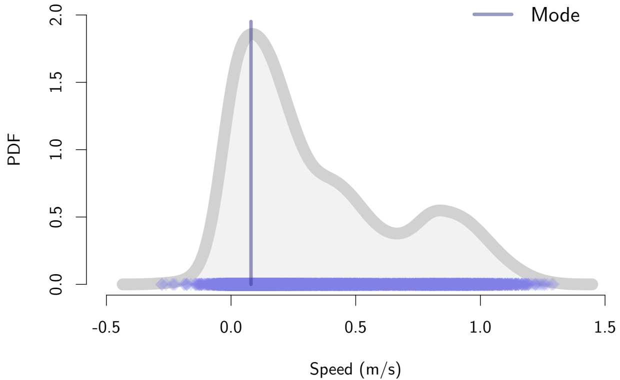
→ If $x_i > 0$ for all i, then we have[2]: $\bar{x}_H \leq \bar{x}_G \leq \bar{x}_A \leq \bar{x}_Q \leq \bar{x}_T$

[2] We have more generally for $x_i > 0$ and $\bar{X}_m = \sqrt[m^{-1}]{\frac{1}{N}\sum_i x_i^m}$, $\bar{X}_m \leq \bar{X}_n$ for all $m \leq n$

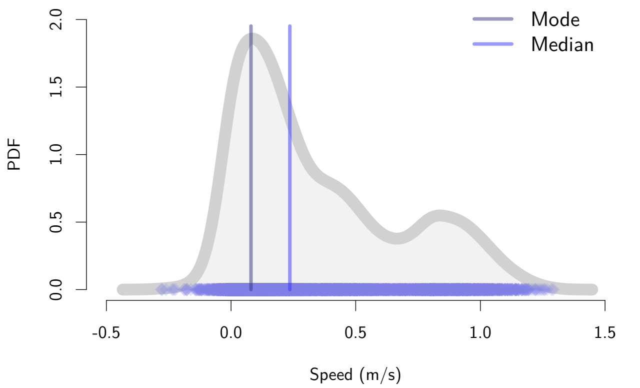
Example



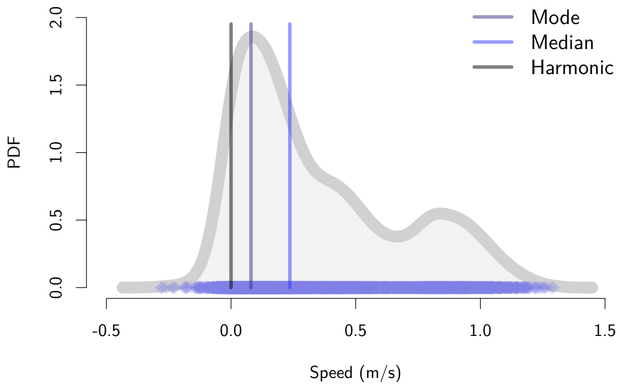
Example



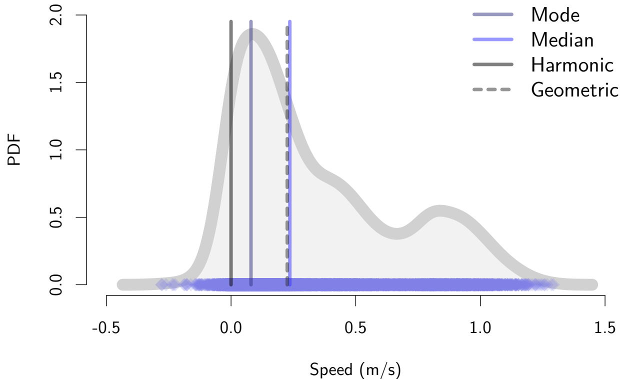
Example



Example

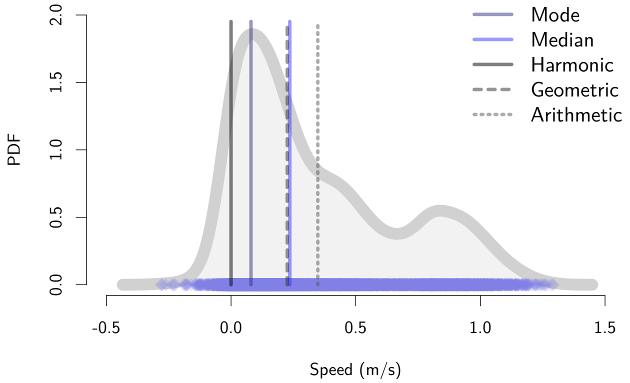


Example

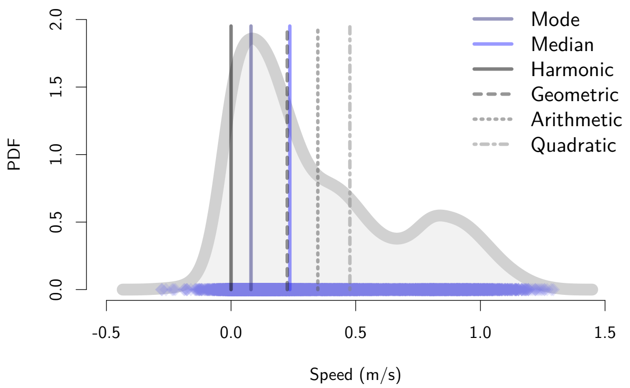


Mode
Median
Harmonic
Geometric
PDF
0.0
0.5
1.0
1.5
2.0
-0.5
0.0
0.5
1.0
1.5
Speed (m/s)

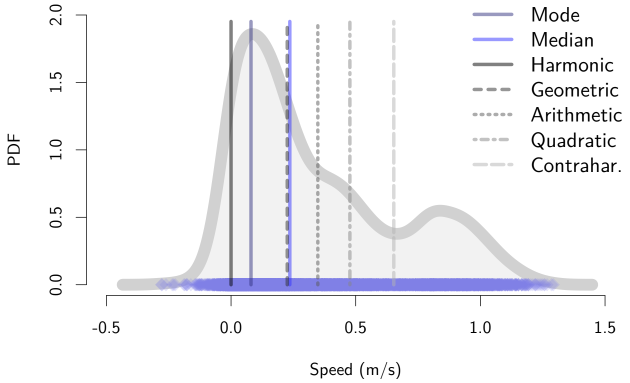
Example


Mode
Median
Harmonic
Geometric
Arithmetic
PDF
0.0
0.5
1.0
1.5
2.0
-0.5
0.0
0.5
1.0
1.5
Speed (m/s)

Example


Mode
Median
Harmonic
Geometric
Arithmetic
Quadratic
PDF
0.0
0.5
1.0
1.5
2.0
-0.5
0.0
0.5
1.0
1.5
Speed (m/s)

Example



PDF
0.0
0.5
1.0
1.5
2.0
-0.5
0.0
0.5
1.0
1.5
Speed (m/s)
Mode
Median
Harmonic
Geometric
Arithmetic
Quadratic
Contrahar.

Scattering statistics — `R : var(x), sd(x), ...`

Main statistics used to measure the variability of $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$

- **Variance** $var_x = \frac{1}{n} \sum_i (x_i - \bar{x})^2$ `R : var(x)`
- **Standard-deviation** $s_x = \sqrt{var_x}$ `sd(x)`
- **Mean absolute error** $abs\,dev_x = \frac{1}{n} \sum_i |x_i - \bar{x}|$ `mean(abs(x-mean(x)))`
- **Inter-quartile range** $IQR_x = q_x(0.75) - q_x(0.25)$ `quantile(x,.75)-quantile(x,.25)`
- **Max–Min difference** $max\,min_x = \max_i x_i - \min_i x_i$ `max(x)-min(x)`

Scattering statistics — R : `var(x)`, `sd(x)`, ...

Main statistics used to measure the variability of $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$

- **Variance** $var_x = \frac{1}{n} \sum_i (x_i - \bar{x})^2$ — R : `var(x)`
- **Standard-deviation** $s_x = \sqrt{var_x}$ — `sd(x)`
- **Mean absolute error** $abs\,dev_x = \frac{1}{n} \sum_i |x_i - \bar{x}|$ — `mean(abs(x-mean(x)))`
- **Inter-quartile range** $IQR_x = q_x(0.75) - q_x(0.25)$ — `quantile(x,.75)-quantile(x,.25)`
- **Max–Min difference** $max\,min_x = \max_i x_i - \min_i x_i$ — `max(x)-min(x)`

* All these statistics are positive and have the units of the data, excepted the variance (unit to the square)
* We have $s_x \geq abs\,dev_x$ and $\max_i x_i - \min_i x_i \geq IQR_x$
* Statistics stable by affine transformation

$$s_{ax+b} = |a|\, s_x, \quad IQR_{ax+b} = |a|\, IQR_x, \quad var_{ax+b} = a^2 var_x$$
$$abs\,dev_{ax+b} = |a|\, abs\,dev_x, \quad max\,min_{ax+b} = |a|\, max\,min_x,$$

Other statistics for the shape of a distribution

Skewness quantifies the symmetry of the distribution

R : `skewness(x)`

$$S_x = \frac{1}{n s_x^3} \sum_i (x_i - \bar{x})^3$$

- $S < 0$: Left asymmetry — Large left tail
- $S = 0$: Symmetric distribution — Similar left and right tails
- $S > 0$: Right asymmetry — Large right tail

Other statistics for the shape of a distribution

Skewness quantifies the symmetry of the distribution

R : `skewness(x)`

$$S_x = \frac{1}{ns_x^3} \sum_i (x_i - \bar{x})^3$$

- $S < 0$: Left asymmetry — Large left tail
- $S = 0$: Symmetric distribution — Similar left and right tails
- $S > 0$: Right asymmetry — Large right tail

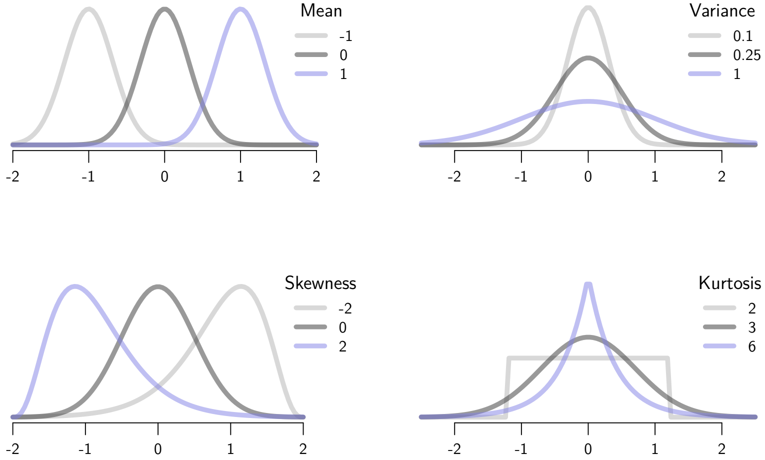
Kurtosis quantifies whether a distribution is straight or centred

R : `kurtosis(x)`

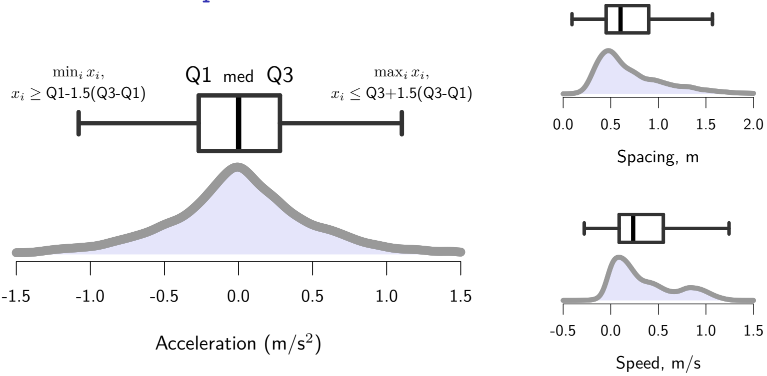
$$K_x = \frac{1}{ns_x^4} \sum_i (x_i - \bar{x})^4$$

- $K < 0$: Tailness distribution — Straight distribution
- $K > 0$: Distribution with tails — Centred distribution

Statistics for the shape of a distribution : illustrative examples



Boxplot — R : boxplot(x)



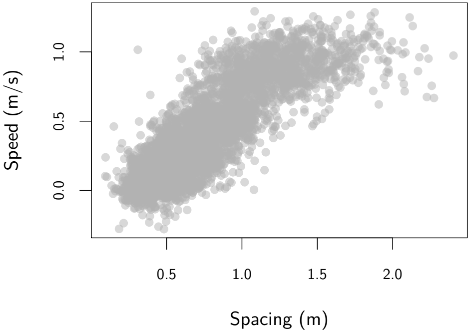
- 50% of the data into the box — 50% right (resp. left) to the median
- Normal distribution : $\geq 95\%$ of the data into the whiskers
- Different definitions for the whiskers exit (0.01/0.99-quantiles, min/max, ...)

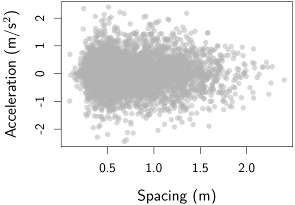
Descriptive statistics for bivariate data

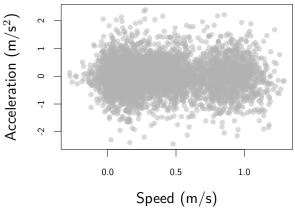
$$\big((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\big) \in \mathbb{R}^{2n}$$

Scatter plot — R : plot(x,y), plot(db)

Scatter plot : The 2D plot of bivariate data







Covariance and correlation — R : cov(x,y), cor(x,y)

One considers $(x, y) = ((x_1, y_1), \ldots, (x_n, y_n))$ some bivariate data

- **The covariance** quantifies how two variables fluctuate together

$$covar_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \in \mathbb{R}$$

- **The correlation** quantifies how two variables *linearly* fluctuate together (linear or Pearson correlation coefficient)

$$cor_{x,y} = \frac{covar_{x,y}}{\sqrt{var_x var_y}} \in [-1, 1]$$

Covariance and correlation — R : cov(x,y), cor(x,y)

One considers $(x, y) = ((x_1, y_1), \ldots, (x_n, y_n))$ some bivariate data

- **The covariance** quantifies how two variables fluctuate together

$$covar_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \in \mathbb{R}$$

- **The correlation** quantifies how two variables *linearly* fluctuate together (linear or Pearson correlation coefficient)

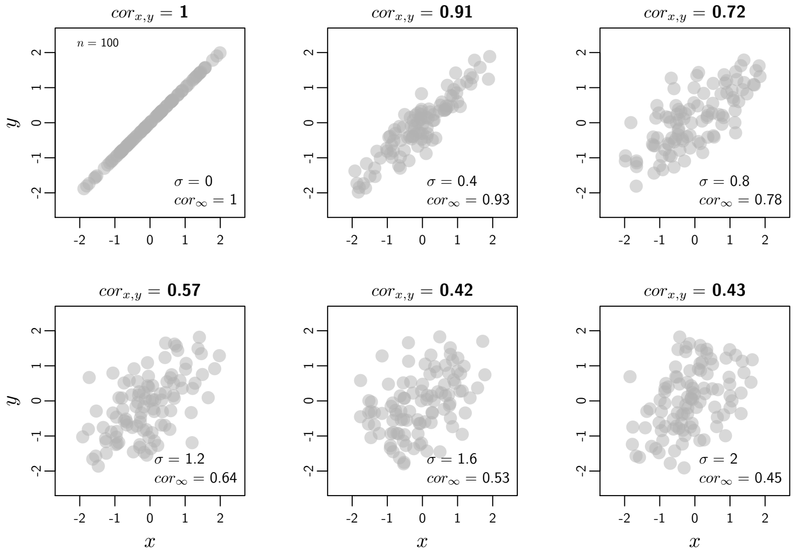
$$cor_{x,y} = \frac{covar_{x,y}}{\sqrt{var_x var_y}} \in [-1, 1]$$

* Covariance and correlation tend to zero as $n \to \infty$ if x and y are independent
* The correlation $cor_{x,y} = |1|$ if and only if x and y are linked by an affine relation
* Symmetric, $covar_{x,x} = var_x$, $covar_{ax+b,cy+d} = ac\ covar_{x,y}$, $cor_{ax+b,cy+d} = \pm cor_{x,y}$

Correlation : Illustrative example

$$y_i = (x_i + \sigma z_i)(1 + \sigma^2)^{-1/2}$$

$cor_{x,y} \rightarrow cor_\infty = (1 + \sigma^2)^{-1/2}$ as $n \rightarrow \infty$



Spearman correlation coefficient — R : `cor(x,y,method='spearman')`

Pearson correlation coefficient allows to assess linear relationships
→ Spearman correlation coefficient extends the assessment to any monotonic relationships

We denote by (rg_x) and (rg_y) the ranks of the variables $(x, y) = ((x_1, y_1), \ldots, (x_n, y_n))$

- **Spearman correlation coefficient**

$$cor^s_{x,y} = cor_{r_x,r_y} = \frac{covar_{r_x,r_y}}{\sqrt{var_{r_x} var_{r_y}}} \in [-1, 1]$$

Spearman correlation coefficient — R : `cor(x,y,method='spearman')`

Pearson correlation coefficient allows to assess linear relationships

→ Spearman correlation coefficient extends the assessment to any monotonic relationships

We denote by (rg_x) and (rg_y) the ranks of the variables $(x, y) = ((x_1, y_1), \ldots, (x_n, y_n))$

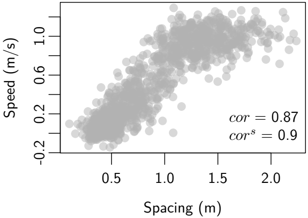
- **Spearman correlation coefficient**

$$cor^s_{x,y} = cor_{r_x,r_y} = \frac{covar_{r_x,r_y}}{\sqrt{var_{r_x} var_{r_y}}} \in [-1, 1]$$

* Stable by any monotonic transformation
* Insensitive to extreme values

$$cor^s_{x,y} = \frac{6 \sum_i d_i^2}{n(n^2-1)} \text{ with } d_i = r_{x_i} - r_{y_i}$$

if all n ranks are distinct integers

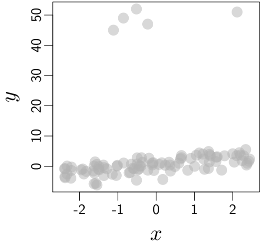


Correlation : Remark 1 — *Low correlation $\not\Rightarrow$ independent variables !*

⚠ Extreme values annihilate Pearson correlation

If $y_i = x_i \ \forall i \neq i'$ and $y_{i'} = \gamma$, then $covar_{x,y} \to 0$ as $\gamma \to \pm\infty$

$cor_{x,y} =$ **0.15 / 0.63**

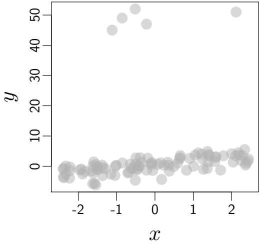


Correlation : Remark 1 — *Low correlation $\not\Rightarrow$ independent variables !*

⚠ Extreme values annihilate Pearson correlation

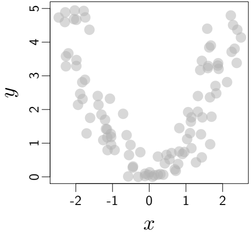
If $y_i = x_i \ \forall i \neq i'$ and $y_{i'} = \gamma$, then $covar_{x,y} \to 0$ as $\gamma \to \pm\infty$

$cor_{x,y} =$ **0.15 / 0.63**



⚠ Symmetric non-linear relations can have correlations nil

$cor_{x,y} =$ **-0.07 / -0.03**



see also `Wikipedia : Correlation`

Correlation : Remark 2 — *Correlation is not causality !*

Simple cause/consequence relationships have high correlation coefficients

⚠ **However high correlation coefficient** $\not\Rightarrow$ **Cause/Consequence relationship**

→ Both variables can be the consequence of the same cause without being linked, or can have just by chance similar trends

Correlation : Remark 2 — *Correlation is not causality !*

Simple cause/consequence relationships have high correlation coefficients

⚠ **However high correlation coefficient $\not\Rightarrow$ Cause/Consequence relationship**

→ Both variables can be the consequence of the same cause without being linked, or can have just by chance similar trends

Illustrative examples

1. Researchers initially believed that electrical towers impact the health because life expectation and living distance to electrical towers are significantly negatively correlated

   ⇝ Further analysis shown that this due to the fact that people living around electrical towers are generally poor, with fewer access to healthcare

2. *Shadoks* scientist found significant correlations between the number of times someone eats his birthday cake and having a long life ...

   ⇝ He deduced that eating his birthday cake is very healthy !

Some useful properties

Mean value

- Mean of a sum is the sum of the means $\overline{x+y} = \bar{x} + \bar{y}$
- Stable for the product if the variables are linearly independent $\overline{xy} = \bar{x}\bar{y}$, if x and y ind.
  In general : $\overline{xy} = \bar{x}\bar{y} + covar(x, y)$

Variance and covariance

- Variance stable by sum when the variables are linearly independent
  In general : $var(x + y) = var(x) + var(y) + 2covar(x, y)$
- Variance of a product is always bigger than the product of the variances
  If x and y are linearly independent : $var(xy) = var(x)var(y) + var(x)\bar{y} + var(y)\bar{x}$
- In general : $var(x) = \overline{x^2} - \bar{x}^2$ and $covar(x, y) = \overline{xy} - \bar{x}\bar{y}$

QQplot — R : `qqplot(x,y)`

Correlations quantify existence of linear or monotonic relationship

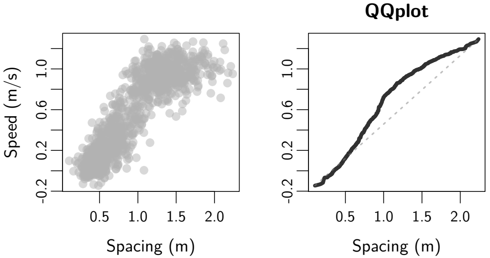
→ More generally, QQplots (quantile/quantile plots) allow to qualitatively compare two distributions

QQplot — R : qqplot(x,y)

Correlations quantify existence of linear or monotonic relationship

→ More generally, QQplots (quantile/quantile plots) allow to qualitatively compare two distributions

- **Affine relationship** if the curve is a straight line
- **Distributions are the same** if the curve is $x \mapsto x$
- **Different distributions** in the other cases

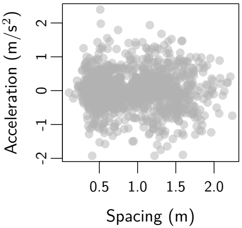


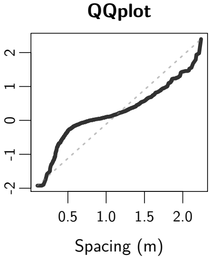
QQplot — R : qqplot(x,y)

Correlations quantify existence of linear or monotonic relationship

→ More generally, QQplots (quantile/quantile plots) allow to qualitatively compare two distributions

- **Affine relationship** if the curve is a straight line
- **Distributions are the same** if the curve is $x \mapsto x$
- **Different distributions** in the other cases



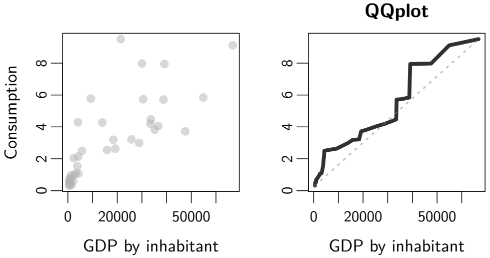


QQplot — R : qqplot(x,y)

Correlations quantify existence of linear or monotonic relationship

→ More generally, QQplots (quantile/quantile plots) allow to qualitatively compare two distributions

- **Affine relationship** if the curve is a straight line
- **Distributions are the same** if the curve is $x \mapsto x$
- **Different distributions** in the other cases

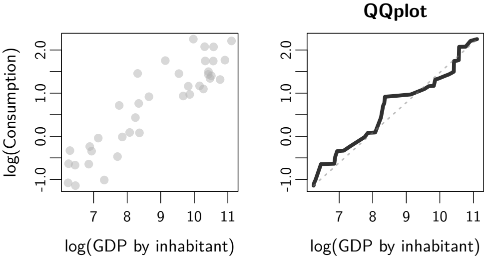


QQplot — R : qqplot(x,y)

Correlations quantify existence of linear or monotonic relationship

→ More generally, QQplots (quantile/quantile plots) allow to qualitatively compare two distributions

- **Affine relationship** if the curve is a straight line
- **Distributions are the same** if the curve is $x \mapsto x$
- **Different distributions** in the other cases



Summary with R

Univariate data

```
# Histogram
hist(x)

# Kernel density
density(x)

# Cumulative distribution function
ecdf(x)

# Quantile, order statistic
quantile(x,0.5);sort(x)

# Mean value, Median
mean(x);median(x)

# Variance, standard deviation
var(x);sqrt(var(x))

# Boxplot
boxplot(x)
```

Bivariate data

```
# Scatter plot
plot(x,y)

# Covariance
cov(x,y)

# Correlation
cor(x,y)

# QQplot
qqplot(y,x)
```

Overview

Part 1 | Descriptive statistics for univariate and bivariate data

Repartition of the data (histogram, kernel density, empirical cumulative distribution function), order statistic and quantile, statistics for location and variability, boxplot, scatter plot, covariance and correlation, QQplot

Part 2 | Descriptive statistics for multivariate data

Least squares and linear and non-linear regression models, principal component analysis, principal component regression, clustering methods (K-means, hierarchical, density-based), linear discriminant analysis, bootstrap technique, artificial neural networks

Part 3 | Parametric statistic

Likelihood, estimator definition and main properties (bias, convergence), punctual estimate (maximum likelihood estimation, Bayesian estimation), confidence and credible intervals, information criteria, test of hypothesis, parametric clustering

Appendix LaTeX plots with R and Tikz

Content

Multivariate data : Large database with observation of several characteristics of individuals

- *Exploring analysis* Analyse of the distribution of the data and correlation of the characteristics (Knowledge discovery and data mining)

  → Database for p characteristics : $(x_i^1, x_i^2, \ldots, x_i^p),\ i = 1, \ldots, n$

- *Prediction analysis* Prediction of certain characteristics (variable to explain) as function of the others (explanatory variable)

  → Database : $(y_i, x_i^1, x_i^2, \ldots, x_i^p),\ i = 1, \ldots, n$

Content

Multivariate data : Large database with observation of several characteristics of individuals

- *Exploring analysis* Analyse of the distribution of the data and correlation of the characteristics (Knowledge discovery and data mining)

  $\rightarrow$ Database for p characteristics : $(x_i^1, x_i^2, \dots, x_i^p),\ i = 1, \dots, n$

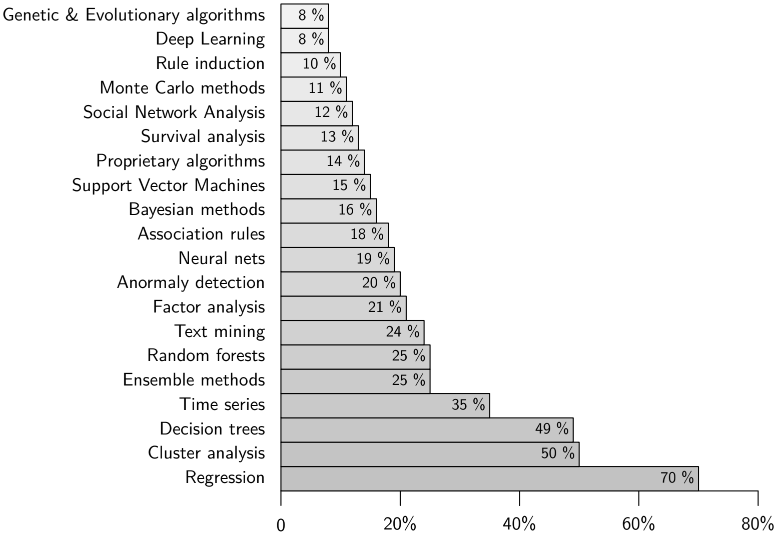
- *Prediction analysis* Prediction of certain characteristics (variable to explain) as function of the others (explanatory variable)

  $\rightarrow$ Database : $(y_i, x_i^1, x_i^2, \dots, x_i^p),\ i = 1, \dots, n$

- **Linear and non-linear regression** — Prediction analysis
- **Principal component analysis** — Exploring analysis
- **Clustering analysis** — Exploring analysis
- **Bootstrap technique** — Exploring and prediction analysis
- **Artificial neural network** — Prediction analysis

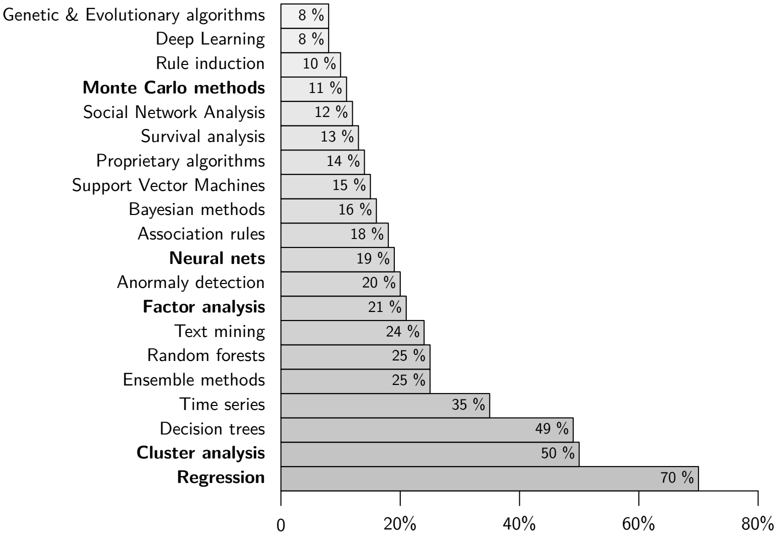
The algorithms data scientists are using

Source : Rexer Analytics, 2016



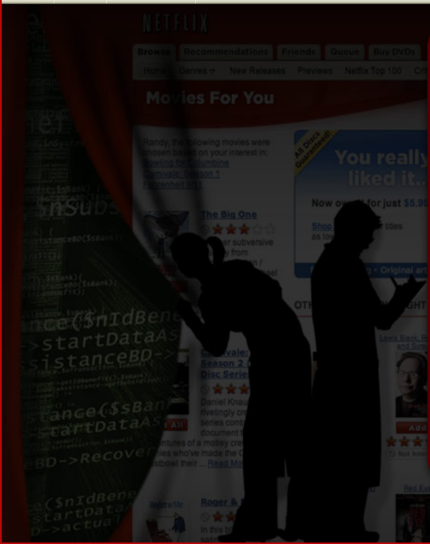
The algorithms data scientists are using

Source : Rexer Analytics, 2016



Netflix Prize

COMPLETED



Congratulations!

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the $1M Grand Prize to team “BellKor’s Pragmatic Chaos”. Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.

We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

Netflix Prize

- **Netflix dataset** : More than 100 million datestamped movie ratings performed by anonymous Netflix customers between Dec 31, 1999 and Dec 31, 2005 (about 480 189 users and 17 770 movies)
- **Training-test set format** : A hold-out set of about 4.2 million ratings was created consisting of the last nine movies rated by each user[3] — Remaining data made up the training set
- **Winner** : "BellKor's Pragmatic Chaos" — Blend of hundreds of different models, Test RMSE : 0.856704 (10.06%)

  "The Ensemble Team" — Blend of 24 prediction models, Test RMSE : 0.856714 (10.06%)

  → BellKor's defeated The Ensemble by submitting just 20 minutes earlier !

[3] or fewer if a user had not rated at least 18 movies over the entire period

DARPA Urban Challenge (2007)

- **Driverless car competition** on a 96 kilometres (60 mi) urban area course, to be completed in less than 6 hours (Nov. 3, 2007 in Victorville, California)

- **Rules** :
  - *Vehicle must be stock or have a documented safety record*
  - *Vehicle must obey the California state driving laws*
  - *Vehicle must be entirely autonomous, using only the information it detects with its sensors and public signals such as GPS*
  - *DARPA will provide the route network 24 hours before the race starts*
  - *Vehicles will complete the route by driving between specified checkpoints*
  - *DARPA will provide a file detailing the checkpoints to 5 minutes before the race start*
  - *Vehicles may "stop and stare" for at most 10 seconds*
  - *Vehicles must operate in rain and fog, with GPS blocked*
  - *Vehicles must avoid collision with vehicles and other objects such as carts, bicycles or traffic barrels*
  - *Vehicles must be able to operate in parking areas and perform U-turns*

DARPA Urban Challenge : Winner

- "Tartan Racing" with Chevrolet Tahoe (Carnegie Mellon University and Pittsburgh Pennsylvania
- Performed the course in 4:10:20 (averaged speed approximately 22.5 kilometre per hour)
- Algorithm is a blend of tens statistical prediction models (regression, neural networks, clustering, etc. . .)

DARPA Urban Challenge : Winner

- "Tartan Racing" with Chevrolet Tahoe (Carnegie Mellon University and Pittsburgh Pennsylvania
- Performed the course in 4:10:20 (averaged speed approximately 22.5 kilometre per hour)
- Algorithm is a blend of tens statistical prediction models (regression, neural networks, clustering, etc. . .)

→ In most of the cases, complex multivariate statistic problems are tackled with combinations of many different statistical algorithms
(Ensemble learning methods)

Regression models

Introduction

Multivariate data $(y_i, x_i^1, \ldots, x_i^p),\ i = 1, \ldots, n$

- $n \times (p+1)$ matrix: n observations of $p+1$ characteristics

| y is the *variable to explain* (output or regressant) — Continuous
| $x^1, \ldots, x^p$ are the p *explanatory variables* (inputs or regressors) — Discrete or continuous

Introduction

Multivariate data $(y_i, x_i^1, \ldots, x_i^p),\ i = 1, \ldots, n$

- $n \times (p+1)$ matrix : n observations of $p+1$ characteristics

y is the *variable to explain* (output or regressant)	Continuous
$x^1, \ldots, x^p$ are the p *explanatory variables* (inputs or regressors)	Discrete or continuous

Model $M_\alpha : \mathbb{R}^p \mapsto \mathbb{R}$ for y as a function of the $(x^1, \ldots, x^p)$

$$y = M_\alpha(x^1, \ldots, x^p) + \sigma\mathcal{E}$$

- α are the parameters and $\sigma\mathcal{E}$ is a *noise* (or an error) with amplitude σ (unexplained part)

Example : Multiple linear model $M_\alpha(x^1, \ldots, x^p) = \alpha_0 + \alpha_1 x^1 + \ldots + \alpha_p x^p$

$\rightarrow$ $p+2$ parameters : $(\alpha_0, \alpha_1, \ldots, \alpha_p)$ and σ — Simple linear regression for $p = 1$

Estimation of the parameters by least squares

Non-parametric estimation of the parameters by least squares

(or ordinary least squares (OLS), or regression model)

$$\tilde{\alpha} = \arg\min_{\alpha} \sum_{i=1}^{n} \left(y_i - M_\alpha\left(x_i^1, \ldots, x_i^j\right) \right)^2$$

Estimation of the parameters by least squares

Non-parametric estimation of the parameters by least squares

(or ordinary least squares (OLS), or regression model)

$$\tilde{\alpha} = \arg\min_{\alpha} \sum_{i=1}^{n} \Big(y_i - M_\alpha(x_i^1, \ldots, x_i^j) \Big)^2$$

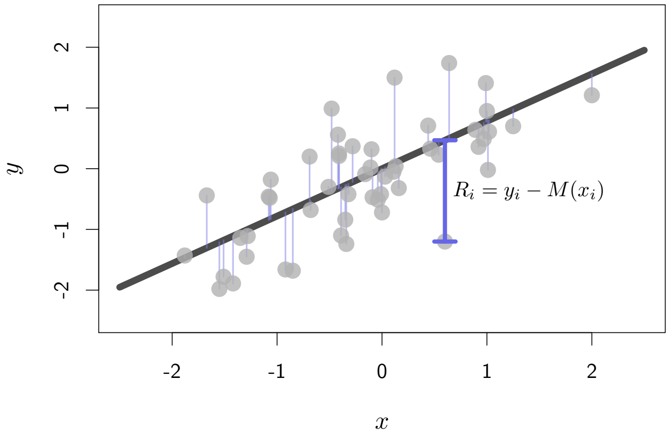
Residuals : $R_\alpha(y, x^1, \ldots, x^p) = y - M_\alpha(x^1, \ldots, x^p)$

- OLS : Minimisation of the variance of the residuals / Sensible to extreme values
- Estimation of the amplitude of the noise using the empirical residual variance

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} R_{\tilde{\alpha}}^2(y_i, x_i^1, \ldots, x_i^p)$$

Estimation of the parameters by least squares

Minimisation of the variance of the residuals



Goodness of the fit

Evaluation of the goodness through the repartition of the variability

- $SST = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$ **Total Sum of Squares**
- $SSM = \sum_{i=1}^{n} \left(\bar{M} - M_{\tilde{\alpha}}(x_i)\right)^2$ **Sum of Squares of the Model**
- $SSR = \sum_{i=1}^{n} \left(y_i - M_{\tilde{\alpha}}(x_i)\right)^2$ **Sum of Squared Residuals**

Residuals centred and linearly independent: $SST = SSM + SSR$

$\rightarrow$ Minimizing the variance of residuals maximizes variance explained by the model

Goodness of the fit

Evaluation of the goodness through the repartition of the variability

- $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ **Total Sum of Squares**
- $SSM = \sum_{i=1}^{n} (\bar{M} - M_{\tilde{\alpha}}(x_i))^2$ **Sum of Squares of the Model**
- $SSR = \sum_{i=1}^{n} (y_i - M_{\tilde{\alpha}}(x_i))^2$ **Sum of Squared Residuals**

Residuals centred and linearly independent : $SST = SSM + SSR$

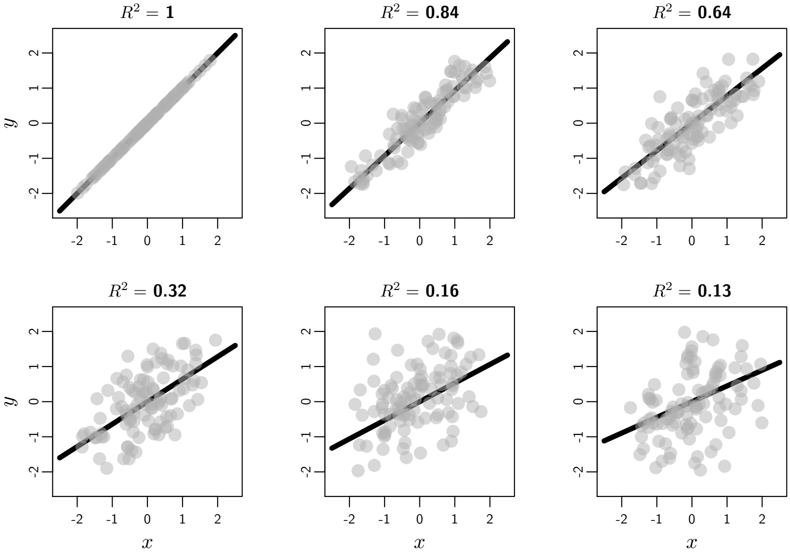
$\rightarrow$ Minimizing the variance of residuals maximizes variance explained by the model

Coefficient of determination *Explained proportion of the variance*

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSR}{SST} \leq 1$$

$\rightarrow$ Good fit if $R^2 \approx 1$ — OLS estimation maximizes the R^2 — If $p = 1$ then $R^2 = cor^2_{x,y}$

R^2 : Example



$R^2 = \mathbf{1}$
$R^2 = \mathbf{0.84}$
$R^2 = \mathbf{0.64}$
$R^2 = \mathbf{0.32}$
$R^2 = \mathbf{0.16}$
$R^2 = \mathbf{0.13}$
y
x
-2
-1
0
1
2

Linear regression — R : `lm(y ∽ x)`

Matrix notations of the multiple linear model :

$$y = X\alpha, \qquad \left| \begin{array}{lcll} y & = & (y_1, \dots, y_n)^t & \text{the variable to explain} \\ X & = & (1_n, x^1, \dots, x^p) & \text{the matrix of the regressors} \\ \alpha & = & (\alpha_0, \dots, \alpha_p)^t & \text{the parameters} \end{array} \right.$$

Linear regression — R : lm(y ∽ x)

Matrix notations of the multiple linear model :

$$y = X\alpha, \qquad \left| \begin{array}{lll} y & = & (y_1, \dots, y_n)^t \\ X & = & (1_n, x^1, \dots, x^p) \\ \alpha & = & (\alpha_0, \dots, \alpha_p)^t \end{array} \right.$$

- y: the variable to explain
- X: the matrix of the regressors
- α: the parameters

OLS estimation of the parameters : $\tilde{\alpha} = (X^tX)^{-1}X^ty$

Formal proof : $\forall j = 1, \dots, p, \ \frac{\partial}{\partial \tilde{\alpha}_j} \sum_i (y_i - \tilde{\alpha}_0 - \tilde{\alpha}_1 x_i^1 - \dots - \tilde{\alpha}_p x_i^p)^2 = 0$

$\Leftrightarrow \quad \forall j = 1, \dots, p, \ \sum_i x_i^j (y_i - \tilde{\alpha}_0 - \tilde{\alpha}_1 x_i^1 - \dots - \tilde{\alpha}_p x_i^p) = 0$

$\Leftrightarrow \quad X^t(y - X\tilde{\alpha}) = 0 \quad \Leftrightarrow \quad \tilde{\alpha} = (X^tX)^{-1}X^ty$

Generalized Least Squares (GLS) estimation : $\tilde{\alpha}^G = (X^t\Omega^{-1}X)^{-1}X^t\Omega^{-1}y$

→ Variance/Covariance matrix Ω for the residuals

Simple linear regression

Bivariate data $(x, y) = ((x_1, y_1), \ldots, (x_n, y_n)) \in \mathbb{R}^2$

The linear regression of y on x is the straight line $y = a_{\text{OLS}} x + b_{\text{OLS}}$

$$(a_{\text{OLS}}, b_{\text{OLS}}) = \arg\min_{a,b} \sum_i (y_i - (ax_i + b))^2 \quad \Rightarrow \quad \begin{cases} a_{\text{OLS}} & = & \frac{covar_{x,y}}{var_x} \\ b_{\text{OLS}} & = & \bar{y} - a_{\text{OLS}} \bar{x} \end{cases}$$

Formal proof: We denote as $F(a, b) = \sum_i (y_i - (ax_i + b))^2$

$\partial F / \partial a = 0$ and $\partial F / \partial b = 0$ is $\begin{cases} \sum_i (-x_i y_i + x_i b + x_i^2 a) & = & 0 \\ \sum_i (y_i + x_i a + b) & = & 0 \end{cases}$

On obtains $a = \frac{cov_{x,y}}{var_x}$ and $b = \bar{y} - a\bar{x}$

→ Regressions y/x and x/y are not the same as soon as $var_x \neq var_y$ but both cross $(\bar{x}_n, \bar{y}_n)$

Linear and non-linear regression

Non-linear regression by invertible (monotone) non-linear transformation of the data

- Linear regression with the variables x and $f(y)$, $f(x)$ and y or $f(x)$ and $f(y)$

> **Example :** Exponential model $M_\alpha = e^{\alpha_0} \cdot (x^1)^{\alpha_1} \dots (x^p)^{\alpha_p}$
> $\rightarrow$ Linear model with $\tilde{x} = \log(x)$ and $\tilde{y} = \log(y)$

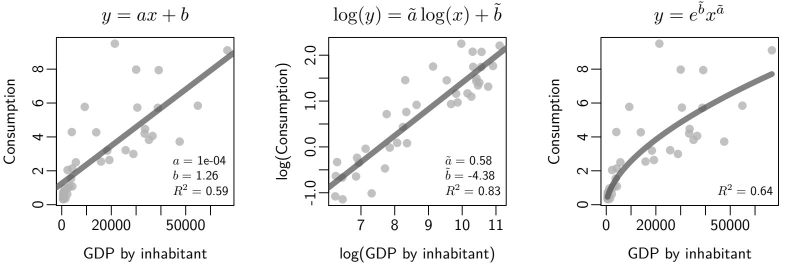
Linear and non-linear regression

Non-linear regression by invertible (monotone) non-linear transformation of the data

- Linear regression with the variables x and $f(y)$, $f(x)$ and y or $f(x)$ and $f(y)$

Example : Exponential model $M_\alpha = e^{\alpha_0} \cdot (x^1)^{\alpha_1} \ldots (x^p)^{\alpha_p}$

$\rightarrow$ Linear model with $\tilde{x} = \log(x)$ and $\tilde{y} = \log(y)$



Linear and non-linear regression

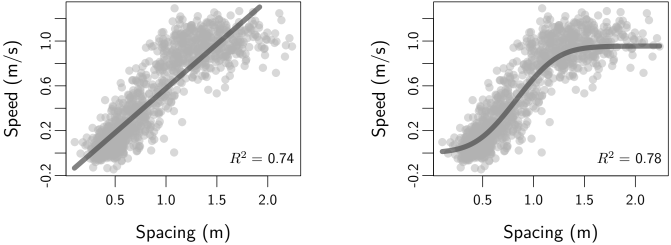
Non-invertible model : Linearisation of the problem and numerical solution

- Iterative algorithms based on the partial derivatives of the model (Jacobian matrix)
- R : `nls(model,data)` Gauss-Newton or Golub-Pereyra algorithms
- Local minima and divergence problems possible

Linear and non-linear regression

Non-invertible model : Linearisation of the problem and numerical solution

- Iterative algorithms based on the partial derivatives of the model (Jacobian matrix)
- R : `nls(model,data)` Gauss-Newton or Golub-Pereyra algorithms
- Local minima and divergence problems possible



Multiple linear and non-linear regression with R

y, x1, x2 and x3 are vectors with the same size

Linear least squares estimate

```
lm(y ∽ x1 + x2 + x3)
```

- Linear regression of y on x1, x2 and x3
- Linear model (with intercept nil) : `lm(y ∽ 0 + x1 + x2 + x3)`

Non-linear least squares estimate

```
nls(y ∽ mod(x,p1,p2,p3,...))
```

- The model must be at least derivable — Default method : Gauss–Newton
- Partial derivative can be given as input or are estimated numerically

Regression models : Summary

- Regression models allow to describe relationships between a variable to explain and explanatory factors
  - Parameter estimations by *least squares method* (sensitivity to extreme values)
  - *Linear* (explicit solution) and *non-linear* (invertible transformation or numerical approximation) models

- The variability of the variable to explain can be decomposed as
  - *Variability explained by the model* (explained part)
  - *Variability of the residuals* (non-explained part)

  $\rightarrow$ The $R^2 \in [0, 1]$ is the proportion of variable explained by the model allowing to compare models and to evaluate the quality of the fit

- Linear and non-linear regression are very easy to implement in R

  $\rightarrow$ `lm(·)` and `nls(·)` functions — `coef(·)` to get the estimations of the coefficients

Principal Component Analysis

Introduction

Multivariate data : Observations of p characteristics of n individuals

$$X = \begin{bmatrix} x_1^1 & x_1^2 & \dots & x_1^p \\ x_2^1 & x_2^2 & \dots & x_2^p \\ \vdots & \vdots & & \vdots \\ x_n^1 & x_n^2 & \dots & x_n^p \end{bmatrix} \in (\mathbb{R}^p)^n, \quad \left| \begin{array}{ll} x_i = (x_i^1, \dots, x_i^p), & i = 1, \dots, n \\ x^j = (x_1^j, \dots, x_n^j)^t, & j = 1, \dots, p \end{array} \right.$$

→ Variables $(x^1, \dots, x^p)$ are correlated (inter-dependence of the characteristics)

Introduction

Multivariate data : Observations of p characteristics of n individuals

$$X = \begin{bmatrix} x_1^1 & x_1^2 & \dots & x_1^p \\ x_2^1 & x_2^2 & \dots & x_2^p \\ \vdots & \vdots & & \vdots \\ x_n^1 & x_n^2 & \dots & x_n^p \end{bmatrix} \in (\mathbb{R}^p)^n, \quad \left| \begin{array}{ll} x_i = (x_i^1, \dots, x_i^p), & i = 1, \dots, n \\ x^j = (x_1^j, \dots, x_n^j)^t, & j = 1, \dots, p \end{array} \right.$$

→ Variables $(x^1, \dots, x^p)$ are correlated (inter-dependence of the characteristics)

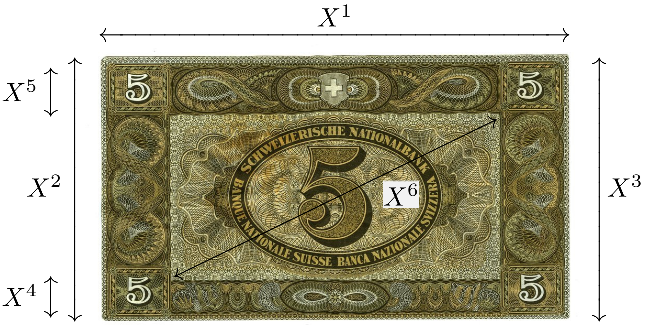
Specific tools for the visualisation and description of multivariate data

- **Scatterplots** — By coupling the variables — $p(p-1)$ plots
- **Parallel plots**, **Andrews plot**, **radar charts** — Different geometrical representations
- **Chernoff faces** — Human face representation
- **Principal component analysis** — Decomposition in principal components

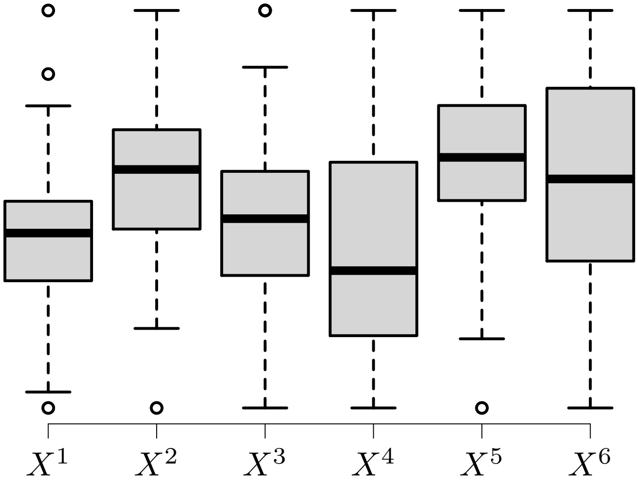
Example

Six measurements of Swiss banknotes ($n = 200$ observations, $p = 6$)

$\rightarrow$ Some are **authentic**, some are **counterfeit**



Boxplot — R : `boxplot(database)` Normed data

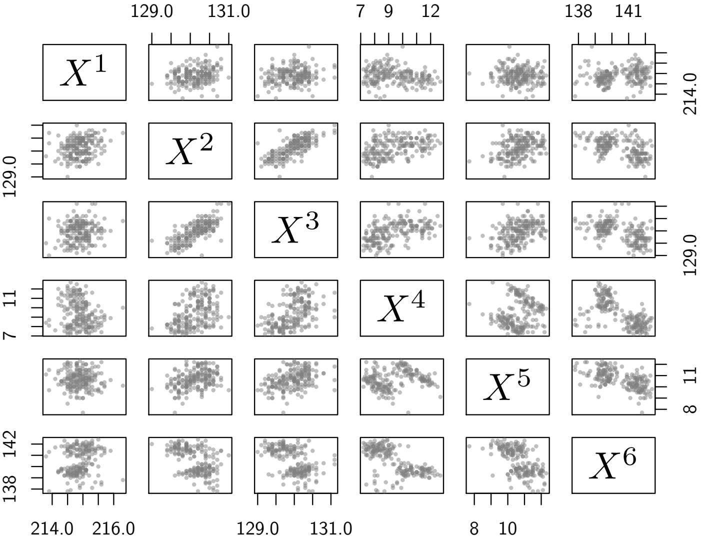


Correlation coefficients

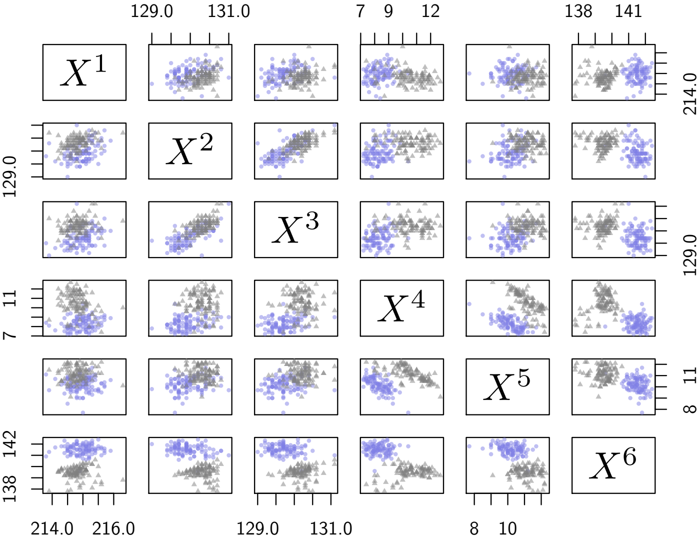
	X^1	X^2	X^3	X^4	X^5	X^6
X^1	1.00	0.23	0.15	-0.19	-0.06	0.19
X^2	0.23	1.00	**0.74**	0.41	0.36	**-0.50**
X^3	0.15	**0.74**	1.00	**0.49**	0.40	**-0.52**
X^4	-0.19	0.41	**0.49**	1.00	0.14	**-0.62**
X^5	-0.06	0.36	0.40	0.14	1.00	**-0.59**
X^6	0.19	**-0.50**	**-0.52**	**-0.62**	**-0.59**	1.00

- X^2 and X^3 are highly correlated
- X^4 and X^5 are highly correlated to X^3
- X^6 is highly correlated to all the variables excepted X^1

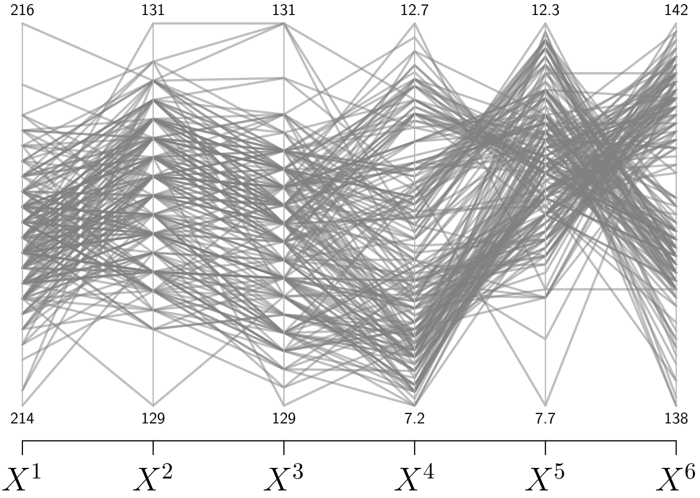
Scatterplot — `R : plot(database)`



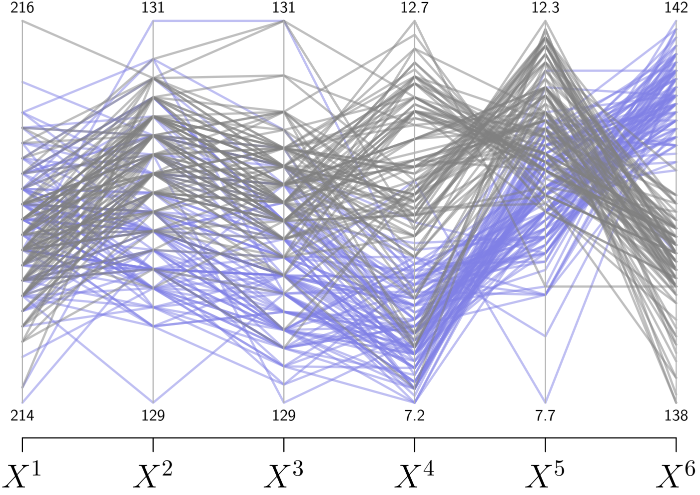
Scatterplot — R : plot(database)



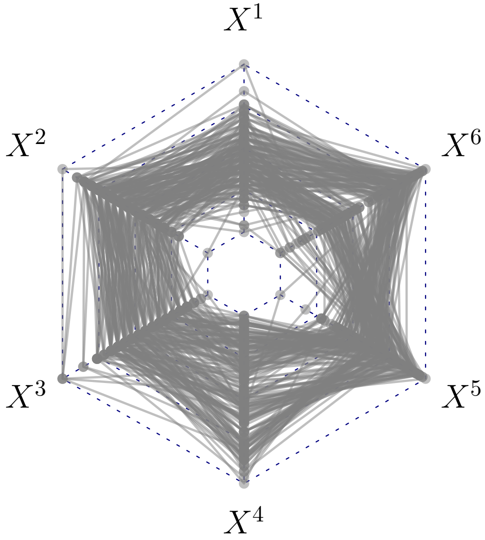
Parallel plots — `R : parcoord(database)` Package : MASS



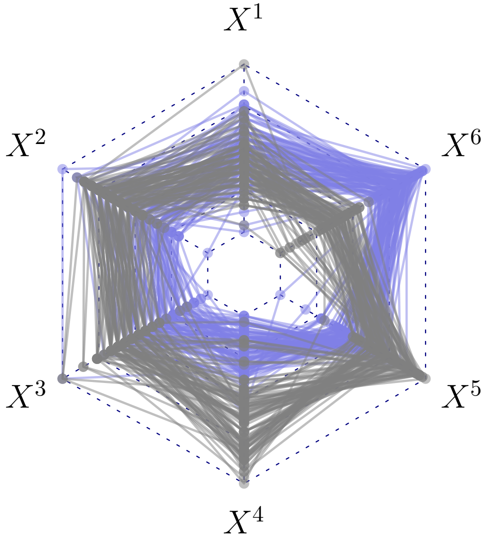
Parallel plots — R : parcoord(database) Package : MASS



Radar charts — R : radarchart(database) Package : fmsb

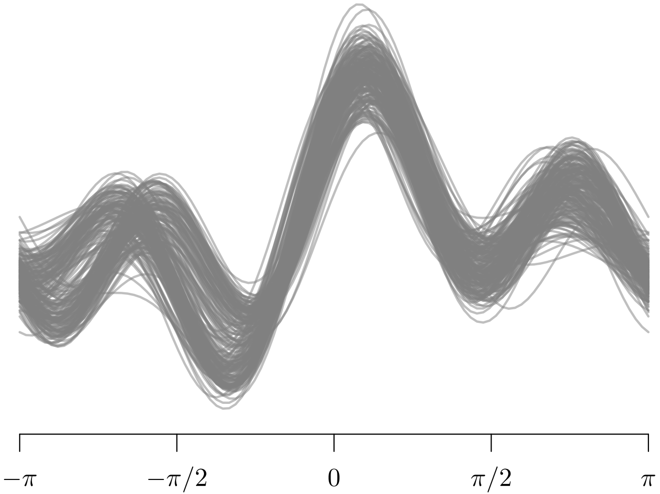


Radar charts — `R : radarchart(database)` Package : `fmsb`



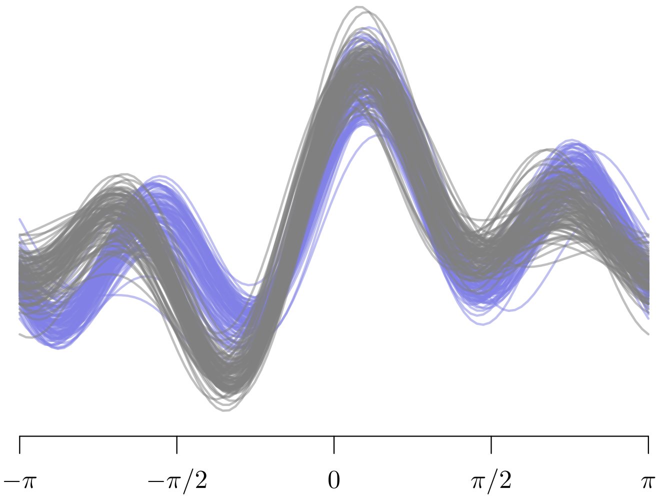
Andrews plots — `R : andrews(database)` Package : andrews

$X^1\cos(t) + X^2\sin(t) + X^3\cos(2t) + X^4\sin(2t) + X^5\cos(3t) + X^6\sin(3t)$



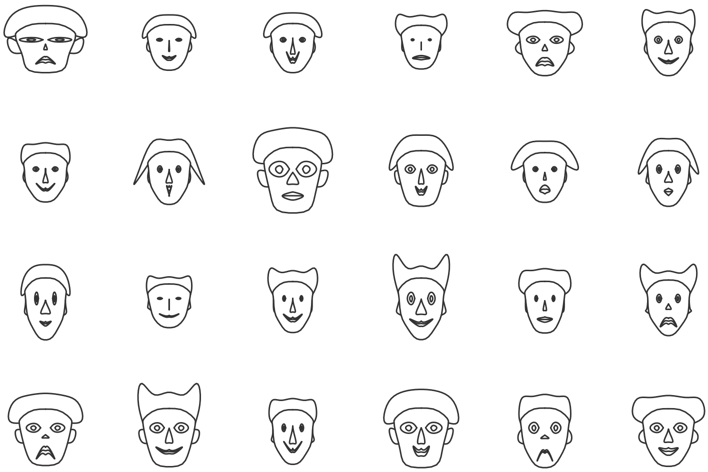
Andrews plots — `R : andrews(database)` `Package : andrews`

$X^1 \cos(t) + X^2 \sin(t) + X^3 \cos(2t) + X^4 \sin(2t) + X^5 \cos(3t) + X^6 \sin(3t)$



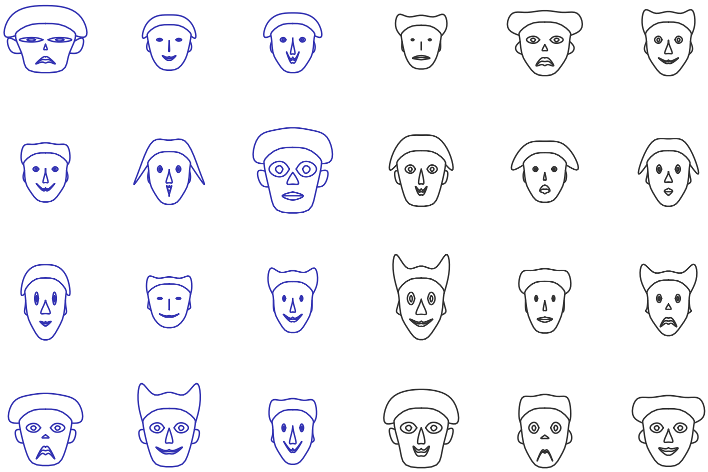
Chernoff faces — R : faces(database) Package : aplpack

$i = 1, \ldots, 24$



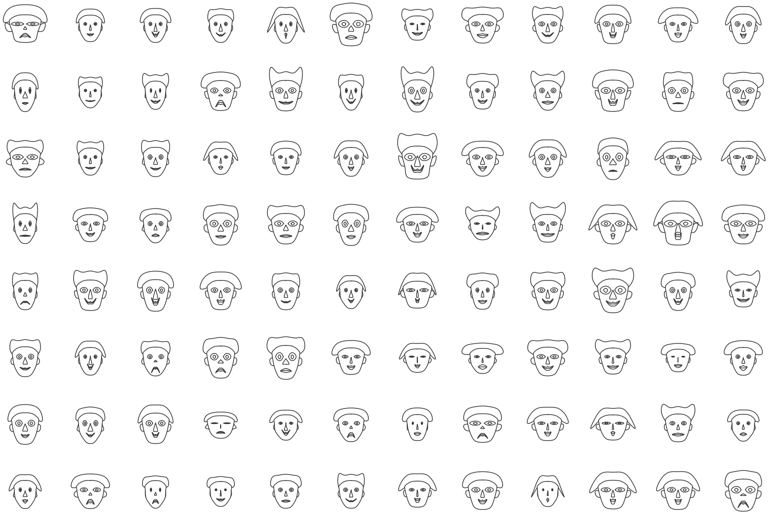
Chernoff faces — R : faces(database) Package : aplpack

$i = 1, \ldots, 24$



Chernoff faces — R : faces(database) Package : aplpack

$i = 1, \ldots, 96$



Chernoff faces — R : faces(database) Package : aplpack

$i = 1, \ldots, 96$

Principal component analysis (PCA)

PCA allows to explore large multivariate data $X = (x_i^1, \ldots, x_i^p)$, $i = 1, \ldots, n$

- The variable $(x^1, \ldots, x^p)$ are dependent (otherwise individual analyse !) and continuous (PCA for categorical data : *Multiple correspondence analysis*)
- The dimension p is high and the visualisation of the global structure of the data is difficult
- Correlated variable bring same information and could be resumed as linear combinations (i.e. principal factors) to reduce the dimension of the database

Principal component analysis (PCA)

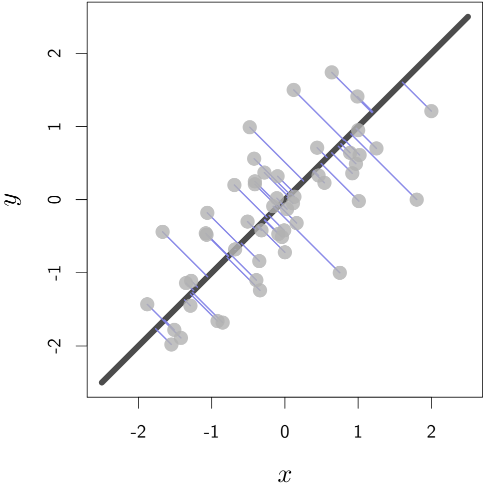
PCA allows to explore large multivariate data $X = (x_i^1, \ldots, x_i^p)$, $i = 1, \ldots, n$

- The variable $(x^1, \ldots, x^p)$ are dependent (otherwise individual analyse !) and continuous (PCA for categorical data : *Multiple correspondence analysis*)
- The dimension p is high and the visualisation of the global structure of the data is difficult
- Correlated variable bring same information and could be resumed as linear combinations (i.e. principal factors) to reduce the dimension of the database

Principle : Reduction of the dimension with uncorrelated linear combinations of $(x^1, \ldots, x^p)$ maximising the variability

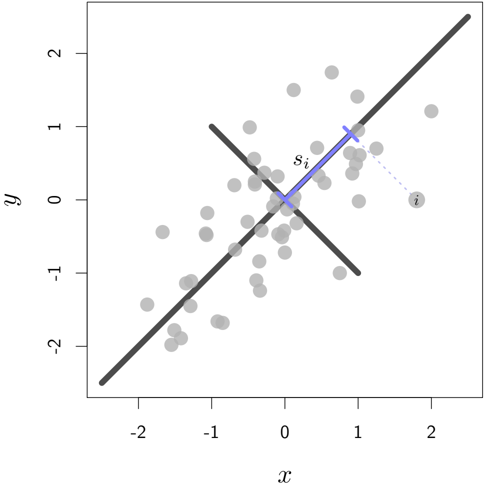
- Geometric interpretation : Projection of the data in orthogonal basis maximising the variance (i.e. the information – other criteria may be used)
- The 1st component is an optimal representation of the data in one dimension, 1st and 2nd components optimal representation of the data in two dimensions, and so on

PCA : Maximisation of the variance



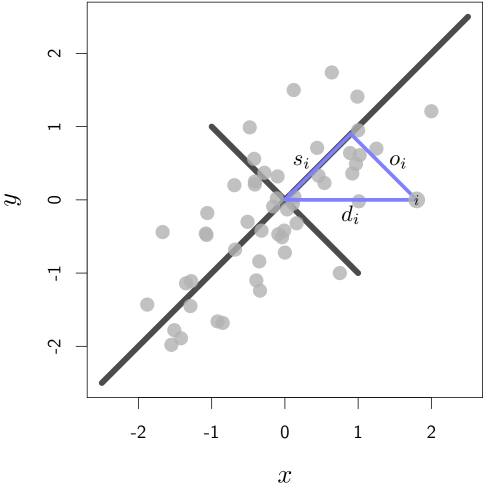
- Orthogonal projection

PCA : Maximisation of the variance



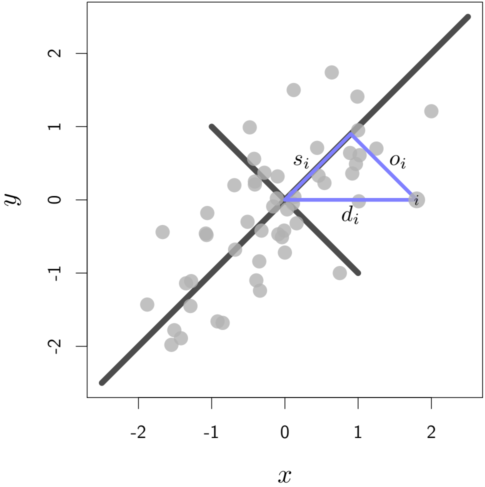
- Orthogonal projection
- Maximisation of the variance $\sum_i s_i^2$

PCA : Maximisation of the variance



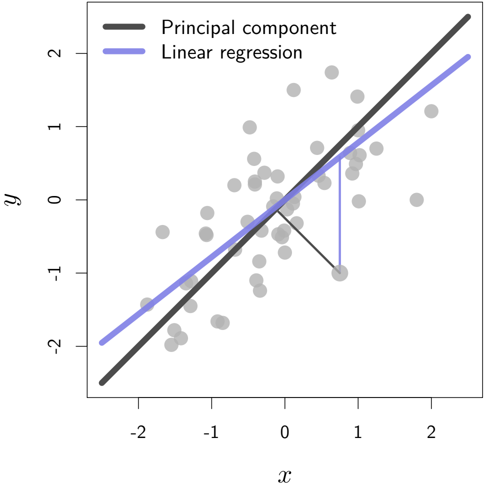
- Orthogonal projection
- Maximisation of the variance $\sum_i s_i^2$
- $\forall i,\ d_i^2 = o_i^2 + s_i^2$ constant in any direction (distance to the center)
  $\Rightarrow \sum_i o_i^2 + \sum_i s_i^2 = C$

PCA : Maximisation of the variance



- Orthogonal projection
- Maximisation of the variance $\sum_i s_i^2$
- $\forall i,\ d_i^2 = o_i^2 + s_i^2$ constant in any direction (distance to the center)
  $\Rightarrow \sum_i o_i^2 + \sum_i s_i^2 = C$
- Maximising the variance $\Leftrightarrow$ Minimising orthogonal squared distances

PCA : Maximisation of the variance

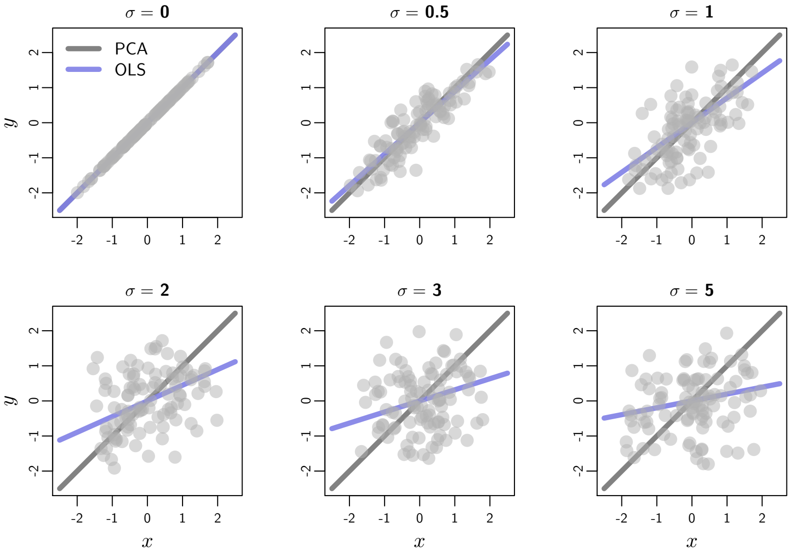


- Orthogonal projection
- Maximisation of the variance $\sum_i s_i^2$
- $\forall i,\ d_i^2 = o_i^2 + s_i^2$ constant in any direction (distance to the center)
  $\Rightarrow \sum_i o_i^2 + \sum_i s_i^2 = C$
- Maximising the variance $\Leftrightarrow$ Minimising orthogonal squared distances
- Principal component $\neq$ linear regression

Example

$$y_i = (x_i + \sigma z_i)(1+\sigma^2)^{-1/2}$$

$a_{\text{PCA}} \to 1$ while $a_{\text{OLS}} \to (1+\sigma^2)^{-1/2}$ as $n \to \infty$



Construction of the components

Centred/Standard score transformation $x_i^j \to \tilde{x}_i^j = x_i^j - \bar{x}^j$ or $x_i^j \to \tilde{x}_i^j = \dfrac{x_i^j - \bar{x}^j}{s_{x^j}}$

- **Total variance** of the dataset

$$var_{\tilde{X}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\tilde{x}_i^j\right)^2 = \sum_{j=1}^{p} s^2_{\tilde{x}^j} \qquad (= p \text{ if std. score})$$

- $P_H \tilde{X}$ is the orthogonal projection of the data on subset H and $\tilde{X} - P_H \tilde{X}$ is the projection on a subset orthogonal to H, then (Pythagore)

$$var_{\tilde{X}} = var_{P_H \tilde{X}} + var_{\tilde{X} - P_H \tilde{X}}$$

- **PCA :** Iterative calculation of orthogonal unidimensional subsets (principal components) maximizing the variance

Construction of the components

Iterative construction of the components $(PC1, PC2, \ldots, PCp)$ as linear combinations of the centred data :

- $PC1 = \tilde{X}u_1$, u_1 such that var_{PC1} maximal
- $PC2 = \tilde{X}u_2$, $u_2 \perp u_1$ and var_{PC2} maximal
- $PC3 = \tilde{X}u_3$, $u_3 \perp (u_1, u_2)$ and var_{PC3} maximal

  $\vdots$
- $PCp = \tilde{X}u_p$, $u_p \perp (u_1, \ldots, u_{p-1})$ (unique)

Construction of the components

Iterative construction of the components $(PC1, PC2, \ldots, PCp)$ as linear combinations of the centred data :

- $PC1 = \tilde{X}u_1$, u_1 such that var_{PC1} maximal
- $PC2 = \tilde{X}u_2$, $u_2 \perp u_1$ and var_{PC2} maximal
- $PC3 = \tilde{X}u_3$, $u_3 \perp (u_1, u_2)$ and var_{PC3} maximal

  $\vdots$
- $PCp = \tilde{X}u_p$, $u_p \perp (u_1, \ldots, u_{p-1})$ (unique)

* The unit vectors $(u_1, u_2, \ldots, u_p)$ form an orthonormal basis of R^p — The last component is fixed
* By construction $var_{PC1} \geq var_{PC2} \geq \ldots \geq var_{PCp}$ and $\sum_j var_{PCj} = var_X$
* The first components contain most of the variability of the data when the initial variables are correlated

Construction with multivariate data

Variance/covariance matrix of the data Γ (diagonalizable $p \times p$ real and symmetric matrix)

$$\Gamma = \frac{1}{n} X^t X \qquad \left| \begin{array}{l} \Gamma_{j,j} = var_{\tilde{x}^j} = \frac{1}{n}\sum_i (\tilde{x}_i^j)^2, \\ \Gamma_{j,j'} = covar_{\tilde{x}^j, \tilde{x}^{j'}} = \frac{1}{n}\sum_i \tilde{x}_i^j \tilde{x}_i^{j'}, \end{array} \right. \qquad \forall j, j' \in \{1, \dots, p\}$$

- **Principal components** $PCj = \tilde{X} u_j$ described by eigenvectors and ordered eigenvalues of Γ

Formal proof: $\tilde{X}_v$ is the projection of the data X on axis subset $v \in \mathbb{R}^p$

$$\begin{aligned} var_{\tilde{X}_v} &= \frac{1}{n}\sum_j \sum_{j'} v_j v_{j'} \sum_i \tilde{x}_i^j \tilde{x}_i^{j'} = v^t \Gamma v \\ &= \sum_j \lambda_j \langle v, u_j \rangle^2 \leq \lambda_1 \sum_j \langle v, u_j \rangle^2 \leq \lambda_1 = var_{PC1} \end{aligned}$$

The axis v for which the variance is maximal is u_1 (and the variance is var_{PC1})

$\rightarrow$ Then for all $v \perp u_1$ (i.e. $\langle v, u_1 \rangle = 0$), the axis maximizing the variance is u_2 etc. . .

Construction with bivariate data

First component $PC1 = u\tilde{x} + \sqrt{1-u^2}\tilde{y}$ is the straight line $y = a_{\text{PCA}} x$ with $a_{\text{PCA}} = \frac{\sqrt{1-u^2}}{u}$ where u is such that

$$var_{\text{PC1}} \propto \sum_i \left(u\tilde{x}_i + \sqrt{1-u^2}\tilde{y}_i\right)^2 \quad \text{is maximal}$$

→ One finds $$a_{\text{PCA}} = \frac{var_y - var_x + \sqrt{\left(var_y - var_x\right)^2 + 4covar_{x,y}^2}}{2covar_{x,y}}$$

Construction with bivariate data

First component $PC1 = u\tilde{x} + \sqrt{1-u^2}\tilde{y}$ is the straight line $y = a_{\text{PCA}}x$ with $a_{\text{PCA}} = \frac{\sqrt{1-u^2}}{u}$ where u is such that

$$var_{\text{PC1}} \propto \sum_i \left(u\tilde{x}_i + \sqrt{1-u^2}\tilde{y}_i\right)^2 \quad \text{is maximal}$$

→ One finds

$$a_{\text{PCA}} = \frac{var_y - var_x + \sqrt{\left(var_y - var_x\right)^2 + 4covar_{x,y}^2}}{2covar_{x,y}}$$

- The slope for linear regression is $a_{\text{OLS}} = \frac{covar_{x,y}}{var_x}$
- If $y_i = ax_i$ for all i, then $a_{\text{PCA}} = a_{\text{OLS}} = a$ (since $covar_{xy} = a\,var_x$ and $var_y = a^2 var_x$)
- If $s_x = s_y$ then $a_{\text{PCA}} = \pm 1$, according to the sign of $covar_{x,y}$ (and $a_{\text{OLS}} = cor_{x,y}$)
- The second component has the slope $-1/a_{\text{PCA}}$

Properties of the components

- **Maximization of the variability** : $PC1$ *best* representation in 1D, $(PC1, PC2)$ *best* representation in 2D, . . .

Properties of the components

- **Maximization of the variability** : $PC1$ *best* representation in 1D, $(PC1, PC2)$ *best* representation in 2D, ...
- The principal components $(PC1, \ldots, PCp)$ are centred :

$$\forall j = 1, \ldots, p, \qquad P\bar{C}j = \frac{1}{n}\sum_{i=1}^{n} PCj_i = 0$$

Properties of the components

- **Maximization of the variability** : $PC1$ *best* representation in 1D, $(PC1, PC2)$ *best* representation in 2D, ...
- The principal components $(PC1, \ldots, PCp)$ are centred :

$$\forall j = 1, \ldots, p, \qquad \bar{PC}j = \frac{1}{n}\sum_{i=1}^{n} PCj_i = 0$$

- The principal components are not correlated, and with variance $(\lambda_1, \ldots, \lambda_p)$:

$$\forall j \neq j', \qquad cov_{PCj,PCj'} = \frac{1}{n}\sum_{i=1}^{n} PCj_i PCj'_i = \lambda_j u_j^t u_{j'} = \begin{cases} \lambda_j & \text{if } j = j' \\ 0 & \text{if } j \neq j' \end{cases}$$

$\rightarrow$ **This does not imply that the principal components are independent**

Only the linear relations are resumed : Observation of non-linear phenomena

Properties of the components

- **Maximization of the variability** : $PC1$ *best* representation in 1D, $(PC1, PC2)$ *best* representation in 2D, ...
- The principal components $(PC1, \dots, PCp)$ are centred :

$$\forall j = 1, \dots, p, \qquad P\bar{C}j = \frac{1}{n}\sum_{i=1}^{n} PCj_i = 0$$

- The principal components are not correlated, and with variance $(\lambda_1, \dots, \lambda_p)$:

$$\forall j \neq j', \qquad cov_{PCj,PCj'} = \frac{1}{n}\sum_{i=1}^{n} PCj_i PCj'_i = \lambda_j u_j^t u_{j'} = \begin{cases} \lambda_j & \text{if } j = j' \\ 0 & \text{if } j \neq j' \end{cases}$$

$\rightarrow$ **This does not imply that the principal components are independent**

Only the linear relations are resumed : Observation of non-linear phenomena

- Interpretation of the components with the correlations to the initial variables

$$\forall j, j' \in \{1, \dots, p\}, \quad cor_{x^j, PCj'} = u^j_{j'} \sqrt{\lambda_{j'}} / s_{x^j}$$

Practical use of PCA

In practice, the PCA consists in :

1. Calculation of the variances of the principal components to select the number of new variables to take in consideration

   → **Plot of the proportions of variance per component** $\tau_j = \lambda_j / \sum_i \lambda_i$

Practical use of PCA

In practice, the PCA consists in :

1. Calculation of the variances of the principal components to select the number of new variables to take in consideration

   → **Plot of the proportions of variance per component** $\tau_j = \lambda_j / \sum_i \lambda_i$

2. Analysis of the correlations of the selected components with the initial variables to interpret the new variables

   → **Circle of the correlations plot**

Practical use of PCA

In practice, the PCA consists in :

1. Calculation of the variances of the principal components to select the number of new variables to take in consideration

   → **Plot of the proportions of variance per component** $\tau_j = \lambda_j / \sum_i \lambda_i$

2. Analysis of the correlations of the selected components with the initial variables to interpret the new variables

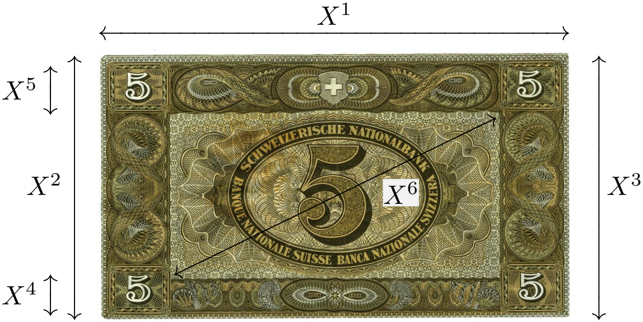
   → **Circle of the correlations plot**

3. Analysis of the components (linear and non-linear phenomena)

   → **Boxplot, scatter plots or clustering analysis of the new variables**

Example of the notes

Six measurements for the notes



Principal components — R : `prcomp(database)`

Rotations — Eigenvectors u_j

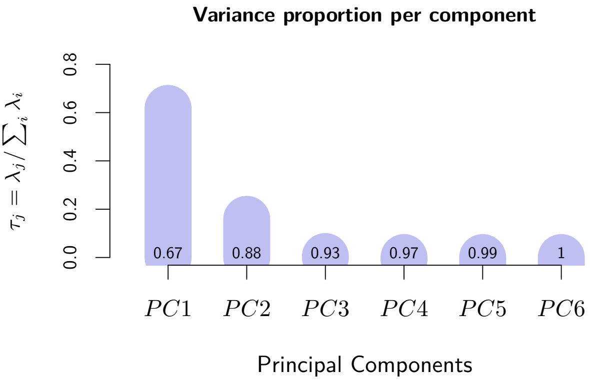
	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	$PC6$
X^1	0.04	-0.01	0.33	-0.56	-0.75	0.10
X^2	-0.11	-0.07	0.26	-0.46	0.35	-0.77
X^3	-0.14	-0.07	0.34	-0.42	0.53	0.63
X^4	-0.77	0.56	0.22	0.19	-0.10	-0.02
X^5	-0.20	-0.66	0.56	0.45	-0.10	-0.03
X^6	0.58	0.49	0.59	0.26	0.08	-0.05

Component variance — Eigenvalues λ_j

	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	$PC6$
λ	3.00	0.94	0.24	0.19	0.09	0.04
τ	0.67	0.21	0.05	0.04	0.02	0.01

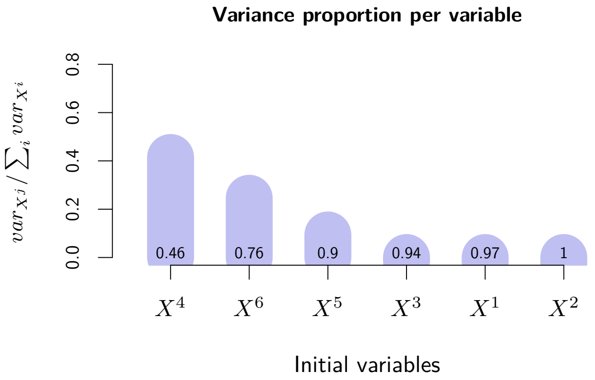
Plot of the proportions of variance per component

Selection of the component number



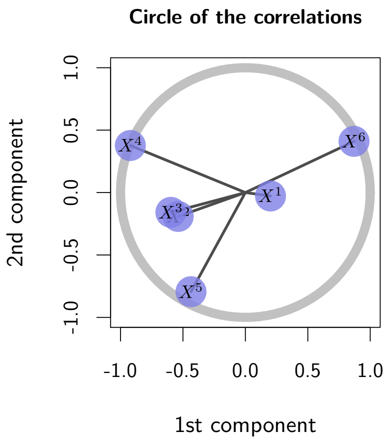
Plot of the proportions of variance per component

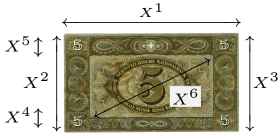
Selection of the component number



Plot of the circle of the correlations

Interpretation of the components

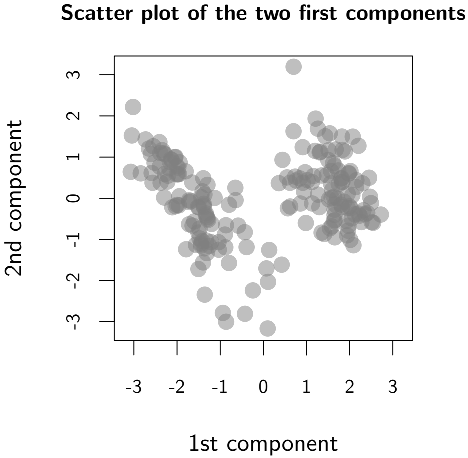




- **PC1** Large flag / Short border — Long / not large note
- **PC2** Large flag and down border / Short up border

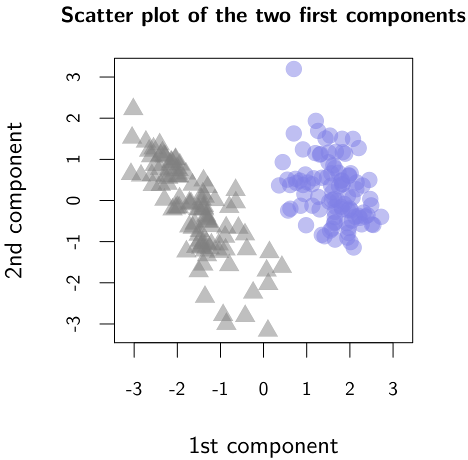
Scatter plot of the components

Analysis of the results



Scatter plot of the components

Analysis of the results



PCA with R

Read of the data : `data=read.table('C/...')`

- **Principal component analysis with R** `prcomp(M)`

  No standard score transformation of the data by default

  `prcomp(M,scale=T)` for PCA on standard scores

- **Basic example :**

```
pca=prcomp(data)
pca$rotations
pca$stddev
summary(pca)
```

Principal component regression

OLS estimation has interesting properties if regressors are linearly independent

→ Regression on the principal components

- **Principal components** : $p \times n$ matrix $PC = \hat{X} S U$

  $\hat{X}$ is the centred data ($\hat{x}_i^j \to x_i^j - \bar{x}^j$ for all i, j)
  $S = Diag(1/s_{x^1}, \ldots, 1/s_{x^p})$ is the diagonal $p \times p$ normalization matrix
  $U = (u_1, \ldots, u_p)$ is the $p \times p$ matrix of unit and orthogonal eigenvectors

- **Regression on the components** : $\hat{y} = \alpha_1^{PC} PC1 + \ldots + \alpha_p^{PC} PCp$

$$\tilde{\alpha}^{PC} = (PC^t PC) PC^t y = (SU)^{-1} (X^t X) X^t y = (SU)^{-1} \tilde{\alpha}$$

Principal component regression

OLS estimation has interesting properties if regressors are linearly independent

$\rightarrow$ Regression on the principal components

- **Principal components** : $p \times n$ matrix $PC = \hat{X} S U$

  $\hat{X}$ is the centred data ($\hat{x}_i^j \rightarrow x_i^j - \bar{x}^j$ for all i, j)

  $S = Diag(1/s_{x^1}, \ldots, 1/s_{x^p})$ is the diagonal $p \times p$ normalization matrix

  $U = (u_1, \ldots, u_p)$ is the $p \times p$ matrix of unit and orthogonal eigenvectors

- **Regression on the components** : $\hat{y} = \alpha_1^{PC} PC1 + \ldots + \alpha_p^{PC} PCp$

$$\tilde{\alpha}^{PC} = (PC^t PC) PC^t y = (SU)^{-1} (X^t X) X^t y = (SU)^{-1} \tilde{\alpha}$$

* The estimation using initial parameters is $\tilde{\alpha} = SU \tilde{\alpha}^{PC}$ and $\tilde{\alpha}_0 = \bar{y} - \frac{1}{n} \hat{X} \tilde{\alpha}$
* By shorting the regressors to the first principal components the model still depends on **all the initial variables**

Principal component analysis : Summary

PCA is a descriptive tool allowing to reduce the dimension of multivariate data

→ Then use of tools for low dimension data (uni- or bivariate)

The principal components are :

- **Linear combinations of the initial variables** — Linear transformation
- **Linearly independent** — By construction
- **Ordered by maximizing the variability** — Best representation in 1D, 2D, ...

Practical use of PCA :

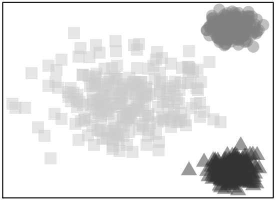
- **Number of components to analyse** — Proportion of variance per component
- **Interpretation of the new variables** — Circle of the correlations
- **Analysis of the components** — Scatter plot of the components

Clustering methods

Introduction

Clustering : Division of heterogeneous data in subsets (clusters)

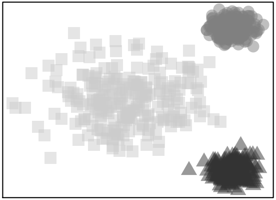
→ Observations in the same cluster are more similar (in some sense) to each other than to those in other subsets



Introduction

Clustering : Division of heterogeneous data in subsets (clusters)

→ Observations in the same cluster are more similar (in some sense) to each other than to those in other subsets



Possible distinctions (among others)

Supervised / unsupervised :	Clusters and cluster number are known / unknown
Strict clustering :	Each observation belongs to exactly one cluster
Strict clustering with outliers :	Observations can also belong to no cluster (outliers)
Overlapping clustering :	Observations may belong to more than one cluster
Fuzzy clustering :	Each observation belongs to each cluster according to a certain degree
Hierarchical clustering :	Observations of a child cluster also belong to the parent cluster
Centroid clustering :	Cluster represented by a centroid (mean value)
Density-based clustering :	Clustering based on empirical PDF estimation

K-means clustering — R : `kmeans(database,K)`

Observation $(x_1, \ldots, x_n)$, partition $S = \{S_1, \ldots, S_K\}$, mean by cluster $(u_1, \ldots, u_K)$

- **K-means :** Unsupervised clustering method based on mean by cluster
  - Clustering for given number of clusters K
  - (*K-medoid* : Clustering based on median by cluster
- **Minimization of the intra-cluster variability**

$$S = \text{arg}\min_S \sum_{j=1}^{K} \sum_{i \in S_j} \|x_i - u_j\|^2$$

K-means clustering — R : `kmeans(database,K)`

Observation $(x_1, \ldots, x_n)$, partition $S = \{S_1, \ldots, S_K\}$, mean by cluster $(u_1, \ldots, u_K)$

- **K-means :** Unsupervised clustering method based on mean by cluster
  - Clustering for given number of clusters K
  - (*K-medoid* : Clustering based on median by cluster
- **Minimization of the intra-cluster variability**

$$S = \text{arg}\min_S \sum_{j=1}^{K} \sum_{i \in S_j} \|x_i - u_j\|^2$$

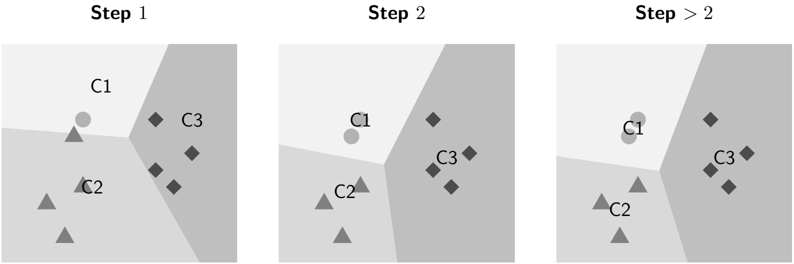
∗ Minimizing the intra-variability ⇔ Maximizing the inter-variability (Pythagore)

∗ Partition based on the Voronoi diagram for the means by cluster

∗ Calculation of the global minimum is a NP-complex problem

→ Iterative numerical algorithms (Hartigan-Wong, Lloyd-Forgy, ...) with convergence to local minima

K-means : Illustrative example with 3 clusters



* Convergence to steady state in 3 steps (the step's number depends on the initial partition / mean values)
* In this example the reached local optimum is the global one

Agglomerative hierarchical method (AHM) — `R : hclust(dist(data))`

Hierarchical method Unsupervised clustering based on tree representations

- Top of the tree : One cluster with all the observations
- Bottom of the tree : each observation is a cluster

Agglomerative hierarchical method (AHM) — `R : hclust(dist(data))`

Hierarchical method Unsupervised clustering based on tree representations

- Top of the tree : One cluster with all the observations
- Bottom of the tree : each observation is a cluster

Agglomerative iterative method Bottom up approach, by opposition to divisive methods

1. Initialization : Each observation is a cluster
2. Definition of the metric (Euclidean, Manhattan, Mahalanobis, maximum, ...)
3. Definition of a distance between two clusters – Linkage (max, min, mean, centroid, ...)
4. Repeat while `Cluster_number > 1` `{Merge_two_closest_clusters}`

Agglomerative hierarchical method (AHM) — R : `hclust(dist(data))`

Hierarchical method Unsupervised clustering based on tree representations

- Top of the tree : One cluster with all the observations
- Bottom of the tree : each observation is a cluster

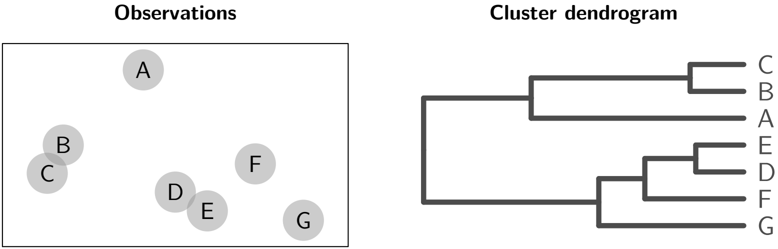
Agglomerative iterative method Bottom up approach, by opposition to divisive methods

1. Initialization : Each observation is a cluster
2. Definition of the metric (Euclidean, Manhattan, Mahalanobis, maximum, ...)
3. Definition of a distance between two clusters – Linkage (max, min, mean, centroid, ...)
4. Repeat while `Cluster_number > 1` {`Merge_two_closest_clusters`}

Dendrogram : Tree with observation in x-coordinate and distances in y-coordinate

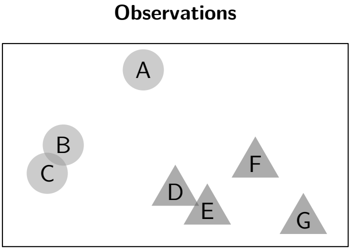
→ Cut of the dendrogram to determinate the number of clusters

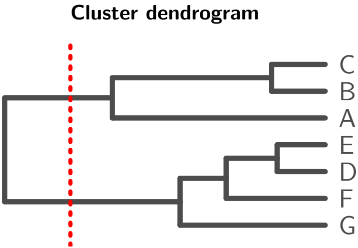
AHM : Illustrative example



- The dendrogram allows to summarize/represent the hierarchical clustering
- Cut of the dendrogram when the branches are long (cut at height h give groups having distance higher than h)

AHM : Illustrative example

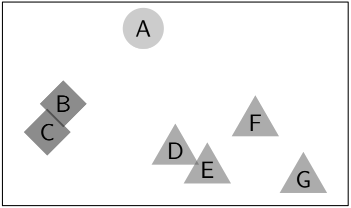




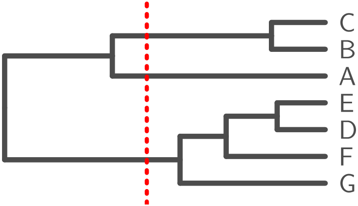
- The dendrogram allows to summarize/represent the hierarchical clustering
- Cut of the dendrogram when the branches are long (cut at height h give groups having distance higher than h)

AHM : Illustrative example

Observations



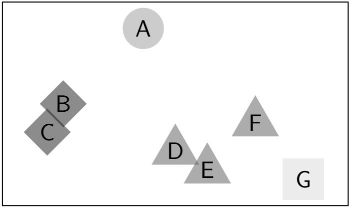
Cluster dendrogram



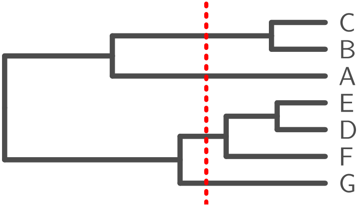
- The dendrogram allows to summarize/represent the hierarchical clustering
- Cut of the dendrogram when the branches are long (cut at height h give groups having distance higher than h)

AHM : Illustrative example

Observations



Cluster dendrogram



- The dendrogram allows to summarize/represent the hierarchical clustering
- Cut of the dendrogram when the branches are long (cut at height h give groups having distance higher than h)

Mean-shift clustering — `R : ms(database)` `Package LPMC`

K-means and AHM based on distances to quantify the similarities

→ Identification of circular cluster (Euclidean distance)

Mean-shift clustering Gradient-method based on kernel density estimate

- Iterative method allowing to detect local maximum of the kernel density
- Method calibrated by a bandwidth (to be given)
- Clustering : Threshold for local maxima (cluster number), kernel density gradient (cluster belonging)

Mean-shift clustering — `R : ms(database)` `Package LPMC`

K-means and AHM based on distances to quantify the similarities

→ Identification of circular cluster (Euclidean distance)

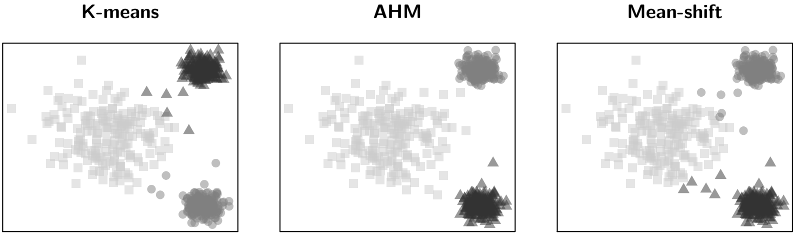
Mean-shift clustering Gradient-method based on kernel density estimate

- Iterative method allowing to detect local maximum of the kernel density
- Method calibrated by a bandwidth (to be given)
- Clustering : Threshold for local maxima (cluster number), kernel density gradient (cluster belonging)

* More flexible method than K-means or AHM, suitable for any type of clusters
* Bandwidth not easy to calibrate, adaptive bandwidth often required

→ See also DBSCAN or OPTICS algorithms

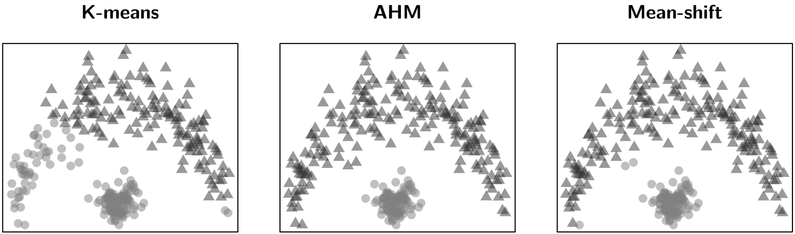
Illustrative examples



Circular clusters : K-means, AHM and mean-shift methods give satisfying results

→ Distance between observations in each clusters smaller than distance between cluster's means

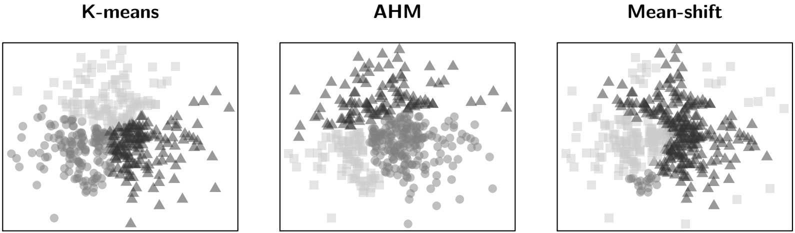
Illustrative examples



Non-circular clusters : K-means not adapted / AHM and mean-shift more robust

→ Distance between observations in each clusters bigger than distance between cluster's means

Illustrative examples

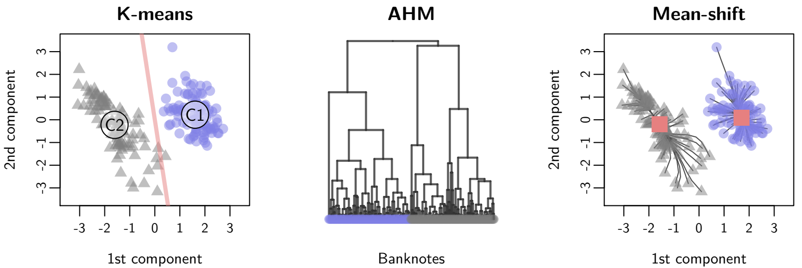


⚠ Clustering methods find clusters even if there is no significant dissimilarities

→ Criteria for significance of inter/intra-variability, dendrogram branch size, bandwidth size, ...

Example of the notes

Detection of the counterfeit notes	Method		
Miss-classification error	K-means	AHM	Mean-shift
Complete sample	0.005%	0	0.005%
Two first components (PCA)	0.005%	0	0%



Linear discriminant analysis — R : `lda(data,cluster)` Package `MASS`

Clustering :	Observations (continuous variables)	→	Clusters (discrete variable)
Discriminant analysis :	Clusters (discrete variable)	→	Observations (discriminant)

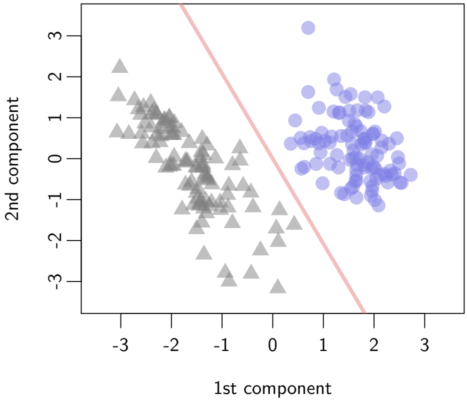
Linear discriminant analysis

- Data :
  - Continuous explanatory variables (regressors) $X^1, \ldots, X^p$
  - Discrete variable to explain (clusters) $Y = 1, \ldots, K$
- Discriminant variable D as linear combination of the regressors minimizing the sum of the variances by cluster $Y = 1, \ldots, K$:

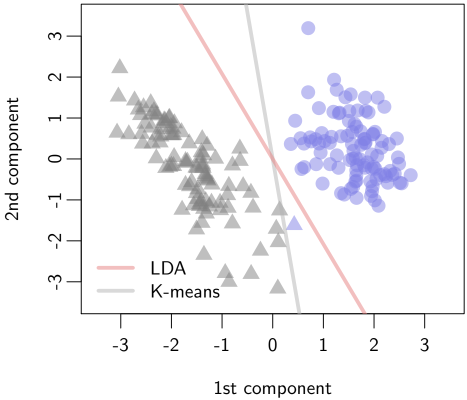
$$D(\alpha_0, \ldots, \alpha_p) = \alpha_0 + \alpha_1 X^1 + \ldots + \alpha_p X^p$$
$$\text{with } (\alpha_0, \ldots, \alpha_p) = \arg\min_\alpha \sum_{j=1}^{K} \sum_{Y_i = j} (D_i - \bar{D}_j)^2$$

→ Best linear combination of the regressors (X^j) for the clustering given by Y

LDA : Example of the notes



LDA : Example of the notes



→ The linear discriminant and the K-means only match when the given clustering in LDA is the one minimizing the intra-variability

Clustering and LDA with R

Clustering methods

- *K-means* `kmean(X,k)`

  with `X` the data (vector or matrix) and `k` the number of clusters

- *AHM* `hclust(dist(X))`

  - Specification of the metric `dist()` (see option `methods`)
  - Specification of the linkage with option `methods` in `hclust()` function
  - Cutting of the dendrogram with `cutree(H,k)`, with `H` a `hclust()`-object and `k` the number of clusters

- *Mean-shift* `ms(X,h)`

  with `X` the data and `h` the bandwidth — Package `LPMC` to install

Linear discriminant analysis `lda(X)` or `fda(X)`

Packages `MASS` or `MDA` to install

Clustering : Summary

Clustering methods allow to partition heterogeneous data in homogeneous clusters

- Optimisation of intra/inter-variability **K-means**
  - → Fixed number of clusters
- Hierarchy between the observations **AHM**
  - → Hierarchical method — Representation with dendrogram
- Cluster based on kernel density estimate **Mean-shift**
  - → Specification of the bandwidth
- Discriminant variable to determine the belonging to a cluster **LDA**
  - → Linear discriminant analysis (linearly separable clusters)

Clustering : Summary

Clustering methods allow to partition heterogeneous data in homogeneous clusters

- Optimisation of intra/inter-variability **K-means**
  - → Fixed number of clusters
- Hierarchy between the observations **AHM**
  - → Hierarchical method — Representation with dendrogram
- Cluster based on kernel density estimate **Mean-shift**
  - → Specification of the bandwidth
- Discriminant variable to determine the belonging to a cluster **LDA**
  - → Linear discriminant analysis (linearly separable clusters)

⚠ **Significance of a clustering to be tested** : Intra/inter-variability difference, branch size of dendrogram, bandwidth size over observation number, ...

Bootstrap technique

Introduction

Regression, PCA and clustering allow analyse data and to define and calibrate models

- **Single (punctual) estimates of the parameters**

  Would the estimations be the same for another sample of observations?

  In other worlds : Does the estimation depend on the specific values of the sample or hold they for the whole population ?

- Evaluation of the precision of the estimation, i.e. estimation of the variability of the estimates

Introduction

Regression, PCA and clustering allow analyse data and to define and calibrate models

- **Single (punctual) estimates of the parameters**

  Would the estimations be the same for another sample of observations?

  In other worlds : Does the estimation depend on the specific values of the sample or hold they for the whole population ?

- Evaluation of the precision of the estimation, i.e. estimation of the variability of the estimates

Bootstrap numerical technique

1. Resampling the observations (independent urn sampling with replacement)
2. Analysing the distribution (and the variability) of the estimates on the bootstrap subsamples

An illustrative example

A machine produces some components

→ Some of them are operational, some others are defective

→ Estimation the probability p that a component is defective

Two sets of observations $p = 0.2$

1. Sample 1 : Among 10 observed components, two are defective
2. Sample 2 : Among 100 observed components twenty two are defective

→ Respective estimates : $\tilde{p}_1 = 0.2$ and $\tilde{p}_2 = 0.22$

Are these estimations precise ?

Bootstraping — R : `sample(data,n,replace=T)`

$p = 0.2$

Sample 1 $(n = 10)$ $\{0, 0, 1, 0, 1, 0, 0, 0, 0, 0\}$, $\tilde{p}_1 = 0.2$

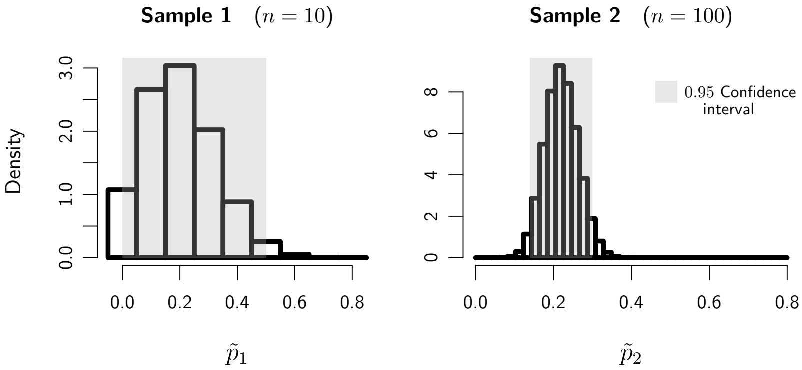
- Bootstrap subsample 1 $\{0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$, $\tilde{p}_1^1 = 0$
- Bootstrap subsample 2 $\{0, 0, 0, 0, 1, 0, 0, 0, 1, 0\}$, $\tilde{p}_1^2 = 0.2$
- Bootstrap subsample 3 $\{0, 0, 0, 0, 0, 0, 1, 0, 0, 0\}$, $\tilde{p}_1^3 = 0.1$
- . . .

Sample 2 $(n = 100)$ $\{0, 0, 0, 0, \ldots, 1, 0, 0, 0\}$, $\tilde{p}_2 = 0.22$

- Bootstrap subsample 1 $\{0, 0, 0, 1, \ldots, 1, 0, 0, 0\}$, $\tilde{p}_2^1 = 0.26$
- Bootstrap subsample 2 $\{0, 0, 0, 0, \ldots, 0, 1, 0, 0\}$, $\tilde{p}_2^2 = 0.25$
- Bootstrap subsample 3 $\{1, 0, 0, 0, \ldots, 0, 1, 1, 0\}$, $\tilde{p}_2^3 = 0.17$
- . . .

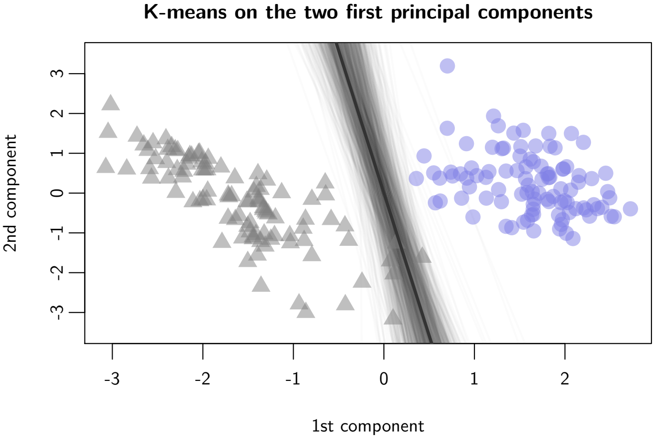
Bootstraping

Histogram of the estimations of the probability $p = 0.2$ for 1e5 bootstrap subsamples



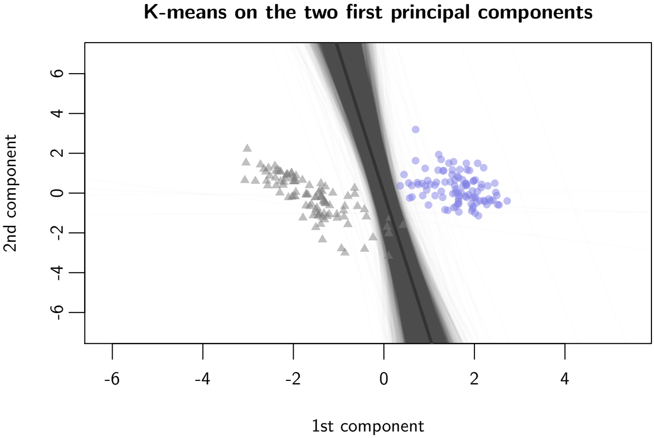
Example of the notes

1e3 bootstrap subsamples



Example of the notes

1e4 bootstrap subsamples



Bootstrap : Summary

- The Bootstrap method is strictly descriptive, **with no assumption on the data and their distribution**
- The method is **purely numerical** and can be **computationally costly**
- Bootstrap **does not improve punctual estimate** but give information on its **variability** (i.e. the precision of estimation)
- The approach can be used for **any type of estimates** (mean, quantile, etc...) but can be imprecise for distribution queue (high or low quantiles)
- **Smooth bootstrap** by adding noise onto each resampled observation (sampling from kernel density estimate of the data)
- Time series : **Moving block bootstrap**
- Bootstrap with random variable generator : **Monte Carlo simulation**

Artificial neural networks

Understanding/Predictive modelling approaches

Statistical models for understanding Identification of underlying mechanisms

- Insights in the nature and physic of the phenomenon of interest
- Model with few parameters that should be interpretable (parsimony principle)
  - → Typically a regression model
  - → Limited model complexity determined by statistical tests

Understanding/Predictive modelling approaches

Statistical models for understanding — Identification of underlying mechanisms

- Insights in the nature and physic of the phenomenon of interest
- Model with few parameters that should be interpretable (parsimony principle)

→ Typically a regression model

→ Limited model complexity determined by statistical tests

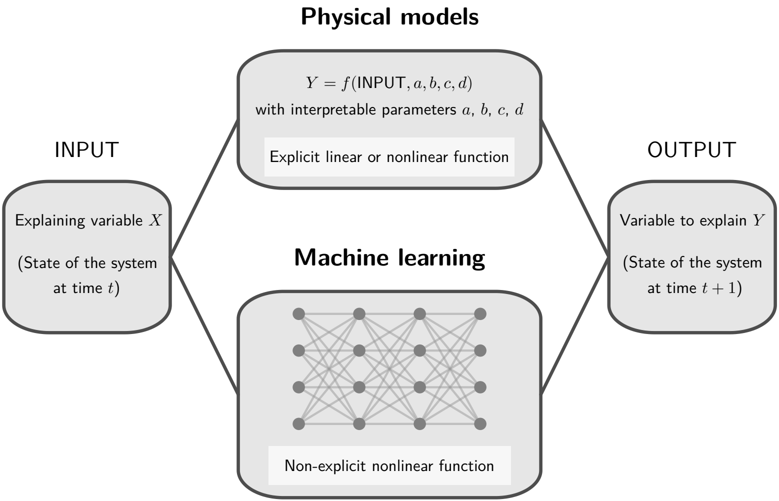
Statistical models for prediction — Machine learning / Data-based algorithms / AI

- Merely an algorithm coming more from the data than from a theory
- Algorithm intentionally complex (very large degrees of freedom/plasticity) with focus on the predictive ability

→ Typically an artificial neural network

→ Algorithm complexity depends on the data (e.g., its size and structure of its distribution)

Understanding/Predictive modelling approaches



Artificial neural network

Artificial neural networks (ANN) are numerical networks of connected cells with weighted activation functions

- The cells are organised as layers (hidden layers) — Generally fully connected
- Important number of parameters (coefficient) — High degrees of freedom
  - → Theoretically large ANN can fit any type of relationship
  - → Trained with e.g. the backpropagation algorithm (right to left error gradient descent)

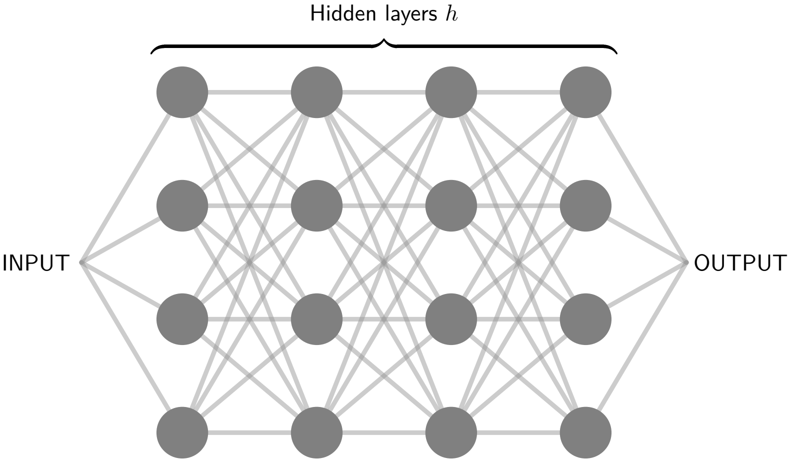
Artificial neural network

Artificial neural networks (ANN) are numerical networks of connected cells with weighted activation functions

- The cells are organised as layers (hidden layers) — Generally fully connected
- Important number of parameters (coefficient) — High degrees of freedom
  - → Theoretically large ANN can fit any type of relationship
  - → Trained with e.g. the backpropagation algorithm (right to left error gradient descent)

- Feedforward (acyclic networks) or recurrent neural networks (RNN) with cycles
- Convolutional neural networks (CNN) with partially overlapping layers
- Deep neural networks (DNN) with multiple hidden layers
- Long short-term memory (LSTM), time delay neural network (TDNN), and many others

Artificial neural network

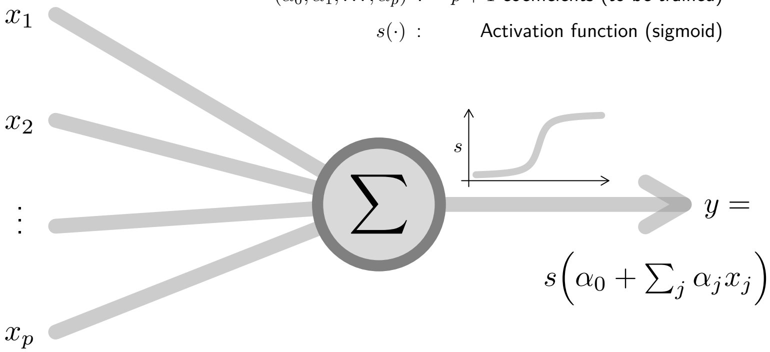


Single node (perceptron)

Settings

$(\alpha_0, \alpha_1, \ldots, \alpha_p)$: $p+1$ coefficients (to be trained)

$s(\cdot)$: Activation function (sigmoid)



Determining the network complexity

The size of the network and its structure depends to the data, its size and its distribution

→ Databased approach by opposition to classical models where the structure and parameters depend on physical consideration

- *Too small networks*: **Limited prediction**, under-use of the data
- *Too large networks*: **Overfitting** (bad prediction of new data) or imprecise calibration

Determining the network complexity

The size of the network and its structure depends to the data, its size and its distribution

→ Databased approach by opposition to classical models where the structure and parameters depend on physical consideration

- *Too small networks*: **Limited prediction**, under-use of the data
- *Too large networks*: **Overfitting** (bad prediction of new data) or imprecise calibration

The network with a single node correspond to a linear regression

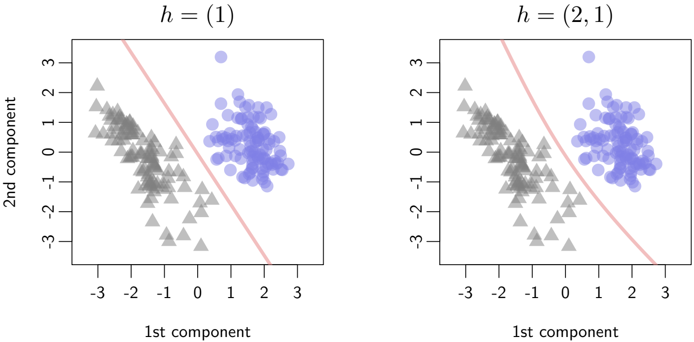
→ Modelling of complex non-linear relationships with large networks

→ However large networks (too various non-linear possibilities) can be superfluous and provide undesired overfitting

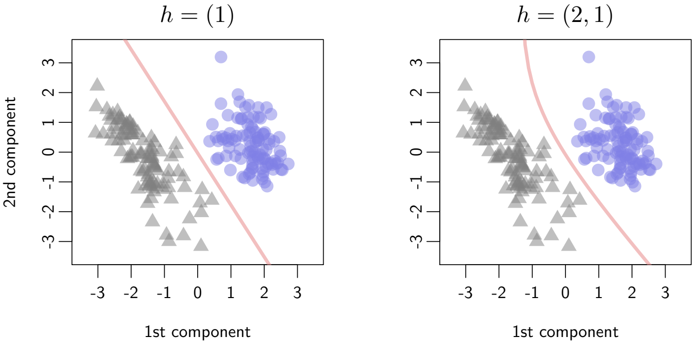
Example of the notes

- Clear *linear* discrimination on the plan of the two first components
- Single node (linear regression) sufficient to discriminate the notes, more complex networks lead to overfitting

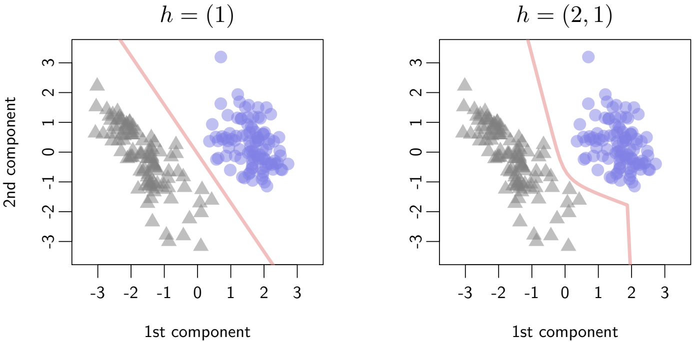
Example of the notes : Subsample 1



Example of the notes : Subsample 2



Example of the notes : Subsample 3



Network complexity : Risk minimization

We denote $f(x_i;\theta)$ the neural networks with parameter θ for prediction of y_i

- **Risk minimization**

  L is a loss function, the risk $R = E(L)$ is the expectation of the loss

  → Empirical risk : $R_{emp} = \frac{1}{n}\sum_i L\big(y_i, f(x_i;\theta)\big)$

  → Vapnik's inequality with proba $1-\alpha$: $R < R_{emp} + \sqrt{\frac{d(\ln(2n/d)+1)-\ln(\alpha/4)}{n}}$

  with d the Vapnik–Chervonenkis dimension (i.e. the cardinality of the largest set of points that the algorithm can shatter — i.e. prediction ability)

- **No distributional assumptions** (only $d \ll n$)

- Selection of the network with minimal bound for R (Ratio d/n of interest)

  → Increase of the complexity and prediction ability d as n increases

Determining the network complexity in practice

Vapnik–Chervonenkis dimension difficult to evaluate in practice

- **Empirical approach** Trade-off between the fit and robustness of a network

  Repeat in a K-Bootstrap loop :

  > S_k is the k-th bootstrap-sampling; partition S_k in two sub-samples S_k^1 and S_k^2
  >
  > S_k^1: **Training set** used to fit the network
  >
  > S_k^2: **Testing set** use to estimate prediction error E_k

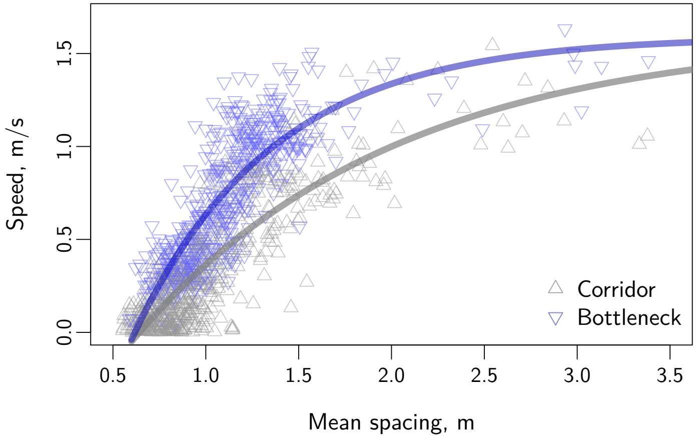
- Cross-validation bootstrap
- Selection of the network with **minimal empirical prediction error**

$$\bar{E}_K = \frac{1}{K} \sum_k E_k$$

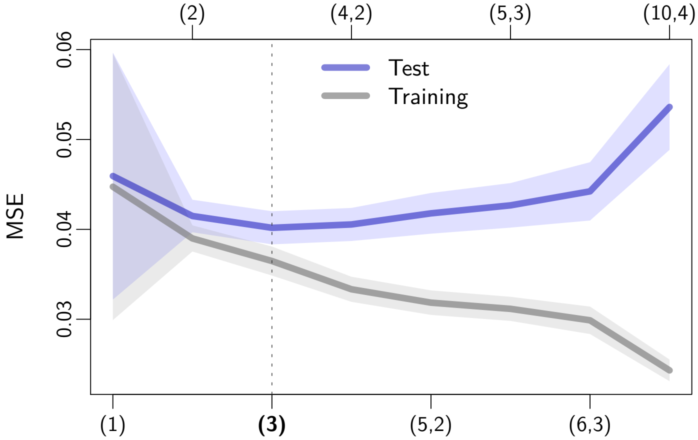
Example : Prediction of pedestrian dynamics

- **Prediction of pedestrian speed** v according to the relative position $(\tilde{x}_j, \tilde{y}_j)$ and distance s_j to the $N = 10$ closest neighbors
- **Data :** Experiments in corridor (C) and bottleneck (B) geometries for various density levels
- **Two modelling approaches**
  1. Physical model (fundamental diagram) with three parameters
  $$\tilde{v} = \mathsf{FD}\big(\bar{s}_N, v_0, T, \ell\big) = v_0\Big(1 - e^{\frac{\ell - \bar{s}_N}{v_0 T}}\Big)$$
  2. Feedforward neural network with hidden layers h
  $$\tilde{v} = \mathsf{NN}\big(h, \bar{s}_N, (\tilde{x}_j, \tilde{y}_j, 1 \geq j \geq N)\big)$$
- **Minimise the mean square error** $\quad \mathsf{MSE} = \frac{1}{n}\sum_i (v_i - \tilde{v}_i)^2$

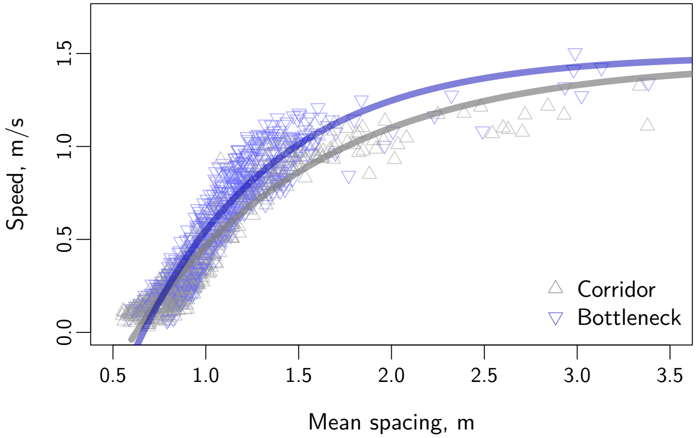
Prediction of pedestrian dynamics: Data



Determining the network complexity

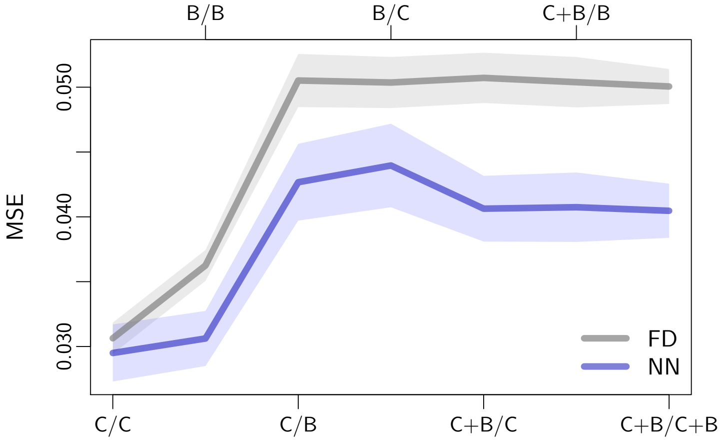


Prediction of pedestrian dynamics



Model comparison

Notation : Training/Testing — E.g., C/B : Trained on the corridor experiment, tested on the bottleneck experiment



Artificial neural networks with R

- Artificial neural networks very easy to train and compute with R
- Package `neuralnet` to install `install.packages('neuralnet')` `require(neuralnet)`
- **Train the network** (backpropagation algorithm)

  `NN=neuralnet(Y∼X1+...+Xp,data=train,hidden=h)`

  Here `Y` is the variable to explain, `X1,...,Xp` are the explanatory variables, `train` are data for the training and `h` are the hidden layers
- **Compute the trained network** `compute(NN,data=test)`

  Here `NN` is a trained network and `test` are data for the testing

Artificial neural networks : Summary

- **Artificial neural network :** Oriented graphs with positive weights
  - Network with nodes as sigmoid activation function
  - Network structure in (hidden) layers — Several types of configurations possible (feedforward, recurrent, convolutional, etc...)
  - Fitting of any transfer function from given input to an output

- **Prediction** of new observations, missing values, dynamics
  - Algorithm coming from the data, trained by backpropagation of a cost or an error
  - No physical investigation of the underlying mechanisms of the studied systems
  - Prediction of complex (non-linear) relationships in high dimension

- **Determining the network complexity**
  - Network complexity depends on size and distribution of the data
  - Empirical setting in training/testing cross-validation

Overview

Part 1 | Descriptive statistics for univariate and bivariate data

Repartition of the data (histogram, kernel density, empirical cumulative distribution function), order statistic and quantile, statistics for location and variability, boxplot, scatter plot, covariance and correlation, QQplot

Part 2 | Descriptive statistics for multivariate data

Least squares and linear and non-linear regression models, principal component analysis, principal component regression, clustering methods (K-means, hierarchical, density-based), linear discriminant analysis, bootstrap technique, artificial neural networks

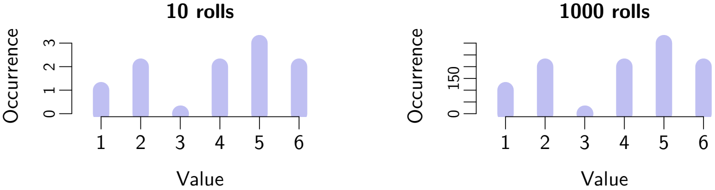
Part 3 | Parametric statistic

Likelihood, estimator definition and main properties (bias, convergence), punctual estimate (maximum likelihood estimation, Bayesian estimation), confidence and credible intervals, information criteria, test of hypothesis, parametric clustering

Appendix LaTeX plots with R and Tikz

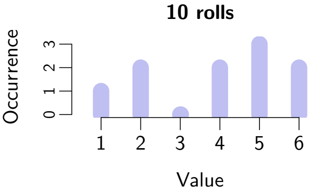
The example of the dices

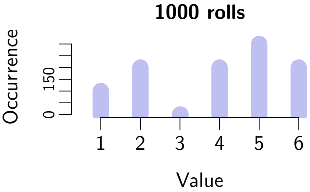
Are my dices biased ??



The example of the dices

Are my dices biased ??





No, well not sure

Observed differences *may be* random

Yes

Observed differences *can not be* random

The example of the machine

A machine produces some components that can be operational or defective

- Estimation of the probability p that a component is defective by mean value

$$\tilde{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \text{with} \quad X_i = \begin{cases} 0 & \text{if the component } i \text{ is operational} \\ 1 & \text{if the component } i \text{ is defective} \end{cases}$$

The estimation from a sample with 100 observations is more precise than the estimation with 10 observations (cf. bootstrap)

Why ? Because the variability of the mean decreases as the observation number increases

- Implicitly this reasoning supposes probabilist assumptions on the convergence of the mean, its distribution or again existence of expected values

$\rightarrow$ **Parametric statistic**

Introduction

Fundamental assumption in parametric (or inference or mathematical) statistic :

> **The observations** $i = 1, \ldots, n$ **are independent random variables with probability distribution function** P_θ**,** $\theta \in \mathbb{R}^k$
>
> $\rightarrow$ Independent and identically distributed (iid) model

- P_θ is general (but can have to satisfy properties) — θ are the parameters of the models
- The data are supposed to be a sample of observations of the distribution P_θ

Introduction

Fundamental assumption in parametric (or inference or mathematical) statistic :

> **The observations $i = 1, \dots, n$ are independent random variables with probability distribution function P_θ, $\theta \in \mathbb{R}^k$**
>
> $\rightarrow$ Independent and identically distributed (iid) model

- P_θ is general (but can have to satisfy properties) — θ are the parameters of the models
- The data are supposed to be a sample of observations of the distribution P_θ

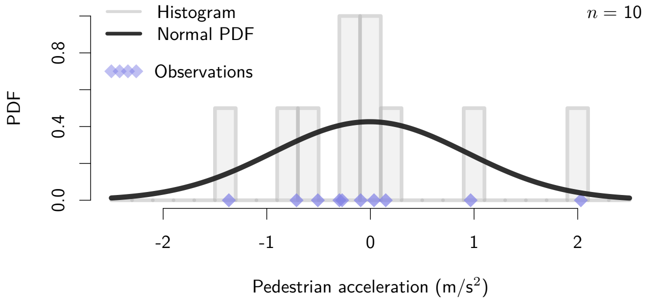
The parametric statistic allows to :

- Fit the parameters θ of a model and evaluate the precision of estimation
- Obtain properties on usual estimators or posterior distribution (Bayesian approach)
- Testing modelling assumptions and compare models

Example 1 : Pedestrian acceleration

Assumption : Normal distribution $\mathcal{N}(\mu, \sigma^2)$ $f(x) = \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\sqrt{2\pi\sigma^2}^{-1}$

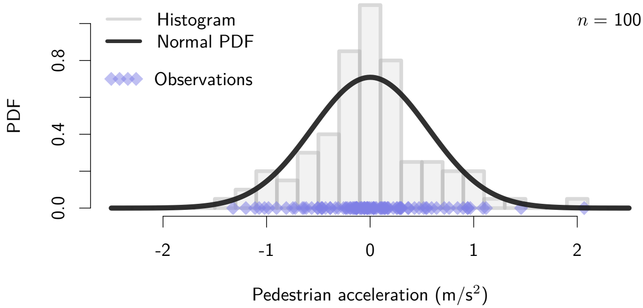
$\rightarrow$ Estimation of μ and σ by $\tilde{\mu}_n = \bar{x}$ and $\tilde{\sigma}_n = s_x$



Example 1 : Pedestrian acceleration

Assumption : Normal distribution $\mathcal{N}(\mu, \sigma^2)$ $f(x) = \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\sqrt{2\pi\sigma^2}^{-1}$

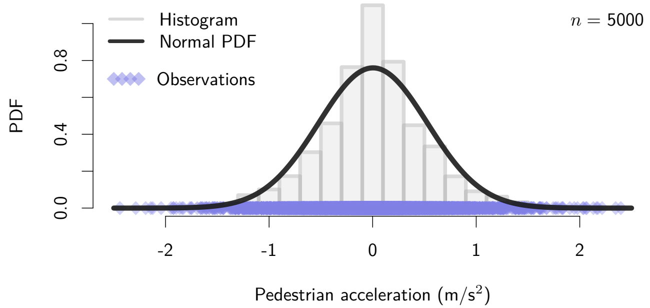
$\rightarrow$ Estimation of μ and σ by $\tilde{\mu}_n = \bar{x}$ and $\tilde{\sigma}_n = s_x$



Example 1 : Pedestrian acceleration

Assumption : Normal distribution $\mathcal{N}(\mu, \sigma^2)$ $f(x) = \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\sqrt{2\pi\sigma^2}^{-1}$

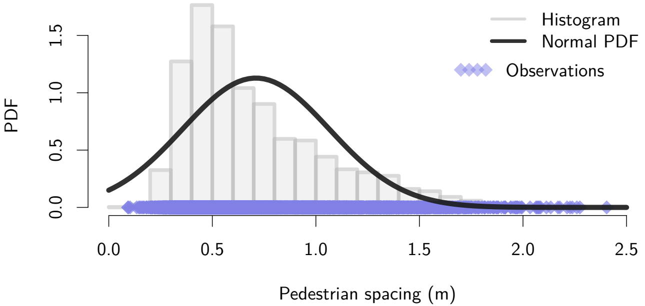
$\rightarrow$ Estimation of μ and σ by $\tilde{\mu}_n = \bar{x}$ and $\tilde{\sigma}_n = s_x$



Example 2 : Pedestrian distance spacing

Assumption : Normal distribution $\mathcal{N}(\mu, \sigma^2)$ $f(x) = \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)\sqrt{2\pi\sigma^2}^{-1}$

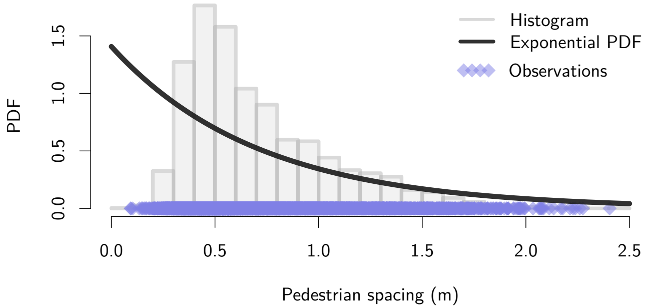
$\rightarrow$ Estimation of μ and σ by $\tilde{\mu}_n = \bar{x}$ and $\tilde{\sigma}_n = s_x$



Example 2 : Pedestrian distance spacing

Assumption : Exponential distribution $\mathcal{E}(\lambda)$ $f(x) = \lambda e^{-\lambda x}$

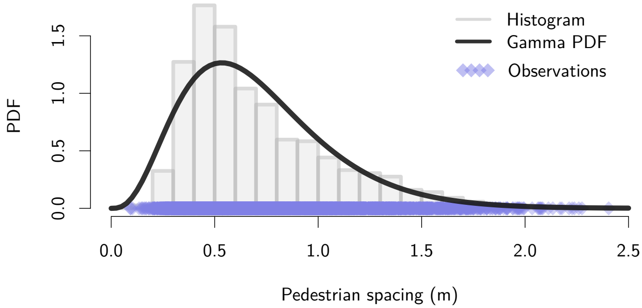
$\rightarrow$ Estimation of expected value λ by $\tilde{\lambda}_n = \bar{x}$



Example 2 : Pedestrian distance spacing

Assumption : Gamma distribution $\mathcal{G}(k,\alpha)$ $f(x) = \frac{x^{k-1}e^{-x/\alpha}}{\Gamma(k)\alpha^k}$

$\rightarrow$ Estimation of k and α by $\tilde{k}_n = \bar{x}^2/var_x$ and $\tilde{\alpha}_n = var_x/\bar{x}$



Convergence of random variables

- **Convergence in distribution** denoted D

  A sequence $X_1, X_2, \ldots$ of real-valued random variables is said to converge in distribution, or converge weakly, or converge in law to a random variable X if

  $$D_n(x) \to D(x) \quad \text{as} \quad n \to \infty \quad \text{for all } x \in \mathbb{R} \text{ at which } F \text{ is continuous}$$

  Here D_n and D are the cumulative distribution functions of X_n and X, respectively.

Convergence of random variables

- **Convergence in distribution** denoted D

  A sequence $X_1, X_2, \ldots$ of real-valued random variables is said to converge in distribution, or converge weakly, or converge in law to a random variable X if

  $$D_n(x) \to D(x) \quad \text{as} \quad n \to \infty \quad \text{for all } x \in \mathbb{R} \text{ at which } F \text{ is continuous}$$

  Here D_n and D are the cumulative distribution functions of X_n and X, respectively.

- **Convergence in probability** denoted P

  $X_1, X_2, \ldots$ converges in probability towards the random variable X if for all $\varepsilon > 0$

  $$P(|X_n - X| \geq \varepsilon) \to 0 \quad \text{as} \quad n \to \infty$$

Convergence of random variables

- **Convergence in distribution** denoted D

  A sequence $X_1, X_2, \ldots$ of real-valued random variables is said to converge in distribution, or converge weakly, or converge in law to a random variable X if

  $$D_n(x) \to D(x) \quad \text{as} \quad n \to \infty \quad \text{for all } x \in \mathbb{R} \text{ at which } F \text{ is continuous}$$

  Here D_n and D are the cumulative distribution functions of X_n and X, respectively.

- **Convergence in probability** denoted P

  $X_1, X_2, \ldots$ converges in probability towards the random variable X if for all $\varepsilon > 0$

  $$P(|X_n - X| \geq \varepsilon) \to 0 \quad \text{as} \quad n \to \infty$$

- **Almost sure convergence** denoted a.s.

  $X_1, X_2, \ldots$ converges almost surely, or almost everywhere, or with probability 1, or strongly towards X if

  $$P\left(X_n \to X \text{ as } n \to \infty\right) = 1$$

Main theorems

Law of large number (LLN)

$(X_1, \ldots, X_n)$ is a iid sample with expected value $E(X_i) = \mu < \infty$. Then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\text{a.s.}}{\rightarrow} E(X_i) = \mu \quad \text{as} \quad n \rightarrow \infty$$

→ Mean value converges to expected value

Main theorems

Law of large number (LLN)

$(X_1, \dots, X_n)$ is a iid sample with expected value $E(X_i) = \mu < \infty$. Then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\text{a.s.}}{\to} E(X_i) = \mu \quad \text{as} \quad n \to \infty$$

→ Mean value converges to expected value

Central limit theorem (CLT)

$(X_1, \dots, X_n)$ is a iid sample with $E(X_i) = \mu < \infty$ and $var_{X_i} = \sigma^2 < \infty$. Then

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \overset{\text{D}}{\to} Z \quad \text{as} \quad n \to \infty, \qquad \text{with } Z \text{ a normal random variable}$$

→ Mean value has asymptotically a normal distribution

Example of the Bernoulli distribution

In the example machine, the state of a component has a Bernoulli distribution with expected value $\mu = p < \infty$ and variance $\sigma^2 = p(1-p) < \infty$

→ **Assumptions of LLN and CLT hold**

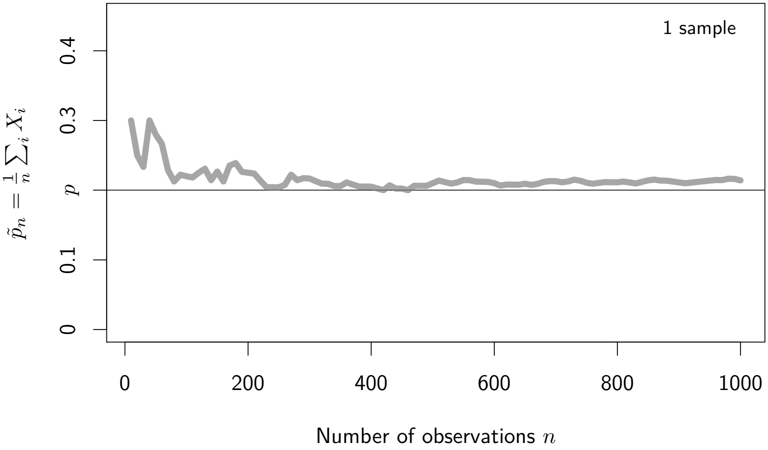
- The estimation $\tilde{p}$ of the probability p that a component is defective is the mean value estimate

$$\tilde{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \text{with} \quad X_i = \begin{cases} 0 & \text{if the component } i \text{ is operational} \\ 1 & \text{if the component } i \text{ is defective} \end{cases}$$

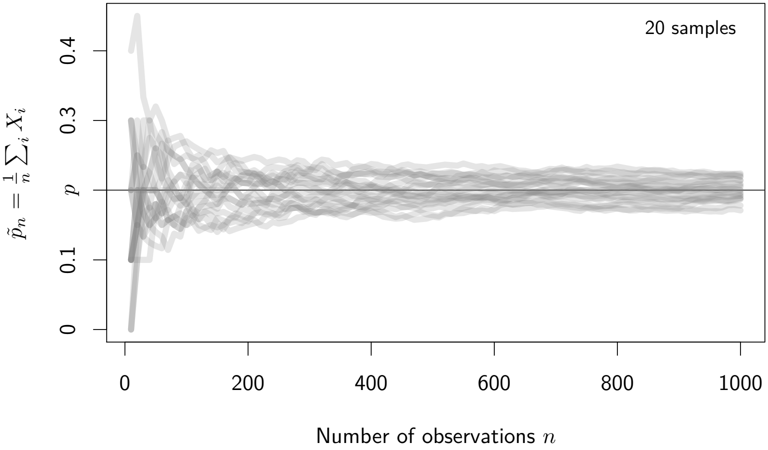
- **LLN** allows to show that the mean $\tilde{p}$ converges to p as $n \to \infty$

- **CLT** allows to describe the distribution of this estimator and to quantify the precision of estimation of p by $\tilde{p}$ for fixed n

Example of the Bernoulli distribution

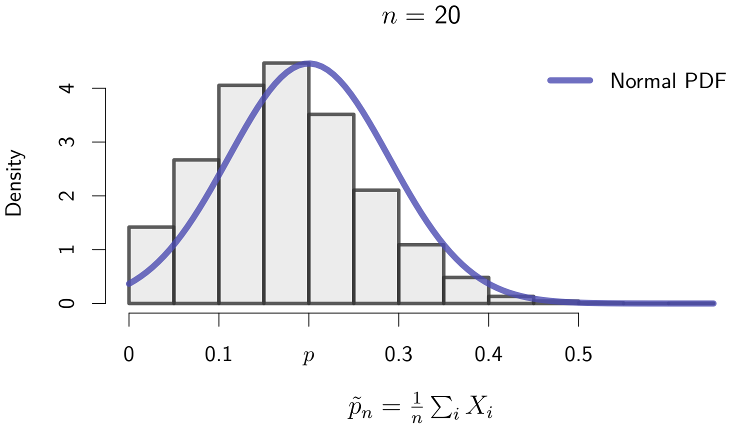


Example of the Bernoulli distribution



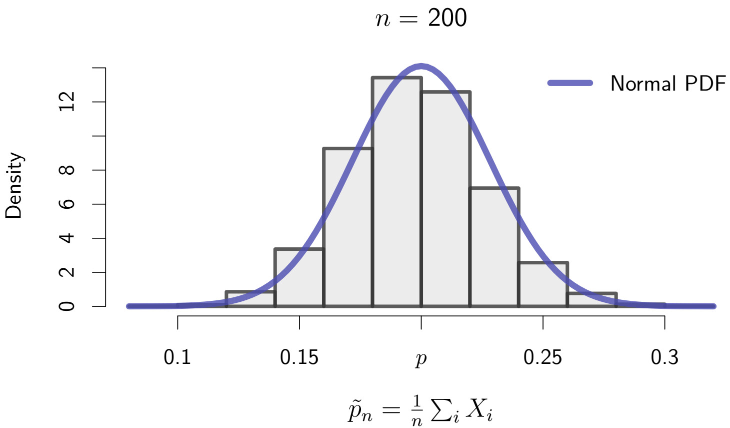
Example of the Bernoulli distribution

Distribution of the mean value — 1e4 samples



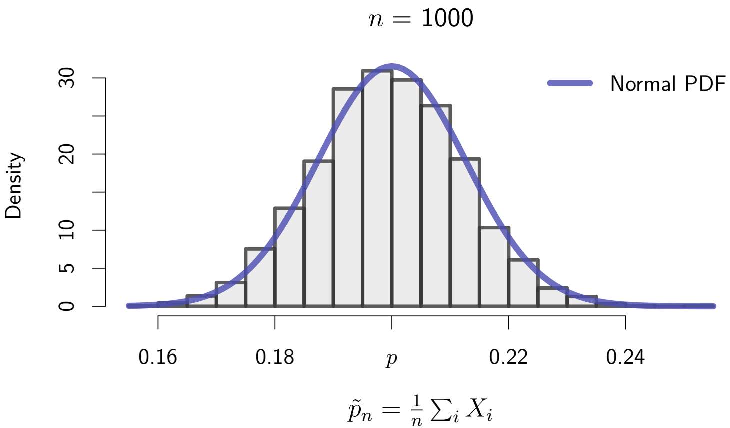
Example of the Bernoulli distribution

Distribution of the mean value — 1e4 samples



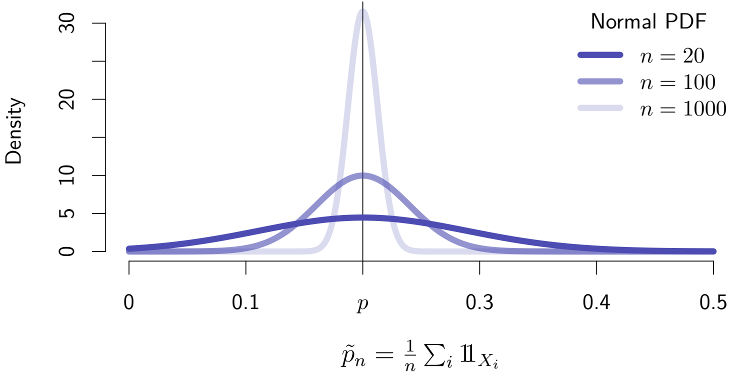
Example of the Bernoulli distribution

Distribution of the mean value — 1e4 samples



Example of the Bernoulli distribution

Distribution of the mean value — 1e4 samples



Example of the Cauchy distribution

Cauchy distribution $\mathcal{C}$ has PDF $f(x) = \left(\pi(1 + x^2)\right)^{-1}$ with no expected value

⚠ Conditions for LLN and CLT are not satisfied — Mean value does not converge !

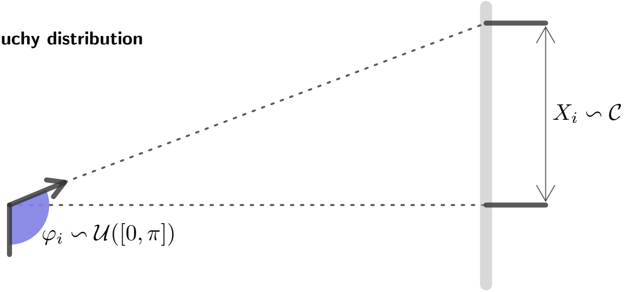
Example of the Cauchy distribution

Cauchy distribution $\mathcal{C}$ has PDF $f(x) = \left(\pi(1+x^2)\right)^{-1}$ with no expected value

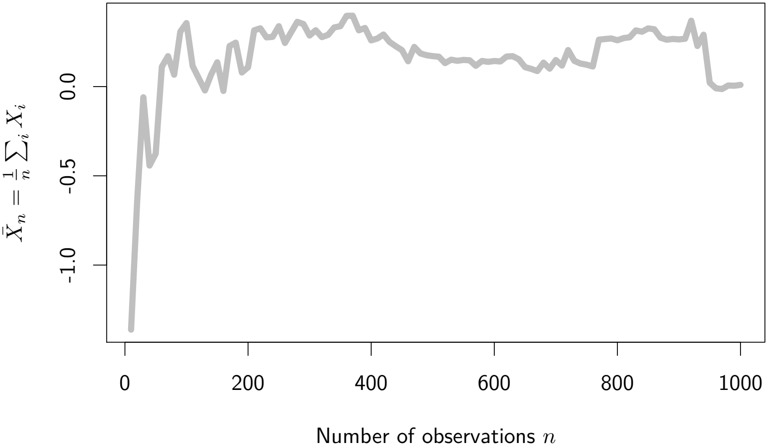
⚠ Conditions for LLN and CLT are not satisfied

Mean value does not converge !

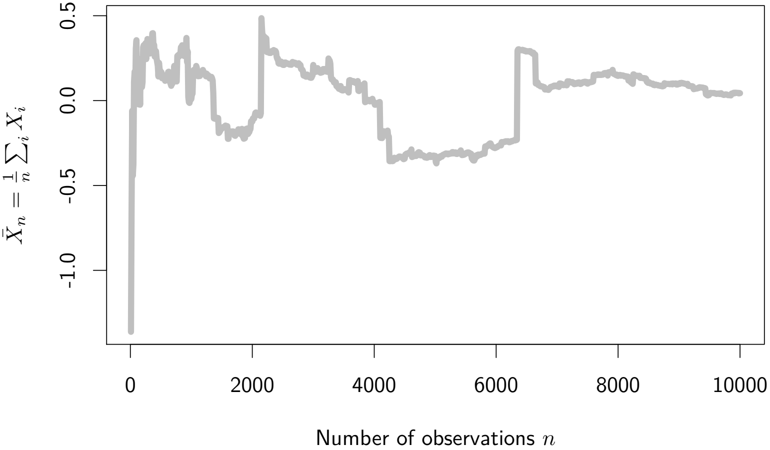
Example of Cauchy distribution



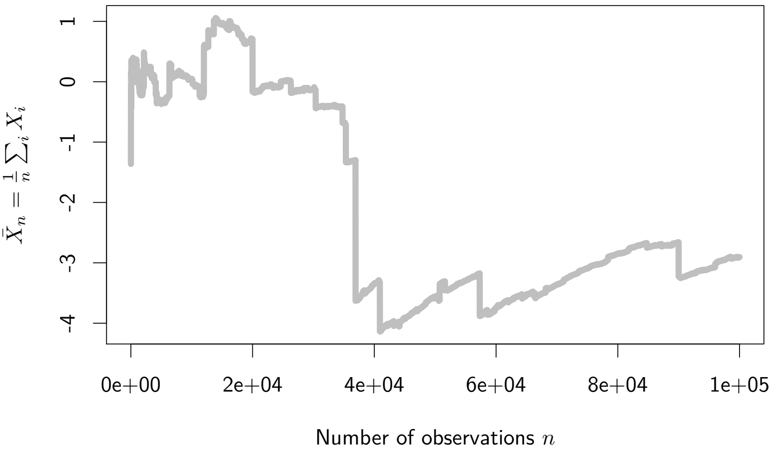
Example of the Cauchy distribution



Example of the Cauchy distribution



Example of the Cauchy distribution



Likelihood function

The likelihood function $L_\theta(x)$ of a set of parameter θ and given data x is

$$L_\theta(x) = P(x \mid \theta) = P(x_1, \ldots, x_n \mid \theta)$$

- The likelihood is a function of θ for a given sample
- Since the observations are *iid*, the likelihood is the *product* with P_θ the family of PDF for the (X_i) $$L_\theta(x) = \prod_{i=1}^{n} P_\theta(x_i)$$
- **Log-likelihood** to manipulate sum instead of product $$\mathcal{L}_\theta(x) = \sum_{i=1}^{n} \log\left(P_\theta(x_i)\right)$$

Likelihood function

The likelihood function $L_\theta(x)$ of a set of parameter θ and given data x is

$$L_\theta(x) = P(x \,|\, \theta) = P(x_1, \ldots, x_n \,|\, \theta)$$

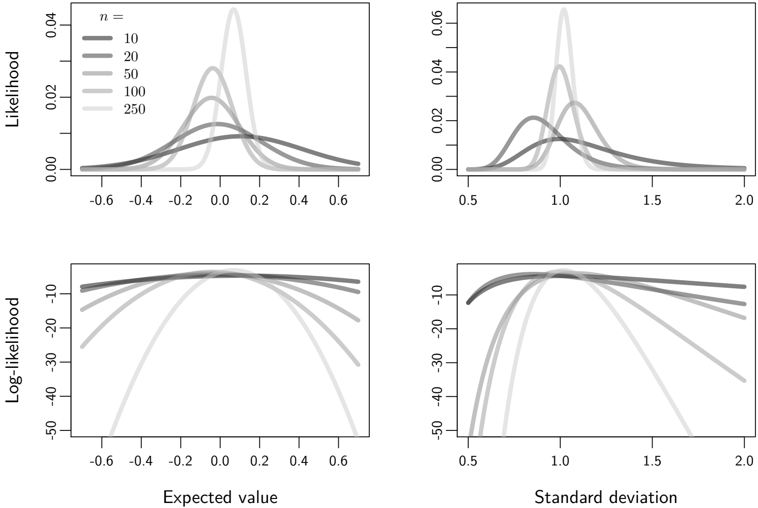
- The likelihood is a function of θ for a given sample
- Since the observations are *iid*, the likelihood is the *product* with P_θ the family of PDF for the (X_i) $$L_\theta(x) = \prod_{i=1}^{n} P_\theta(x_i)$$
- **Log-likelihood** to manipulate sum instead of product $$\mathcal{L}_\theta(x) = \sum_{i=1}^{n} \log\left(P_\theta(x_i)\right)$$

Normal model :

$$L_\theta(x) = \exp\left(-\frac{\sum_i (x_i-\mu)^2}{2\sigma^2}\right)(2\pi\sigma^2)^{-\frac{n}{2}}$$

$$\mathcal{L}_\theta(x) = -\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2 - \frac{n}{2}\log\left(2\pi\sigma^2\right)$$

Normalised likelihood and log-likelihood for the normal distribution



PDF and random number generation with R

`d`{*distrib_name*}*(x)*	Density function
`p`{*distrib_name*}*(q)*	Distribution function
`q`{*distrib_name*}*(p)*	Quantile function
`r`{*distrib_name*}*(n)*	Random number generator

More than 20 distributions available in R

Examples

`dnorm(), pnorm(), qnorm(), rnorm()`	Normal distribution
`dunif(), punif(), qunif(), runif()`	Uniform distribution
`dpois(), ppois(), qpois(), rpois()`	Poisson distribution

...

Estimator

Estimator

The parameters θ are calibrated using estimators

→ **An estimator** $\tilde{\theta}_n$ is a statistic, i.e. a function of the data

$$\begin{array}{cccccc} \tilde{\theta} & : & \mathbb{R}^n & \mapsto & \mathbb{R}^k \\ & & x & \mapsto & \tilde{\theta}_n(x) \end{array} \quad \text{with} \quad \left| \begin{array}{l} n \text{ the number of observations} \\ k \text{ the number of parameters} \\ x = (x_1, \ldots, x_n) \text{ the observations} \end{array} \right.$$

Estimator

The parameters θ are calibrated using estimators

$\rightarrow$ **An estimator** $\tilde{\theta}_n$ is a statistic, i.e. a function of the data

$$\begin{array}{cccc} \tilde{\theta} & : & \mathbb{R}^n & \mapsto & \mathbb{R}^k \\ & & x & \mapsto & \tilde{\theta}_n(x) \end{array} \quad \text{with} \quad \left| \begin{array}{l} n \text{ the number of observations} \\ k \text{ the number of parameters} \\ x = (x_1, \ldots, x_n) \text{ the observations} \end{array} \right.$$

- An estimator $\tilde{\theta}_n$ is a random variable (with mean value, variance, etc. . .)
- The distribution of $\tilde{\theta}_n$ depends on the distribution of the data (and so on θ and on n)
- An estimator $\tilde{\theta}_n$ must have specific properties to well estimate the parameter

Bias of an estimator

$E_\theta \tilde{\theta}_n = \int_{\mathbb{R}^n} \tilde{\theta}_n(x) \prod_i \mathrm{d}P_\theta(x_i)$ is the expected value of the estimator $\tilde{\theta}_n$

- **The bias** B of an estimator $\tilde{\theta}_n$ of θ is the quantity

$$B_\theta(\tilde{\theta}_n) = \theta - E_\theta(\tilde{\theta}_n)$$

- An estimator is called *unbiased* if

$$E_\theta(\tilde{\theta}_n) = \theta \qquad \forall \theta \in \mathbb{R}^k$$

- An estimator is *asymptotically unbiased* if

$$E_\theta(\tilde{\theta}_n) \to \theta \quad \text{as} \quad n \to \infty \qquad \forall \theta \in \mathbb{R}^k$$

Bias : Examples

Bias for the mean value

- The mean $\bar{X} = \frac{1}{n}\sum_i X_i$ is a unbiased estimate of the expected value $E_\mu(X_i) = \mu$

$$E_\mu(\bar{X}) = E_\mu\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n}\sum_i E_\mu X_i = \mu \qquad \forall\mu$$

Bias : Examples

Bias for the mean value

- The mean $\bar{X} = \frac{1}{n}\sum_i X_i$ is a unbiased estimate of the expected value $E_\mu(X_i) = \mu$

$$E_\mu(\bar{X}) = E_\mu\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n}\sum_i E_\mu X_i = \mu \qquad \forall\mu$$

Bias for the variance

- The empirical variance $s_X^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$ is asymptotically an unbiased estimate of the variance $var_\sigma(X_i) = \sigma^2$

$$E_\sigma(s_X^2) = E_\sigma\left(\frac{1}{n}\sum_i (X_i - \bar{X})^2\right) = \frac{1}{n}\sum_i E_\sigma(X_i^2) - E_\sigma(\bar{X}^2) = \frac{n-1}{n}\sigma^2 \qquad \forall\sigma$$

$\rightarrow$ $\tilde{s}_X^2 = \frac{n}{n-1} s_X^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$ is an unbiased estimate of the variance

Error and mean squared error

The error e of an estimator $\tilde{\theta}_n$ of θ is the quantity

$$e_\theta(\tilde{\theta}_n) = \tilde{\theta}_n - \theta$$

- The error is a random variable for which the variability is the one of the estimator
- The error is centred if the estimator is unbiased

Error and mean squared error

The error e of an estimator $\tilde{\theta}_n$ of θ is the quantity

$$e_\theta(\tilde{\theta}_n) = \tilde{\theta}_n - \theta$$

- The error is a random variable for which the variability is the one of the estimator
- The error is centred if the estimator is unbiased

The mean squared error (MSE) of an estimator $\tilde{\theta}_n$ of θ is the quantity

$$\text{MSE}_\theta(\tilde{\theta}_n) = E_\theta((\tilde{\theta}_n - \theta)^2) = var_\theta(\tilde{\theta}_n) + B_\theta^2(\tilde{\theta}_n)$$

- The mean squared error is a deterministic quantity (variance of the error)
- Compromise between bias and variance of the estimator

Convergence properties

Consistency An estimator $\tilde{\theta}_n$ of θ is called consistent if

$$\tilde{\theta}_n \to \theta \quad \text{as} \quad n \to \infty \qquad \forall \theta \in \mathbb{R}^k$$

- Necessary $\text{MSE}_\theta(\tilde{\theta}_n) \to 0$ for a consistent estimator, i.e. at least asymptotic unbiased and with asymptotic variance nil
- Property generally obtained from the law of large numbers

Convergence properties

Consistency An estimator $\tilde{\theta}_n$ of θ is called consistent if

$$\tilde{\theta}_n \to \theta \quad \text{as} \quad n \to \infty \qquad \forall \theta \in \mathbb{R}^k$$

- Necessary $\text{MSE}_\theta(\tilde{\theta}_n) \to 0$ for a consistent estimator, i.e. at least asymptotic unbiased and with asymptotic variance nil
- Property generally obtained from the law of large numbers

The speed of convergence of a consistent estimator $\tilde{\theta}_n$ of θ is $\gamma > 0$ such that

$$n^\gamma(\tilde{\theta}_n - \theta) \to Z \quad \text{as} \quad n \to \infty \qquad \forall \theta \in \mathbb{R}^k$$

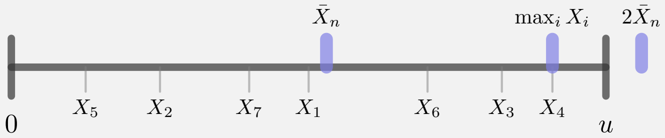
- Higher the convergence speed, better is the estimator
- Asymptotic convergence speed of $1/2$ given by the central limit theorem

Example of the uniform distribution

$(X_1, \ldots, X_n)$ **uniform random variables** on $[0, u]$ PDF : $f(x) = \frac{1}{u} \mathbb{1}_{[0,u]}(x)$

→ Two estimators for u

$$\tilde{u}_1 = 2\bar{X}_n \qquad \text{and} \qquad \tilde{u}_2 = \max_i X_i$$



Example of the uniform distribution

Estimator $\tilde{u}_1 = 2\bar{X}_n = \frac{2}{n}\sum_i X_i$

- Expected value : $E(\tilde{u}_1) = \frac{2}{n}\sum_i E(X_i) = u$ since $E(X_i) = u/2$ *Unbiased estimator*
- Convergence speed : $\gamma = 1/2$ CLT : $n^{1/2}(\tilde{u}_1 - u) \to Z$ as $n \to \infty$

Example of the uniform distribution

Estimator $\tilde{u}_1 = 2\bar{X}_n = \frac{2}{n}\sum_i X_i$

- Expected value : $E(\tilde{u}_1) = \frac{2}{n}\sum_i E(X_i) = u$ since $E(X_i) = u/2$ *Unbiased estimator*
- Convergence speed : $\gamma = 1/2$ CLT : $n^{1/2}(\tilde{u}_1 - u) \to Z$ as $n \to \infty$

Estimator $\tilde{u}_2 = \max_i X_i$

- $P(\tilde{u}_2 \leq x) = P(\cap_i \{X_i \leq x\}) = (x/u)^n$ therefore a PDF for $\tilde{u}_2$ is $f_2(x) = nx^{n-1}u^{-n}$
  Expected value : $E(\tilde{u}_2) = \int x f_2 \, \mathrm{d}x = \frac{n}{n+1}u$ *Asymptotically unbiased estimator*
- $P\big(n^\gamma(\tilde{u}_2 - u) \geq \varepsilon\big) = 1 - (1 + \varepsilon n^{-\gamma}/u)^n \sim 1 - e^{\varepsilon n^{1-\gamma}/u} \to 0$ as $n \to \infty$ if $\gamma > 1$
  Convergence speed : $\gamma = 1$

Example of the uniform distribution

Estimator $\tilde{u}_1 = 2\bar{X}_n = \frac{2}{n}\sum_i X_i$

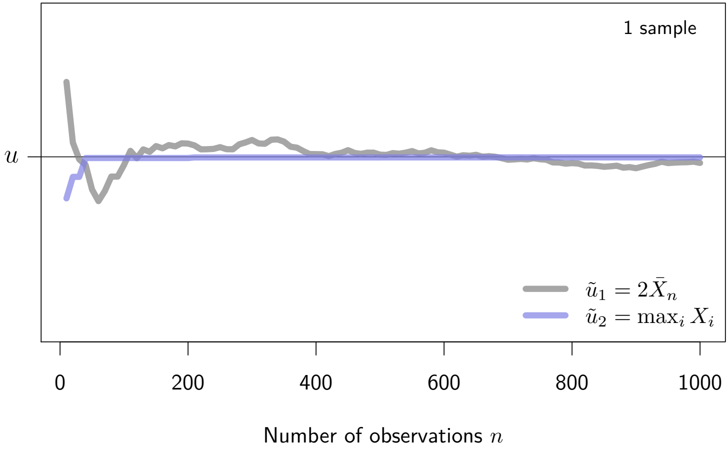
- Expected value : $E(\tilde{u}_1) = \frac{2}{n}\sum_i E(X_i) = u$ since $E(X_i) = u/2$ *Unbiased estimator*
- Convergence speed : $\gamma = 1/2$ CLT : $n^{1/2}(\tilde{u}_1 - u) \to Z$ as $n \to \infty$

Estimator $\tilde{u}_2 = \max_i X_i$

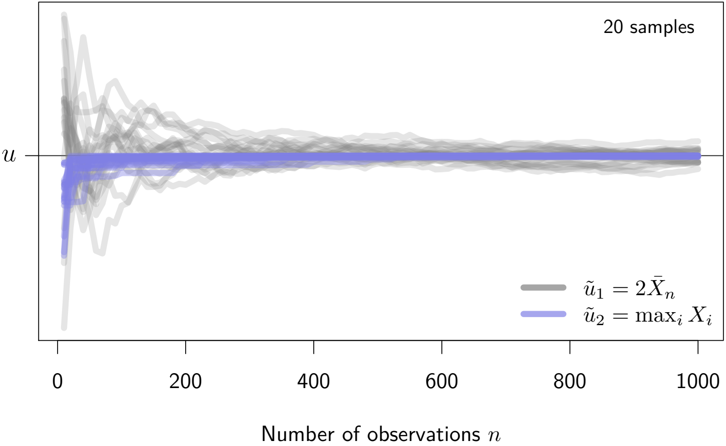
- $P(\tilde{u}_2 \leq x) = P(\cap_i \{X_i \leq x\}) = (x/u)^n$ therefore a PDF for $\tilde{u}_2$ is $f_2(x) = nx^{n-1}u^{-n}$
  Expected value : $E(\tilde{u}_2) = \int x f_2 \, dx = \frac{n}{n+1}u$ *Asymptotically unbiased estimator*
- $P\big(n^\gamma(\tilde{u}_2 - u) \geq \varepsilon\big) = 1 - (1 + \varepsilon n^{-\gamma}/u)^n \sim 1 - e^{\varepsilon n^{1-\gamma}/u} \to 0$ as $n \to \infty$ if $\gamma > 1$
  Convergence speed : $\gamma = 1$

$\tilde{u}_2$ better than $\tilde{u}_1$

Example of the uniform distribution

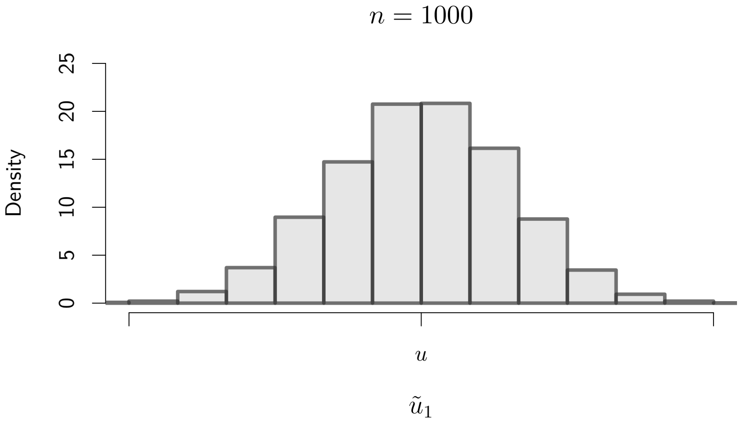


Example of the uniform distribution



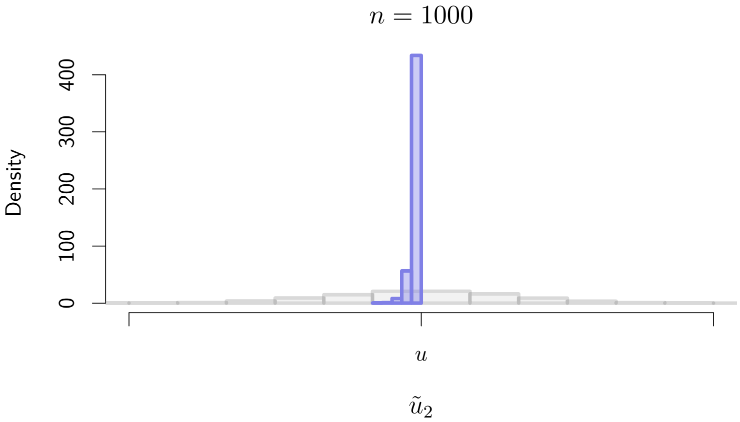
Example of the uniform distribution

Distribution of the estimators — 1e4 samples



Example of the uniform distribution

Distribution of the estimators — 1e4 samples



Sufficient statistic, Fisher Information and efficient estimate

A statistic $\tilde{\theta}_n^s(x)$ is sufficient (or exhaustive) with respect to an unknown parameter θ if

No other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter (Ronald Fisher)

- **Fisher–Neyman factorization criterion** : $\tilde{\theta}_n$ sufficient for θ iff $\exists g,h,\ L_\theta(x) = h(x) g_\theta(\tilde{\theta}_n(x))$

  Example of the uniform distribution on $[0, u]$: $L_u(x) = u^{-n} \mathbb{1}_{\min_i x_i \geq 0} \mathbb{1}_{\max_i x_i \leq u}$
  $\rightarrow$ $\tilde{u}_2 = \max_i x_i$ is a sufficient statistic for u but $\tilde{u}_1 = 2\bar{x}_n$ is not

Sufficient statistic, Fisher Information and efficient estimate

A statistic $\tilde{\theta}_n^s(x)$ is sufficient (or exhaustive) with respect to an unknown parameter θ if

No other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter (Ronald Fisher)

- **Fisher–Neyman factorization criterion** : $\tilde{\theta}_n$ sufficient for θ iff $\exists g,h,\ L_\theta(x) = h(x) g_\theta(\tilde{\theta}_n(x))$

  > Example of the uniform distribution on $[0, u]$: $L_u(x) = u^{-n} \mathbb{1}_{\min_i x_i \geq 0} \mathbb{1}_{\max_i x_i \leq u}$
  > $\rightarrow$ $\tilde{u}_2 = \max_i x_i$ is a sufficient statistic for u but $\tilde{u}_1 = 2\bar{x}_n$ is not

- **Blackwell–Rao theorem** : For any estimate $\tilde{\theta}_n$ of θ, $var_\theta\big(E(\tilde{\theta}_n | \tilde{\theta}_n^s)\big) \leq var_\theta(\tilde{\theta}_n)$

Sufficient statistic, Fisher Information and efficient estimate

A statistic $\tilde{\theta}_n^s(x)$ is sufficient (or exhaustive) with respect to an unknown parameter θ if

No other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter (Ronald Fisher)

- **Fisher–Neyman factorization criterion** : $\tilde{\theta}_n$ sufficient for θ iff $\exists g,h,\ L_\theta(x) = h(x) g_\theta(\tilde{\theta}_n(x))$

> Example of the uniform distribution on $[0, u]$: $L_u(x) = u^{-n} \mathbb{1}_{\min_i x_i \geq 0} \mathbb{1}_{\max_i x_i \leq u}$
> $\rightarrow$ $\tilde{u}_2 = \max_i x_i$ is a sufficient statistic for u but $\tilde{u}_1 = 2\bar{x}_n$ is not

- **Blackwell–Rao theorem** : For any estimate $\tilde{\theta}_n$ of θ, $var_\theta\big(E(\tilde{\theta}_n | \tilde{\theta}_n^s)\big) \leq var_\theta(\tilde{\theta}_n)$
- **Fisher information** : $I_x(\theta) = E[(\partial ln(L_\theta(x))/\partial\theta)^2]$ quantifies information on θ given by x

$\rightarrow$ We have in general $I_{\tilde{\theta}(x)}(\theta) \leq I_x(\theta)$ and $I_{\tilde{\theta}^s(x)}(\theta) = I_x(\theta)$ for a sufficient statistic

Sufficient statistic, Fisher Information and efficient estimate

A statistic $\tilde{\theta}_n^s(x)$ is sufficient (or exhaustive) with respect to an unknown parameter θ if

No other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter (Ronald Fisher)

- **Fisher–Neyman factorization criterion** : $\tilde{\theta}_n$ sufficient for θ iff $\exists g, h,\ L_\theta(x) = h(x) g_\theta(\tilde{\theta}_n(x))$

  Example of the uniform distribution on $[0, u]$: $L_u(x) = u^{-n} \mathbb{1}_{\min_i x_i \geq 0} \mathbb{1}_{\max_i x_i \leq u}$
  $\rightarrow$ $\tilde{u}_2 = \max_i x_i$ is a sufficient statistic for u but $\tilde{u}_1 = 2\bar{x}_n$ is not

- **Blackwell–Rao theorem** : For any estimate $\tilde{\theta}_n$ of θ, $var_\theta\left(E(\tilde{\theta}_n | \tilde{\theta}_n^s)\right) \leq var_\theta(\tilde{\theta}_n)$

- **Fisher information** : $I_x(\theta) = E[(\partial ln(L_\theta(x))/\partial\theta)^2]$ quantifies information on θ given by x
  $\rightarrow$ We have in general $I_{\tilde{\theta}(x)}(\theta) \leq I_x(\theta)$ and $I_{\tilde{\theta}^s(x)}(\theta) = I_x(\theta)$ for a sufficient statistic

- **Cramer–Rao bound** : *Under regularity assumptions* $1/I_x(\theta) \leq var_\theta(\tilde{\theta}_n)$, $\forall \tilde{\theta}_n$ unbiased
  $\rightarrow$ An estimate is called efficient iff $var_\theta(\tilde{\theta}_n) = 1/I_x(\theta)$
  $\rightarrow$ An efficient statistic is necessary sufficient

Punctual estimation

Introduction

Punctual estimations of parameters are mathematically non-linear optimisation problems for an *objective function*

$f_x(\theta)$: Function to optimize

x are the data (given)

θ are the parameters (to optimize over $\mathbb{R}^k$)

→ Hard problem when f is not regular (discontinuous, multi-modal, noisy, ...)

→ Convergence to local minima

Introduction

Punctual estimations of parameters are mathematically non-linear optimisation problems for an *objective function*

$f_x(\theta)$: Function to optimize

x are the data (given)

θ are the parameters (to optimize over $\mathbb{R}^k$)

→ Hard problem when f is not regular (discontinuous, multi-modal, noisy, ...)

→ Convergence to local minima

Formulation of the objective function f by

- **Least squares** — Non-parametric approach
- **Likelihood** — Maximum likelihood estimate
- **Bayesian approach** — Posterior distribution for some given prior on the parameters

Optimisation with R

Punctual estimations (Least squares, MLE and posterior PDF) are optimisation problems for functions $f : \mathbb{R}^k \mapsto \mathbb{R}$

- **Optimisation with R (general case)** `optim(par,f)`

  with `par` the initial values for the parameters and `f` the function to optimize

  > Exist different optimisation methods (Nelder-Mead, quasi-Newton, ...)
  > Quasi-Netwon method `''L-BFGS-B''` allows box constraints for the parameter

Least-squares optimisation with R

- Multilinear models `lm(f,X)`
- Non-linear models `nls(f,X,par)`

Maximum likelihood estimation

Maximum Likelihood Estimation (MLE)

$$\tilde{\theta}^{\mathsf{MLE}}(x) = \arg\max_{\theta \in \mathbb{R}^k} L_\theta(x)$$

- Most probable estimation knowing the data of parameter θ for the distribution family
- MLE can be determined by maximizing the log-likelihood

Maximum likelihood estimation

Maximum Likelihood Estimation (MLE)

$$\tilde{\theta}^{\mathsf{MLE}}(x) = \arg\max_{\theta \in \mathbb{R}^k} L_\theta(x)$$

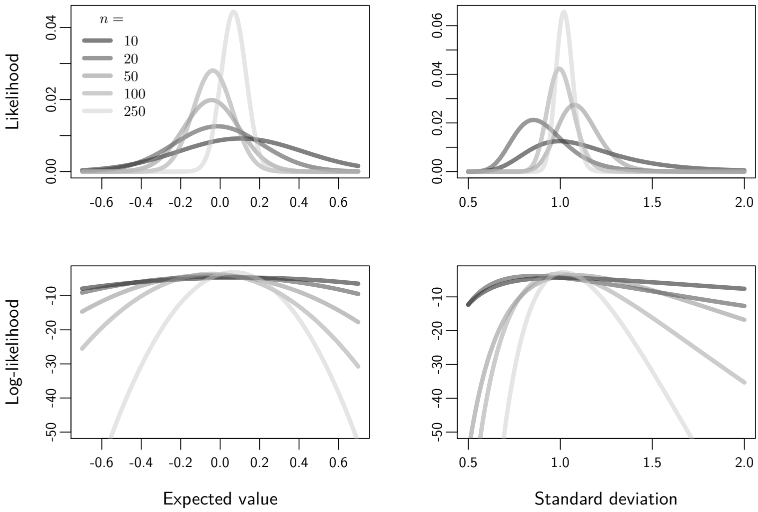
- Most probable estimation knowing the data of parameter θ for the distribution family
- MLE can be determined by maximizing the log-likelihood

Properties

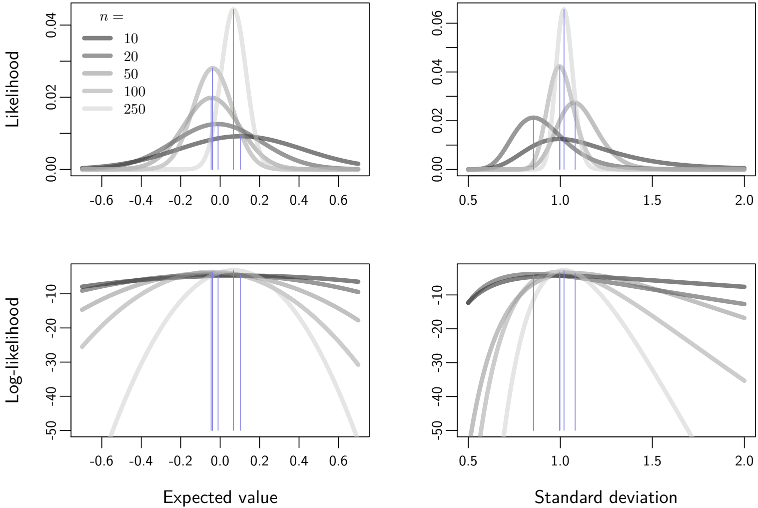
- MLE not necessary unbiased but is in general asymptotically unbiased
- If it exits a sufficient statistic then MLE depends on it (but MLE not necessary sufficient)
- If it exits a efficient statistic then it is the MLE (regularity assumptions of Cramer-Rao th.)

$\rightarrow$ MLE generally better than least squares or moment methods (cf. uniform distribution)

MLE for the normal distribution



MLE for the normal distribution



MLE for different distributions

- **Normal distribution**

  The likelihood of the Gaussian model is $L_\theta(x) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\big(-\sum_i (x_i - \mu)^2/2\sigma^2\big)$

  MLE of μ and σ solution of $\frac{\partial L_\theta}{\partial \mu} = \frac{\partial L_\theta}{\partial \sigma} = 0$ are: $\tilde{\mu}_n^{\text{MLE}} = \bar{x}$ and $\tilde{\sigma}_n^{\text{MLE}} = s_x$

$\rightarrow$ Arithmetic mean and empirical variance are the MLE for parameters μ and σ^2 of the normal distribution

MLE for different distributions

- **Normal distribution**

  The likelihood of the Gaussian model is $L_\theta(x) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_i (x_i - \mu)^2 / 2\sigma^2\right)$

  MLE of μ and σ solution of $\frac{\partial L_\theta}{\partial \mu} = \frac{\partial L_\theta}{\partial \sigma} = 0$ are: $\tilde{\mu}_n^{\text{MLE}} = \bar{x}$ and $\tilde{\sigma}_n^{\text{MLE}} = s_x$

$\rightarrow$ Arithmetic mean and empirical variance are the MLE for parameters μ and σ^2 of the normal distribution

- **Exponential distribution**

  The likelihood of the exponential model is $L_\lambda(x) = \lambda^n \exp\left(-\lambda \sum_i x_i\right)$

  MLE of λ solution of $\frac{\partial L_\lambda}{\partial \lambda} = 0$ is: $\tilde{\lambda}_n^{\text{MLE}} = (\bar{x})^{-1}$

$\rightarrow$ Inverse of arithmetic mean is the MLE for the exponential distribution parameter λ

MLE for different distributions

- **Normal distribution**

  The likelihood of the Gaussian model is $L_\theta(x) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\big(-\sum_i (x_i - \mu)^2/2\sigma^2\big)$

  MLE of μ and σ solution of $\frac{\partial L_\theta}{\partial \mu} = \frac{\partial L_\theta}{\partial \sigma} = 0$ are: $\tilde{\mu}_n^{\text{MLE}} = \bar{x}$ and $\tilde{\sigma}_n^{\text{MLE}} = s_x$

  → Arithmetic mean and empirical variance are the MLE for parameters μ and σ^2 of the normal distribution

- **Exponential distribution**

  The likelihood of the exponential model is $L_\lambda(x) = \lambda^n \exp\big(-\lambda \sum_i x_i\big)$

  MLE of λ solution of $\frac{\partial L_\lambda}{\partial \lambda} = 0$ is: $\tilde{\lambda}_n^{\text{MLE}} = (\bar{x})^{-1}$

  → Inverse of arithmetic mean is the MLE for the exponential distribution parameter λ

- **Uniform distribution**

  The likelihood of the uniform model on $[0, u]$ is $L_u(x) = \begin{cases} 1/u^n & \text{if } \min_i x_i \geq 0,\ \max_i x_i \leq u \\ 0 & \text{otherwise} \end{cases}$

  MLE of u is: $\tilde{u}_n^{\text{MLE}} = \max_i x_i$ (but $\frac{\partial L_u}{\partial u}$ not defined for $u = \max_i x_i$)

  → The maximum is the MLE of u for the uniform distribution on $[0, u]$

MLE and the linear regression

Linear model with Gaussian noise

$$y_i = (ax_i + b) + \sigma\mathcal{E}_i, \qquad \text{with } (\mathcal{E}_i) \text{ iid } \mathcal{N}(0,1)$$

- Residuals $R_i(a,b) = y_i - (ax_i + b)$ are supposed normally distributed

MLE and the linear regression

Linear model with Gaussian noise

$$y_i = (ax_i + b) + \sigma\mathcal{E}_i, \qquad \text{with } (\mathcal{E}_i) \text{ iid } \mathcal{N}(0,1)$$

- Residuals $R_i(a,b) = y_i - (ax_i + b)$ are supposed normally distributed

Likelihood of the Gaussian linear model is

$$L_\theta(x) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{\sum_i (y_i - (ax_i + b))^2}{2\sigma^2}\right)$$

- Likelihood maximal if $\sum_i (y_i - (ax_i + b))^2$ is minimal

MLE and the linear regression

Linear model with Gaussian noise

$$y_i = (ax_i + b) + \sigma\mathcal{E}_i, \qquad \text{with } (\mathcal{E}_i) \text{ iid } \mathcal{N}(0,1)$$

- Residuals $R_i(a,b) = y_i - (ax_i + b)$ are supposed normally distributed

Likelihood of the Gaussian linear model is

$$L_\theta(x) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(- \frac{\sum_i (y_i - (ax_i + b))^2}{2\sigma^2} \right)$$

- Likelihood maximal if $\sum_i (y_i - (ax_i + b))^2$ is minimal

$\rightarrow$ OLS estimates is MLE when the residuals are Gaussian
(and the empirical standard deviation is the MLE of noise amplitude σ)

The Bayesian approach

Bayesian approach consists in using prior distributions for the parameters and to analyse posterior distributions conditionally to the data

- Data x are observable random variables with distribution (likelihood) $P(x \mid \theta)$
- Parameters θ are latent (unknown) random variables with prior distribution $P(\theta)$

The Bayesian approach

Bayesian approach consists in using prior distributions for the parameters and to analyse posterior distributions conditionally to the data

- Data x are observable random variables with distribution (likelihood) $P(x \mid \theta)$
- Parameters θ are latent (unknown) random variables with prior distribution $P(\theta)$

Bayes Theorem assuming $P(x), P(\theta) > 0$

$$P_x(\theta) = P(\theta \mid x) = \frac{P(x, \theta)}{P(x)} = \frac{P(\theta)P(x \mid \theta)}{P(x)}$$

$$posterior \propto prior * likelihood$$

- Punctual estimations of θ by mode, median or mean of posterior distribution $P_x(\theta)$
- Posterior distribution = (normalized) likelihood when prior is uniform
  → MLE is the mode of posterior with non-informative prior

Algorithms to calculate MLE and posterior PDF

MLE or posterior PDF are complex optimization problems having in general no explicit solutions

$\rightarrow$ Approximation by iterative algorithms (starting from initial value $\tilde{\theta}_n^{(0)}$ for the parameters)

- **Gibbs sampling** — Randomized algorithm – MCMC

  Simulation of $\tilde{\theta}_n^{(i)}$ as random variables with distribution $P\left(\tilde{\theta}_n^{(i-1)}\right)P\left(x \mid \tilde{\theta}_n^{(i-1)}\right)$ (convergence to posterior distribution)

- **Expectation-Maximization (EM)** — Deterministic algorithm

  Iterations of maximisation of the parameters $\tilde{\theta}_n^{(i)}$ of the expected log-likelihood conditionally to the data and values $\tilde{\theta}_n^{(i-1)}$ of the parameters at previous step

- **Variational Bayesian (VB)** — Deterministic algorithm

  Estimation of posterior distribution by minimizing the Kullback-Leibler divergence measure with parameter previous values $\tilde{\theta}_n^{(i-1)}$ over a partition of their domain

Comparing Bayesian, MLE and OLS approaches

- OLS and MLE are close when residuals have compact (normal) distributions
- Bayesian estimate and MLE are close when prior bring few information (straight distribution) or data is large (concentrated likelihood)
- Bayesian estimate and MLE are different when prior are strong (concentrated distribution) or data is few (straight likelihood)

Comparing Bayesian, MLE and OLS approaches

- OLS and MLE are close when residuals have compact (normal) distributions
- Bayesian estimate and MLE are close when prior bring few information (straight distribution) or data is large (concentrated likelihood)
- Bayesian estimate and MLE are different when prior are strong (concentrated distribution) or data is few (straight likelihood)

In general, MLE or OLS should be substituted by Bayesian estimates when :

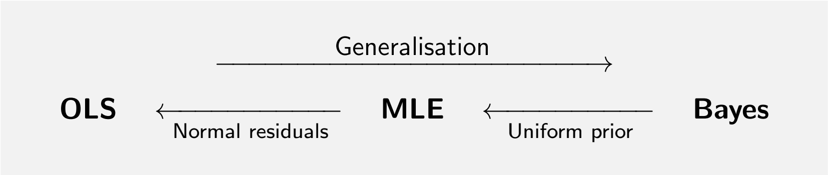
- The dataset is small
- Models are complex (many parameters)
- There are priori on the parameter values
- Dynamical integration of new data

Summary

Approach	**Advantage**	**Inconvenient**
OLS	Easy to use	Sensible to extreme values
MLE	Many strong and useful properties	Asymptotic theory (valid if enough data)
Bayes	Flexible / Valid for any sample size	Can strongly depend on prior

Summary

Approach	**Advantage**	**Inconvenient**
OLS	Easy to use	Sensible to extreme values
MLE	Many strong and useful properties	Asymptotic theory (valid if enough data)
Bayes	Flexible / Valid for any sample size	Can strongly depend on prior



Precision of estimation

Introduction

Punctual estimates give no indication on the precision of estimation

A fitting can be insignificant when it changes from a sample to another (cf. bootstrap)
Significance of the differences between different populations to statute

→ Evaluation of the precision of estimation with confidence intervals

Introduction

Punctual estimates give no indication on the precision of estimation

| A fitting can be insignificant when it changes from a sample to another (cf. bootstrap)
| Significance of the differences between different populations to statute

$\rightarrow$ Evaluation of the precision of estimation with confidence intervals

CI $= [i_-, i_+]$ is a **confidence interval for θ at the confidence level** $1 - \alpha$ if

$$P_\theta(\theta \in \text{CI}) \geq 1 - \alpha, \qquad \forall \theta \in \mathbb{R}^k$$

$\rightarrow$ Parameter θ belongs to CI in more than $1 - \alpha$ % of the cases

- Interval of values with a confidence level instead of punctual estimation
- Precision of estimation of deterministic quantities : Size of the CI reduces as $n \rightarrow \infty$
- Distinct from prediction intervals taking into account the noise to predict new observations

Construction of a confidence interval

The construction of a confidence interval is based on knowledge on the distribution (variability), or on the asymptotic distribution, of an estimator

> If $q_\theta(u)$ is the quantile of the estimator $\tilde{\theta}_n$, then by construction
> $$P_\theta\left(\tilde{\theta}_n(x) \in [q_\theta(\alpha/2), q_\theta(1-\alpha/2)]\right) \geq 1-\alpha, \qquad \forall\theta \in \mathbb{R}^k, \quad \alpha \in (0,1)$$

→ Construction of a CI by extracting θ in the inequalities $\quad \tilde{\theta}_n(x) \in [q_\theta(\alpha/2), q_\theta(1-\alpha/2)]$

Construction of a confidence interval

The construction of a confidence interval is based on knowledge on the distribution (variability), or on the asymptotic distribution, of an estimator

> If $q_\theta(u)$ is the quantile of the estimator $\tilde{\theta}_n$, then by construction
> $$P_\theta\left(\tilde{\theta}_n(x) \in [q_\theta(\alpha/2), q_\theta(1-\alpha/2)]\right) \geq 1-\alpha, \qquad \forall \theta \in \mathbb{R}^k, \quad \alpha \in (0,1)$$

→ Construction of a CI by extracting θ in the inequalities $\tilde{\theta}_n(x) \in [q_\theta(\alpha/2), q_\theta(1-\alpha/2)]$

⚠ Situation generally not accessible since estimator distribution is unknown

- Use of sufficient conditions — Tchebychev inequality
- Asymptotic distribution — Central limit theorem
- Posterior distribution — Bayes approach

Confidence interval with the Tchebychev inequality

Assumption : $x = (X_1, \dots, X_n)$ is a iid P_θ-sample, $\theta = E(X_i)$, for which exists unbiased estimator $\tilde{\theta}_n$ of θ such that $var_\theta(\tilde{\theta}_n) \leq K_n < \infty$

- **Tchebychev inequality** : $$P_\theta\left(|\theta - \tilde{\theta}_n| > \epsilon\right) \leq \frac{K_n}{\epsilon^2}, \qquad \forall \epsilon > 0, \quad \theta \in \mathbb{R}$$
- For $\epsilon = \sqrt{K_n/\alpha}$, $\alpha \in (0,1)$, we get the symmetric CI for θ :

$$P_\theta\Big(\theta \in \underbrace{\Big[\tilde{\theta}_n \pm \sqrt{K_n/\alpha}\Big]}_{\text{CI level } \alpha}\Big) \geq 1 - \alpha$$

Confidence interval with the Tchebychev inequality

Assumption : $x = (X_1, \dots, X_n)$ is a iid P_θ-sample, $\theta = E(X_i)$, for which exists unbiased estimator $\tilde{\theta}_n$ of θ such that $var_\theta(\tilde{\theta}_n) \leq K_n < \infty$

- **Tchebychev inequality** : $$P_\theta\left(|\theta - \tilde{\theta}_n| > \epsilon\right) \leq \frac{K_n}{\epsilon^2}, \qquad \forall \epsilon > 0, \quad \theta \in \mathbb{R}$$
- For $\epsilon = \sqrt{K_n/\alpha}$, $\alpha \in (0,1)$, we get the symmetric CI for θ :

$$P_\theta\Big(\theta \in \underbrace{\left[\tilde{\theta}_n \pm \sqrt{K_n/\alpha}\right]}_{\text{CI level } \alpha}\Big) \geq 1 - \alpha$$

* CI tends to punctual estimator if variability bound K_n tends to zero
* CI tends to $\mathbb{R}$ if $\alpha \to 0$ (θ trivially always belong to CI)
* Tchebychev inequality very large : Parameter belongs to the CI in more than $1 - \alpha$ % of the cases

$\longrightarrow$ *Confidence interval for excess*

Asymptotic confidence intervals

Assumption : $x = (X_1, \ldots, X_n)$ is a iid P_θ-sample, $\theta = E(X_i)$ and $\sigma^2 = var(X_i) < \infty$

- **CLT** : $$P_\theta\big(\sqrt{n}\frac{1/n\sum_i X_i - \theta}{\sigma} \in [q_{\mathcal{N}}(\alpha/2), q_{\mathcal{N}}(1-\alpha/2)]\big) \underset{n\to\infty}{\overset{D}{\to}} 1-\alpha$$
- **Asymptotic symmetric confidence interval** for θ :

$$P_\theta\Big(\theta \in \underbrace{\Big[\frac{1}{n}\sum_i X_i \pm q_{\mathcal{N}}(\alpha/2)\frac{\sigma}{\sqrt{n}}\Big]}_{\text{asymptotic CI level } \alpha}\Big) \to 1-\alpha \qquad \text{as} \quad n \to \infty$$

Asymptotic confidence intervals

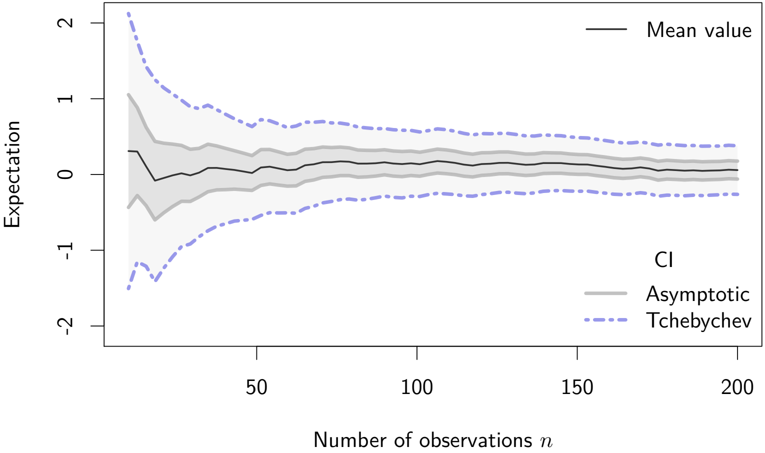
Assumption : $x = (X_1, \ldots, X_n)$ is a iid P_θ-sample, $\theta = E(X_i)$ and $\sigma^2 = var(X_i) < \infty$

- **CLT** : $$P_\theta\Big(\sqrt{n}\frac{1/n\sum_i X_i - \theta}{\sigma} \in [q_{\mathcal{N}}(\alpha/2), q_{\mathcal{N}}(1-\alpha/2)]\Big) \underset{n\to\infty}{\overset{D}{\rightarrow}} 1-\alpha$$
- **Asymptotic symmetric confidence interval** for θ :

$$P_\theta\Bigg(\theta \in \underbrace{\Big[\frac{1}{n}\sum_i X_i \pm q_{\mathcal{N}}(\alpha/2)\frac{\sigma}{\sqrt{n}}\Big]}_{\text{asymptotic CI level } \alpha}\Bigg) \rightarrow 1-\alpha \quad \text{as} \quad n \rightarrow \infty$$

- CI tends to mean value if $\sigma^2 = var(X_i) \rightarrow 0$ or if $n \rightarrow \infty$
- CI tends to $\mathbb{R}$ if $\alpha \rightarrow 0$
- Asymptotic CI still valid substituting σ by empirical estimator σ_x (exact CI : Student)

CI for the expected value of normal distribution



Bayesian credible interval using posterior PDF

Assumption : $x = (X_1, \ldots, X_n)$ is a iid P_θ-sample and $P(\theta)$ is a prior distribution on the parameters such that $P(\theta) > 0$

- **Bayesian credible interval** CI^B of θ given by the quantiles q_x^B of posterior PDF

$$P_\theta\big(\theta \in \underbrace{\big[q_x^B(\alpha/2), q_x^B(1-\alpha/2)\big]}_{\text{Bayesian CI}^B \text{ level } \alpha}\big) \geq 1 - \alpha$$

Bayesian credible interval using posterior PDF

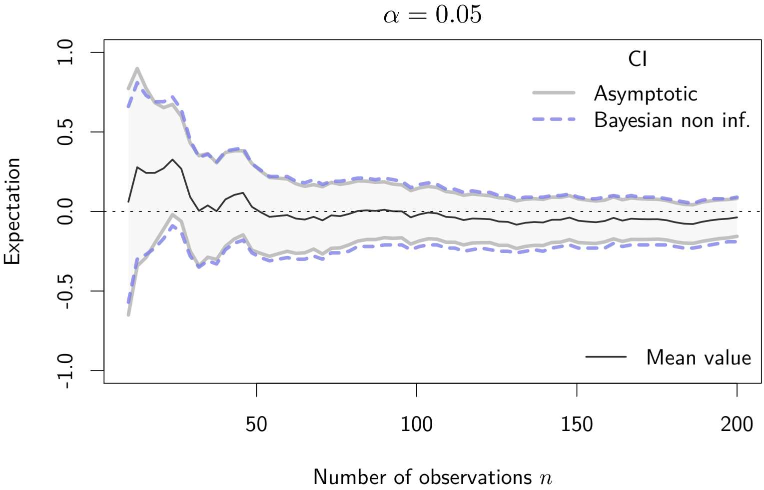
Assumption : $x = (X_1, \ldots, X_n)$ is a iid P_θ-sample and $P(\theta)$ is a prior distribution on the parameters such that $P(\theta) > 0$

- **Bayesian credible interval** CI^B of θ given by the quantiles q_x^B of posterior PDF

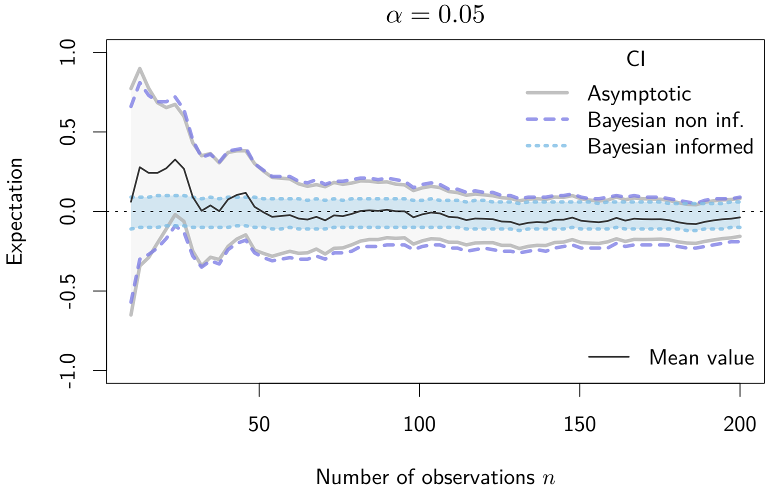
$$P_\theta\left(\theta \in \underbrace{\left[q_x^B(\alpha/2), q_x^B(1-\alpha/2)\right]}_{\text{Bayesian CI}^B \text{ level } \alpha}\right) \geq 1 - \alpha$$

* The size and symmetry of CI^B depends on the posterior distribution that depends on the prior and likelihood
* Asymptotic CI converges to the uninformed Bayes CI^B with uniform prior

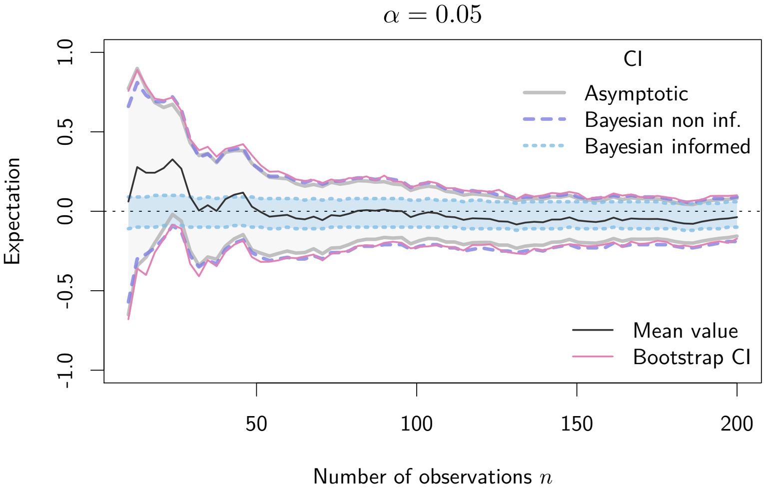
CI for the expected value of normal distribution



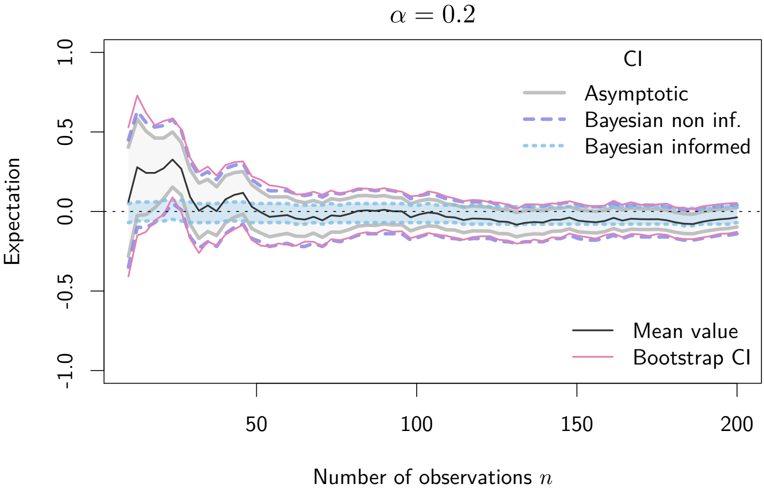
CI for the expected value of normal distribution



CI for the expected value of normal distribution



CI for the expected value of normal distribution



Asymptotic confidence interval for the variance

- **Central limit theorem** :

$$\frac{1}{\sigma^2}\sum_i (x_i - \bar{x}_n)^2 = \frac{(n-1)s_\star^2}{\sigma} \underset{n\to\infty}{\overset{D}{\to}} \chi^2(n-1)$$

with $\chi^2(n-1)$ the Chi-square distribution with $n-1$ degrees of freedom

- **Asymptotic confidence interval for the variance** parameter σ^2

$$P\Big(\sigma^2 \in \underbrace{\Big[\frac{(n-1)s_\star^2}{q_{\chi^2}(1-\alpha/2)}, \frac{(n-1)s_\star^2}{q_{\chi^2}(\alpha/2)}\Big]}_{\text{asymptotic CI level } \alpha}\Big) \underset{n\to\infty}{\to} 1-\alpha$$

Asymptotic confidence interval for the variance

- **Central limit theorem** :
$$\frac{1}{\sigma^2}\sum_i (x_i - \bar{x}_n)^2 = \frac{(n-1)s_\star^2}{\sigma} \underset{n\to\infty}{\overset{D}{\to}} \chi^2(n-1)$$
with $\chi^2(n-1)$ the Chi-square distribution with $n-1$ degrees of freedom

- **Asymptotic confidence interval for the variance** parameter σ^2
$$P\Big(\sigma^2 \in \underbrace{\Big[\frac{(n-1)s_\star^2}{q_{\chi^2}(1-\alpha/2)}, \frac{(n-1)s_\star^2}{q_{\chi^2}(\alpha/2)}\Big]}_{\text{asymptotic CI level } \alpha}\Big) \underset{n\to\infty}{\to} 1-\alpha$$

* Do not require to know the expected value
* Asymmetric CI since Chi-square distribution is asymmetric

Asymptotic confidence interval for linear regressions

Data $(x, y) = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$ Linear model $y_i = ax_i + b + \varepsilon_i$

OLS estimates : $\tilde{a} = a + \frac{\sum_i x_i \varepsilon_i}{\sum (x_i - \bar{x}_n)^2}$ and $\tilde{b} = b + \bar{x}_n \frac{\frac{1}{n}\sum_i x_i \varepsilon_i}{\sum (x_i - \bar{x}_n)^2}$

- The statistics $$\frac{\tilde{a} - a}{s_{\tilde{a}}} \quad \text{and} \quad \frac{\tilde{b} - b}{s_{\tilde{b}}}$$

  with $s_{\tilde{a}} = \sqrt{\frac{1}{n} \sum_i \varepsilon_i^2 / \sum_i (x_i - \bar{x}_n)^2}$ and $s_{\tilde{b}} = \sqrt{\frac{1}{n} \sum_i \varepsilon_i^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{\sum_i (x_i - \bar{x}_n)^2} \right)}$

  have asymptotically a Student distribution t_{n-2} with $n - 2$ degrees of freedom (CLT)

- **Asymptotic confidence interval** with risk level α for the coefficients a and b of the linear regression :
$$\tilde{a} \pm q_{t_{n-2}}(\alpha/2) s_{\tilde{a}} \quad \text{and} \quad \tilde{b} \pm q_{t_{n-2}}(\alpha/2) s_{\tilde{b}}$$

Confidence and prediction bands for linear regressions

Confidence band R : `predict(object,x,'confidence',level)`

Interval of estimation with confidence level $1-\alpha$ for the mean at a given abscissa $x^\star$

$$\tilde{a}\,x^\star + \tilde{b} \pm q_{t_{n-2}}(\alpha/2)\tilde{\sigma}\sqrt{\frac{1}{n} + \frac{(x^\star - \bar{x}_n)^2}{\sum_i (x_i - \bar{x}_n)^2}}$$

Confidence and prediction bands for linear regressions

Confidence band R : `predict(object,x,'confidence',level)`

Interval of estimation with confidence level $1-\alpha$ for the mean at a given abscissa $x^\star$

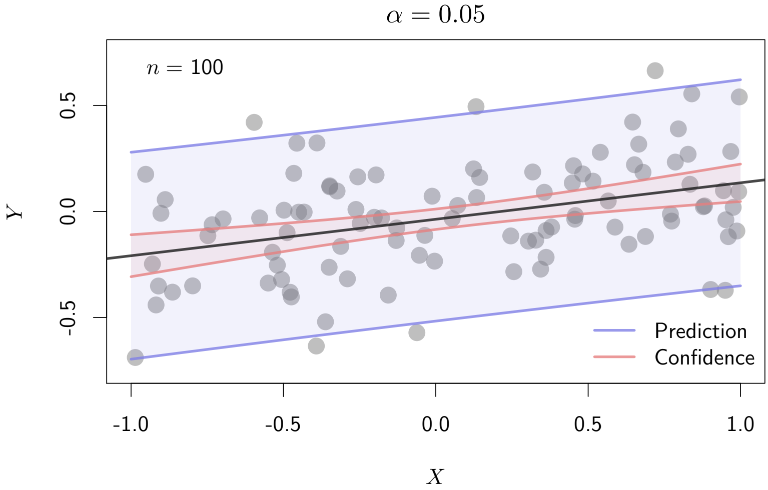
$$\tilde{a}\,x^\star + \tilde{b} \pm q_{t_{n-2}}(\alpha/2)\tilde{\sigma}\sqrt{\frac{1}{n} + \frac{(x^\star - \bar{x}_n)^2}{\sum_i (x_i - \bar{x}_n)^2}}$$

Prediction band R : `predict(object,x,'predict',level)`

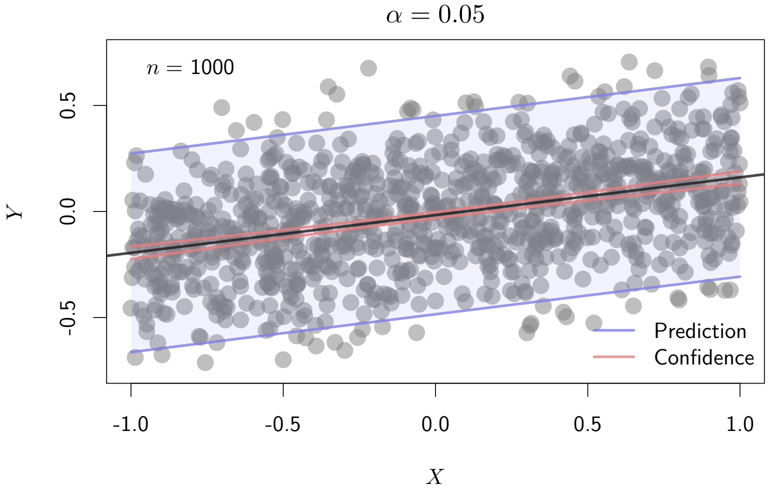
Interval of prediction of a new observation at $x^\star$ with confidence level $1-\alpha$

$$\tilde{a}\,x^\star + \tilde{b} \pm q_{t_{n-2}}(\alpha/2)\tilde{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x}_n)^2}{\sum_i (x_i - \bar{x}_n)^2}}$$

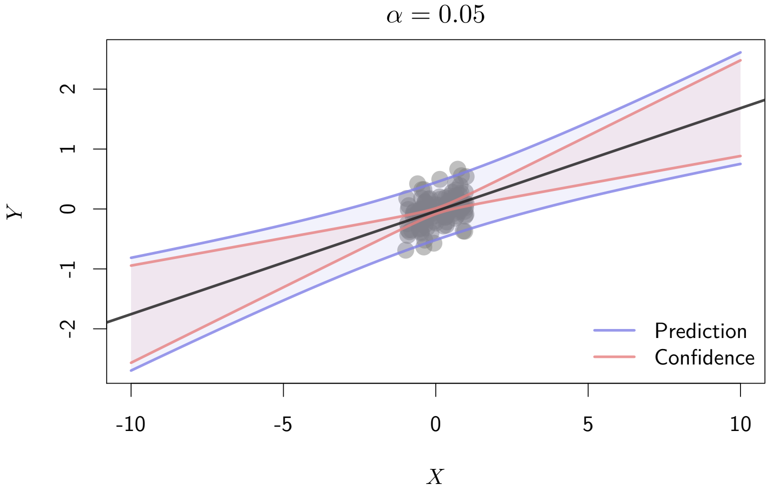
Confidence and prediction bands for a linear regression



Confidence and prediction bands for a linear regression



Confidence and prediction bands for a linear regression



Confidence interval with R

- **Confident interval** `confint(object,level)`
- **Confident band** `predict(object,x,'confidence',level)`
- **Prediction band** `predict(object,x,'predict',level)`

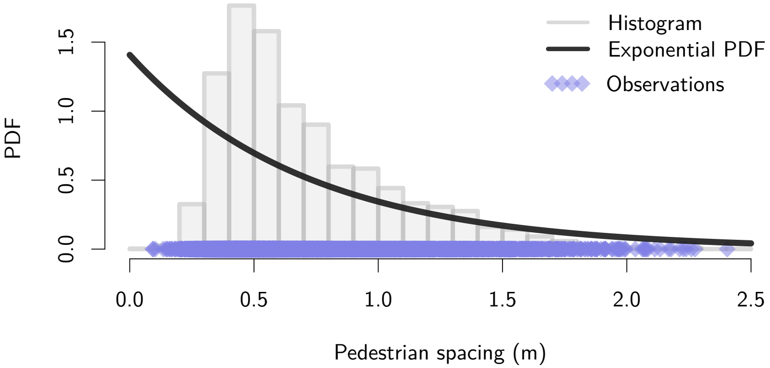
> Generic function for any fitted model object
>
> `level` is the confidence level
>
> Default method assume asymptotic normal distribution for the residuals (asymptotic CI)

Example

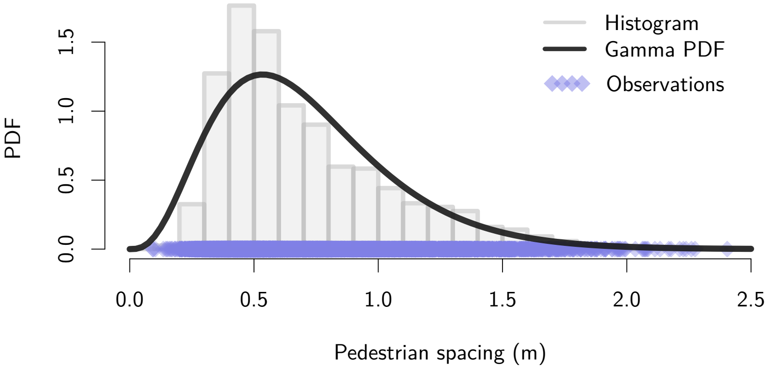
```
object=lm(y~x)
confint(object,0.95)
predict(object,data.frame(1:100),interval='confidence',0.95)
```

Information criteria

Fit of the spacing with exponential distribution



Fit of the spacing with gamma distribution



Comparison of models

MLE and posterior PDF allow to find an optimal fit of the parameters
CI allows to evaluate the precision of this fit

$\rightarrow$ No indication on the quality of description of the data using the optimal fit

Example : Better fit of pedestrian spacing using gamma distribution than exponential

Comparison of models

MLE and posterior PDF allow to find an optimal fit of the parameters
CI allows to evaluate the precision of this fit

$\rightarrow$ No indication on the quality of description of the data using the optimal fit

Example: Better fit of pedestrian spacing using gamma distribution than exponential

Quality of a model evaluated by information criteria

Akaike Information Criterion (AIC)

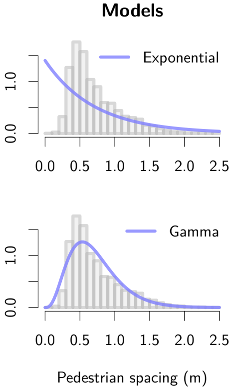
$$\text{AIC} = 2k - 2\ln(L)$$

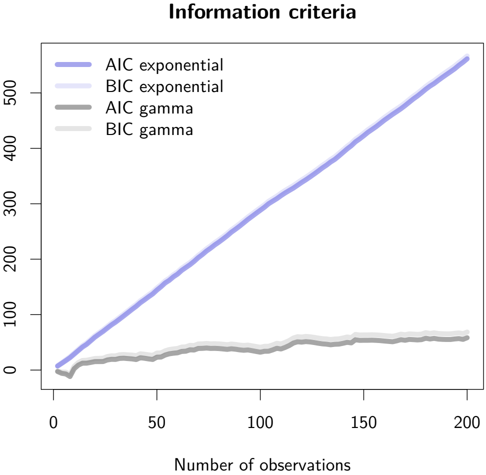
Bayesian Information Criterion (BIC)

$$\text{BIC} = k\ln(2\pi n) - 2\ln(L)$$

- Compromise between goodness of the fit through maximum likelihood L and the complexity of the model through the parameter number k
- Better model minimizes criteria

Information criteria for the fit of the spacing





Likelihood ratio and Bayes factor

- **Likelihood ratio** D : Ratio of the maximum likelihood

$$D = \frac{\max_{\theta_1} L_1(\theta_1)}{\max_{\theta_2} L_2(\theta_2)}$$

$\rightarrow$ Better fit of the model 1 compared to model 2 if $D > 1$ or $\log D > 0$

Likelihood ratio and Bayes factor

- **Likelihood ratio** D : Ratio of the maximum likelihood

$$\mathsf{D} = \frac{\max_{\theta_1} L_1(\theta_1)}{\max_{\theta_2} L_2(\theta_2)}$$

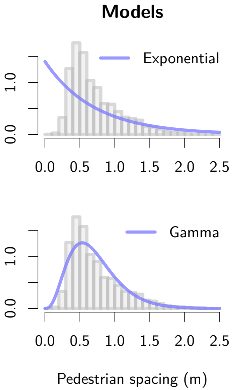
$\rightarrow$ Better fit of the model 1 compared to model 2 if $D > 1$ or $\log D > 0$

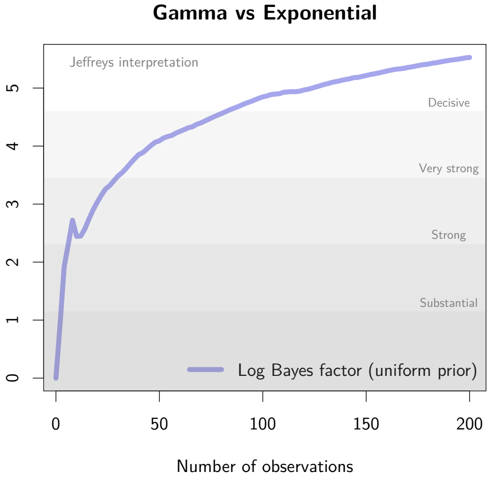
- **Bayes factor** BF : Ratio of the mean likelihood over given prior f_1 and f_2

$$\mathsf{BF} = \frac{\int L_1(\theta) f_1(\theta)\,\mathsf{d}\theta}{\int L_2(\theta) f_2(\theta)\,\mathsf{d}\theta}$$

$\rightarrow$ Better fit of the model 1 when $BF > c$ or $\log BF > \log c$
(cf. Jeffreys interpretation)

Likelihood ratio and Bayes factor for the fit of the spacing





Test of hypothesis

Neyman Pearson statistical test

Statistical test : Test of a null hypothesis H_0 against an alternative hypothesis on a sample of iid data

- → The goal is to test the validity of H_0 (and not H_1 — asymmetric approach)
- → In general, hypothesis are $H_0 : \{\theta \in \Theta_0\}$ vs $H_1 : \{\theta \notin \Theta_0\}, \quad \Theta_0 \in \mathbb{R}^k$

Neyman Pearson statistical test

Statistical test : Test of a null hypothesis H_0 against an alternative hypothesis on a sample of iid data

→ The goal is to test the validity of H_0 (and not H_1 — asymmetric approach)

→ In general, hypothesis are $H_0 : \{\theta \in \Theta_0\}$ vs $H_1 : \{\theta \notin \Theta_0\}, \quad \Theta_0 \in \mathbb{R}^k$

Four possible configurations :

Test \ Reality	H_0 is true	H_0 is false
Reject of H_0	**Error1**	OK
No reject of H_0	OK	**Error2**

- The probability of occurrence of Error1 is $\alpha \in (0, 1)$ — Valid for any number of observations
- The probability of occurrence of Error2 tends to zero as $n \to \infty$ — Power of the test

Construction and usage of a test

A test is based on a statistic S for which the distribution | is known under H_0 / diverges under H_1

- Construction of a region of rejection R_α of H_0

$$P_{H_0}(R_\alpha(S)) = P(\text{Error1}) \leq \alpha$$

- Binary response of a test for given α | Reject of H_0 if $S \in R_\alpha$ / No reject otherwise

Construction and usage of a test

A test is based on a statistic S for which the distribution | is known under H_0 / diverges under H_1

- Construction of a region of rejection R_α of H_0

$$P_{H_0}(R_\alpha(S)) = P(\text{Error1}) \leq \alpha$$

- Binary response of a test for given α | Reject of H_0 if $S \in R_\alpha$ / No reject otherwise

P-value : Critical level $\alpha^\star$ such that

$\alpha > \alpha^\star$:	Reject of H_0
$\alpha < \alpha^\star$:	No Reject of H_0

$\alpha^\star$ is the probability to observe the value for S under H_0 — It is not the probability of H_0

Construction and usage of a test

A test is based on a statistic S for which the distribution | is known under H_0 / diverges under H_1

- Construction of a region of rejection R_α of H_0

$$P_{H_0}(R_\alpha(S)) = P(\text{Error1}) \leq \alpha$$

- Binary response of a test for given α | Reject of H_0 if $S \in R_\alpha$ / No reject otherwise

P-value : Critical level $\alpha^\star$ such that

$\alpha > \alpha^\star$:	Reject of H_0
$\alpha < \alpha^\star$:	No Reject of H_0

$\alpha^\star$ is the probability to observe the value for S under H_0 — It is not the probability of H_0

Reject of H_0 if $\alpha^\star$ small (e.g. $\alpha^\star < 0.01$) — **No conclusion otherwise**

Example of the machine

$(X_1, \ldots, X_n)$ is a iid sample of Bernoulli distribution with distribution $p = 0.2$

$\rightarrow$ $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$, $E(X_i) = p$ and $var(X_i) = p(1 - p)$

Test of the hypothesis $H_0 : \{p = 0.2\}$ VS $H_1 : \{p \neq 0.2\}$

Example of the machine

$(X_1, \dots, X_n)$ is a iid sample of Bernoulli distribution with distribution $p = 0.2$

$\rightarrow$ $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$, $E(X_i) = p$ and $var(X_i) = p(1-p)$

Test of the hypothesis $H_0 : \{p = 0.2\}$ VS $H_1 : \{p \neq 0.2\}$

LLN and **TCL**

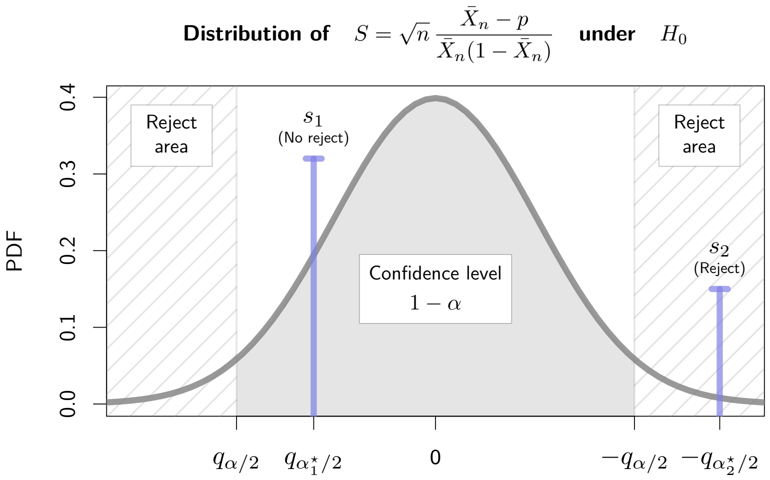
$$S_n = \sqrt{n}\frac{\bar{X}_n - p}{\bar{X}_n(1 - \bar{X}_n)} \rightarrow \begin{cases} \mathcal{N}(0,1) & \text{under } H_0 \\ \pm\infty & \text{under } H_1 \end{cases} \quad \text{as} \quad n \rightarrow \infty$$

Rejection region $R_\alpha(S_n) = |S_n| > \xi_\alpha$ such that $P_{H_0}(|S_n| > \xi_\alpha) \leq \alpha$

- $\xi_\alpha = -q_{\alpha/2}$ i.e. $R_\alpha(S_n) = |S_n| > -q_{\alpha/2}$ with q quantile of normal distribution
- P-value: $\alpha^\star = P(|S_n| > s_n) = \begin{cases} 0.5 \text{ (in average) if } H_0 \text{ is true} \\ 0 \text{ as } n \rightarrow \infty \text{ if } H_1 \text{ is true} \end{cases}$

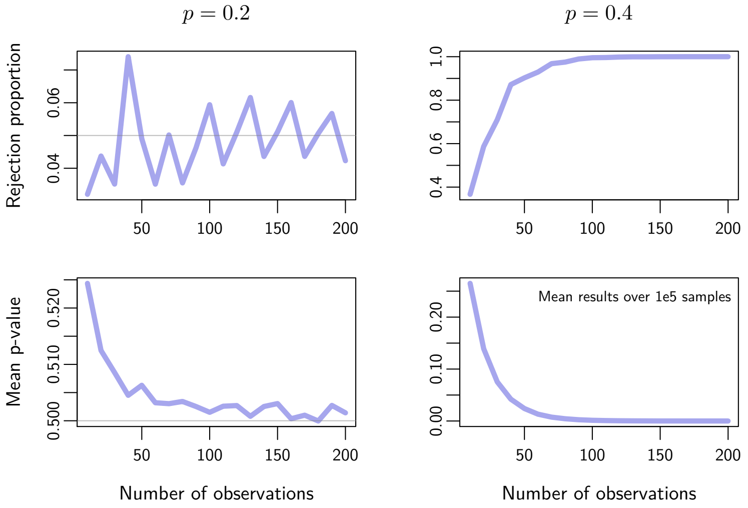
Example of the machine

$H_0 : \{ p = 0.2 \}$ VS $H_1 : \{ p \neq 0.2 \}$ at level $\alpha = 0.05$



Example of the machine

$H_0 : \{p = 0.2\}$ VS $H_1 : \{p \neq 0.2\}$ at level $\alpha = 0.05$



Some tests with R

Test for	Statistic	Distribution	R
Mean value $\{\mu = \mu_0\}$	$\sqrt{n}\,\frac{\bar{x}-\mu_0}{s_x}$	Student	`t.test(x,mu0)`
Variance $\{\sigma = \sigma_0\}$	$(n-1)\,\frac{s_x^2}{\sigma_0^2}$	Chi–squared	—
Mean equality $\{\mu_1 = \mu_2\}$	$\frac{\bar{x}-\bar{y}}{\left(s_x^2/n_1+s_y^2/n_2\right)^{1/2}}$	Student	`t.test(x,y)`
Variance equality $\{\sigma_1 = \sigma_2\}$	s_x^2/s_y^2	Fisher	`var.test(x,y)`
Adequacy of discrete distribution	$\frac{\sum_i (E_i-O_i)^2}{E_i}$	Chi–squared	`chisq.test(x,p)`
Adequacy of continuous distribution	$\sup_z \lvert D_x(z) - D_y(z)\rvert$	Kolmogorov	`ks.test(x,y)`
Normality	$\frac{\left(\sum_i a_i x^{(i)}\right)^2}{n\, s_x^2}$	Shapiro-Wilk	`shapiro.test(x)`
Independence	$\frac{\sum_i (nE_{i,j}-E_iE_j)^2}{nE_iE_j}$	Chi–squared	`chisq.test(x,y)`

Parametric clustering

Parametric clustering (density- or distribution-based clustering)

Assumption : Observations as mixture of identical models with different parameter values

Gaussian mixture model Multivariate normal distribution

- Observables : Data x supposed to be iid observations of a multivariate normal distribution f
- Parameters : $\theta_k = (\mu_k, \sigma_k)$ of the Gaussian mixture and the proportions of observations per cluster π_k, $k = 1, \ldots, K$

$\rightarrow$ Log-likelihood :

$$\mathcal{L}_\theta(x) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k f(x_i, \theta_k) \right)$$

Parametric clustering (density- or distribution-based clustering)

Assumption : Observations as mixture of identical models with different parameter values

Gaussian mixture model Multivariate normal distribution

- Observables : Data x supposed to be iid observations of a multivariate normal distribution f
- Parameters : $\theta_k = (\mu_k, \sigma_k)$ of the Gaussian mixture and the proportions of observations per cluster π_k, $k = 1, \ldots, K$

→ Log-likelihood :

$$\mathcal{L}_\theta(x) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f(x_i, \theta_k)\right)$$

Likelihood maximisation according to parameters $(\mu_k, \sigma_k, \pi_k),\ k = 1, \ldots, K$

1. Local optimum for fixed K through iterative algorithms EM, Gipps sampling, VB, ...
2. Selection of the cluster number K with information criteria AIC, BIC, likelihood ratio, ...

Gaussian mixture model with R : `Mclust(data)` Package : `mclust`

`Mclust(data,modelNames)` : Gaussian mixture for multivariate dataset fitted via

EM algorithm and **BIC criterion**

Gaussian mixture model with R : `Mclust(data)` Package : `mclust`

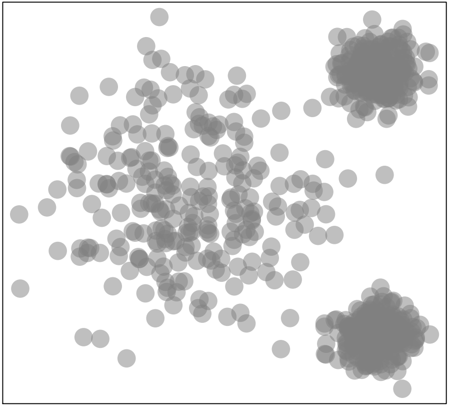
`Mclust(data,modelNames)` : Gaussian mixture for multivariate dataset fitted via

EM algorithm and **BIC criterion**

Several shapes for the cluster can be used Option : `modelNames`

- `EII` : Spherical, equal volume
- `VII` : Spherical, varying volume
- `EEV` : Ellipsoidal, equal volume & shape
- `VEV` : Ellipsoidal, equal shape
- `EVV` : Ellipsoidal, equal volume
- `VVV` : Ellipsoidal, varying volume & shape

Observations

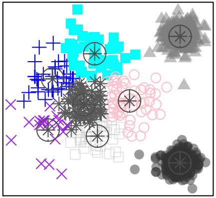


Mclust : Example 1

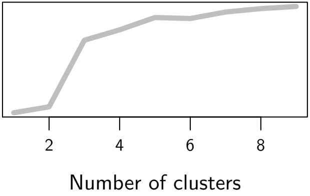
Spherical clusters

EII : Spherical, equal volume

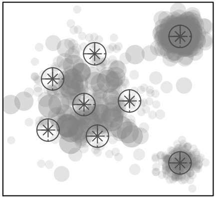
Classification



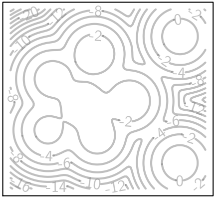
BIC criterion



Uncertainty



log Density Contour Plot

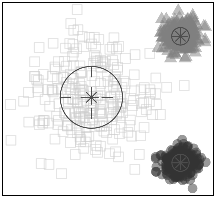


Mclust : Example 1

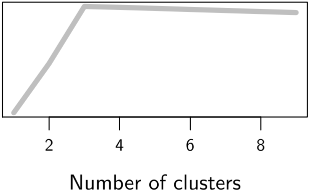
VII : Spherical, varying volume

Spherical clusters

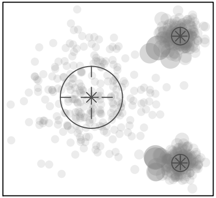
Classification



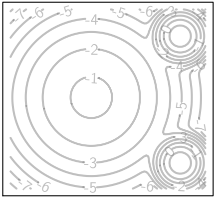
BIC criterion

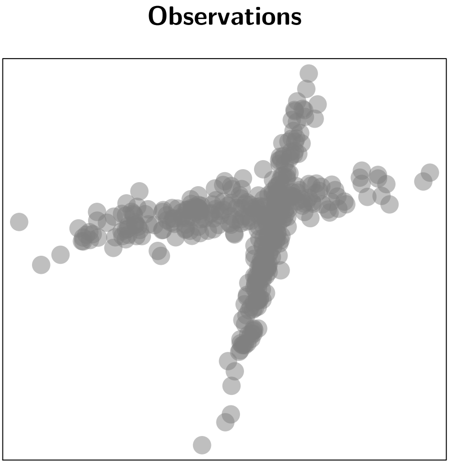


Uncertainty



log Density Contour Plot



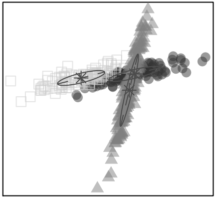


`Mclust` : Example 2

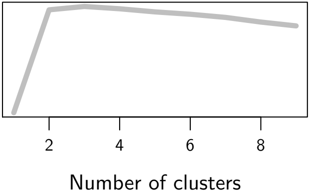
`EVV` : Ellipsoidal, equal volume

Linear clusters

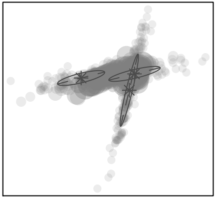
Classification



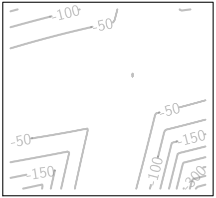
BIC criterion



Uncertainty



log Density Contour Plot

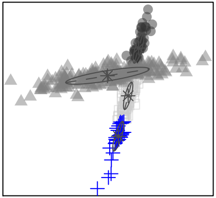


`Mclust` : Example 2

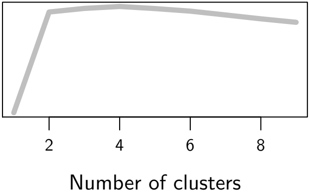
Linear clusters

`VEV` : Ellipsoidal, equal shape

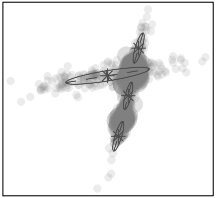
Classification



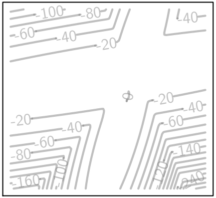
BIC criterion



Uncertainty



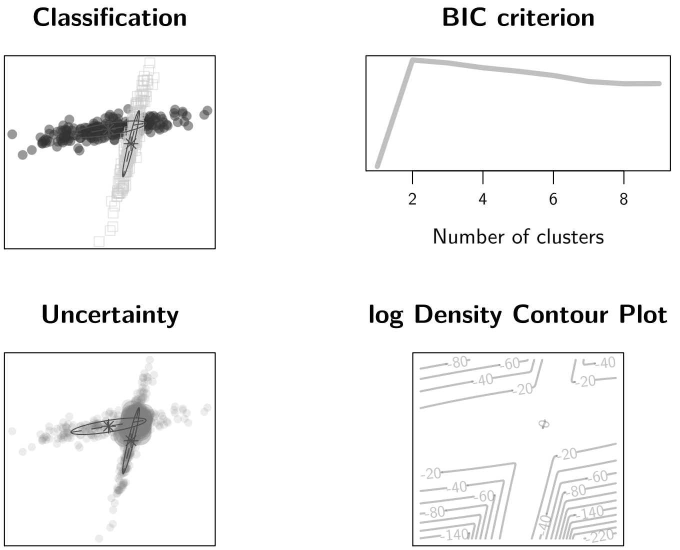
log Density Contour Plot



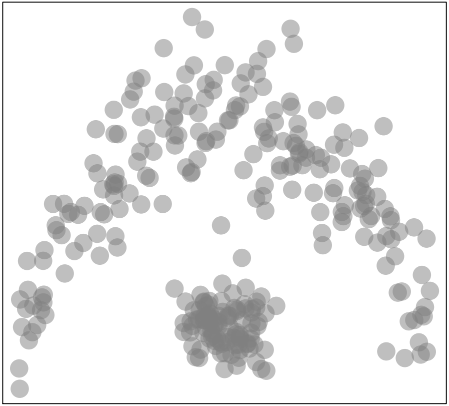
Mclust : Example 2

VVV : Ellipsoidal, varying volume & shape

See also mixture of linear models here



Observations

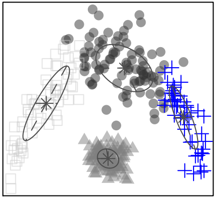


`Mclust` : Example 3

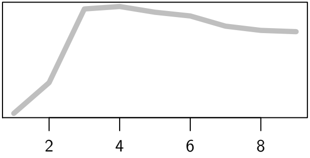
Irregular clusters : Non-parametric clustering

`VVV` : Ellipsoidal, varying volume & shape

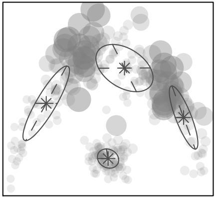
Classification



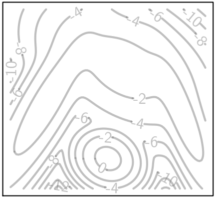
BIC criterion



Uncertainty



log Density Contour Plot



Parametric statistic : Summary

- In parametric statistic, the data are supposed to be samples of independent and identically distributed (iid) random variables

  $\rightarrow$ Estimation of the parameters of the distributions

  - **Punctual estimation** (Maximizing the likelihood or posterior distribution)
  - **Precision of the estimation** (confidence and credible intervals)
  - **Goodness of the fit and test of hypothesis** (AIC, BIC, Bayes factor, test for mean value, variance, independence, adequacy to distributions etc. . .)

- **The likelihood** is a fundamental function in parametric statistic

- **Bayesian approaches** are useful when we have prior on the parameters, the size of the sample are small or the models are complex

- Statistics based on square error are accurate when observations are distributed on 'compact' supports (like normal ones)

  ⚠ **High extreme values** can bring disproportionate weights

Summary

Descriptive statistic allows to describe data without modelling assumptions

→ Exploration of the data — Knowledge database discovery, data mining, big data

→ Elaboration of data-based models — Senseless parameters

Parametric statistic allows to obtain precise assessments on statistical models

→ Parameter estimation, confidence interval, information criteria, test of hypothesis

→ Assumptions on the distribution of the data — Meaningful parameters

Summary

Descriptive statistic allows to describe data without modelling assumptions

→ Exploration of the data — Knowledge database discovery, data mining, big data

→ Elaboration of data-based models — Senseless parameters

Parametric statistic allows to obtain precise assessments on statistical models

→ Parameter estimation, confidence interval, information criteria, test of hypothesis

→ Assumptions on the distribution of the data — Meaningful parameters

R and its numerous packages and help forums is a practical software for both descriptive and parametric data analysis

References and links

Books

- T.W. Anderson & J.D. Finn *The statistical analysis of data* Springer 1996
- D. Montgomery & G. Runger *Applied Statistics and Probability for Engineers* Wiley 2010
- P. Congdon *Bayesian statistical modelling* (2nd edition) Wiley 2006

Websites

- The R project for statistical computing `r-project.org`
- Wikipedia : Statistics `wikipedia.org/Statistics`
- Online courses `statistics.com`
- Python & R codes for common machine learning algorithms `analyticsvidhya.com`

Videos

- R vs Python `blog.dominodatalab.com`
- R statistics tutorials `youtube.com`

Integrated development environments for R

- RStudio, Jupyter (online), Rattle, Red-R, R Commander, JGR, RKWard, Deducer, ...

Abbreviations

PDF	*Probability Density Function*
ECDF	*Empirical Cumulative Distribution Function*
iff	*If and only if*
th.	*Theorem*
ind.	*Independent*
iid	*Independent and identically distributed*
OLS	*Ordinary Least Squares*
PCA	*Principal Component Analysis*
lc	*Linear combination*
D	*Distribution*
P	*Probability*
a.s.	*Almost surely*
LLN	*Law of Large Numbers*
CLT	*Central Limit Theorem*
MSE	*Mean Squared Error*
MLE	*Maximum Likelihood Estimator*

Overview

Part 1 Descriptive statistics for univariate and bivariate data

Repartition of the data (histogram, kernel density, empirical cumulative distribution function), order statistic and quantile, statistics for location and variability, boxplot, scatter plot, covariance and correlation, QQplot

Part 2 Descriptive statistics for multivariate data

Least squares and linear and non-linear regression models, principal component analysis, principal component regression, clustering methods (K-means, hierarchical, density-based), linear discriminant analysis, bootstrap technique

Part 3 Parametric statistic

Likelihood, estimator definition and main properties (bias, convergence), punctual estimate (maximum likelihood estimation, Bayesian estimation), confidence and credible intervals, information criteria, test of hypothesis, parametric clustering

Appendix LaTeX plots with R and Tikz

Appendix 1 : Plotting with R

R is not only a software for data analysis and mathematical modelling, it is also a software to get graphics[4]

→ Basically R allows to produce figures in Metafile, Postscript, PDF, Png, Bmg, TIFF, jpg

→ `tikzDevice` package allows to get LaTeX file (.tex)

Simple plot `plot(x,y)`

- Options `xlab, ylab, main, ...`
- Legends `legend('topright', ...)`
- Specification of the axis label `axis(1, ...)`

Multiplot

- Figures with 2 lines of 3 plots `par(mfrow=c(2,3));plot()...`
- Customized position of the plots `split.screen(rbind(...));screen(1)...`
- Scatterplot of a database `plot(data_base)`

[4] See `demo(graphics)`, package 'ggplot2', CRAN Task View, Google image : R graphics

LaTeX plot with R

Script

```
require(tikzDevice)
tikz('exemple.tex',width=5,height=3,standAlone=T)
curve(sin(x)/x,xlim=c(0,20),xlab='$x$',ylab='$f(x)$',lwd=7,col=rgb(.5,.5,.5))
legend('topright',c('$f(x)=\\frac1x\\sin(x)$'),lwd=7,col=rgb(.5,.5,.5))
dev.off()
```