Institut de Mathématiques de Toulouse, INSA Toulouse

Introduction, Exploration, Unsupervised classification

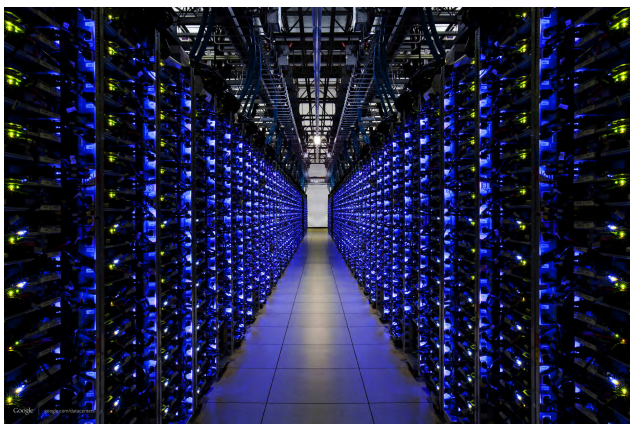
Data Mining
October, 2022

Béatrice Laurent - Philippe Besse - Olivier Roustant

Outline

Introduction

- Statistical learning plays a key role in many fields of sciences, medicine, industry, marketing, finance ..
- The development of data storage and computing resources leads to the storage of a huge amount of data from which the data scientist will try to learn crucial informations to better understand the underlying phenomena or to provide predictions.



- Many fields are impacted, here are some examples of learning problems :
  - **Signals :** Aerospace industry produces a huge amont of signal measurements obtained from thousand of on-board sensors. It is particularly important to detect possible anomalies before launching the satellite. Similarly, many sensors are involved in planes and it is important to detect a abnormal behavior on a sensor.
    $\hookrightarrow$ curve clustering or classification and anomaly detections in a set of curves for predictive maintenance purposes.

  - **Images :** Convolutional neural networks and deep learning lead to important progresses for image classification. Many fields are concerned : medical images (e.g. tumor detection), earth observation satellite images, photos, video surveillance images, handwritten text images ...

  - **Geolocalisation data :** Machine learning based on geolocalisation data has also many potential applications : targeted advertising, road traffic forecasting, monotoring the behavior of fishing vessels ...

- **Consumers preferences data :** Websites and supermarkets collect a huge amount of data on the behavior of consumers. Machine learning algorithms are used to valorize these data (gathered sometimes with personal data such as age, sex, job, adress ..) for recommandation systems ..

- **Microarray data :** DNA microarrays allow to measure the expression of thousands of genes simultaneously on a single individual. It is, for example, a challenge to try to infer from those kind of data which genes are involved in a certain type of cancer. The number p of genes measured on a microarray is generally much larger than the number n of individuals in the study.
  $\hookrightarrow$ Variable selection in high dimension.

From Statistic to AI through *Data Science*

1930-70 h-Octets Statistical inference

1950 Beginnings of Artificial Intelligence : Allan Turing

1970s kO Data analysis and *exploratory data analysis*

1980s MO Neural networks, functional data analysis

1990s GO *Data mining* : pre-acquired data

2000s TO Bioinformatics : $p \gg n$, *Machine Learning*

2008 Data Science

2010s PO Big Data p and n very large

2012 *Deep Learning*

2016 Artificial Intelligence (IA) : AlphaGo

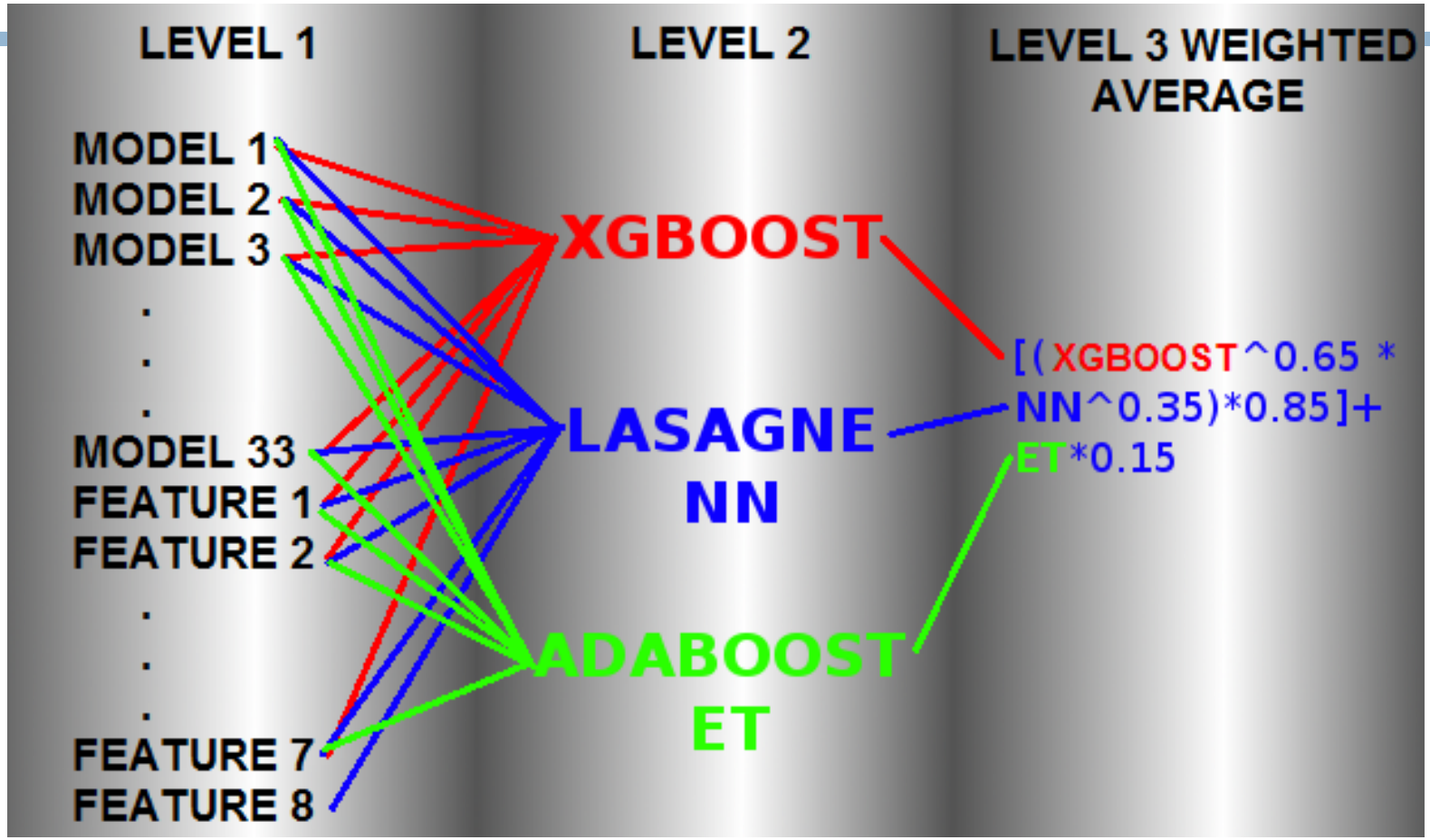
VVV...VV : Volume, Variety, Velocity... Veracity, Valorization

Objectives

- Exploration : description, visualisation, clustering (taxonomy)
- Explanation $Y = f(\boldsymbol{X})$ (supervised learning)
- Prediction and selection, explanable and interpretable models
- "Raw" forecasting (black box models)
- Anomaly Detection

Aim

- Academic publication (*Benchmarks — UCI repository*)
- Valorisation, Industrial solutions
- Kaggle type competition.



Kaggle competition : Identify people who have a high degree of Psychopathy based on Twitter usage.

Usecase Ozone

Aim : Prediction of the ozone concentration for the next day at 5 PM (max. of the day) from a learning sample composed of the explanatory variables $X^1, \ldots X^p$:

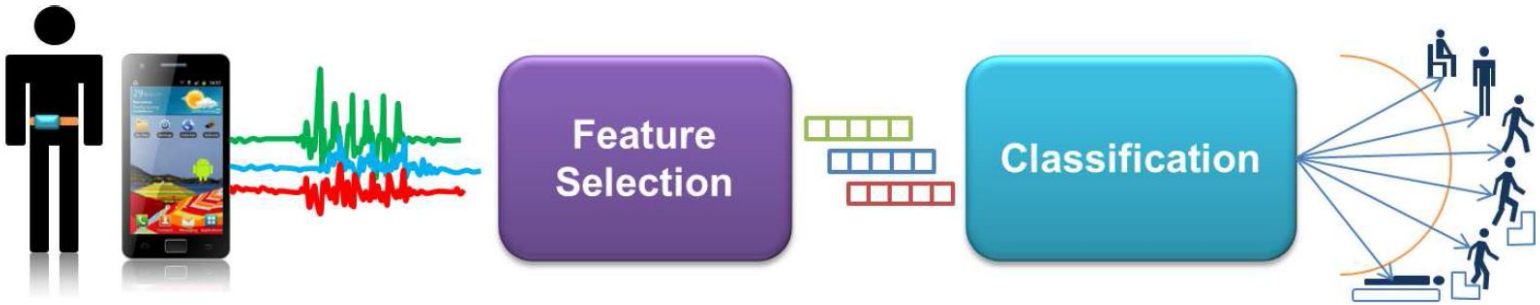
- MOCAGE (deterministic model of Meteo France),
- NO2, NO3,
- H2O,
- Temperature,
- Wind speed and orientation,
- Station,
- Type of day (holiday or not)

and the variable to explain :

- Y : Ozone concentration

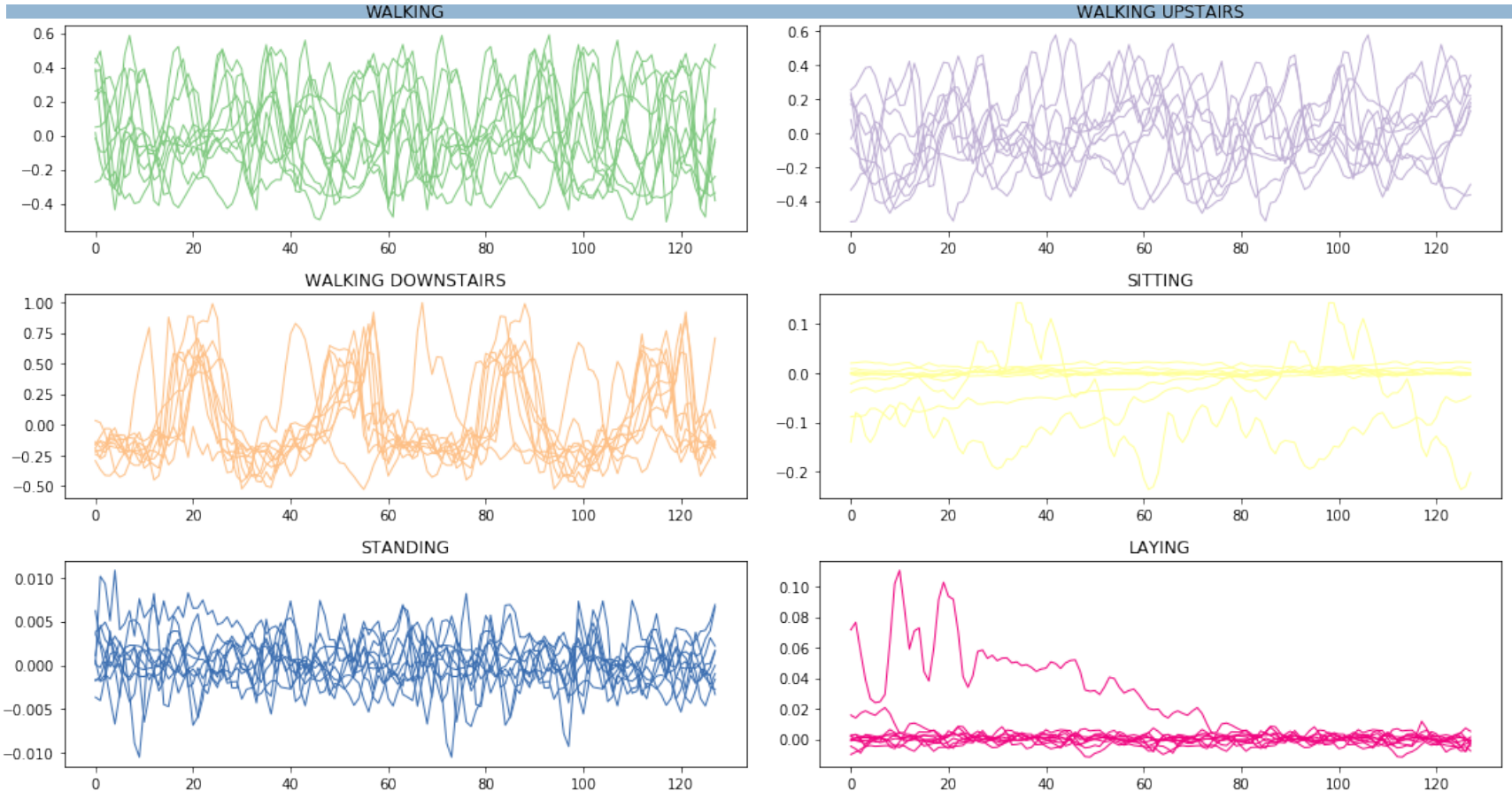
$\hookrightarrow$ **Statistical adaptation**

Usecase HAR



Human activity recognition HAR

- Public data available on *UCI repository*
- 9 signals per individual : accelerations in x, y, z, those by subtracting the natural gravity and angular accelerations in x, y, z obtained from the gyroscope.
- Each signal contains $p = 128$ measures sampled at 64Hz during 2s.
- 7352 samples for learning and 2947 for testing.
- Objectives : Activity recognition (6 classes) standing, sitting, lying, walking, walking upstairs or walking downstairs.



Human activity recognition : acceleration in y by class

HAR First step : "features" variables obtained from signal processing

- $p = 561$ new variables (*features*)
  - Time domain : min, max, means, variances, correlations...
  - Frequency domain : largest, mean, energy per frequency band...

HAR ... to be continued

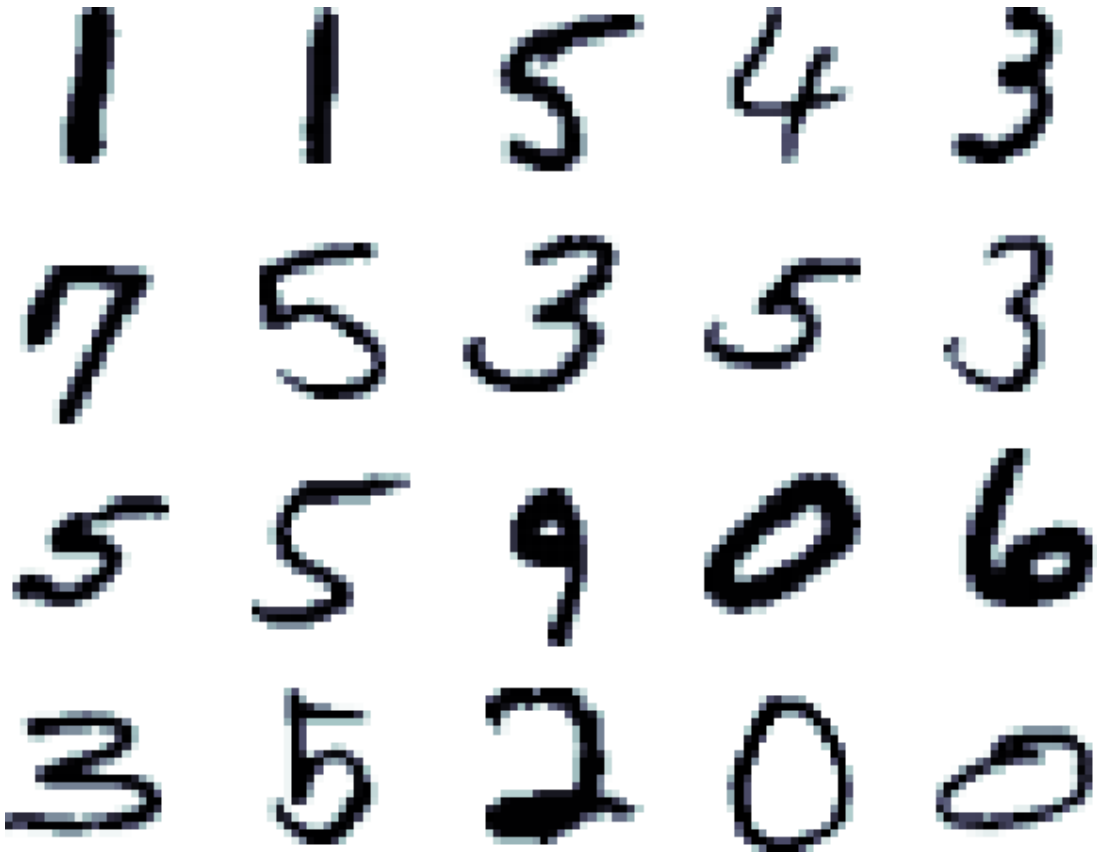
- raw signals and *deep learning*

Usecase MNIST

MNIST dataset

- Yann le Cun website
- 60 000 handwritten digits, $28 \times 28 = 784$ pixels
- Test : 10 000 images
- Classical methods (k-nn, Random Forests)
- Preprocessing : normalisation of the images
- Specific Distance with invariance properties
- Deep learning : *TensorFlow, Keras*

Usecase MNIST



MNIST : some examples of handwritten digits

Unsupervised vs Supervised learning

- In the framework of **supervised learning**, we have a **learning sample** composed of observation data of the type **input/output** :

$$d_1^n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$$

where $\boldsymbol{x}_i = (x_i^1, x_i^2, \ldots, x_i^p)$ is a p-dimensional variable (quantitative or qualitative or both), $y_i \in \mathcal{Y}$ for $i = 1 \ldots n$ is the variable to explain (label).

Objectives : From the learning sample, we want to

- **Estimate** the link between the input vector $\boldsymbol{x}$ (explanary variables) and the output y :

$$y = f(x^1, x^2, \ldots, x^p)$$

- **Predict** the output y associated to a new entry $\boldsymbol{x}$,
- **Select** the important explanatory variables among $x^1, \ldots, x^p$.

Unsupervised vs Supervised learning

- **Supervised learning**,

quantitative output	qualitative output
$\mathcal{Y} \subset \mathbb{R}^p$	$\mathcal{Y}$ finite
$\downarrow$	$\downarrow$
regression	**classification**

- In the framework of **unsupervised learning**, we only observe $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$.

  Objectives :

  - Find underlying structures in these unlabeled data.
  - Clustering.

Steps of Data Science

1. Data management
2. Exploration
3. Modeling, supervised or unsupervised learning
4. Scaling up, industrialization

Step 0 : Data management

Data management

- Acquisition
- Archiving
- Extractions
- Fusion
- Quantification : frequency of occurence, discretization ...

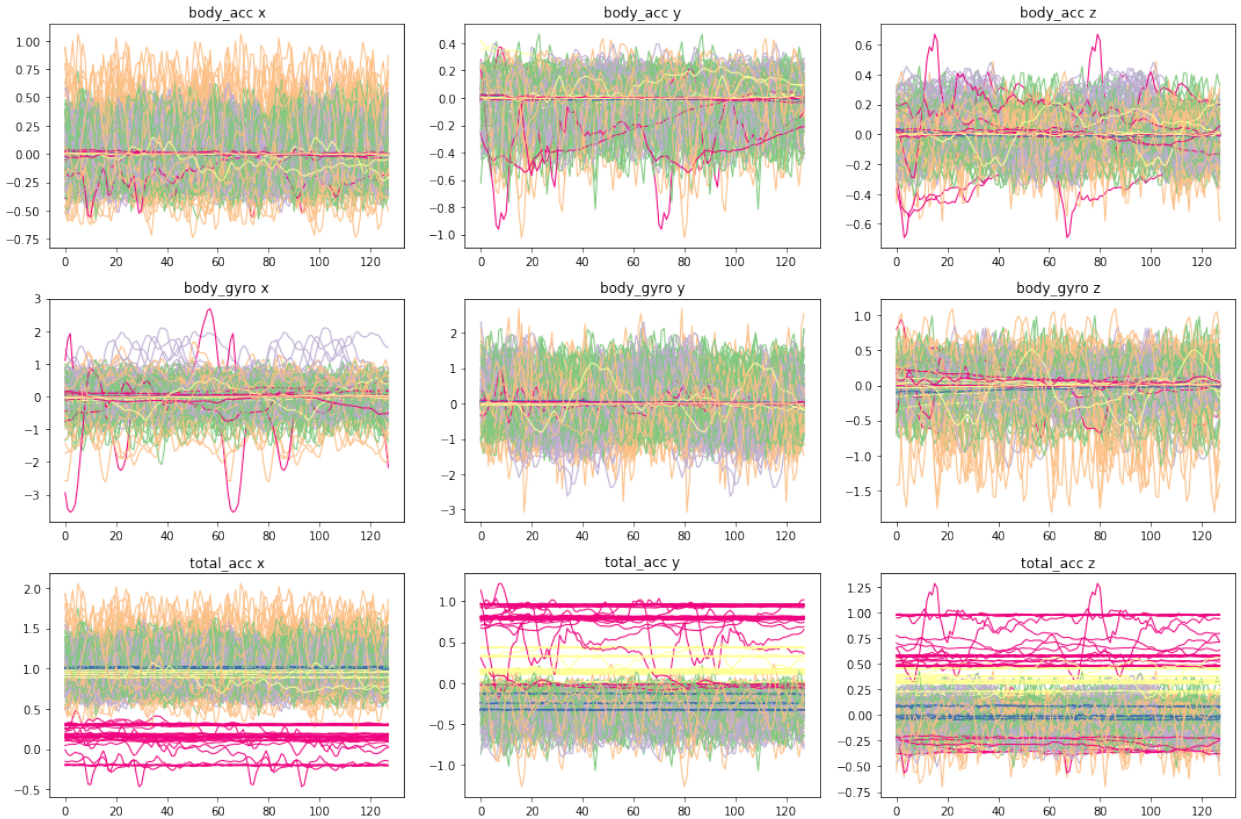
Data (matrix)

- p variables $X = (X^1, \ldots, X^p)$
- n individuals or observations
- Y variable to explain

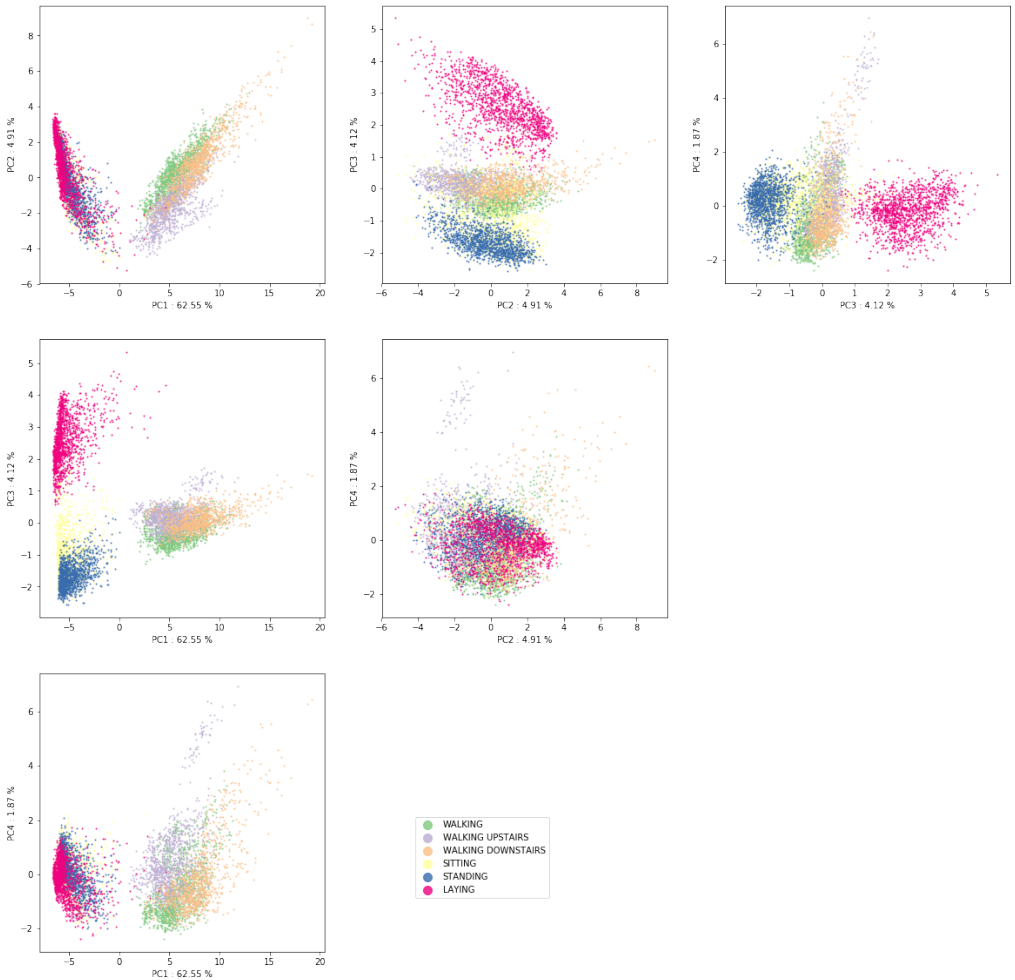
Step 1 : Exploration

Data munging

- Extraction, cleaning, consistency check
- Elementary statistics : univariate, bivariate
- Distribution, transformation of the variables
- New variables or features
- Outlier detection
- Missing data (imputation ?)
- Dimension reduction : Principal Component Analysis, Factorial Discriminant Analysis
- Visualisation



HAR : raw signals



HAR : Principal component analysis of the signal processing features

Step 2 : Modeling, supervised or unsupervised learning

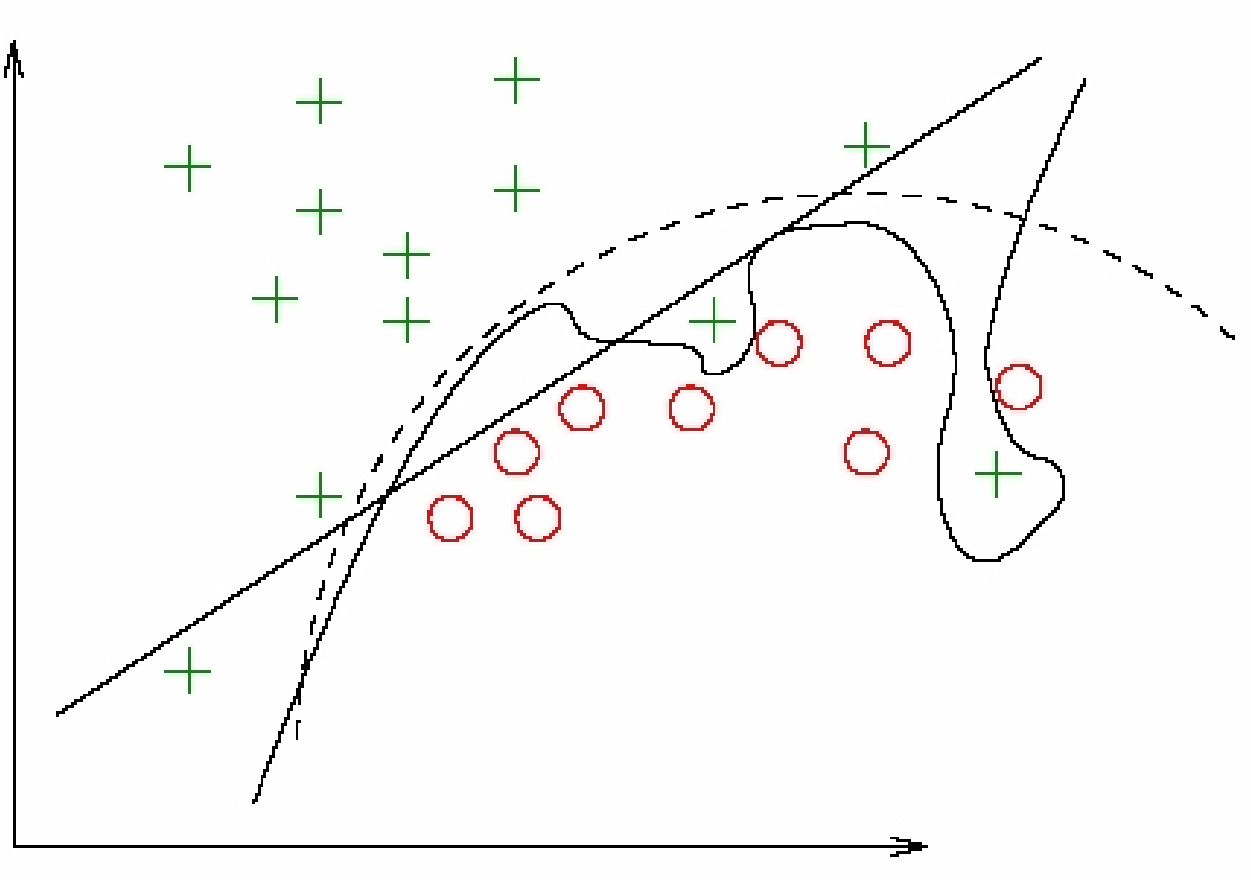
Unsupervised Classification (*clustering*)

- HCA : Hierarchical Cluster Analysis
- *k*-means, *k*-medoids (PAM)
- Gaussian mixtures
- DBSCAN, Self Organizing Map

Supervised Learning (Classification)

- Logistic Regression
- Discriminant analysis, k-nearest neighbours
- Support Vector Machine
- Classification And regression Trees (CART)
- Bagging, Random Forests
- Boosting
- Neural Networks, Deep Learning

Supervised learning : risk of overfitting

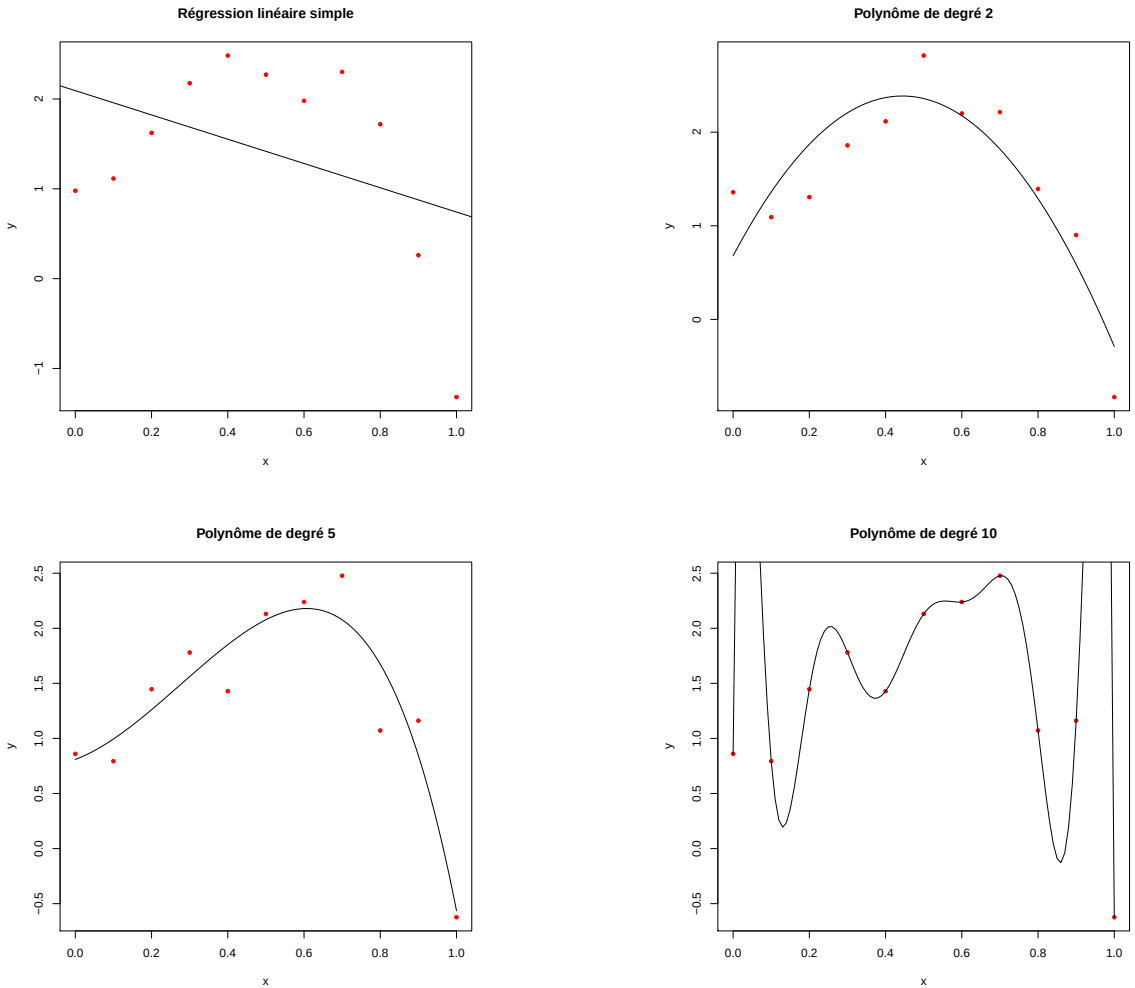


Model complexity in supervised classification

Supervised Learning (Regression)

- Linear regression
- k nearest neighbours
- Support Vector Regression
- Classification And Regression Trees (CART)
- Bagging, Random Forests
- Boosting
- Neural Networks, Deep Learning

Supervised learning : risk of overfitting



Model complexity in polynomial regression

Strategy for supervised *Learning*

1. Random Partition of the sample : learning, (validation), test
2. **For** each method that we consider :
   - Learning (estimation) depending on θ (complexity)
   - Optimization of θ : validation set or cross-validation with the learning set
3. Comparison of the methods : prediction error on the test sample
4. Eventual Iteration (*Monte Carlo*)
5. Choice of the method (prevision *vs.* interpretability).
6. Estimation of the selected model with all the sample, exploitation

Possibly : Aggregation of several models

Outline

- Introduction
- Multidimensional Exploration
  - Basic tools
  - Principal components analysis
  - Factorial discriminant analysis
- Unsupervised classification

Multidimensional Exploration

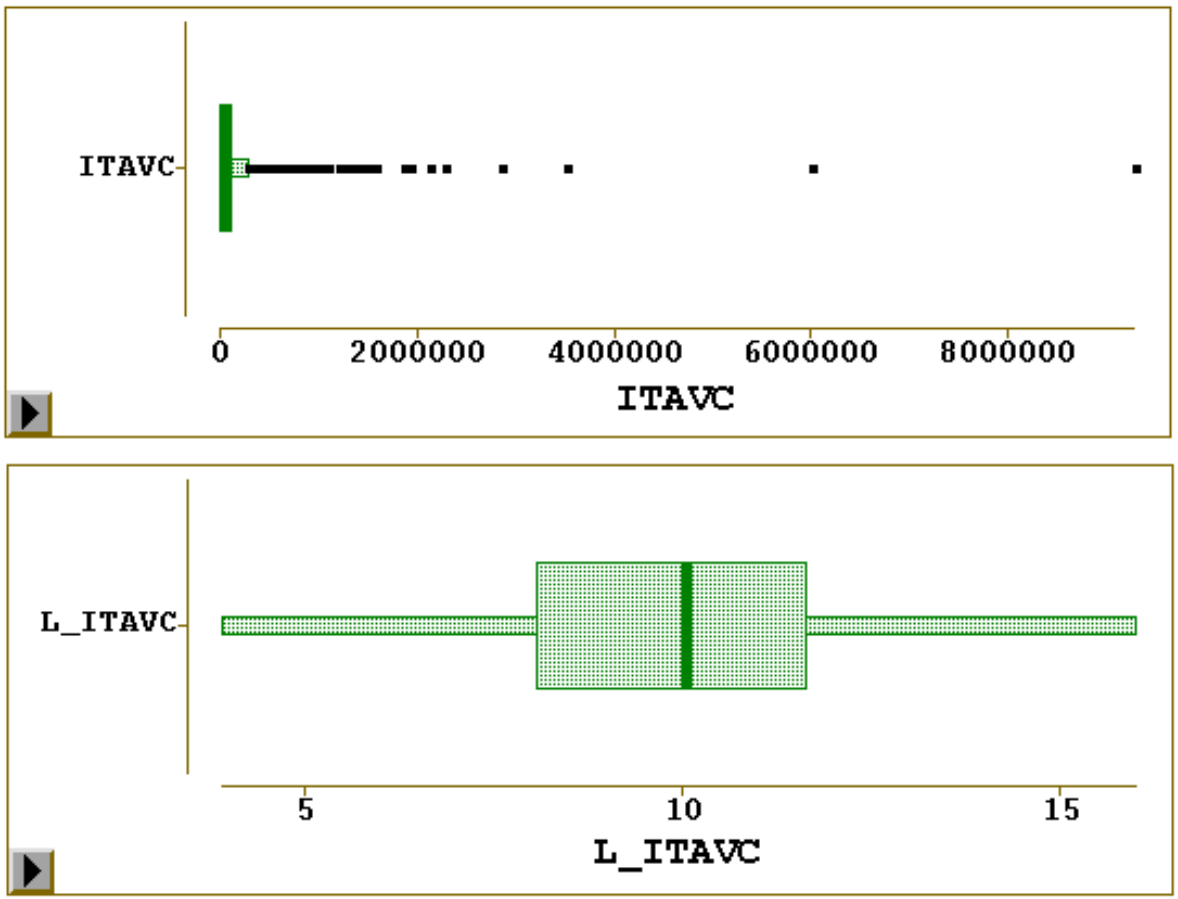
Some basic tools

- Univariate and bivariate description
- Principal components analysis

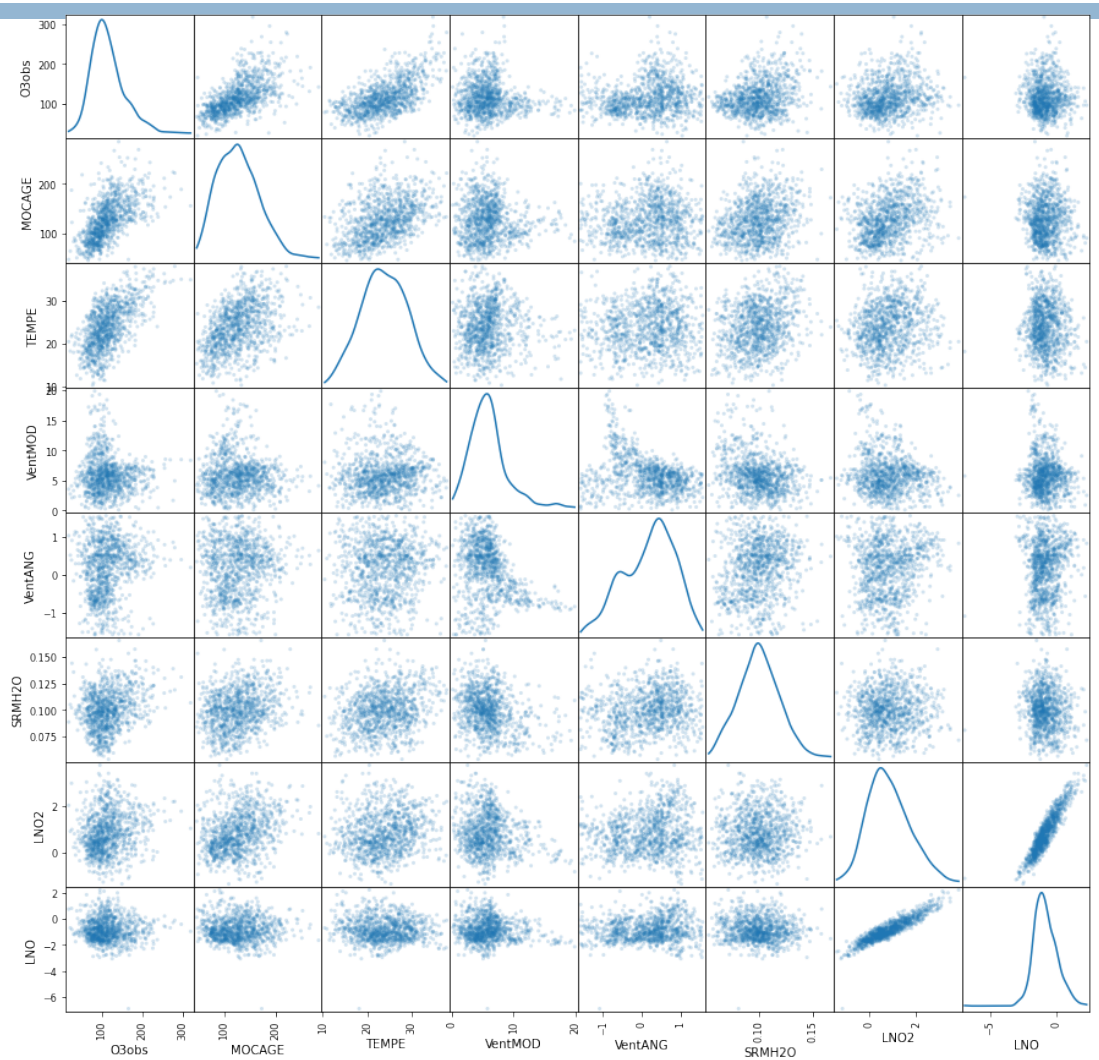
Diagnosis

- Error detection, inconsistency
- Detection of outliers
- Normality of the distributions

Log-normal variable transformation



ScatterPlot



Ozone data set

Outline

- Introduction
- Multidimensional Exploration
  - Basic tools
  - Principal components analysis
  - Factorial discriminant analysis
- Unsupervised classification

Principal Component Analysis : objectives

- We have a dataset composed with n individuals on which we observed p variables.
- If $p > 2$ (or 3), it is impossible to represent the individuals.

Objectives of PCA :

- Perform a dimension reduction to visualize the data.
- Identify trends on individuals (profiles) and "typical individuals" (e.g. in marketing, or in clinical trials)
- Have a synthetic visualization of the main links between the variables.
- Represent the data in smaller dimensions, avoiding "redundancies".

PCA only works on quantitative variables.

Illustrative example

	Mathematics	Physics	French	English
Nathan	6	6	5	5.5
Emma	8	8	8	8
Lola	6	7	11	9.5
Baptiste	14,5	14,5	15,5	15
Mathilde	14	14	12	12.5
Ines	11	10	5.5	7
Lucas	5.5	7	14	11.5
Aymeric	13	12.5	8.5	9.5
Chloe	9	9.5	12.5	12

TABLE – *Notes (fictitious) of a "panel" of 9 students..*

Data preprocessing

Raw data

$$X = \begin{pmatrix} X_{11} & \dots & X_{1j} & \dots & X_{1p} \\ \vdots & & \vdots & & \vdots \\ X_{i1} & \dots & X_{ij} & \dots & X_{ip} \\ \vdots & & \vdots & & \vdots \\ X_{n1} & \dots & X_{nj} & \dots & X_{np} \end{pmatrix}$$

Data preprocessing

Raw data

$$X = \begin{pmatrix} X_{11} & \dots & X_{1j} & \dots & X_{1p} \\ \vdots & & \vdots & & \vdots \\ X_{i1} & \dots & X_{ij} & \dots & X_{ip} \\ \vdots & & \vdots & & \vdots \\ X_{n1} & \dots & X_{nj} & \dots & X_{np} \end{pmatrix}$$

The i^{th} row $X_i := (X_{i1}, \dots X_{ip})$ corresponds to individual i.

Data preprocessing

Raw data

$$X = \begin{pmatrix} X_{11} & \dots & X_{1j} & \dots & X_{1p} \\ \vdots & & \vdots & & \vdots \\ X_{i1} & \dots & X_{ij} & \dots & X_{ip} \\ \vdots & & \vdots & & \vdots \\ X_{n1} & \dots & X_{nj} & \dots & X_{np} \end{pmatrix}$$

The i^{th} row $X_i := (X_{i1}, \dots X_{ip})$ corresponds to individual i.
The j^{th} column $X^{(j)} := (X_{1j}, \dots X_{nj})'$ corresponds to variable j.

Data preprocessing

Raw data

$$X = \begin{pmatrix} X_{11} & \dots & X_{1j} & \dots & X_{1p} \\ \vdots & & \vdots & & \vdots \\ X_{i1} & \dots & X_{ij} & \dots & X_{ip} \\ \vdots & & \vdots & & \vdots \\ X_{n1} & \dots & X_{nj} & \dots & X_{np} \end{pmatrix}$$

The i^{th} row $X_i := (X_{i1}, \dots X_{ip})$ corresponds to individual i.
The j^{th} column $X^{(j)} := (X_{1j}, \dots X_{nj})'$ corresponds to variable j.

Reference individual : $\Omega = (0, \dots, 0) \in \mathbb{R}^p$.

Data preprocessing

Centered data

$$X_c = \begin{pmatrix} X_{11} - \bar{X}^{(1)} & \dots & X_{1j} - \bar{X}^{(j)} & \dots & X_{1p} - \bar{X}^{(p)} \\ \vdots & & \vdots & & \vdots \\ X_{i1} - \bar{X}^{(1)} & \dots & X_{ij} - \bar{X}^{(j)} & \dots & X_{ip} - \bar{X}^{(p)} \\ \vdots & & \vdots & & \vdots \\ X_{n1} - \bar{X}^{(1)} & \dots & X_{nj} - \bar{X}^{(j)} & \dots & X_{np} - \bar{X}^{(p)} \end{pmatrix}$$

where $\bar{X}^{(j)} := \frac{1}{n} \sum_{i=1}^{n} X_{ij}$

Reference individual : $\Omega = \bar{X} = \left(\bar{X}^{(1)}, \dots, \bar{X}^{(j)}, \dots, \bar{X}^{(p)}\right) \in \mathbb{R}^p$.

Data preprocessing

Reduced and centered data

$$X_{cr} = \begin{pmatrix} \frac{X_{11} - \bar{X}^{(1)}}{s^{(1)}} & \cdots & \frac{X_{1j} - \bar{X}^{(j)}}{s^{(j)}} & \cdots & \frac{X_{1p} - \bar{X}^{(p)}}{s^{(p)}} \\ \vdots & & \vdots & & \vdots \\ \frac{X_{i1} - \bar{X}^{(1)}}{s^{(1)}} & \cdots & \frac{X_{ij} - \bar{X}^{(j)}}{s^{(j)}} & \cdots & \frac{X_{ip} - \bar{X}^{(p)}}{s^{(p)}} \\ \vdots & & \vdots & & \vdots \\ \frac{X_{n1} - \bar{X}^{(1)}}{s^{(1)}} & \cdots & \frac{X_{nj} - \bar{X}^{(j)}}{s^{(j)}} & \cdots & \frac{X_{np} - \bar{X}^{(p)}}{s^{(p)}} \end{pmatrix}$$

where

$$\bar{X}^{(j)} := \frac{1}{n} \sum_{i=1}^{n} X_{ij} \qquad s^{(j)} := \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(X_{ij} - \bar{X}^{(j)} \right)^2}.$$

(if different units or very different scales)

Reference individual : $\Omega = \bar{X} = \left(\bar{X}^{(1)}, \ldots, \bar{X}^{(j)}, \ldots, \bar{X}^{(p)} \right) \in \mathbb{R}^p$.

Global Inertia

In the following, we denote by X, the matrix of centered data (eventually centered and reduced data).
The ith individual corresponds to

$$X_i = (X_{i1}, X_{i2}, \ldots, X_{ip}) \in \mathbb{R}^p.$$

Definition

The inertia of the set of points relative to the reference individual Ω is defined as

$$I_\Omega := \frac{1}{n} \sum_{i=1}^{n} \|X_i\|^2. \tag{1}$$

Global Inertia

Proposition

— Let

$$\Gamma = \frac{1}{n} X'X.$$

The inertia is equal to

$$I_\Omega = Tr(\Gamma),$$

where $Tr()$ represents the trace of the matrix.

- For centered PCA , Γ is the variance-covariance matrix, and

$$I_\Omega = \sum_{j=1}^{p} \left(s^{(j)}\right)^2.$$

- For centered and reduced PCA , Γ is the correlation matrix, hence

$$I_\Omega = p.$$

In practice

Framework : Centered PCA

$$\Gamma = \begin{pmatrix} 11.39 & 9.92 & 2.66 & 4.82 \\ 9.92 & 8.94 & 4.12 & 5.48 \\ 2.66 & 4.12 & 12.06 & 9.29 \\ 4.82 & 5.48 & 9.29 & 7.91 \end{pmatrix}.$$

$$I_\Omega = 11.39 + 8.95 + 12.06 + 7.91 = 40.3,$$

measures the dispersion of the set of points, i.e. the heterogeneity of the group of students over all four subjects.

Axial inertia

- Let a be a normalized vector in $\mathbb{R}^p$, i.e. such that $\|a\|^2 = 1$.
- The orthogonal projection of $x \in \mathbb{R}^p$ on the line generated by a is equal to

$$P_a(x) = \langle x, a \rangle a.$$

- We want to quantify the dispersion of the orthogonal projection of the set of points on this axis.

Axial inertia

Definition

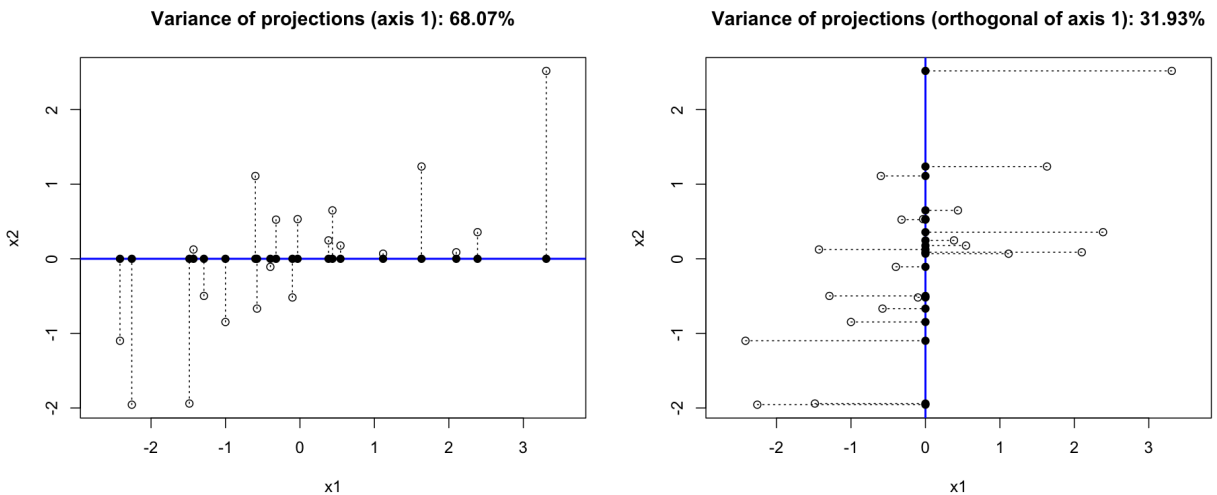
Let $a \in \mathbb{R}^p$ a normalized vector. The inertia on the axis a is defined by

$$I_\Omega(a) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\langle X_i, a\rangle^2}_{\|P_a(X_i)\|^2} = \frac{1}{n}\|Xa\|^2.$$

Remark : Xa is a vector containing the coordinates of the orthogonal projections of the n individuals on the axis a.

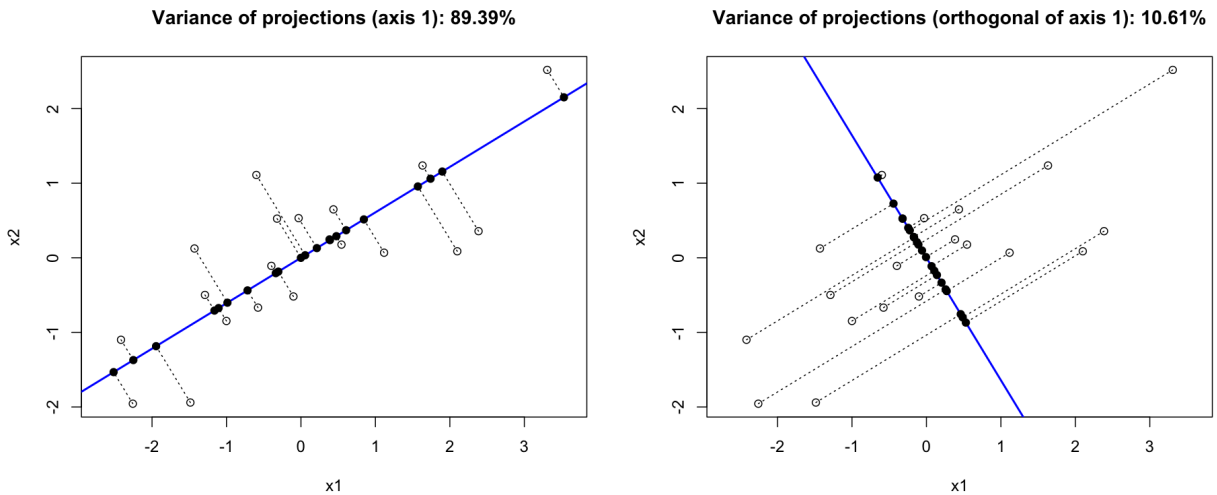
Search for principal components

- We want to project the observations in smaller dimension, while keeping as much as information as possible, i.e. maximizing the inertia of the projected set of points.
- First step : projection in dimension 1. How to obtain the axis with maximal axial inertia ?



Search for principal components

- Solution :



- To complete your intuition on PCA, play with the applet : https://roustant.shinyapps.io/pca-app/

Principal components decomposition

Proposition

— *Let $\lambda_1 > \ldots > \lambda_p$ the eigenvalues (supposed all distinct) of the variance-covariance matrix $\Gamma = \frac{1}{n} X'X$.*

Let $a_1, \ldots, a_p$ the corresponding normalized eigenvectors.

1. *For all $j \neq k$,*
$$\langle a_j, a_k \rangle = 0.$$
2. *Moreover, for all j in $\{1, \ldots, p\}$,*
$$I_\Omega(a_j) = \lambda_j.$$

Principal components decomposition

- a_1 is the axis maximizing axial inertia.
- a_2 is the axis maximizing axial inertia among all axes orthogonal to a_1.
- a_3 is the axis maximizing axial inertia among all axes orthogonal to a_1 and a_2.
- etc.

The principal components decomposition thus amounts to find the eigenvalues and eigenvectors $(\lambda_j, a_j)_{j=1\ldots p}$ of the variance-covariance matrix Γ.

Principal components decomposition

Definition

Let $(\lambda_j, a_j)_{j=1\dots p}$ the eigenvalues and normalized eigenvectors of Γ :

- *$a_j \in \mathbb{R}^p$ are called principal (or factorial) axes.*
- *$C_j := Xa_j$ are called principal components.*
- *For $k, j \in \{1, \dots p\}$, $k \neq j$, the plane generated by the vectors a_j and a_k is called factorial plane (a_j, a_k).*

Principal components decomposition

The jth meta-variable $C_j = Xa_j = \begin{pmatrix} \langle X_1, a_j \rangle \\ \vdots \\ \langle X_n, a_j \rangle \end{pmatrix} \in \mathbb{R}^n$.

In particular, by Proposition 2,

$$\|C_j\|^2 = \|Xa_j\|^2 = I_\Omega(a_j) = \lambda_j.$$

- $(a_1, \ldots, a_p)$: new orthonormal system.
- $(C_1, \ldots, C_p)$: projection of the set of points in this new orthonormal system $\Longrightarrow$ meta-variables.
- New representation in possibly smaller dimension, maximizing the inertia.

In practice

Variance-covariance matrix :

$$\Gamma = \begin{pmatrix} 11.39 & 9.92 & 2.66 & 4.82 \\ 9.92 & 8.94 & 4.12 & 5.48 \\ 2.66 & 4.12 & 12.06 & 9.29 \\ 4.82 & 5.48 & 9.29 & 7.91 \end{pmatrix}.$$

Eigen values :

$$\lambda_1 = 28.23, \quad \lambda_2 = 12.03, \quad \lambda_3 = 0.03 \quad \lambda_4 = 0.01.$$

Principal (factorial) axes :

$$a_1 = \begin{pmatrix} 0.52 \\ 0.51 \\ 0.49 \\ 0.48 \end{pmatrix}, \quad a_2 = \begin{pmatrix} -0.57 \\ -0.37 \\ 0.66 \\ 0.33 \end{pmatrix}, \quad a_3 = \begin{pmatrix} 0.19 \\ -0.45 \\ -0.46 \\ 0.74 \end{pmatrix}, \quad a_4 = \begin{pmatrix} -0.61 \\ 0.63 \\ -0.34 \\ 0.33 \end{pmatrix}.$$

In practice

Principal components : $C_j = Xa_j$ (meta-variables)

	C_1	C_2	C_3	C_4
Nathan	-8.61	-1.41	0.07	0.07
Emma	-3.88	-0.50	0.01	-0.07
Lola	-3.21	3.47	-0.17	0.01
Baptiste	9.85	0.60	0.04	-0.15
Mathilde	6.41	-2.05	-0.08	0.19
Ines	-3.03	-4.92	0.08	-0.14
Lucas	-1.03	6.38	-0.16	-0.03
Aymeric	1.95	-4.20	-0.20	0.04
Chloe	1.55	2.63	0.42	0.07

What is the meaning of the meta-variables ?

Interpretation of the results

Objective : select a "small" number of meta-variables.

Proposition

. — *Let* $\lambda_1 > \cdots > \lambda_p$ *the eigenvalues of the matrix* Γ. *Then*

$$I_\Omega = \sum_{j=1}^{p} \lambda_j.$$

Choice of the number of components

Idea : Keep only the axes carrying a significant part of the inertia

We define for all j in $\{1, \dots, p\}$,

- Percent of inertia : λ_j / I_Ω,
- Cumulative percentage of inertia :

$$PC(j) = \sum_{l=1}^{j} \frac{\lambda_l}{I_\Omega} = \frac{\sum_{l=1}^{j} \lambda_l}{\sum_{l=1}^{p} \lambda_l}.$$

In general

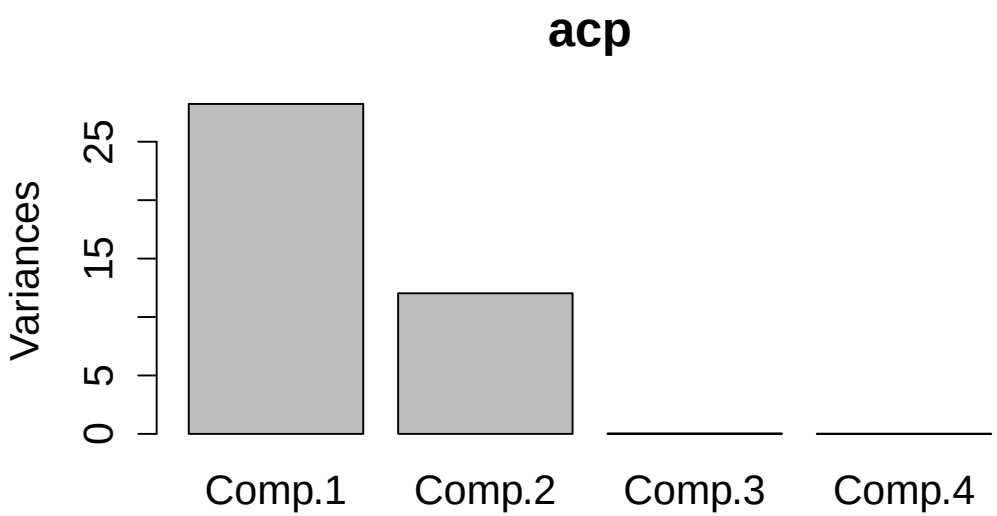
- For $\alpha \approx 10\%, 20\%$, we keep j_0 principal components where

$$j_0 = \min\{j \in \{1, \dots, p\} \ ; \ PC(j) \geq 1 - \alpha\}.$$

- More recommended : "clear" break in the decrease in eigenvalues.

In practice

Component	% *Inertia*	% *cumulated inertia*
c_1	70.05%	70.05 %
c_2	29.84%	99.89 %
c_3	0.08%	99.97 %
c_4	0.03%	100 %



$\Longrightarrow$ we keep 2 principal components.
What are their meaning ?

Links between individuals and principal axes

We are looking for individuals with a particular "profile" (to refine the interpretation), or homogeneous groups of individuals.
Reminder :

$$(C_k)_i = \langle X_i, a_k \rangle.$$

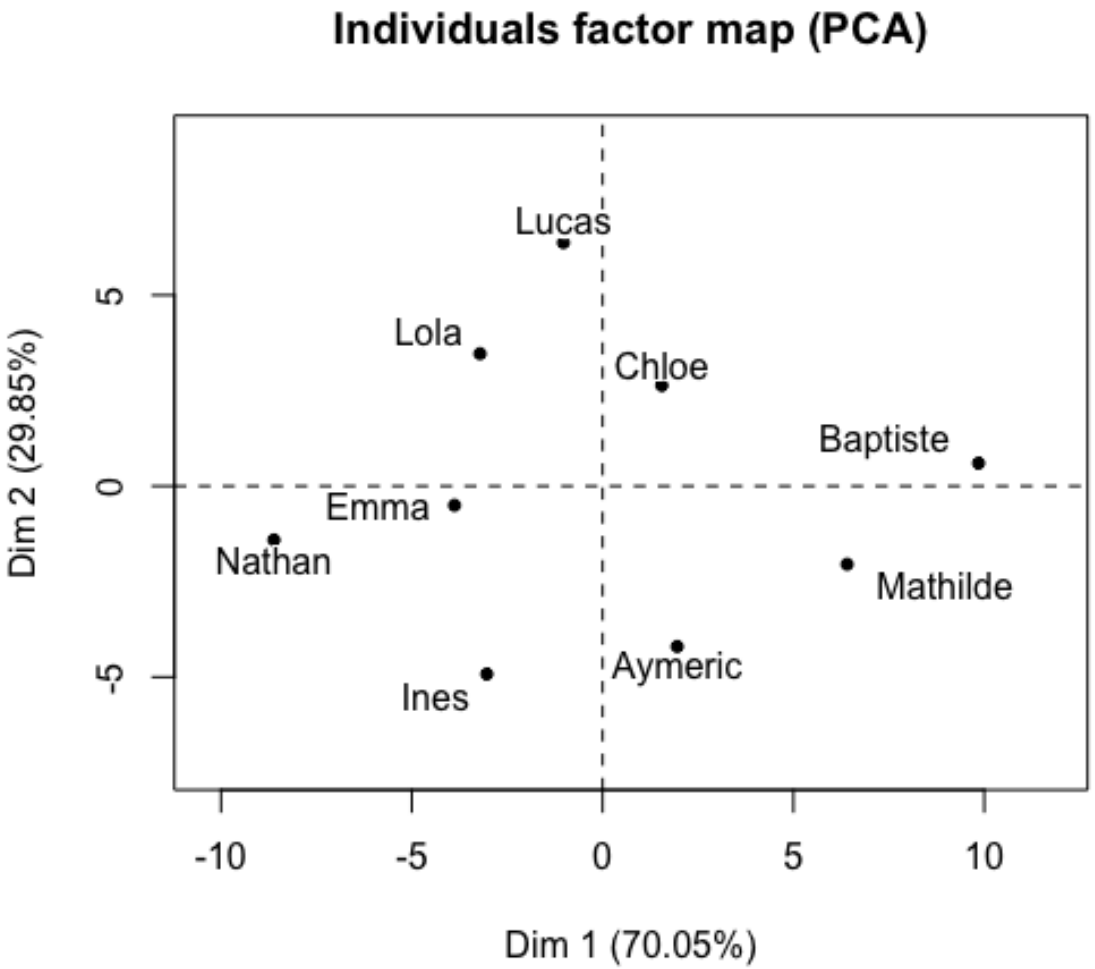
Graphically, we represent the orthogonal projections of the set of points (in $\mathbb{R}^p$) on the factorial plane Vect$< a_1, a_2 >$.

The ith individual is represented by the point with coordinates :

$$\Big(\langle X_i, a_1 \rangle, \langle X_i, a_2 \rangle \Big) = \Big((C_1)_i, (C_2)_i \Big).$$

It is possible to generalize to other factorial planes Vect$< a_k, a_l >$.

In practice



	c_1	c_2
Nathan	-8.61	-1.41
Emma	-3.88	-0.50
Lola	-3.21	3.47
Baptiste	9.85	0.60
Mathilde	6.41	-2.05
Ines	-3.03	-4.92
Lucas	-1.03	6.38
Aymeric	1.95	-4.20
Chloe	1.55	2.63

Links between variables and principal components

- Initial variables $X^{(j)} \in \mathbb{R}^n$.
- Principal components $C_k \in \mathbb{R}^n$.

Proposition

— *Principal components are linear combinations of the initial variables. For all* $j \in \{1, \ldots, p\}$,

$$C_j = \sum_{k=1}^{p} a_{j,k} X^{(k)}.$$

Since X *has been centered, the principal components are centered.*

What is their meaning ?

Correlation coefficient

Remark : For centered variables x and y in $\mathbb{R}^n$,

$$\mathrm{Var}(x) = \frac{1}{n}\|x\|^2 \qquad \text{and} \qquad \mathrm{Cov}(x,y) = \frac{1}{n}\langle x, y\rangle.$$

Let X a centered matrix, for all $1 \leq j,k \leq p$,

$$r_{j,k} = \frac{\mathrm{Cov}\left(X^{(j)}, C_k\right)}{\sqrt{\mathrm{Var}\left(X^{(j)}\right)\mathrm{Var}(C_k)}} = \frac{\langle X^{(j)}, C_k\rangle}{\|X^{(j)}\|\|C_k\|}.$$

If the variables are **reduced**, then $r_{j,k}$ is the coordinate of $X^{(j)}$ on the normalized principal component $C^k/\|C^k\|$:

$$r_{j,k} = \langle X^{(j)}, C_k/\|C_k\|\rangle$$

Circle of correlation

If the variables are **reduced**, then using Pythagorus theorem :

$$1 = \|X^{(j)}\|^2 = \sum_{k=1}^{p} r_{j,k}^2$$

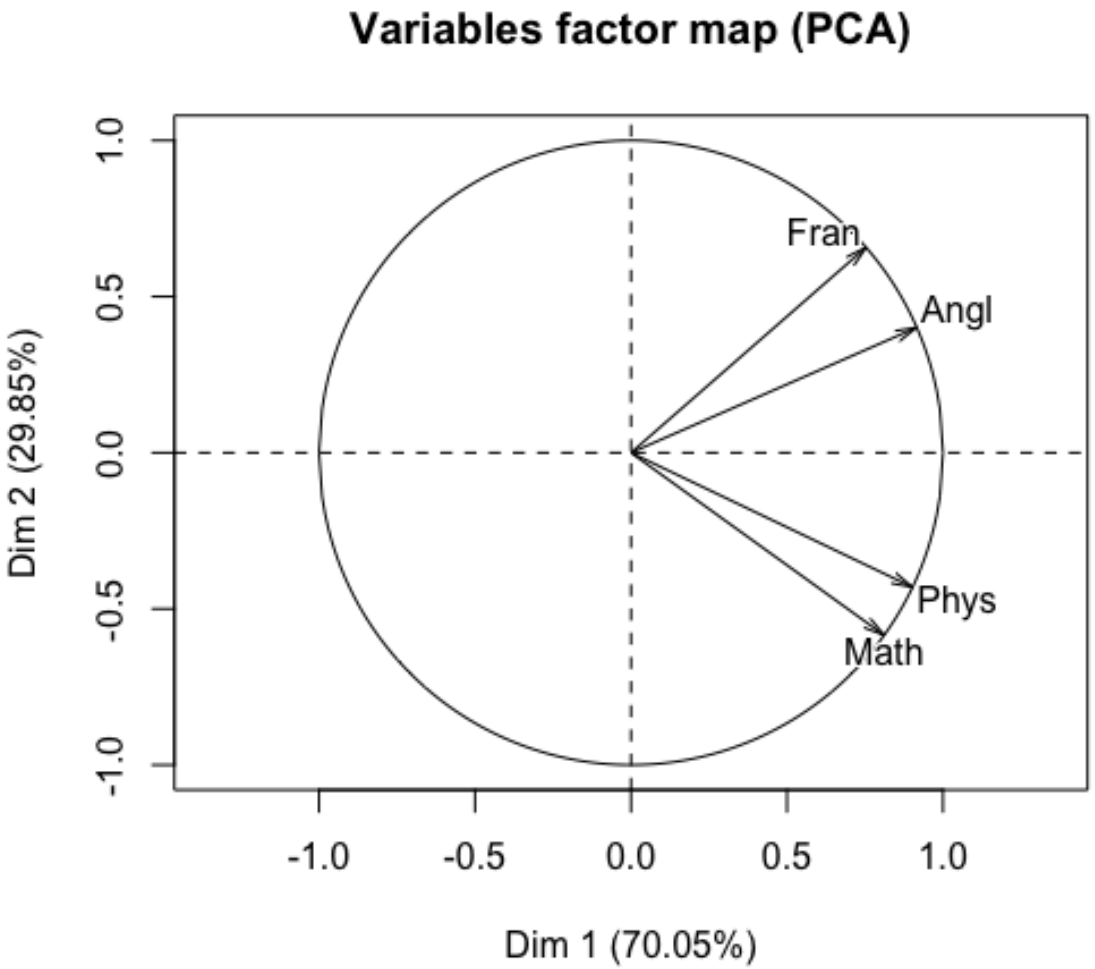
Thus, the vector of correlations $r_{j,1}, \ldots, r_{j,p}$ belong to a ***p*-dimensional sphere**. Its projection onto a 2D plane of coordinates (k_1, k_2) is a circle, called **circle of correlations**.

- $X^{(j)}$ is well explaned by components k_1, k_2 if and only if

$$1 = \|X^{(j)}\|^2 \approx r_{j,k_1}^2 + r_{j,k_2}^2$$

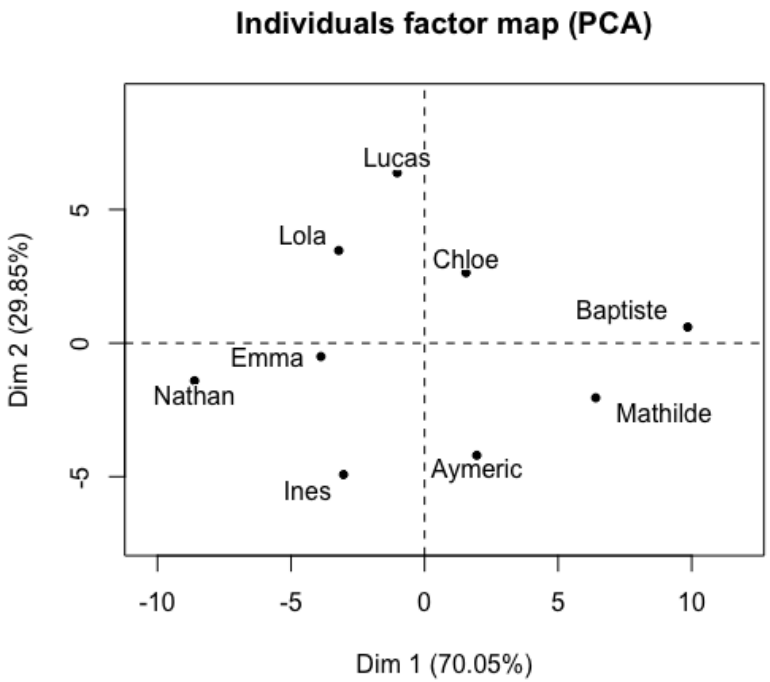
  i.e. the projection of $X^{(j)}$ is closed to the circle.
- The coordinates of all the variables on the axis k_1 (resp. k_2) are proportional to the principal component k_1 (resp. k_2)

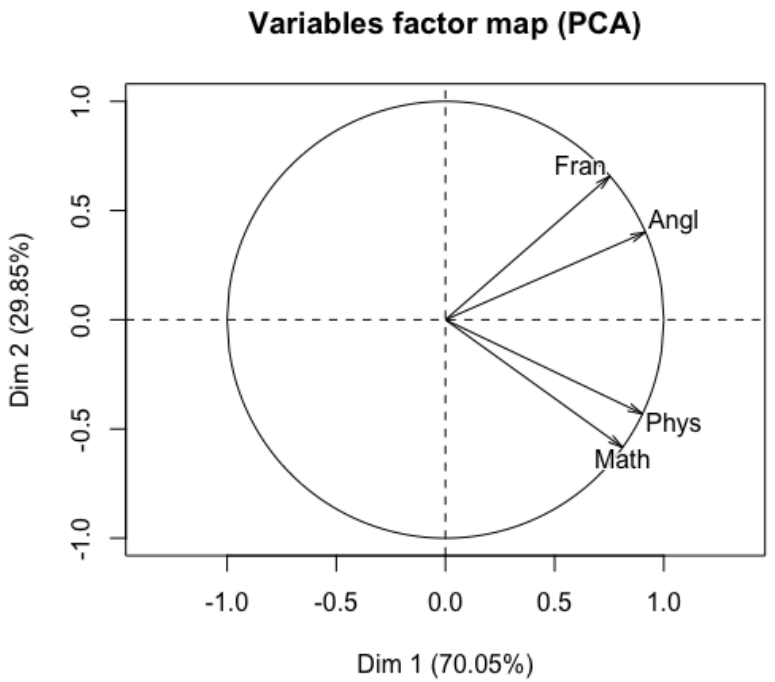
In practice



	c_1	c_2
Math	0.81	-0.58
Physique	0.90	-0.43
Français	0.75	0.66
Anglais	0.91	0.40

In practice





We can summarize the set of points with two meta-variables :

- The first components represents the global level.
- The second one the profile scientific / literary.

Outline

- Introduction
- Multidimensional Exploration
  - Basic tools
  - Principal components analysis
  - Factorial discriminant analysis
- Unsupervised classification
  - Hierarchical Cluster Analysis
  - k-means

Factorial Discriminant Analysis

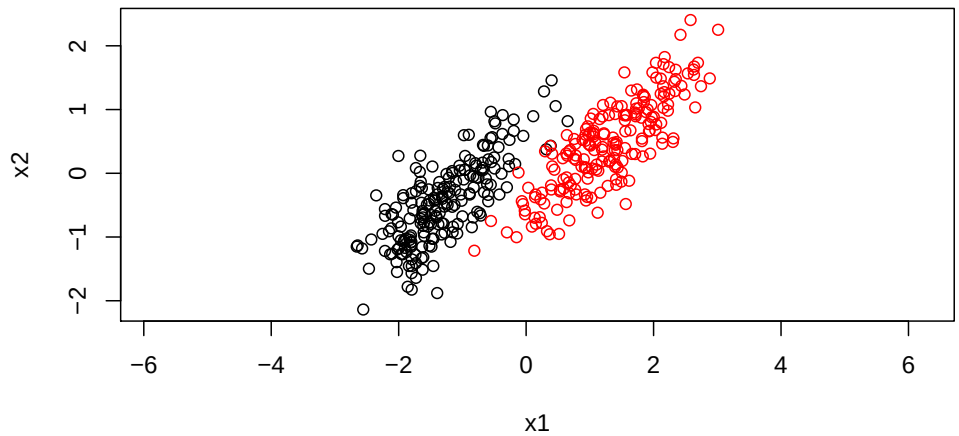


FIGURE – This is a cloud of points, with two classes, in dimension 2 (higher in general)

Can you find two 1D axis 'suitable' to identify classes ?

FDA, as an exploratory tool

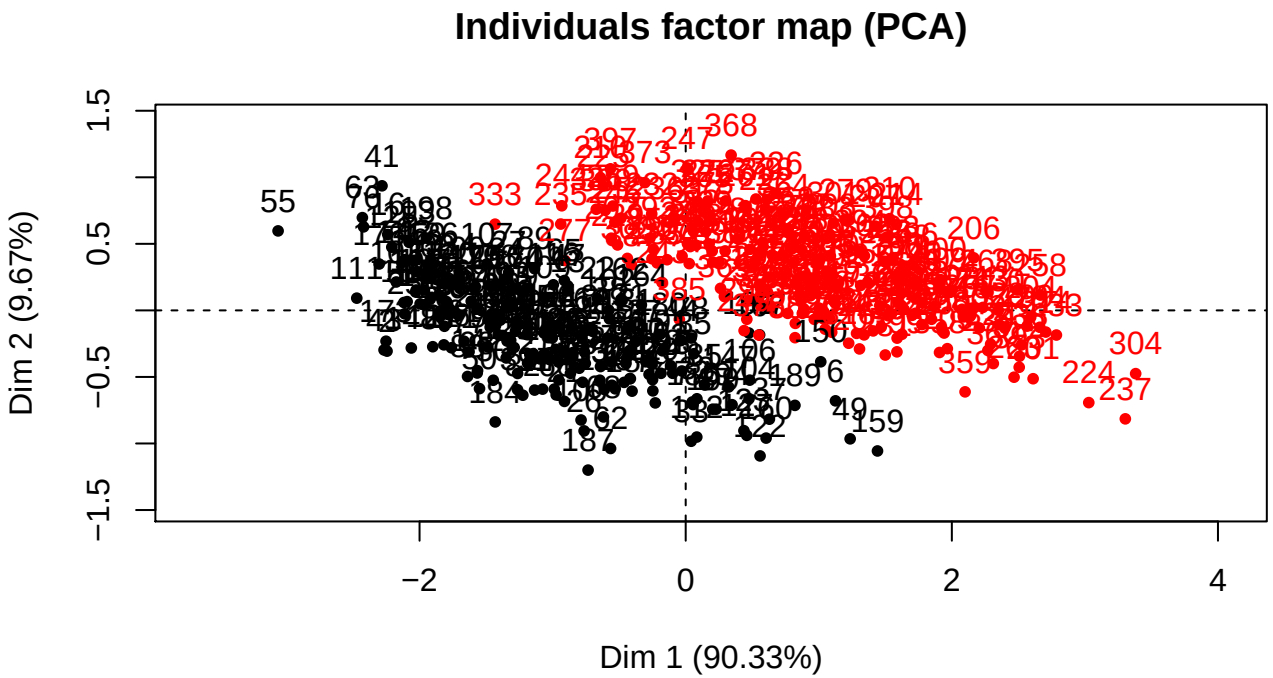


FIGURE – Result of the PCA analysis. Can we do better ?

FDA, as an exploratory tool

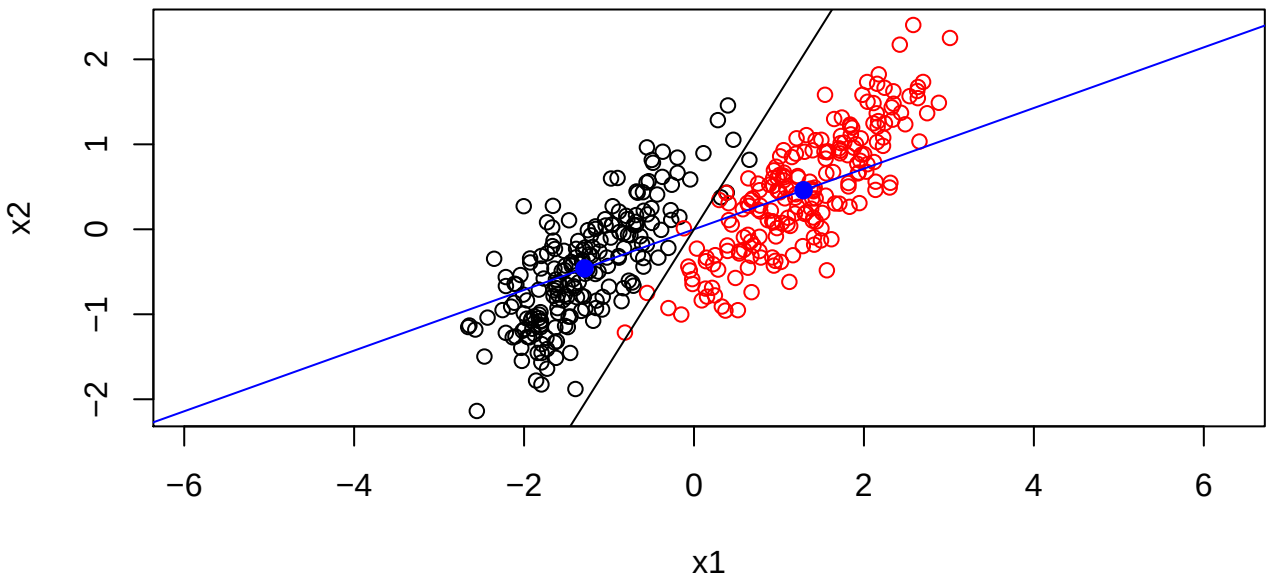


FIGURE – Result of the FDA analysis, actually PCA for the centroids : two data only ! The two axes are orthogonal... for a specific ('Mahalanobis') metric !

FDA, as an exploratory tool

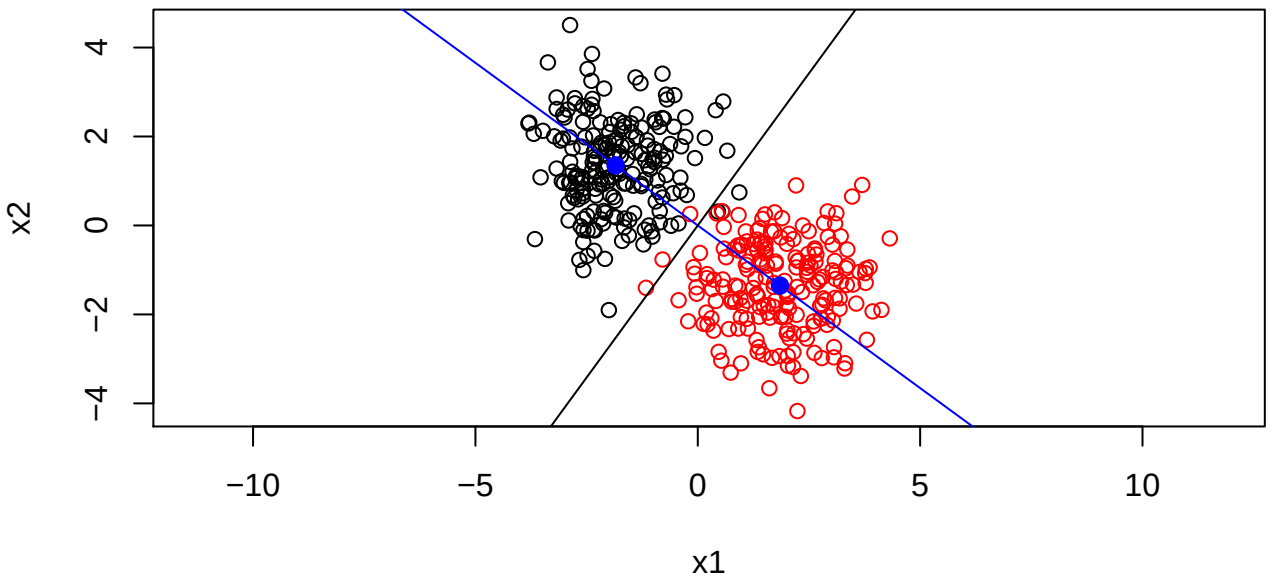


FIGURE – Result of the FDA analysis : visualization for tranformed data. The two axes are orthogonal for the usual metric.

Mahalanobis metric and 'sphered' data

The Mahalanobis metric is such that the covariance matrix is identity. This is equivalent to (matricially) reduce or 'sphere' the data :

$$\boldsymbol{x} \mapsto \mathrm{Cov}^{-1/2}\boldsymbol{x}$$

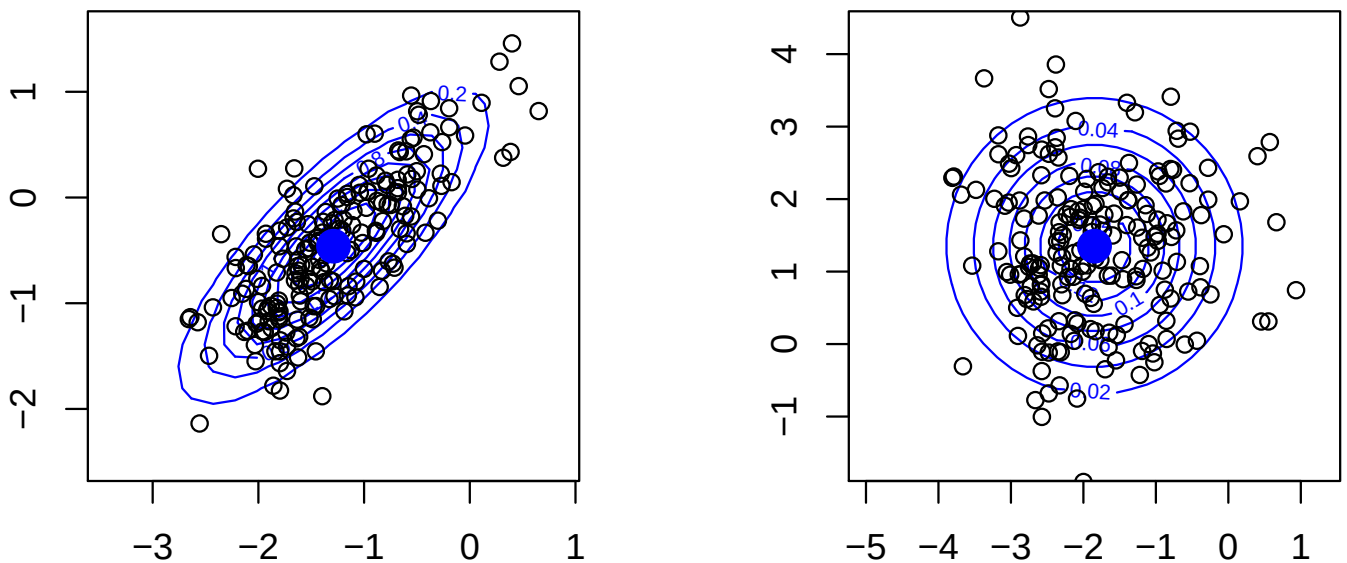
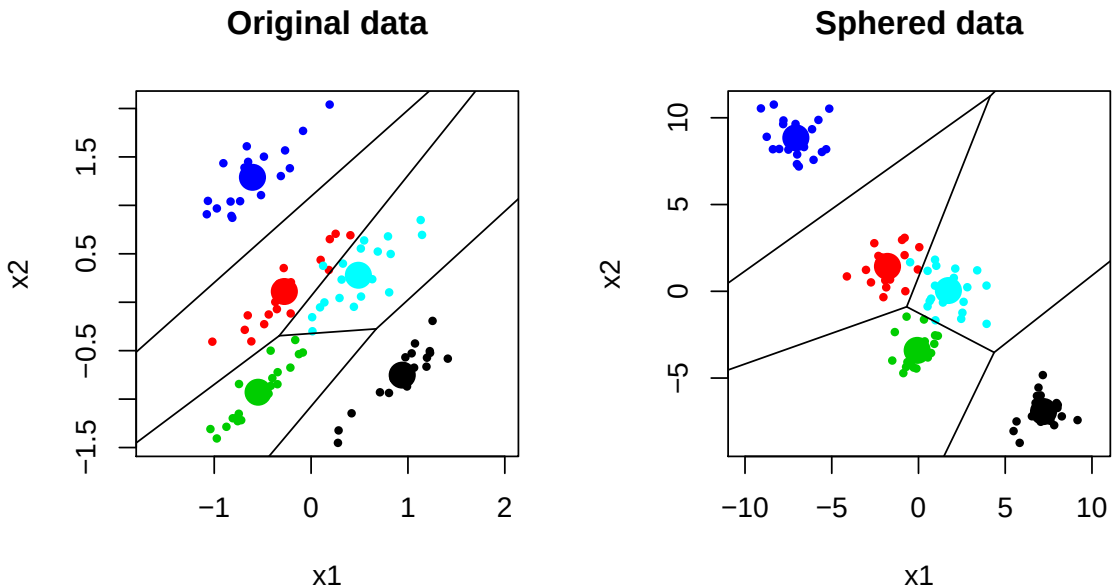


FIGURE – Left : Original data. Right : Reduced data. Level sets are for the multinormal distribution with corresponding covariance matrix.

Remark : FDA for supervised classification

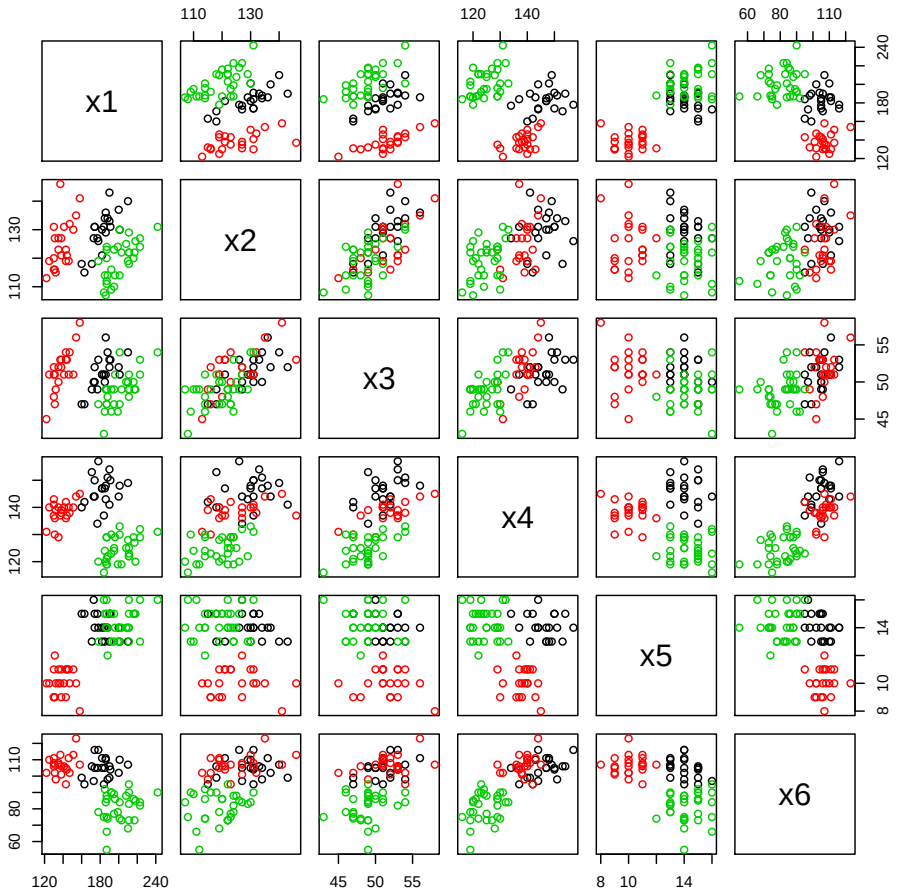
- FDA can be used for classification. When classes have equal sizes, one chooses the class corresponding to the closest centroid. This gives linear frontiers.



- Test other configurations with the applet : https://roustant.shinyapps.io/lda-app/

A six dimensional example

We consider the Lubitsch data for insects. There are 74 data, 6 variables, and 3 classes.



A six dimensional example

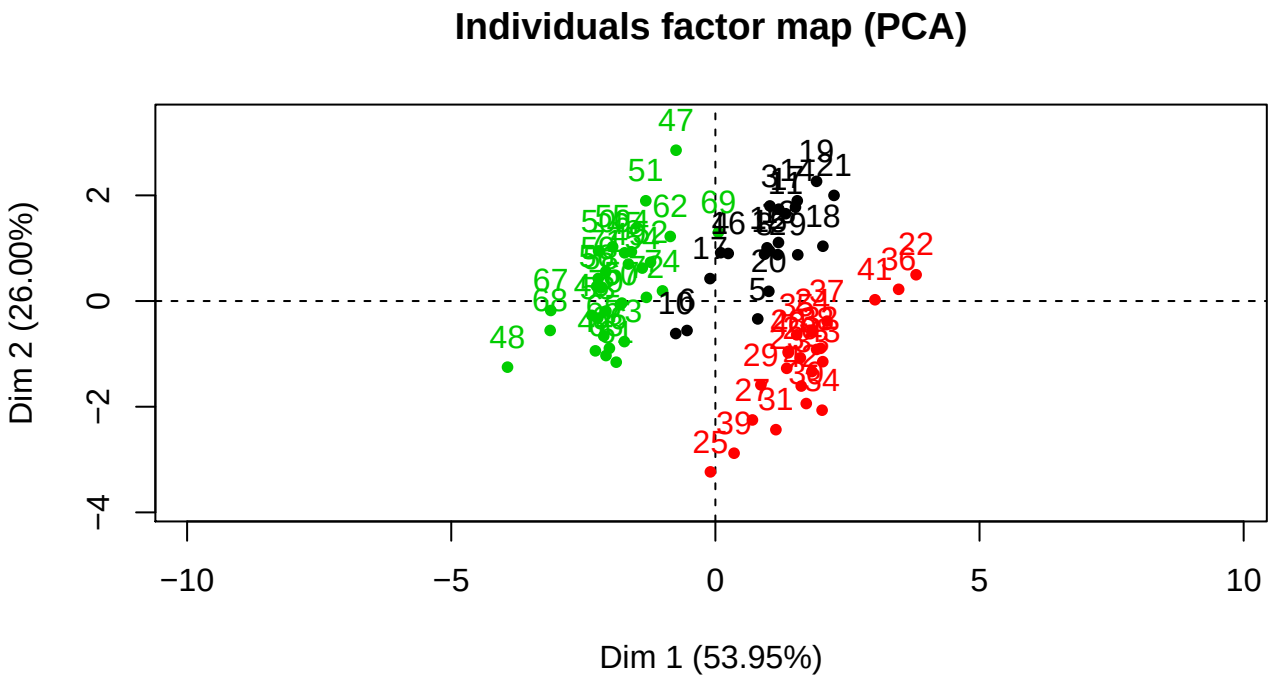


FIGURE – Insect dataset. Result of the PCA analysis.

A six dimensional example

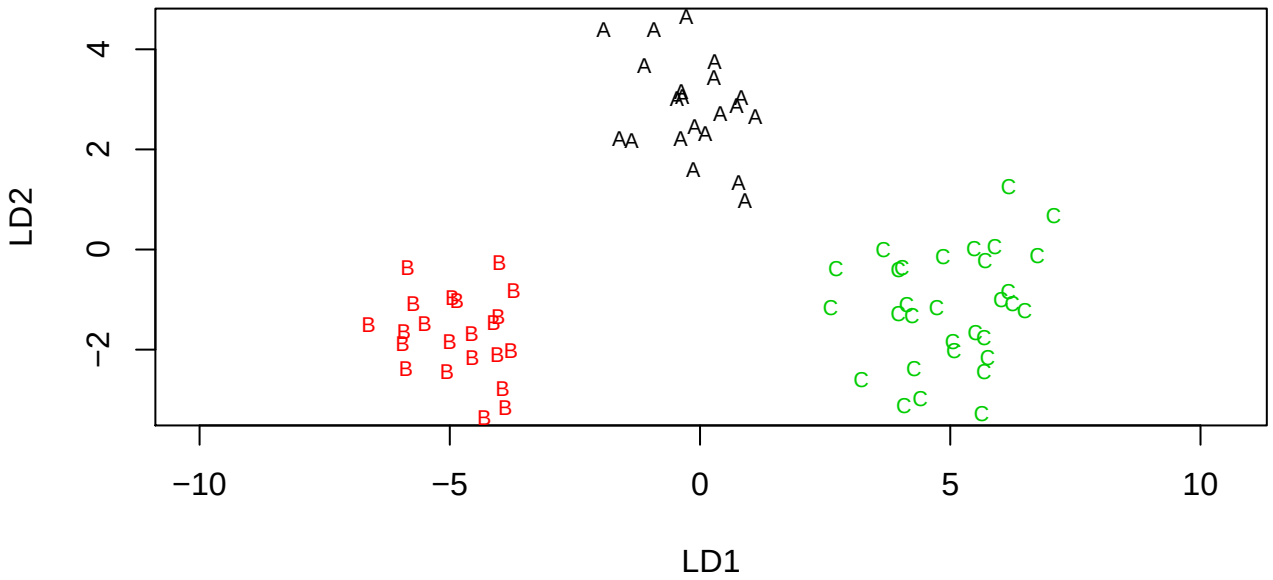
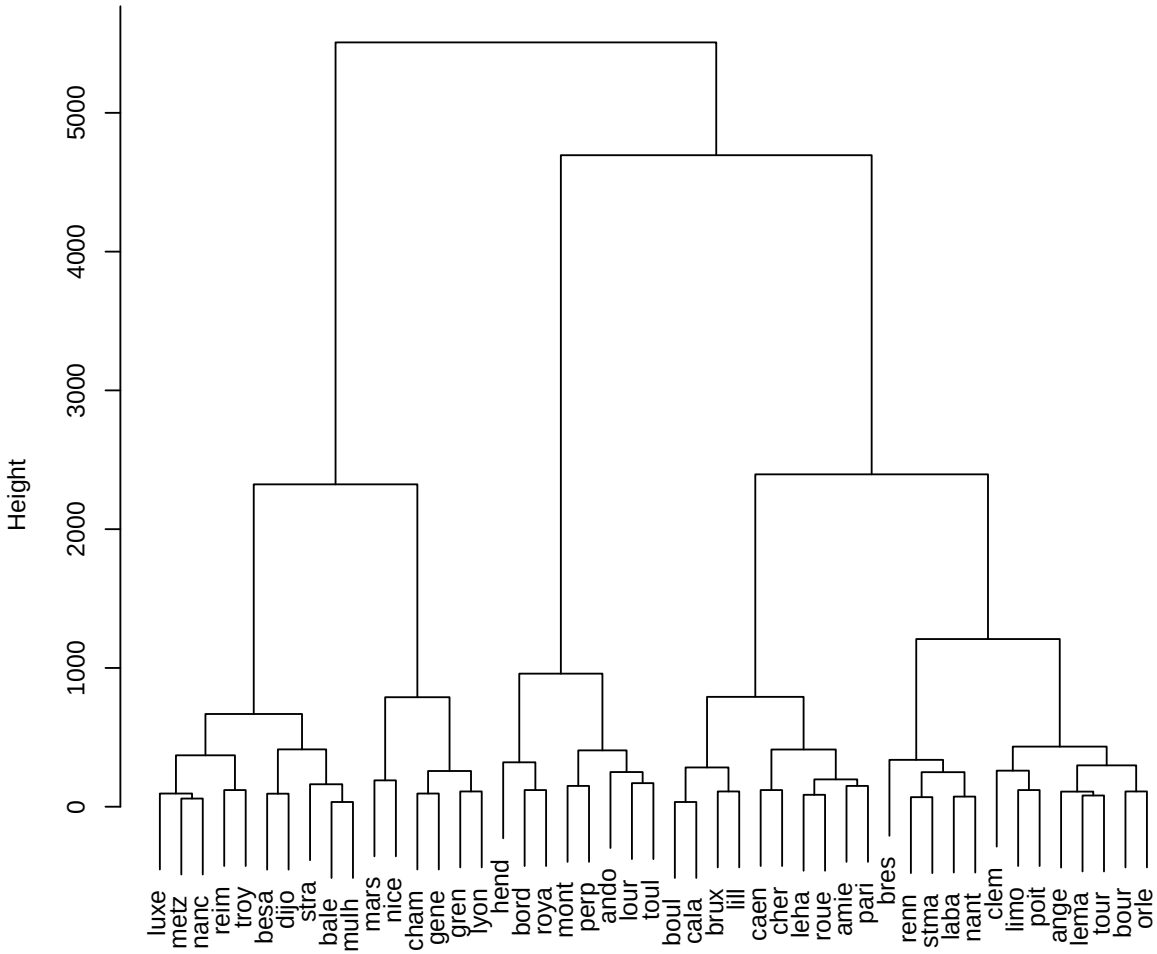


FIGURE – Insect dataset. Result of the FDA analysis.

Outline

- Introduction
- Multidimensional Exploration
  - Basic tools
  - Principal components analysis
  - Factorial discriminant analysis
- Unsupervised classification
  - Hierarchical Cluster Analysis
  - k-means

Cities : dendrogram example



Algorithm for Hierarchical Cluster Analysis

- Initialization : singletons, distances between pairs of singletons
- Iterate until aggregation in a single class :
  1. gather the two closest classes within the meaning of the chosen "distance" between groups
  2. update the distance table by replacing the two classes with the new one and by calculating its "distance" with the other classes.

Distance between groups

Linkage functions

The distances between two sets A and B, can be computed in various way. Here are some examples, denoting by g_A, g_B the centroids of A, B :

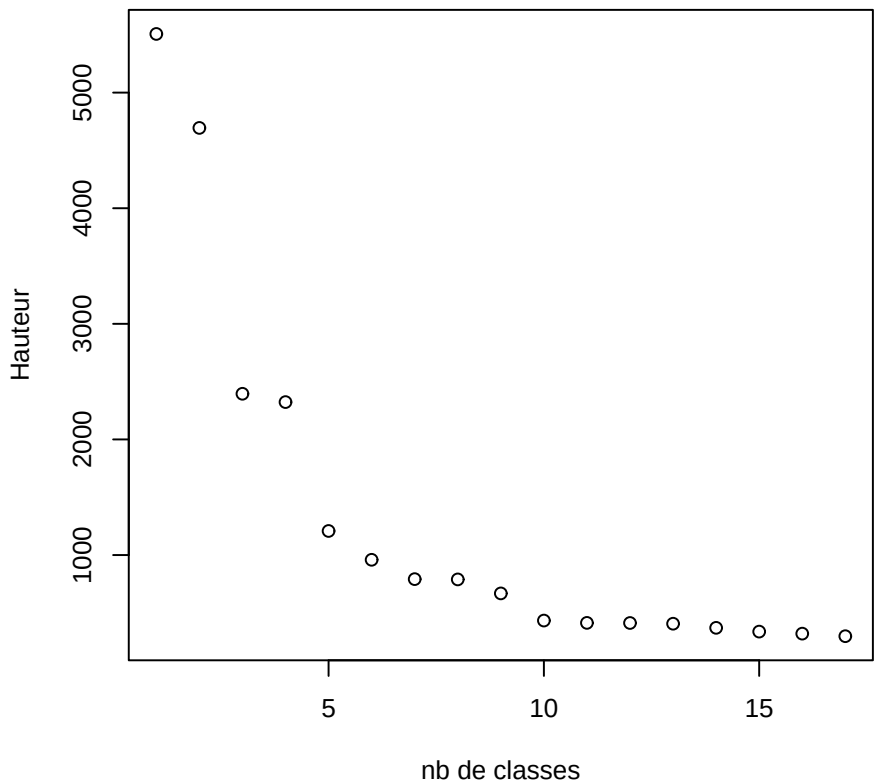
linkage name	linkage function $d(A,B)$
single	$\min_{i\in A,\, j\in B} d_{ij}$
complete	$\sup_{i\in A,\, j\in B} d_{ij}$
average	$\frac{1}{\lvert A\rvert.\lvert B\rvert}\sum_{i\in A,\, j\in B} d_{ij}$
centroid	$d(g_A, g_B)$
Ward	$\frac{\lvert A\rvert.\lvert B\rvert}{\lvert A\rvert+\lvert B\rvert} d(g_A, g_B)^2$

Ward's criterion is equal to the **between-class inertia**,

$$|A|d(g_A, g)^2 + |B|d(g_B, g)^2$$

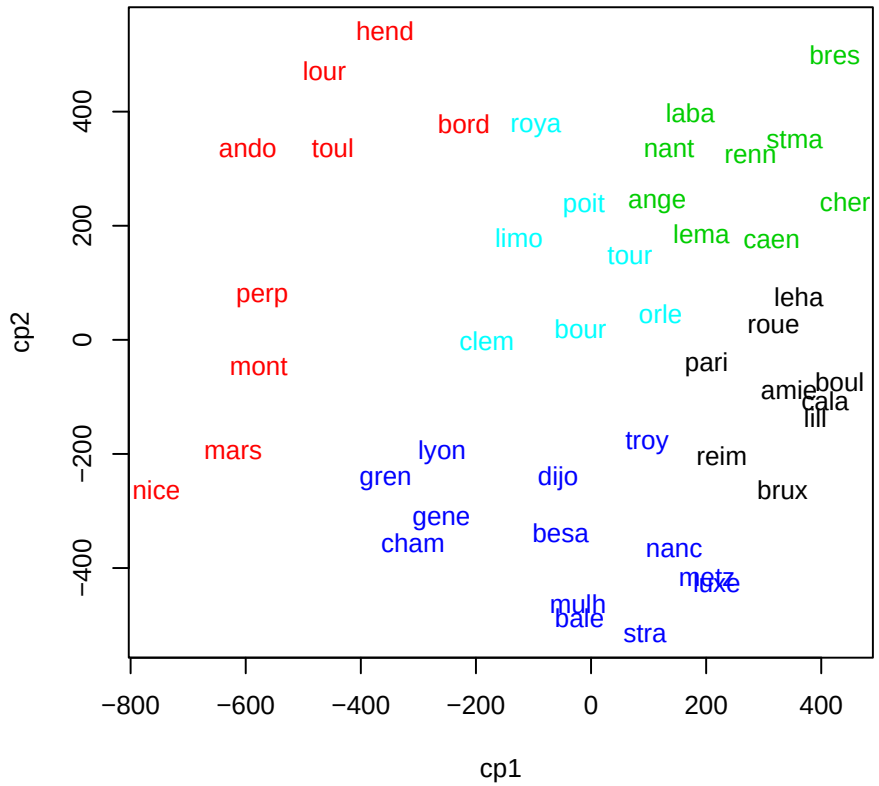
where g is the centroid of the set $\{A, B\}$ obtained after merging A, B.

Number of classes



Cities : Decrease of the between-class inertia

Representation of the classes



Outline

- Introduction
- Multidimensional Exploration
  - Basic tools
  - Principal components analysis
  - Factorial discriminant analysis
- Unsupervised classification
  - Hierarchical Cluster Analysis
  - k-means

Principle of moving center

- dynamic reallocation of individuals to classes
- The number of classes k is fixed *a priori*

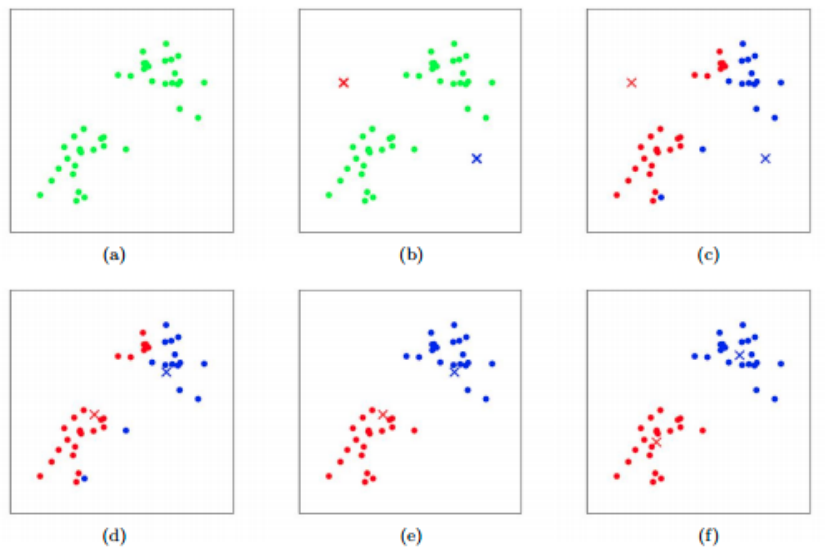
Principle of Forgy algorithm

- Initialisation Draw - or randomly select - k points in the space of individuals called centers.
- Iterate until stabilization of the between-class inertia
  1. Allocate each individual to a center, i.e. to a class
  2. Compute the centroid of each class, it becomes the new center.

Variants

- k-means Algorithme : the center of the classes, here centroids, are recalculated at each allocation of a point to a class : more efficient algorithm.
- Partitionning Around Medoïds (PAM) or k-medoïds.
  - Medoïd : individual minimizing the mean of the distances to the other points (more robust algorithm).

Illustration of k-means



Conclusion

- We have seen several tools to visualize data.
- Concerning unsupervised classification or clustering, many other algorithms are available.
- The choice among all possible algorithms is often difficult.
- The next steps will be devoted to supervised classification where an estimation of the prediction error is possible.
- To go further... read the textbooks on wikistat.fr and execute the notebooks on the Github wikistat.