Big data project
Master's Degree in Computer Engineering

Dr. Carles Comas

Departament de Matemàtica
Escola Tècnica Superior d'Enginyeria Agrària, despatx 4.2.11
Universitat de Lleida

carles.comas@udl.cat

1st Semester 2023-2024

Lesson 1: General setting

- In this course we shall consider two general techniques to analyse large data sets
- Principal Component Analysis (PCA)
- The Expectation-Maximization algorithm (EM)
- The PCA is a statistical procedure defined to describe the data under analysis (descriptive statistics), whilst
- The EM algorithm is a statistical technique used to estimate parameters in statistical models (statistical inference and statistical classification)

Exemple 1: A first large data set

- 53 Blood and urine measurements (wet chemistry) from 65 people (33 alcoholics, 32 non-alcoholics).

	M1	M2	M3	⋯	M53
Subject1	0.03	0.24	2.29	⋯	10.45
Subject2	0.04	0.35	3.56	⋯	9.76
⋮	⋮	⋮	⋮		⋮
Subject65	0.014	0.17	1.75	⋯	11.72

Possible scientific questions

- Can these blood and urine measurements discern between alcoholic and non-alcoholic persons?
- Are some of these measurement more relevant than other to reveal the effects of alcoholism?
- Do you think that some of these variables are correlated between each other?
- If this is so, then, probably we do not need the 53 variables to explain the results of this table...
- A first tentative option is to delete correlated variables
- But in this case, we are losing information

Some difficulties

- Difficult to represent a data set with 53 variables
- Usually, we can represent univariate, bivariate and trivariate data
- Thus, if we assume that some of the 53 variables are correlated, we may consider to reduce the number of dimensions in data
- How to find the best low dimension space that conveys maximum useful information?
- PCA can be considered to obtain this dimensionality reduction
- In highly correlated variables, it is possibles that a 20% of the variables explain more than the 80% of the variability of the original data

Example 2: A second large data set

- 30 questions (items) from 2000 people (a survey of attitudes toward immigration, likert scale).

	Item1	Item2	Item3	⋯	Item30
Subject1	1	5	3	⋯	3
Subject2	2	4	—	⋯	1
⋮	⋮	⋮	⋮		⋮
Subject2000	—	4	—	⋯	1

Possible scientific questions

- We want to estimate the value of a parameter of this data set
- For instance, as we have a likert scale, we can estimate the value of the overall attitude of people toward immigration
- Values of 1 implies negative attitudes and values close to 5 suggests very positive attitudes toward immigration

Some difficulties

- Unfortunately, we have some missing values.
- Some individuals have not answered all the questions
- This is a typical problem in survey analysis
- A tentative way to solve this problem is to avoid individuals with missing answers
- Although this can permits to obtain estimated values from the sample, this clearly reduce the sample size and it can reduce the quality of our estimation
- What can we do to use all the sample size (in this case the 2000 individuals) even the individuals with missing values?

- A tentative approximation to that problem could be:
  1. Estimate the parameter of the model from the complete data (avoiding individual with missing answers)
  2. Use this estimated values to predict the missing values
  3. Replace the missing values for their predictions and obtain again an estimation of the parameters of the model
  4. Repeat steps 2 and 3 until convergence of the estimated parameters
- The above algorithm is the basis of the Expectation-Maximization algorithm (EM)
- The EM algorithm can also be used to classify and recognise data from several and distinct statistical populations

R statistical package

- In general, we shall not consider any specific programming language and software environment for statistical computing
- All the mathematical derivation will be done by basic R scripts (specific commands and short codes)
- Obviously, we can find specific statistical software devoted to these two methodologies

Lesson 2: Principal Component Analysis

2.1 The Multivariate data matrix

- Example of a multivariate table: “Heredity of Head Form in Man”

	1st. son	
	$\mathbf{x}_1$	$\mathbf{x}_2$
$\mathbf{e}_1$	191	155
$\mathbf{e}_2$	195	149
$\mathbf{e}_3$	181	148
$\mathbf{e}_4$	183	153
$\mathbf{e}_5$	176	144
$\mathbf{e}_6$	208	157
$\mathbf{e}_7$	189	150
$\mathbf{e}_8$	197	159
$\mathbf{e}_9$	188	152
$\mathbf{e}_{10}$	192	150
$\mathbf{e}_{11}$	179	158
$\mathbf{e}_{12}$	183	147
$\mathbf{e}_{13}$	174	150
$\mathbf{e}_{14}$	190	159
$\mathbf{e}_{15}$	188	151
$\mathbf{e}_{16}$	163	137
$\mathbf{e}_{17}$	195	155
$\mathbf{e}_{18}$	186	153
$\mathbf{e}_{19}$	181	145
$\mathbf{e}_{20}$	175	140
$\mathbf{e}_{21}$	192	154
$\mathbf{e}_{22}$	174	143
$\mathbf{e}_{23}$	176	139
$\mathbf{e}_{24}$	197	167
$\mathbf{e}_{25}$	190	163

- Here $\mathbf{e}_i$ denotes individuals
- And $\mathbf{x}_1$ and $\mathbf{x}_2$ are head length and width (mm)
- Source: Frets, G. P. (1921), "Heredity of Head form in Man," Genetica, **3**, 193–384
- Are these two variables correlated?
- If so, if they provide the some information, could we consider a single independent new variable to convey information about head form?

- Let us now write this multivariate data in a matrix notation

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{pmatrix}$$

- Where $x_{j,k}$ is the value of the j individual from the k variable

2.2 Data transformation

- The first thing we need to do is to obtain a multivariate matrix where each element is adjusted by the mean of each variable
- If the variables considered in the analysis are measured in distinct units, it is advisable to divide each value by its respective standard deviation
- This can also be done if the resulting variables have very different variability
- This centred and reduced $\mathbf{X}$ matrix is adimensional (in the raw matrix $\mathbf{X}$ each variable can be given in distinct units)
- Moreover, variables are considered in the analysis independently of their dispersion (variability)
- Using the raw $\mathbf{X}$ matrix or the centred and reduced $\mathbf{X}$ matrix to obtain the components is a controversial issue

- For each variable (column), we can obtain its mean and standard deviation values and adjust each matrix value to obtain a centrered and reduced matrix **Y**

$$\mathbf{Y} = \begin{pmatrix} (x_{1,1} - \bar{x}_1)/S_1 & \cdots & (x_{1,m} - \bar{x}_m)/S_m \\ \vdots & \ddots & \vdots \\ (x_{n,1} - \bar{x}_1)/S_1 & \cdots & (x_{n,m} - \bar{x}_m)/S_m \end{pmatrix}$$

- Where $\bar{x}_j = \sum_{i=1}^{n} x_{i,j}/n$ and $S_j^2 = \sum_{i=1}^{n}(x_{i,j} - \bar{x}_j)^2/n$ and $S_j = \sqrt{S_j^2}$, $j = 1, \dots, m$ variables

The **Y** matrix

- Obtain the **Y** matrix
- This is based on matrix notation, although this matrix can be also obtained just by obtaining the mean and standard deviation of each column and using them in the raw matrix.

$$\mathbf{B} = (\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathsf{T}}\frac{1}{n}\mathbf{I})\mathbf{X}$$

and

$$\mathbf{Y} = \mathbf{B}\mathbf{D}_{1/s}$$

- where **I** is the identity matrix, $\mathbf{1}^{\mathsf{T}} = [1, \ldots, 1] \in \mathbb{R}^n$ (matrix transpose) and

The $\mathbf{Y}$ matrix

$$\mathbf{D}_{1/s} = \begin{pmatrix} 1/s_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1/s_m \end{pmatrix}$$

R code; The **Y** matrix

```
  ### Assuming that we have the X matrix ###
print(X)
n <- nrow(X); m <- ncol(X)    # number of rows and columns

## obtain the 1 matrix####
mx1<-matrix(rep(1,n),nrow=n, ncol=1)
#### 1 transpose matrix ####
mx1T<-t(mx1)
#### Identity matrix ####
Id <- diag(n)
#### Weight matrix D ###
D <- Id/n
#### Compute the  B matrix ####

B<-(Id-mx1 %*% mx1T %*% D) %*% X
```

R code; The **Y** matrix

```
### Compute the S matrix (variance-covariance matrix of X)

S<-(t(B) %*% B)/(n)

### Obtain a diagonal D_1s matrix of order m

D_1s<- diag(sqrt(1/diag(S)), nrow=m, ncol=m)

### Obtain the Y matrix ###

Y<-B %*% D_1s
```

2.3 The Correlation matrix

$$\mathbf{R} = \begin{pmatrix} 1 & r_{1,2} & \cdots & r_{1,m} \\ r_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_{m-1,m} \\ r_{m,1} & \cdots & r_{m,m-1} & 1 \end{pmatrix}$$

where $r_{u,v} = S_{u,v}/(S_u S_v)$ is a correlation coeficient

- This matrix defines the correlation structure of the original $\mathbf{X}$ data
- If variables are correlated we expect to obtain values of these coefficients distinct than zero
- This can be tested by using the Bartlett's Sphericity test
- If this matrix contains a large amount of zeros then we should not consider a PCA

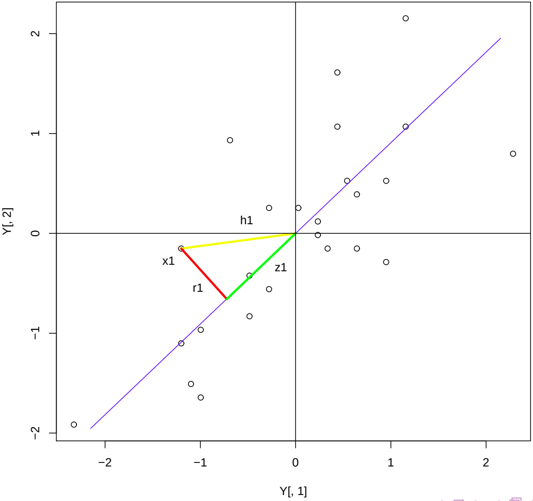
R code; The **R** matrix

```
### Compute the  R matrix ###
### We can obtain this matrix directly in R by using ###

R<-cor(X,method="pearson")
```

2.4 Data representation and Orthogonal projection

- Now the **Y** matrix provides the adjusted values of **X**



- Our objective is to find a line that goes through the points and conveys maximum information
- For a given point (individual) i, $(Y_{i,1}\ \ Y_{i,2})$, we want to obtain a new value Z_i, where $Z_i = a_1 Y_{i,1} + a_2 Y_{i,2}$
- So we need to estimate the a_1 and a_2 coeficients for this new line
- This line reduces the dimensionality of the data from 2 to 1
- Thus the new line (blue line) is a line defined by the lineal combination of variables Y_1 and Y_2, i.e. $Z = f(Y_1, Y_2)$
- PCA will help us to obtain the coeficients of this new line under some restrictions.

- It can be proved that the line that goes through the majority of points and provides maximum information is a line that results in minimun orthogonal distance between points and this line (r distances, red line)
- This can be expressed as $\sum_{i=1}^{n} r_i^2$ has to be minimum
- By the Pythagorean theorem we can write for a given point i

$$h_i^2 = r_i^2 + z_i^2$$

and summing up for all the points we obtain

$$\sum_{i=n}^{n} h_i^2 = \sum_{i=n}^{n} r_i^2 + \sum_{i=n}^{n} z_i^2$$

- $\sum_{i=n}^{n} h_i^2$ is constant and independent of the line
- Then, minimise $\sum_{i=n}^{n} r_i^2$ is equivalent to maximise $\sum_{i=n}^{n} z_i^2$, i.e. maximise the sum of the square values of point projections
- Point proyections z_i are the values of the new random variable
- Thus maximise $\sum_{i=n}^{n} z_i^2$ is equivalent to find the maximum of the variance of the new random variable, $\sum_{i=n}^{n}(z_i - 0)^2$
- Remember: the variance of a given random variable is $S^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$
- Note that the new random variable also has mean equal to zero

- Then the line that goes through the points that maximise the variance of the projected variables (the new variable) will define the *first component* of our analysis
- For a given individual i the value of its projection over the line can be obtained via

$$\begin{aligned} z_i &= a_1 y_{i1} + \ldots + a_m y_{im} \\ &= \left(y_{1,1} \; y_{1,2} \; \ldots \; y_{1,m} \right) \times \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} \\ &= \mathbf{Y}_i \mathbf{a} \end{aligned}$$

where $\mathbf{a}^{\mathbf{T}} = (a_1, \ldots, a_m)$ is a vector that defines the coeficients of the line for the m variables.

2.5 The Variance-Covariance matrix of **Y**

- Now, we know that the first principal component of our data is a lineal combination of the original variables in the direction where there is greatest variation in the ellipse that form the cloud of points
- This first principal component is the new related variable that explain more variabily of the complet data set
- Let us now obtain the projection of our transformed data over a given line, thus

$$\mathbf{Z} = \mathbf{Ya}$$

- Then, the variance of these projections can be obtained as

$$\begin{aligned} Var[\mathbf{Z}] &= \frac{1}{n}\mathbf{z}^{\mathbf{T}}\mathbf{z} \\ &= \mathbf{a}^{\mathbf{T}}\mathbf{Ra} \end{aligned}$$

where **R** is the variance-covariance matrix of the transformed variable **Y**, and also the correlation matrix of the raw data **X**

- Remember the projected variable $\mathbf{Z}$ has also zero mean
- It can be proved that the $\mathbf{R}$ matrix is the variance-covariance matrix of the centered and reduced matrix $\mathbf{Y}$
- Note that the variance of the projected values $\frac{1}{n}\mathbf{Z}^{\mathbf{T}}\mathbf{Z}$ is a scalar value

2.6 Variance maximization

- If we maximise $Var[\mathbf{Z}]$ we obtain that the set of coeficients $\mathbf{a}$ satisfies

$$\mathbf{R}\mathbf{a} = \lambda\mathbf{a}$$

- This implies that $\mathbf{a}$ is an eigenvector of the square matrix $\mathbf{R}$, and λ its corresponding eigenvalue (or characteristic value)
- Remember: An eigenvector is a vector which direction is not modified when it is multiplied by a matrix, although its length can be modified; the value of this modification is the scalar λ (the eigenvalue).
- Note that for a given matrix, we can have as many eigenvectors and eigenvalues as rows or columns (square matrix).

- Then we only need to find the eigenvectors and eigenvalues of the square matrix **R** to obtain coeficients of the linear combinations of our data set
- Also the variance of the new random variable **Z** (projected value) (based on a given eigenvector) is equal to its corresponding eigenvalue λ
- Thus the maximum eigenvalue of the matrix **R** will be the maximum explained variance
- Then its corresponding eigenvector determines the coefficients of each variable in the first principal component (a new random variable)
- The second largest eigenvalue determines the second principal component (defined as its eigenvector), and so on.

2.7 Eigenvalues and Eigenvectors

- To obtain the eigenvalues of the (square) matrix **R** we need to solve
$$|\mathbf{R} - \lambda \mathbf{I}| = 0$$
- This equation results in a polynomial (the characteristic polynomial) of degree m, where each root (solution) is an eigenvalue of **R**
- To obtain the corresponding eigenvector $\mathbf{a}^*$, for a given eigenvalue λ^*, we solve
$$\mathbf{R}\mathbf{a}^* = \lambda^* \mathbf{a}^*$$
- The first principal component will be defined by the eigenvector corresponding to the maximum eigenvalue

R code; Eigenvalues and eignevector of the **R** matrix

```
### Obtain Eigenvalues and Eigenvectors from matrix R ###

Eig<-eigen(R)

### Eigenvalues ###

Eig$values

### Eigenvectors ###

Eig$vectors
```

2.8 The first principal component

- Consider that λ_F is the maximum eigenvalue of $\mathbf{R}$, and $\mathbf{a}_F = (a_{F1}, a_{F2}, \ldots, a_{Fm})^{\mathbf{T}}$ its corresponding eigenvector
- The first principal component projection for the first individual (say) Y_{1j} $(j = 1, \ldots, m)$ is obtained as

$$Z_{F1} = a_{F1}Y_{11} + a_{F2}Y_{12} + \ldots + a_{Fm}Y_{1m}$$

- where Z_{F1} is the new value of the first individual using the lineal combination provided by the first principal component
- Note that essentially the vector $\mathbf{a}_F$ weights the effect of each variable in the new variable Z_F
- We can also write (in matrix notation)

$$\mathbf{Z_F} = \mathbf{Y a_F}$$

- where $\mathbf{Z_F} = (Z_{F1}, Z_{F2}, \ldots, Z_{Fn})^{\mathbf{T}}$

2.9 The second principal component

- It can be proved that the second largest eigenvalue provides the second principal component
- Consider now that λ_S is the second largest eigenvalue of $\mathbf{R}$, and $\mathbf{a}_S = [a_{S1}, a_{S2}, \ldots, a_{Sm}]^{\mathbf{T}}$ its corresponding eigenvector
- Then the second principal component projection for the first individual can be obtained via

$$Z_{S1} = a_{S1} Y_{11} + a_{S2} Y_{12} + \ldots + a_{Sm} Y_{1m}$$

- This second component is orthogonal to the first component (they are independent)
- It explains the part of remain variability which is not explained by the first component
- We can also write (in matrix notation)

$$\mathbf{Z_S} = \mathbf{Y a_S}$$

- where $\mathbf{Z_S} = (Z_{S1}, Z_{S2}, \ldots, Z_{Sn})^{\mathbf{T}}$

- The first and second principal components are descriptives tools
- Note that the description is now done using the centered and reduced new variables
- This can help to identify the hidden structure of the data

R code; To obtain the first and second components

```
### First component ###
### Obtain the eigenvector for the first component ###

FC<-matrix(Eig$vector[,1], nrow=m, ncol=1, byrow=T)

### Obtain the vector Z1 with the values of new variable
### (first component)
Z1<-Y %*% (+1*FC) ##if you want to change the sign
```

```
### Similarly for the second component ###

SC<-matrix(Eig$vector[,2], nrow=m, ncol=1, byrow=T)

Z2<-Y %*% (+1*SC) ##if you want to change the sign
```

2.10 Perpendicular (orthogonal) regression line

- An ordinary regression line of X_2 on X_1 minimizes the sum of squares of vertical lines from points to the line
- The first principal component line represents a perpendicular "regression" line between the other two
- Then the first principal component is the vector direction $\mathbf{a}_F$ of a line that goes through point $(\bar{Y}_1, \bar{Y}_2)$ (for the modified variables, and assuming two variables)
- In other words, the first principal component is a line that goes through point (a_{F1}, a_{F2}) and point $(\bar{Y}_1, \bar{Y}_2)$
- And the second principal component is the vector direction $\mathbf{a}_S$ of a line that also goes through point $(\bar{Y}_1, \bar{Y}_2)$
- Notice that as they are orthogonal, $\mathbf{a}_F^T \mathbf{a}_S = 0$

2.11 Properties of the components

- The trace of the matrix **R** (correlation matrix) provides the total variance of the modified data

$$tr(\mathbf{R}) = Var(Y_1) + \ldots + Var(Y_m) = m$$

- where m is the total number of variables
- Moreover, this total variance can be also obtained as the sum of the resulting eigenvalues associated to **R**

$$\lambda_1 + \ldots + \lambda_m = m$$

- Each component explains part of the variability associated to the original data
- The proportion of the variability explained by each component is equal to its corresponding eigenvalue.

- Thus, the proportion of variability explained by the first component is

$$\lambda_F/tr(\mathbf{R}) = \lambda_F/m$$

- Obviously, as the first component is the component with the largest eigenvalue, this is the component that explains more variability, and so on.
- The proportion of variability explained by the second component is

$$\lambda_S/m$$

and so on.

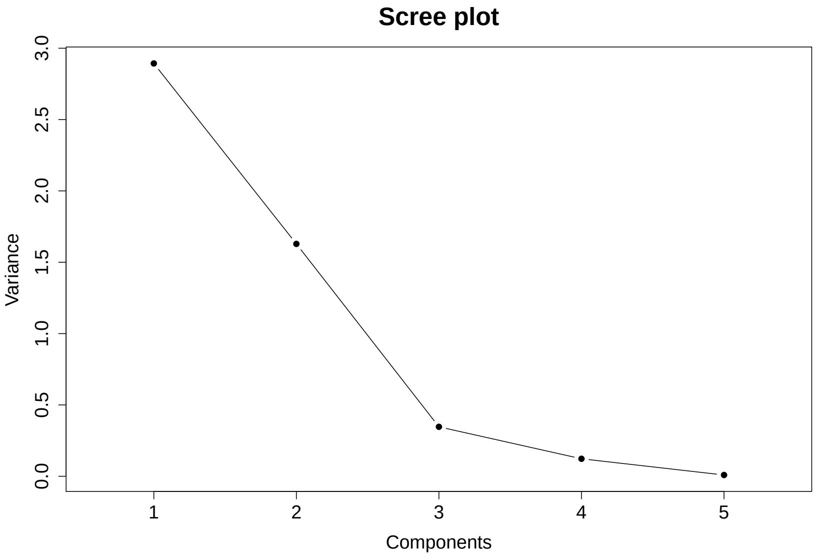
```
### Obtain the proportion of the variability ###
### explained for each component ###

Z_var<-round(Eig$values/m*100,2)

### where Z_var is vector with proportion of ###
### variability for each component ###
```

2.12 Selection and number of components

- A priory, we can consider all the associated eigenvectors to matrix **R** to describe our data
- However, in practice, we shall only consider the resulting eigenvectors that explain most of the total variability (for instance more than the 80% of variability)
- As a rule of thumb, we can consider the components with eigenvalues larger than 1
- We can consider a Scree plot to decide which components to retain
- A Scree plot is a graphical tool where the eigenvalues on the y-axis and the number of factors on the x-axis are shown
- It always displays a downward curve and it is used to decided the number of components to retain


Scree plot
Variance
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Components
1
2
3
4
5

2.13 Biplots and Correlation circles

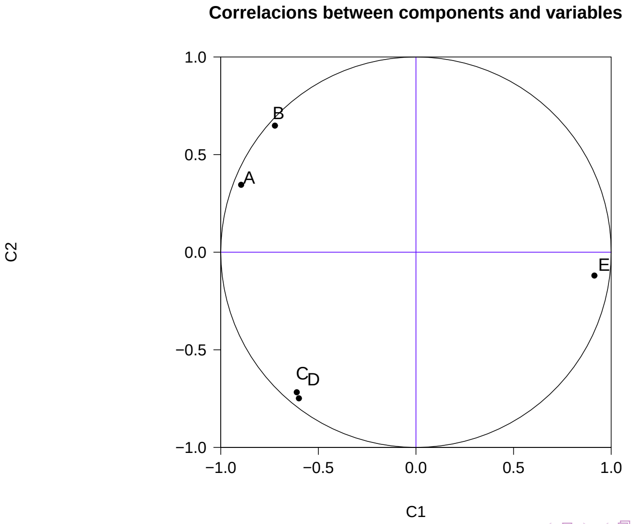
- The observation of the resulting components in bivariate plots (biplots) can help to identify and describe the inner structure of the data
- Each variable is correlated with each component and this correlation mainly determines the interpretation of the PCA
- The correlation between the ith component and m variables is obtained via

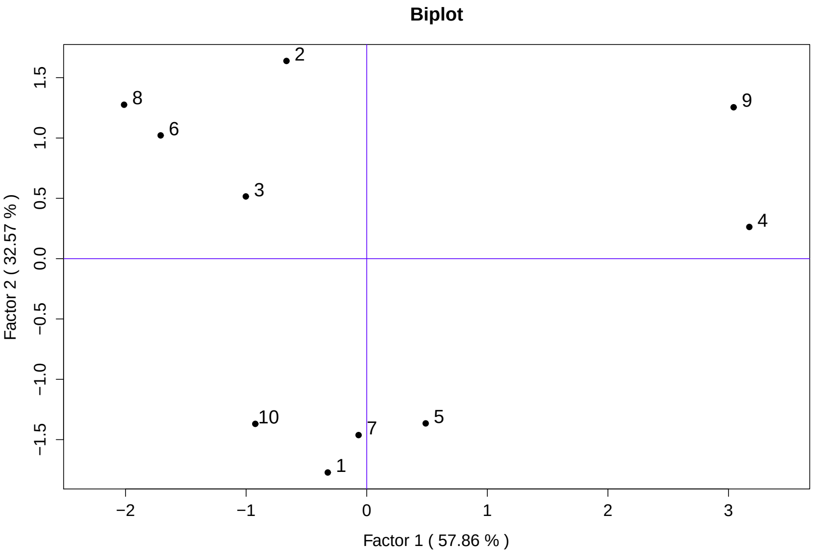
$$\sqrt{\lambda_i}\mathbf{a}_i^{\mathbf{T}} = \sqrt{\lambda_i}(a_{i1}, a_{i2}, \ldots, a_{im})$$

- For the first component the related correlations with the m variables can be obtained as

$$\sqrt{\lambda_F}(a_{F1}, a_{F2}, \ldots, a_{Fm})$$

- These correlations usually are shown via the Correlation circle




Biplot
Factor 1 (57.86 %)
Factor 2 (32.57 %)
−2
−1
0
1
2
3
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
1
2
3
4
5
6
7
8
9
10

```
### To obtain the matrix of the correlation ###
### between components and variables ###
### Here we consider only 3 components ###
Rcx <- diag(0, nrow=3, ncol=m)
for (j in 1:3){
  for (k in 1:m) {
    Rcx[j,k] <- sqrt(Eig$values[j])*Eig$vector[k,j]
  }
}
```

2.14 Components interpretation

- The interpretation of the components can be difficult and it can require a lot of expertise
- One may expect that components summarise information from sets of variables with some similar characteristics
- For this analysis to be easily understandable, variables should have large correlation values only with a single component

Lesson 3: Expectation Maximization algorithm

3.1 Introduction

- This technique was initially proposed by Dempster, Laird and Rubin; JRSS, B (1977)
- This is an iterative method for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables.
- However, this method can be easily adapted to analyse mixture models
- In this way, this approach can be considered in cluster classification and recognition

3.2 The EM algorithm; general setting

- Let us consider that we have a sample with size n with variables $\mathbf{x}$ where several observation are missing
- The EM algorithm allows the estimation of the model parameters $\boldsymbol{\theta}$ assuming the presence of missing values
- This methods is an iterative procedure where an initial value of the model parameters is considered to obtain iteratively values of model parameters, and values of the missing values.
- This method is based on a maximization of the likelihood of the model parameters and the expectation with respect to the variables of the missing values
- This Expectation maximization procedure provides estimated values for the model parameters, and values for the missing values
- The iteration procedure finishes once the estimated values of the model parameters converge

- The capability of this procedure to obtain (estimated) missing values permits to use it as a method to classify data
- We can generate artificial new variables that allow us to detect and recognise distinct clusters
- This method estimates the values of these new variables and finally allow us to detect distinct cluster structures.

3.3 The Normal distribution

- A random variables X follows a Normal distribution with $E[X] = \mu$ and $Var[X] = \sigma^2$ if it has the density function

$$f_X(s) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(s-\mu)^2}{2\sigma^2}\right), -\infty < s < \infty$$

- $X \sim N(\mu, \sigma)$, where $\sigma = \sqrt{\sigma^2}$
- Remember: a probability density function is a function that describes the relative likelihood for a given random variable to take on a given value
- The Normal distribution is the statistical model more used for continuous variables, for instance, the height or the weight of a human population

3.4 Mixture models; Gaussian Mixture Model

- Consider now the height, h, of a human population, then $h \sim N(\mu, \sigma)$
- Let's assume that we have men and women, so we can write $h_m \sim N(\mu_1, \sigma_1)$ for man, and $h_w \sim N(\mu_2, \sigma_2)$ for woman, altogether in the same general population
- Usually these sub-populations can be defined by distinct statistical models

- If the probabilistic models that rule the mixture are Normal distributions, this mixture is defined as a Gaussian Mixture Model (GMM)
- Then a Gaussian Mixture Model can be defined as a parametric probability density function represented as a weighted sum of Gaussian (Normal) component densities
- A mixture model is a probabilistic model for representing a population composed by several sub-populations, without requiring that an observed data set should identify the sub-population to which an individual observation belongs
- The EM algorithm can be used to classify and recognise these Gaussian sub-populations

- Let us now assume that we take a sample of n individuals of the real population of height h

Indiv.	h	Sex
e_1	1.89	M?
e_2	1.55	W?
e_3	1.62	M or W?
e_4	1.84	M?
$\vdots$	$\vdots$	$\vdots$
e_n	1.70	M or W?

- A tentative model for this data set can be defined as

$$\pi_1 f_1(h) + (1 - \pi_1) f_2(h)$$

- where $f_i(h) \sim N(\mu_i, \sigma_i)$, and $0 \leq \pi_1 \leq 1$ being the proportion of individual of the sub-population 1, for $i = 1, 2$.

3.4 Maximum likelihood estimation (MLE) method

- One step in the EM algorithm is to obtain Maximum Likelihood Estimators (MLE) of a given statistical model
- The MLE method is a statistical method for estimating the parameters of a statistical model given observations
- This method maximizes the likelihood of making the observations given the parameters of a statistical model

3.5 MLE for a Gaussian mixture

- Now using the MLE method we want to obtain estimators of parameters of the Gaussian mixture model, i.e. μ_1, μ_2, σ_1, σ_2 and π_1, where π_1 is new parameter defined in the Gaussian mixture model
- This new π_1 parameter should be estimated based on a new variable defined for each individual from the sample,
- So now for each individual we define new variables, $0 \leq z_{ij} \leq 1$ for $i = 1, 2$ and $j = 1, \ldots, n$.

- These new variables are the probability that the value of interest is from the sub-population 1 or from sub-population 2
- For instance, if $h_{23} = 1.89$, then z_1 could the probability that this height value corresponds to a man, so, likely z_1 will be very close to 1
- So in any case we do not know the values of π_1 and $\pi_2 = 1 - \pi_1$
- Note that $z_2 = 1 - z_1$ is the probability of being a woman (for our example)

- Now our initial data set has two new variables z_1 and z_2, and thus we have $2 \times n$ missing values

Indiv.	h	Sex	z_1	$z_2 = 1 - z_1$
e_1	1.89	M?	0.9?	0.1
e_2	1.55	W?	0.15?	0.85
e_3	1.62	M or W?	0.45?	0.55
e_4	1.84	M?	0.86?	0.14
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
e_n	1.70	M or W?	0.58?	0.42

- Note that we do not know the values of z_1 and z_2

- Now using the MLE method we can obtain estimates of the parameters of the Gaussian mixture model

$$\hat{\mu}_1^{(1)} = \frac{\sum z_{1j}^{(1)} h_j}{\sum z_{1j}^{(1)}}$$

$$\hat{\sigma}_1^{2^{(1)}} = \frac{1}{\sum z_{1j}^{(1)}} \sum_{j=1}^{n} z_{1j}^{(1)} (h_j - \hat{\mu}_1^{(1)})^2$$

$$\text{and} \quad \hat{\pi}_1^{(1)} = \frac{\sum z_{1j}^{(1)}}{n}$$

- Similar expressions can be obtained for $\hat{\mu}_2^{(1)}$ $\hat{\sigma}_2^{2^{(1)}}$ and $\hat{\pi}_2^{(1)}$
- Note that superscript $^{(1)}$ stands for the first iteration

- So to obtain the first iteration of $\hat{\mu}_1$, we need a first value of the known variables z_1 and z_2
- Then we need to give a tentative first value for these variables to start the iteration
- For instance, we can consider $z_{1j} = z_{2j} = 0.5$ or $z_{1j} = 1.0$ and $z_{2j} = 0.0$, for $j = 1, \ldots, n$
- Obviously, the choice of these initial values can affect the estimation of these parameters
- Also, we can just consider directly initial values for $\hat{\mu}_i^{(1)}$, $\hat{\sigma}_i^{2^{(1)}}$ and $\hat{\pi}_i^{(1)}$, for $i = 1, 2$

3.6 Expectation algorithm; the estimate of a parameter is a random variable

- Our objective is to obtain the best estimates for the new variables z_1 and z_2
- These estimates will help us to classify our data set
- An estimated value is a random variable defined by a probability distribution
- For instance, consider the population mean of the height of total population of $(10,000)$ students of the UdL
- Assume that this population mean is 1.68 metres

- If we take several samples of $n = 200$ from this population, we can obtain distinct estimated values for this population parameter, say, 1.65, 1.72, 1.69 1.59, 1.72.
- So the resulting estimator is itself a random value
- In this case, a good approximation of this population parameter could obtained as the average of all these values, thus a good approximation for the real value 1.68 metres
- So, $(1.65 + 1.72 + 1.69 + 1.59 + 1.72)/5 = 1.674$ metres
- The following step of the EM algorithm is based on the same idea
- From all the possible values that z_{ij} can take from distinct re-samplings of our real population we want to obtain the "average" one

- It can be proved that the expectation of z_{1j}, for $j = 1, \ldots, n$, conditioned to $\hat{\boldsymbol{\theta}}_1^{(1)} = (\hat{\mu}_1^{(1)}, \hat{\sigma}_1^{2^{(1)}}, \hat{\pi}_1^{(1)})$, and the data set h_j is

$$E[z_{1j}|\hat{\boldsymbol{\theta}}_1^{(1)}, h_j] = \frac{\hat{\pi}_1^{(1)} f_1(h_j|\hat{\boldsymbol{\theta}}_1^{(1)})}{\hat{\pi}_1^{(1)} f_1(h_j|\hat{\boldsymbol{\theta}}_1^{(1)}) + (1 - \hat{\pi}_1^{(1)}) f_2(h_j|\hat{\boldsymbol{\theta}}_1^{(1)})}$$

- where $f_1(h_j|\hat{\boldsymbol{\theta}}_1^{(1)})$ for a Normal distribution is

$$f_1(h_j|\hat{\boldsymbol{\theta}}_1^{(1)}) = \frac{1}{\hat{\sigma}_1^{(1)}\sqrt{2\pi}} \exp\left(\frac{-(h_j - \hat{\mu}_1^{(1)})^2}{2\hat{\sigma}_1^{2^{(1)}}}\right)$$

- Similarly, we can obtain $E[z_{2j}|\hat{\boldsymbol{\theta}}_2^{(1)}, h_j]$

- These resulting expected values $E[z_{1j}|\hat{\boldsymbol{\theta}}_1^{(1)}, h_j]$ and $E[z_{2j}|\hat{\boldsymbol{\theta}}_2^{(1)}, h_j]$ will be for us the values of z_{1j} and z_{2j} that we will used for the second iteration

$$z_{1j}^{(2)} = E[z_{1j}|\hat{\boldsymbol{\theta}}_1^{(1)}, h_j] \qquad \text{and} \qquad z_{2j}^{(2)} = E[z_{2j}|\hat{\boldsymbol{\theta}}_2^{(1)}, h_j]$$

receptively

- and then obtain for the $z_{1j}^{(2)}$ (similar results can be obtained for $z_{2j}^{(2)}$)

$$\hat{\mu}_1^{(2)} = \frac{\sum z_{1j}^{(2)} h_j}{\sum z_{1j}^{(2)}} \qquad \hat{\sigma}_1^{2^{(2)}} = \frac{1}{\sum z_{1j}^{(2)}} \sum_{j=1}^{n} z_{1j}^{(2)} (h_j - \hat{\mu}_1^{(2)})^2$$

$$\text{and} \quad \hat{\pi}_1^{(2)} = \frac{\sum z_{1j}^{(2)}}{n}$$

- This iteration procedure can be finished once

$$\|\hat{\boldsymbol{\theta}}_1^{(i)} - \hat{\boldsymbol{\theta}}_1^{(i+1)}\| < \epsilon$$

  for a convenience small value of ϵ

- Similar results can be obtained if we consider

$$\|\hat{\boldsymbol{\theta}}_2^{(i)} - \hat{\boldsymbol{\theta}}_2^{(i+1)}\| < \epsilon$$