

Multivariate description
Visualisation
Reduction of dimensionality

Data Mining course
Master in Information Technologies
Enginyeria Informàtica

Tomàs Aluja

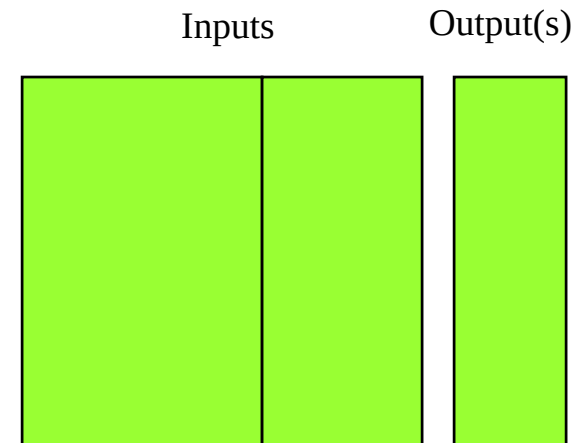
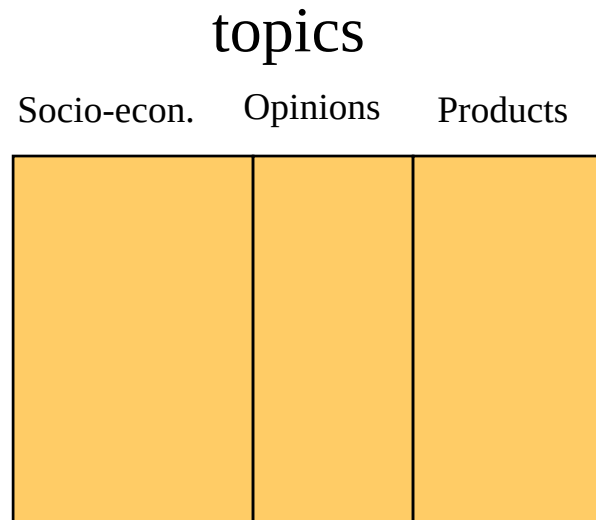


UNIVERSITAT POLITÈCNICA
DE CATALUNYA

Two types of datasets to analyze

Data in *Data Mining*:

massive, secondary, not random, with errors and missing values

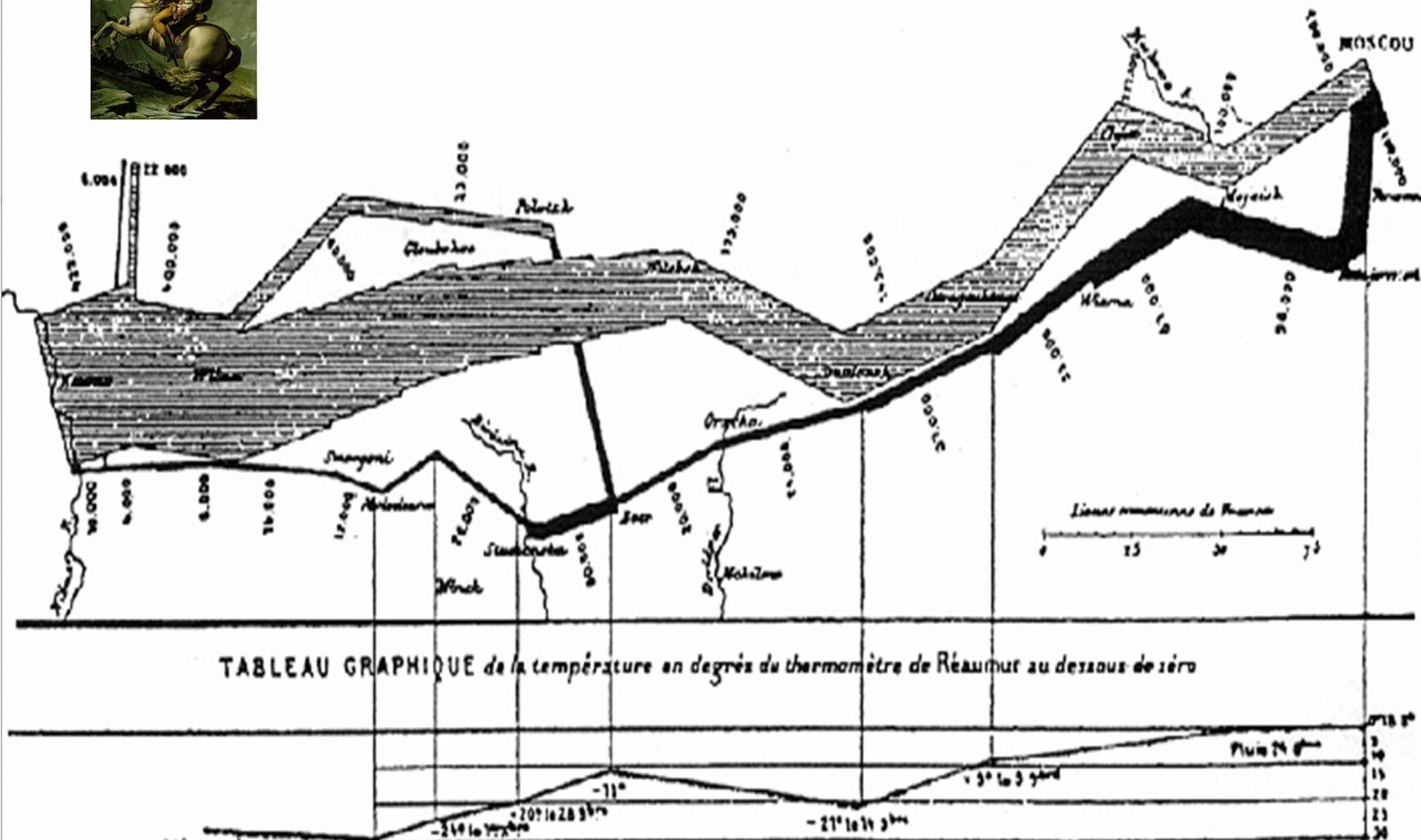


Data to explore

Data to modelize

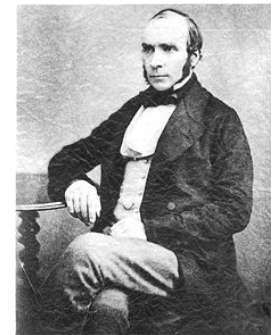
CARTE FIGURATIVE des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

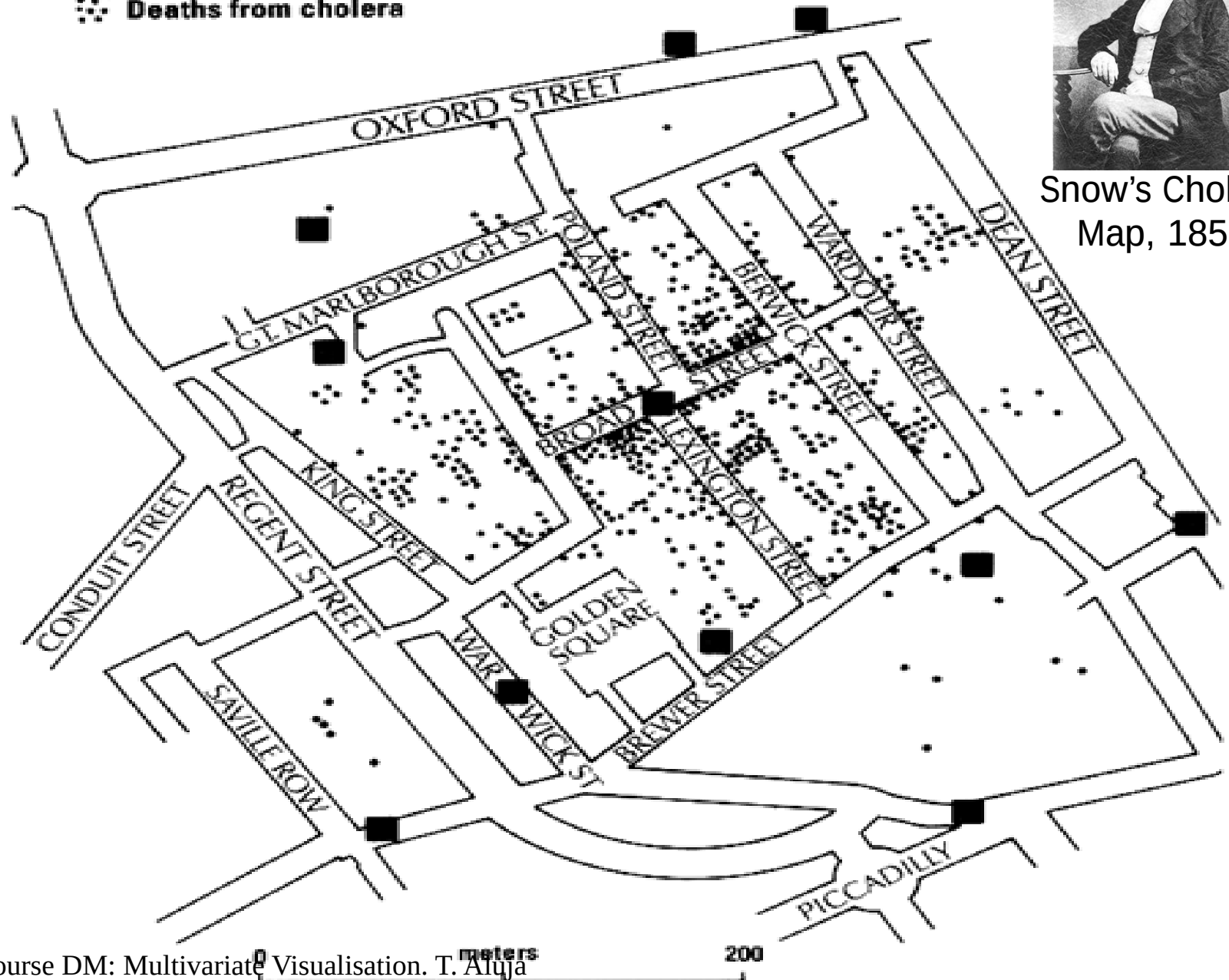


■ Pump sites

• Deaths from cholera



Snow's Cholera Map, 1855



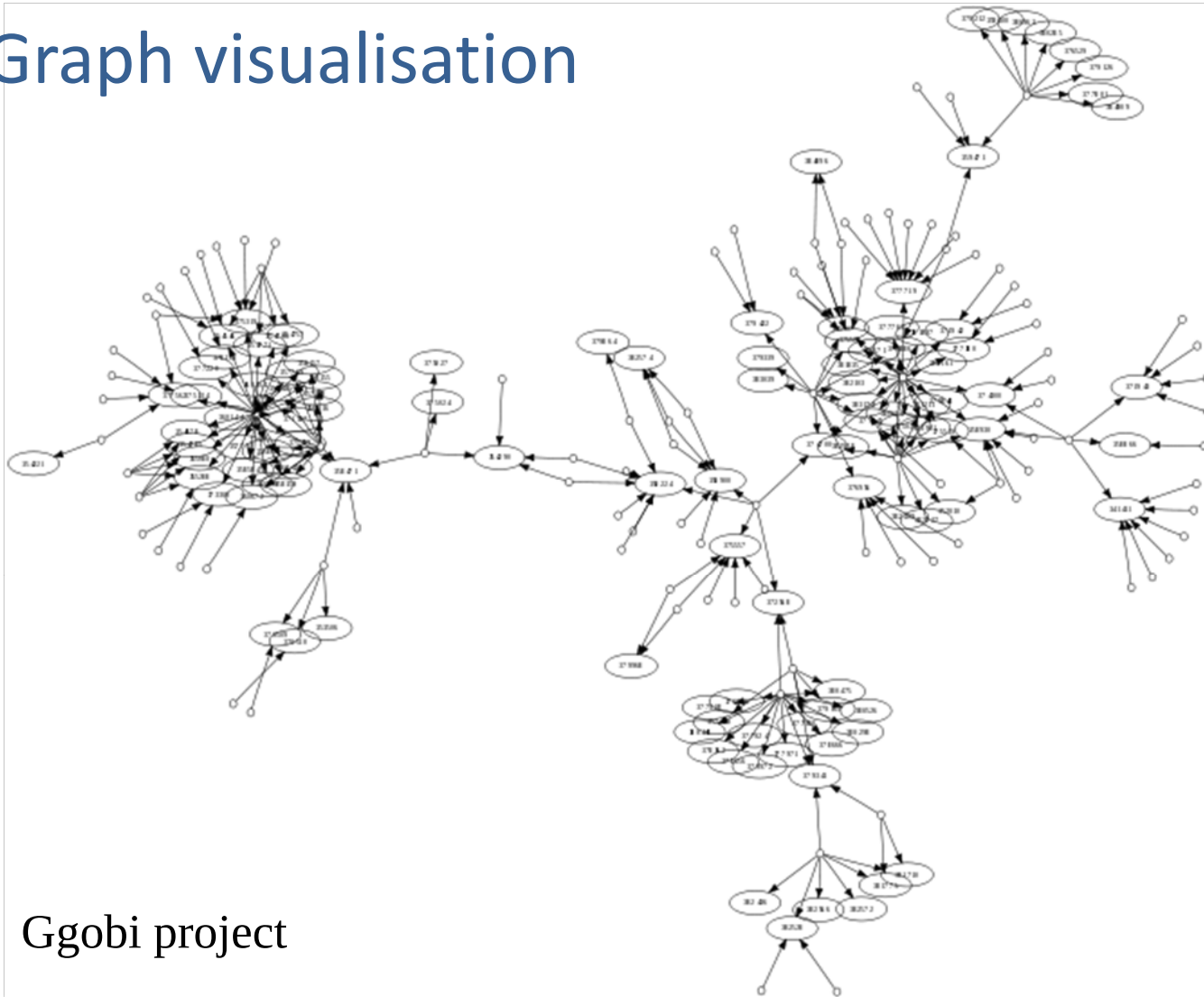
South and North Korea at night

North Korea
Notice how dark it is

South Korea,
Guess where is Seoul?

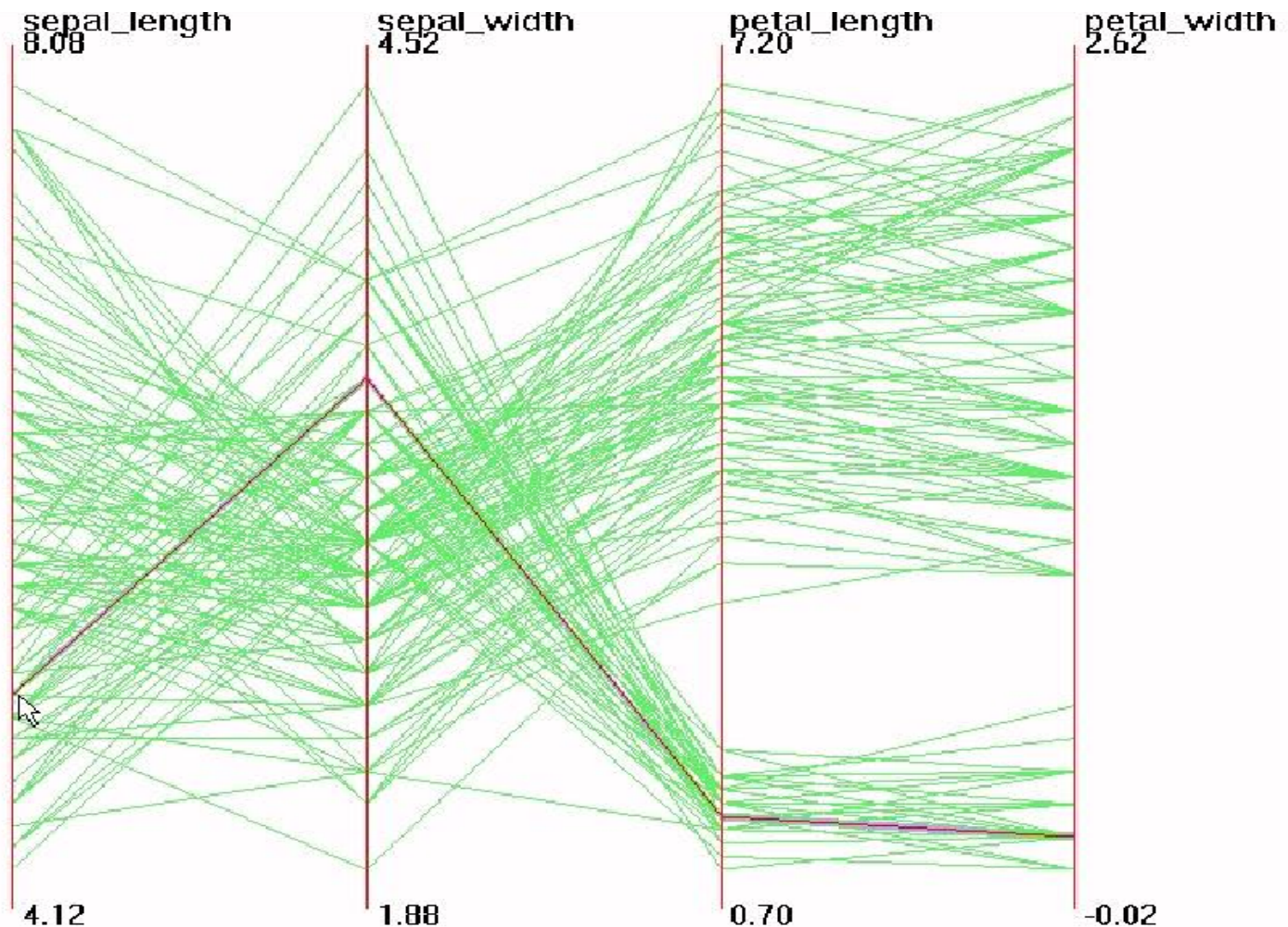


Graph visualisation

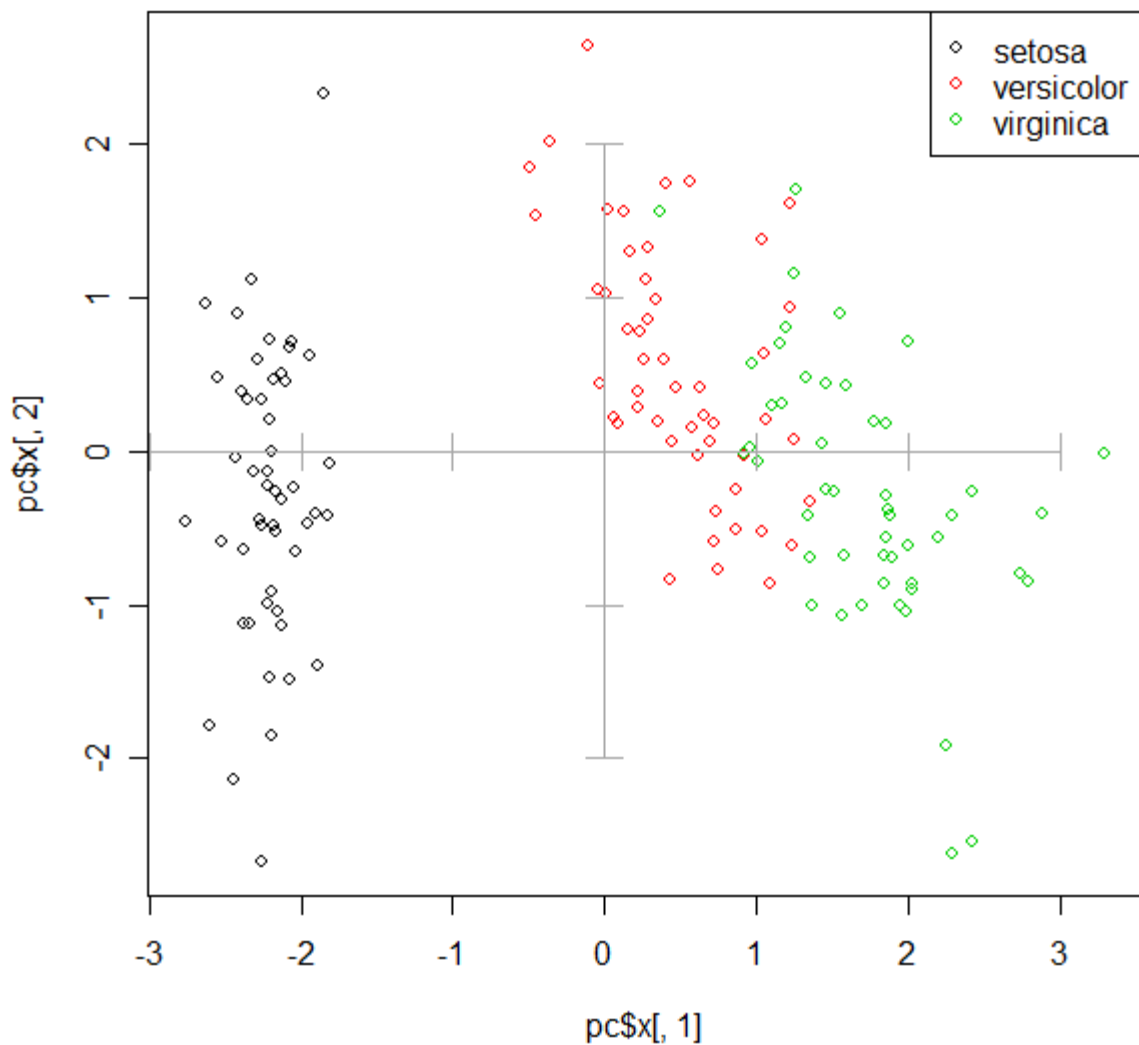


Ggobi project

Parallel coordinates of IRIS data



PCA of IRIS dataset



Iris setosa

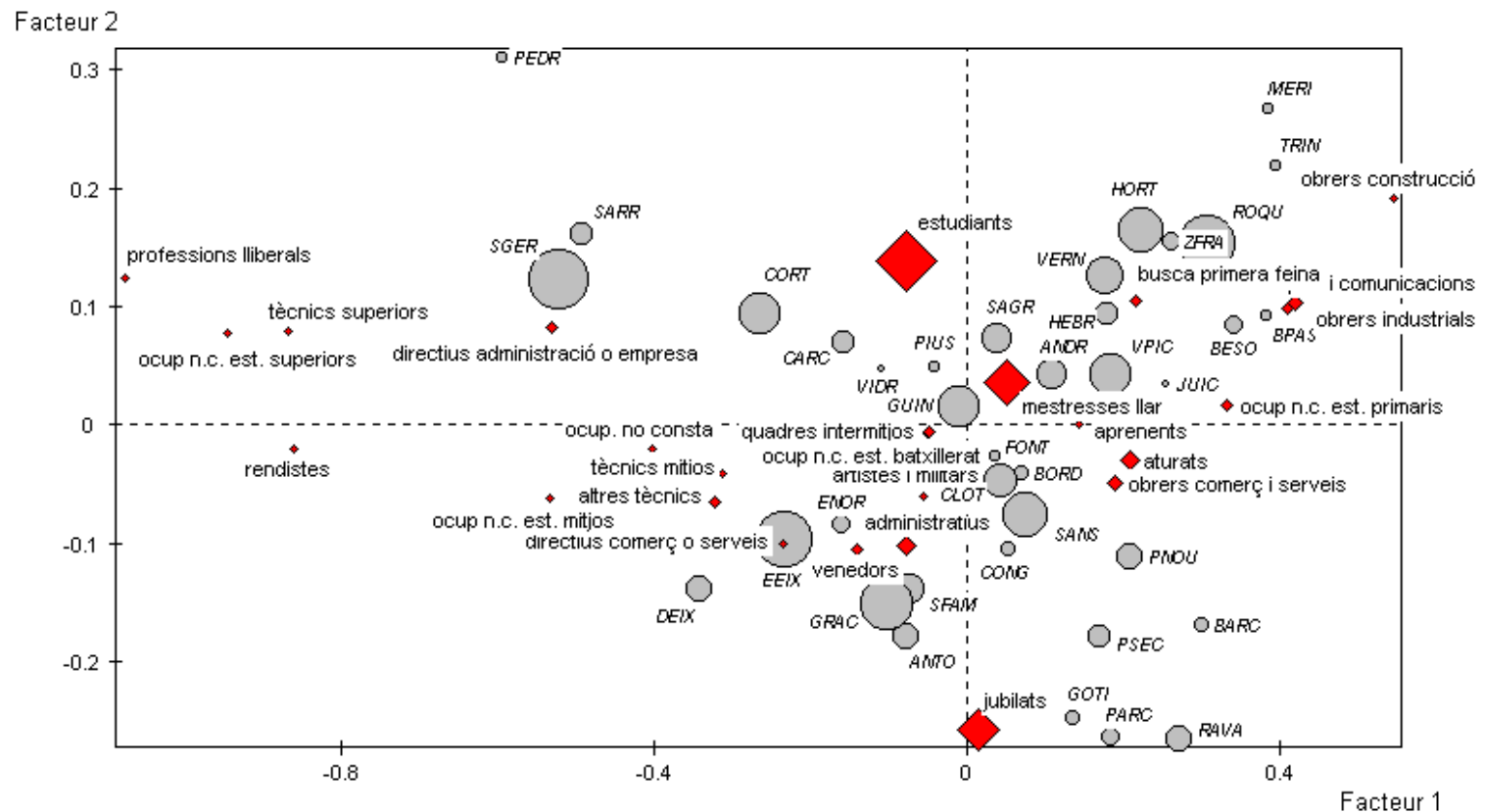


Iris versicolor



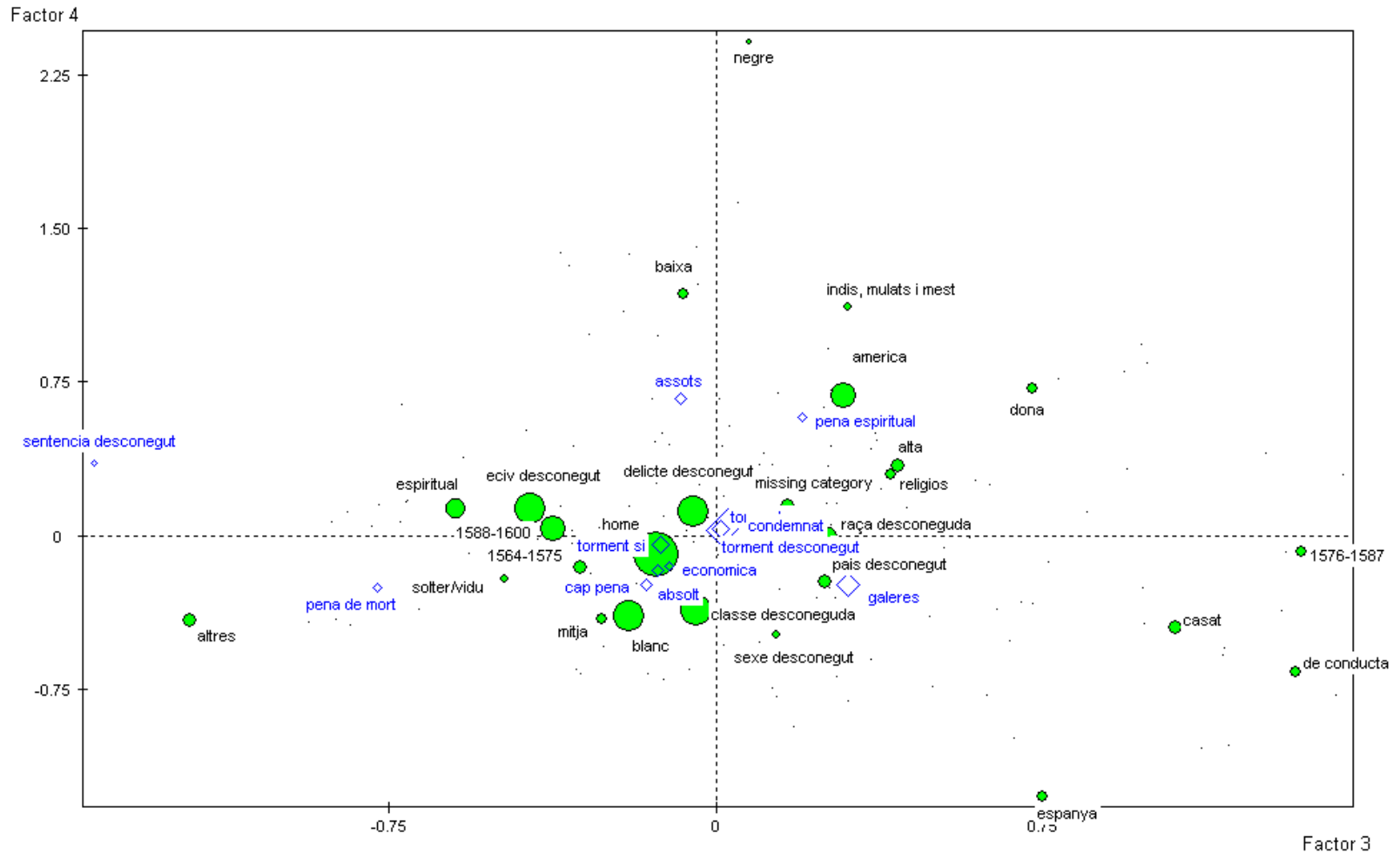
Iris virginica

Visualization of the table *BCN Quarters x Profession of inhabitants*

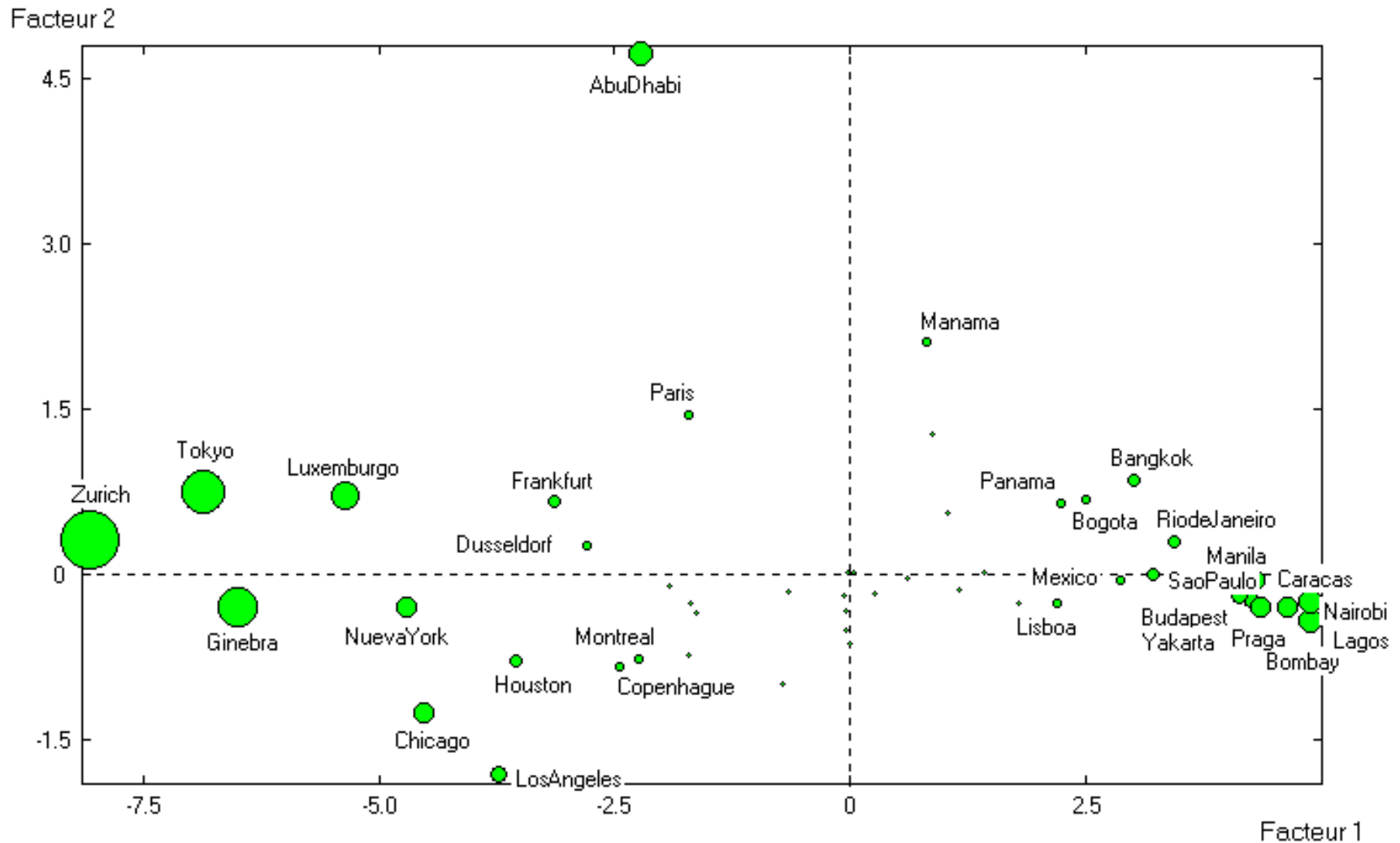


Spanish inquisition 1567-1600

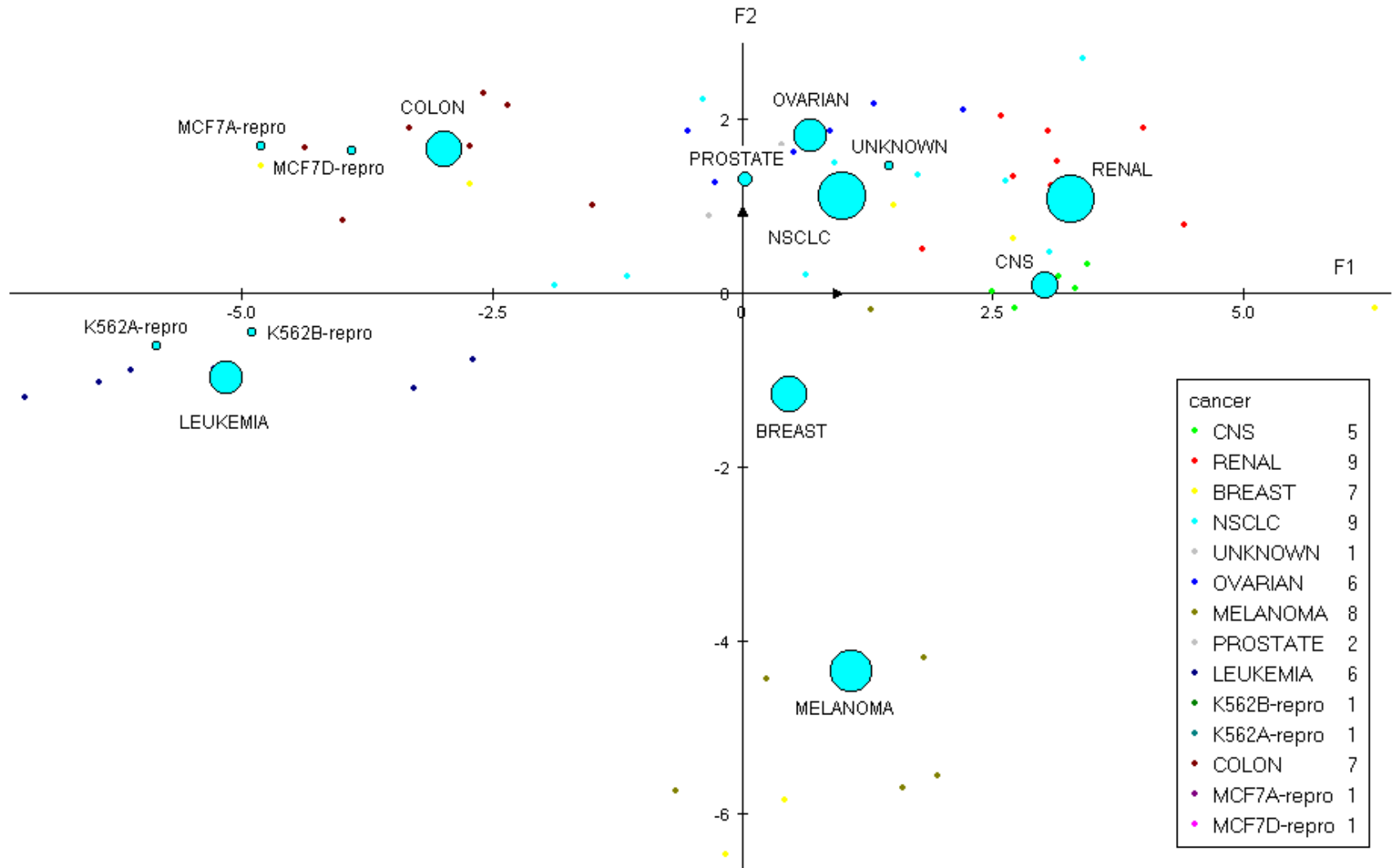
sentences & crimes



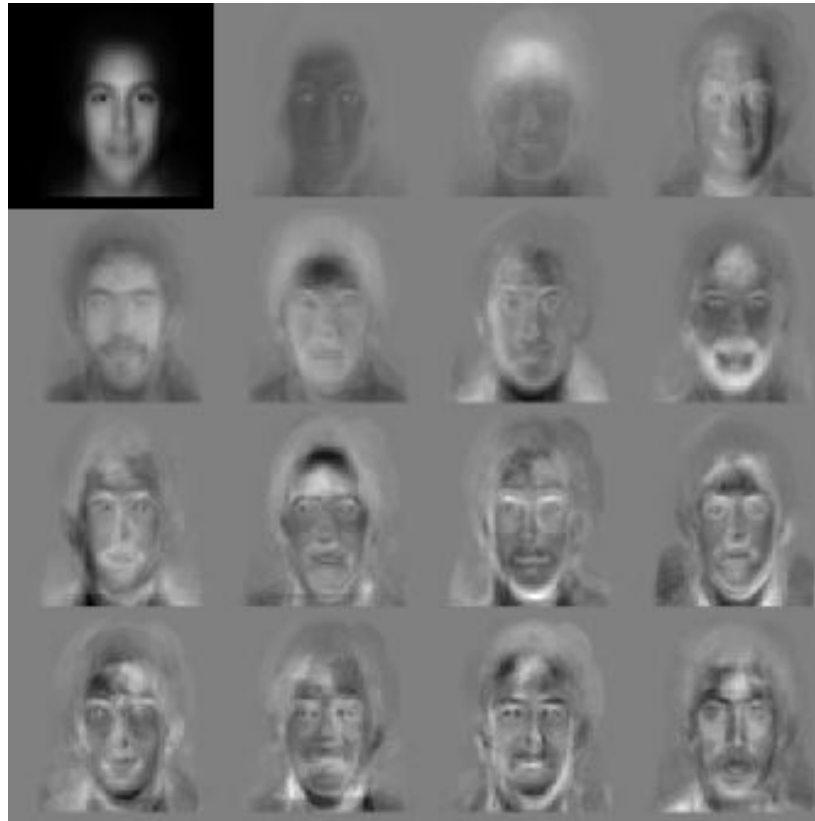
Visualisation of international cities according their salaries. USB 1994.



Microarray data: 64 cancers 6830 genes

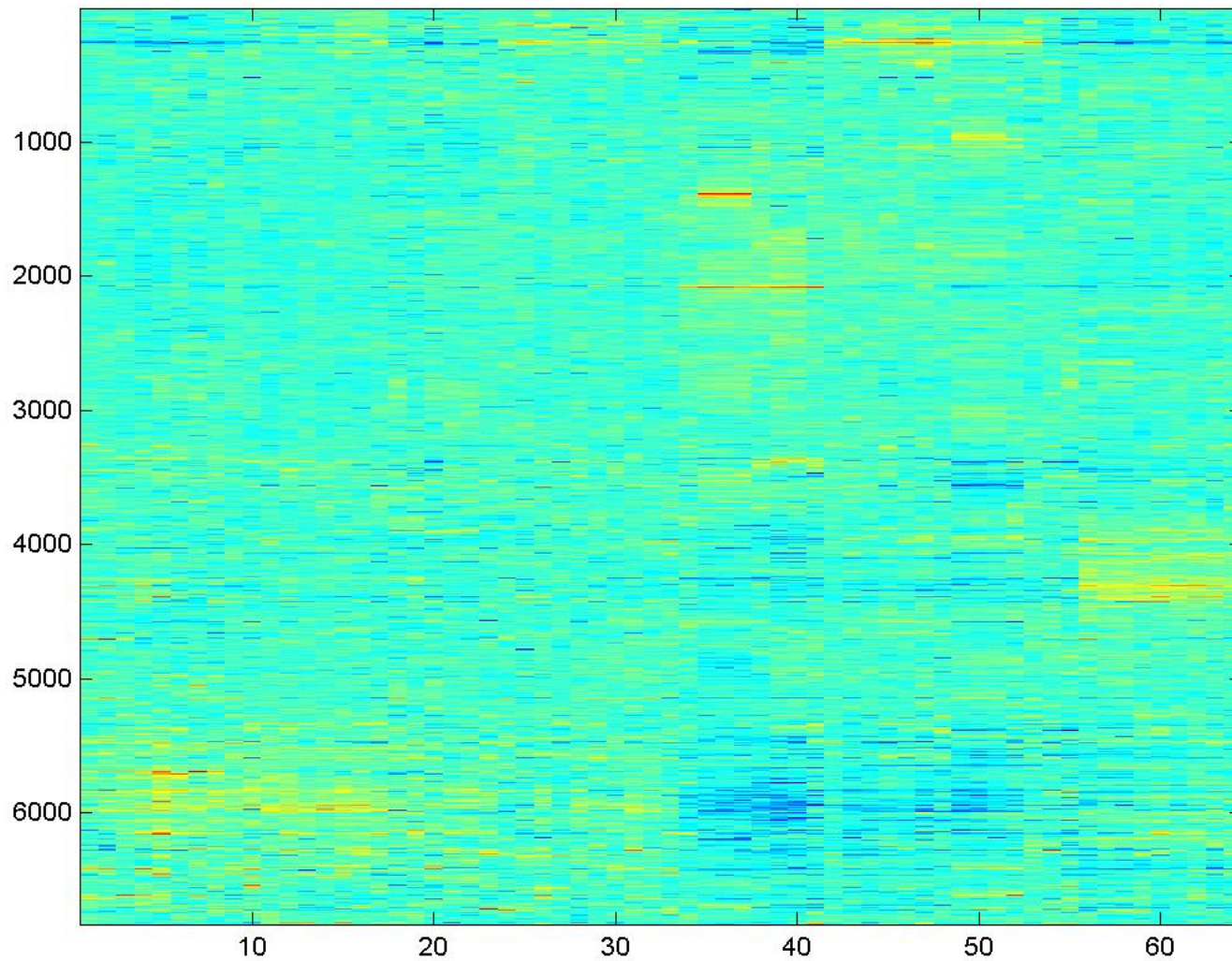


Reconstitution of images

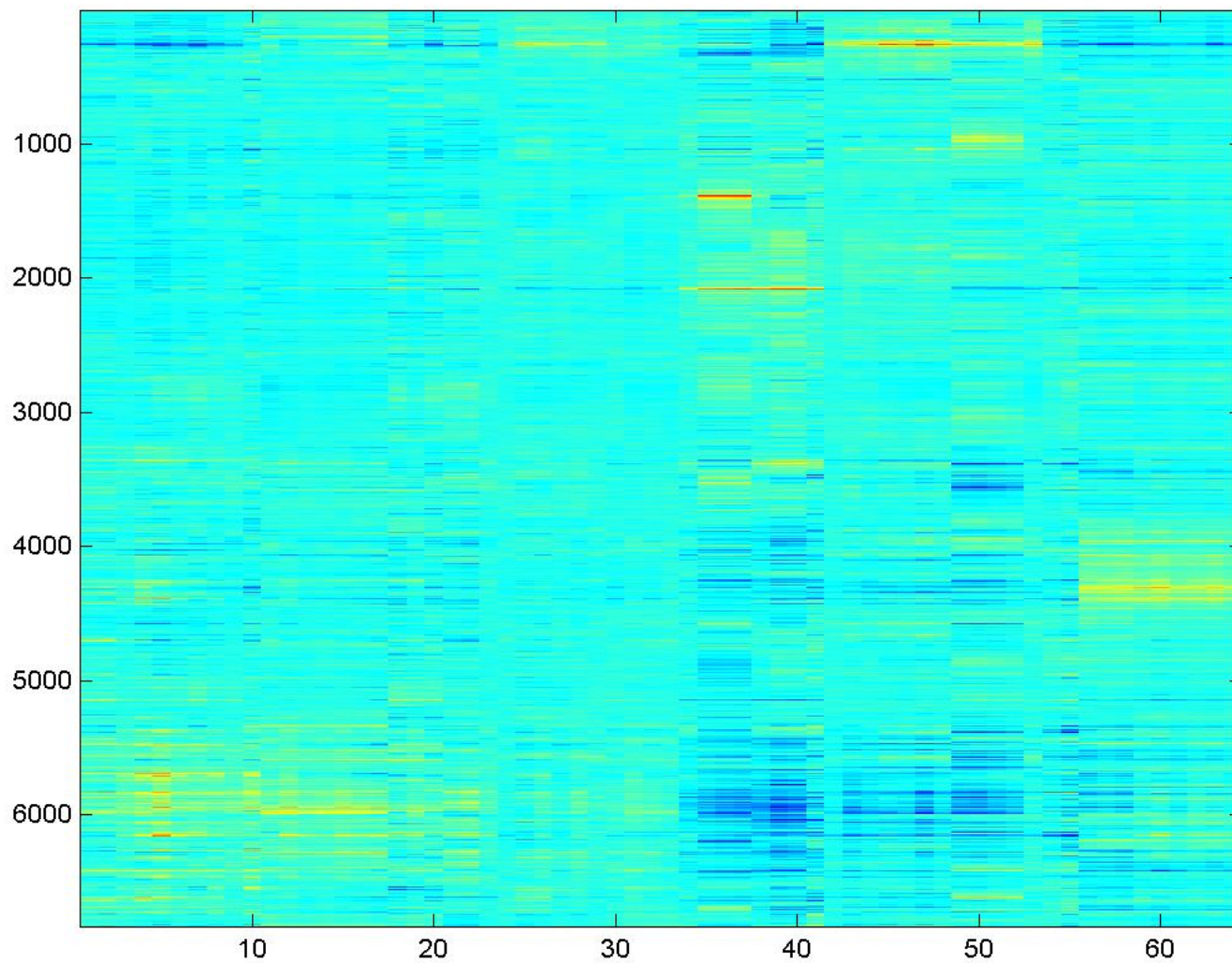


M.Turk and A.Pentland. Eigen Faces for Recognition.
Journal of Cognitive Neuroscience, 3(1), 1991.

Actual image

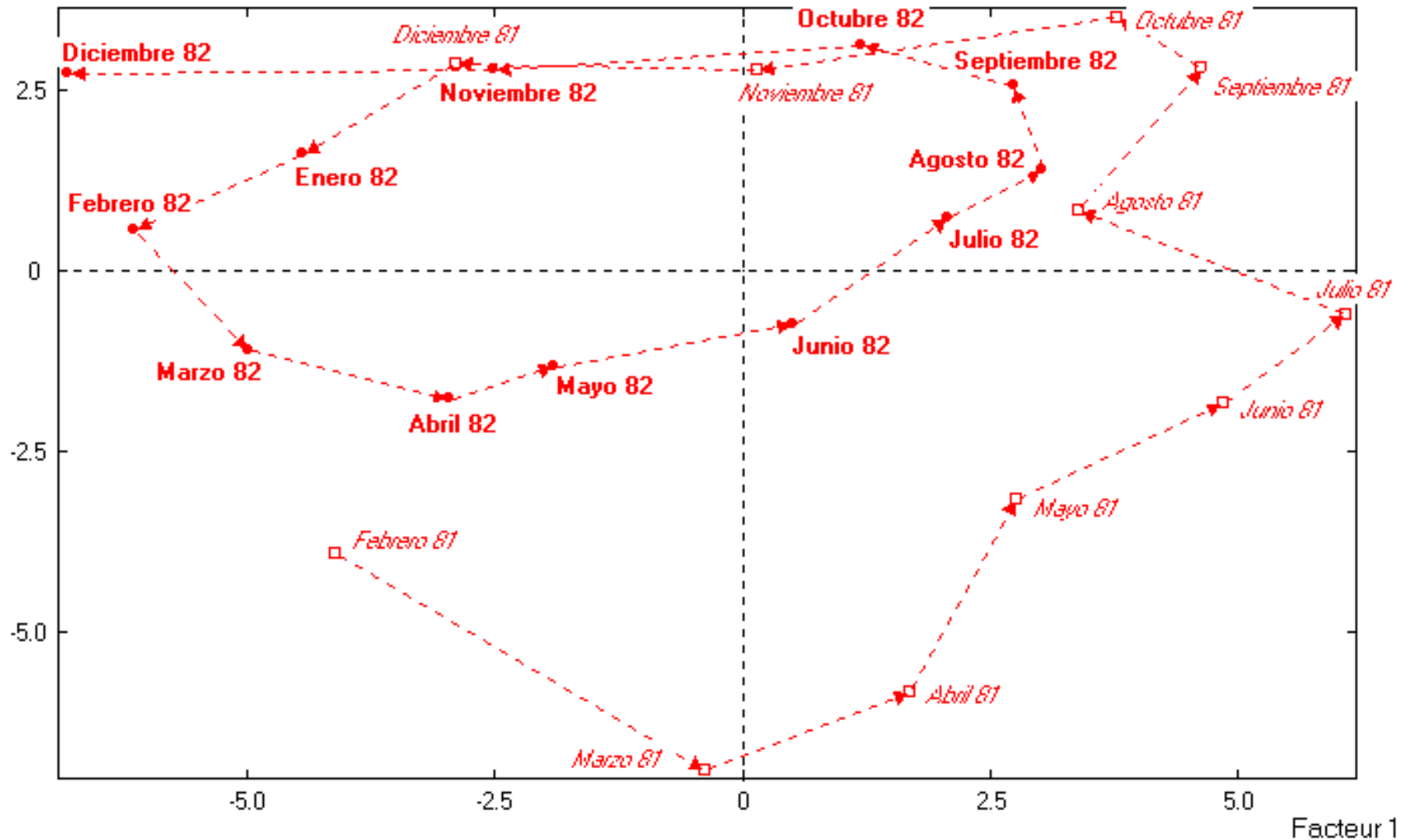


Reconstituted image



Monitoring of the inner temperatures of Lascaux cave (France):

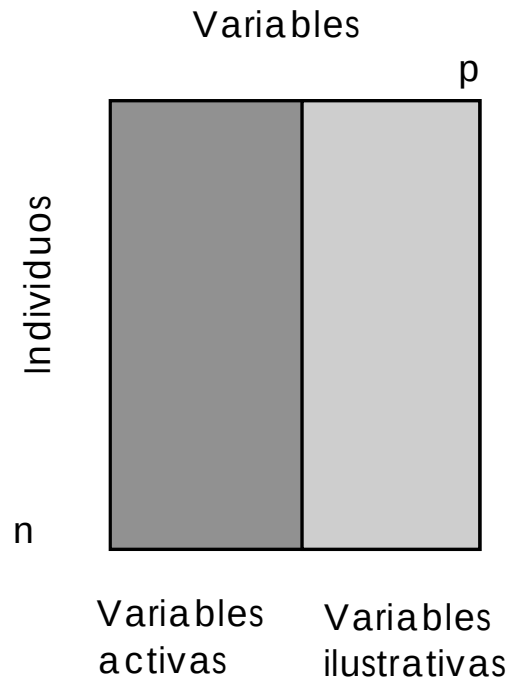
Facteur 2



Multivariate Visualization

Selection of the active topic

- Exploratory situation (without response variable but with illustrative variables).



Active topic	Multivariate technique
Continuous variables	PCA - Principal Component Analysis
Count variables	CA - (Simple) Correspondence Analysis
Categorical variables	MCA - Multiple Correspondence Analysis

PCA, CA, MCA can be useful for ...

- Visualisation of the information contained in a data matrix
- Detection of “outliers”
- Reduction of the dimensionality (feature selection)
- Image compression
- Extraction of new derived variables (latent), “feature extraction”
- Smoothing of data (error reduction, avoiding collineality)
- First phase of the explanatory variables for modeling

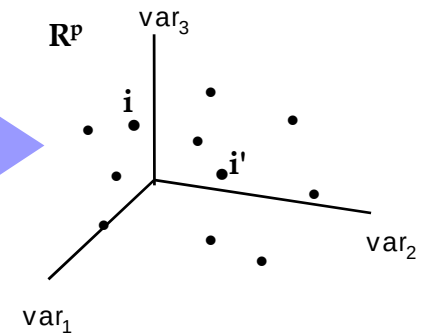
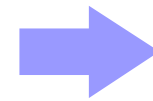
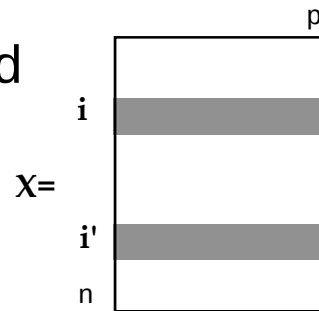


Principal Component Analysis

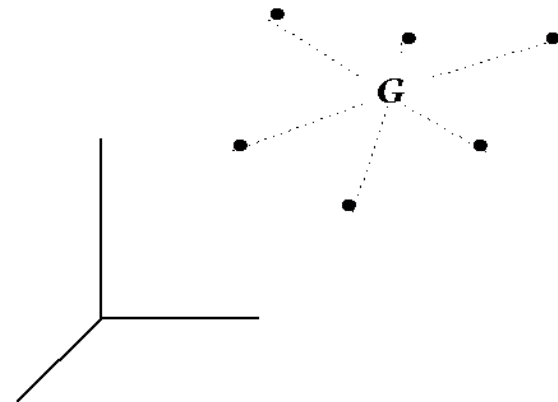
Harold Hotelling, 1895-1973

American statistician

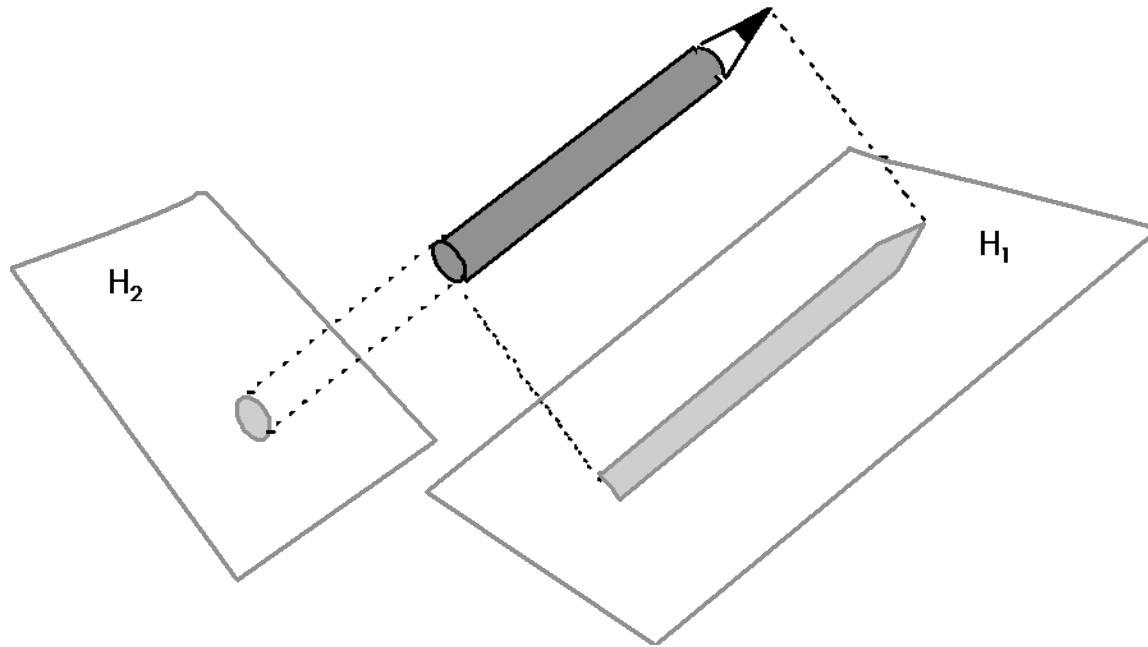
- Cloud of points associated to the rows of the data matrix



- Total information contained in the cloud of points: *the inertia respect G*

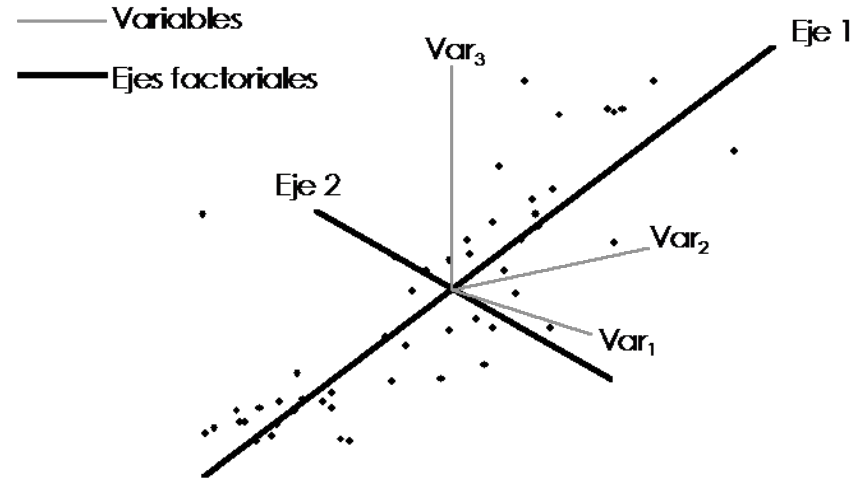


- Purpose:
 - To project the cloud of points upon a subspace (a plan) to retain the maximum of the original cloud information.



Principal Component Analysis

- Fitness Criterion
 - Find the subspace maximizing the projected inertia.



- Decomposition of inertia in orthogonal directions (factorial axes)

$$I_{total} = I_1 + I_2 + \dots + I_p$$

$$I_1 > I_2 > \dots > I_p$$

Fit in R^p

Let X be the data matrix: **centered** or **standardized**

Let call $u \in R^p$ the unit vector defining the direction maximizing the projected inertia

$$\psi = X u$$

$$\underset{u}{Max} \sum_{i=1}^n p_i \psi_i^2 = \psi' N \psi = u' X' N X u$$

$$\begin{aligned} Max u_1' X' N X u_1 &= \lambda_1 \\ u' u &= 1 \end{aligned}$$



$$X' N X u = \lambda u$$

$$\begin{aligned} &\rightarrow \lambda_1, \dots, \lambda_r \quad r = \text{rang}(X) \\ &u_1, \dots, u_r \end{aligned}$$

$$\text{diag}(X' N X) = \begin{cases} \text{Cov}(X) \\ \text{Cor}(X) \end{cases}$$

Diagonalization of the correlation matrix (or covariance)

Direction maximizing the projected inertia: u_1 .

Direction maximizing the projected inertia orthogonal to u_1 : u_2

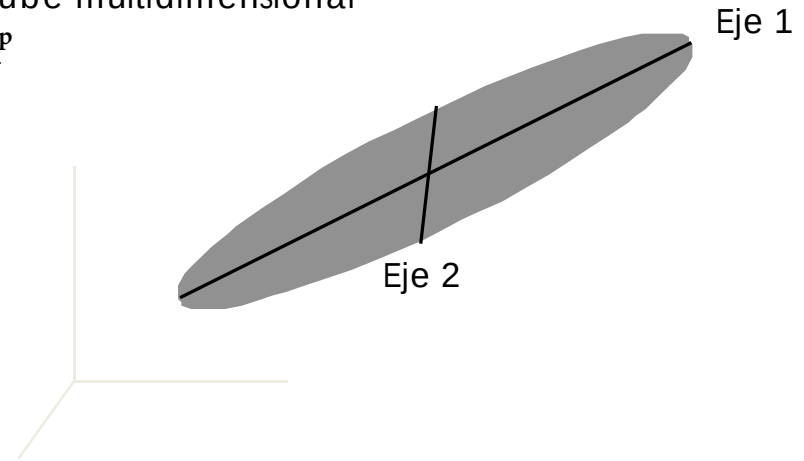
...

$$\psi_\alpha = Xu_\alpha$$

Principal Components
(derived latent variables)

Factors, ...

nube multidimensional
 $\mathbb{R}^p$

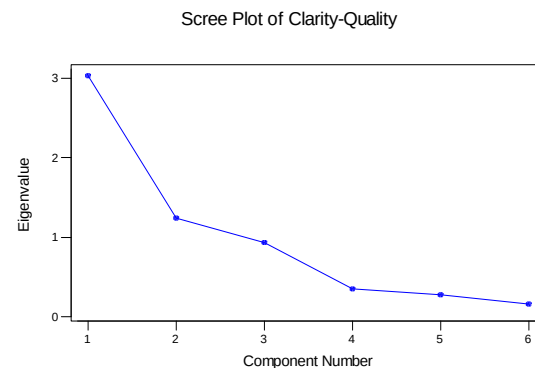


Inertia of a PC

$$\psi'_\alpha N \psi_\alpha = \lambda_\alpha$$

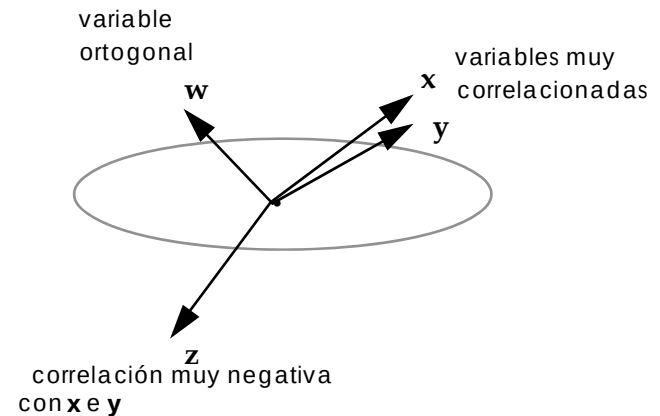
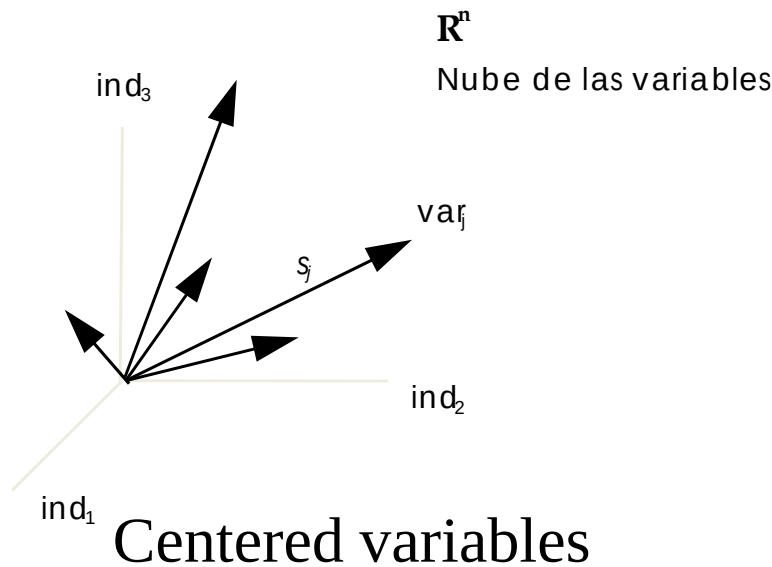
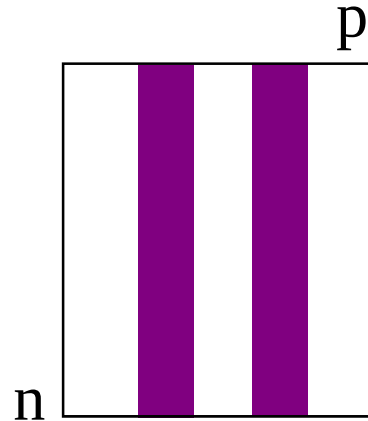
Assessing the importance of
orthogonal directions

Scree plot of eigenvalues:



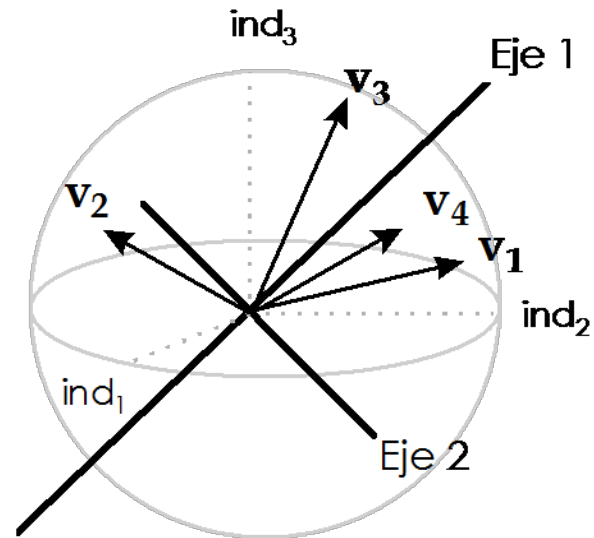
Associated cloud of points to the columns of a data matrix in $\mathbb{R}^n$

$X =$

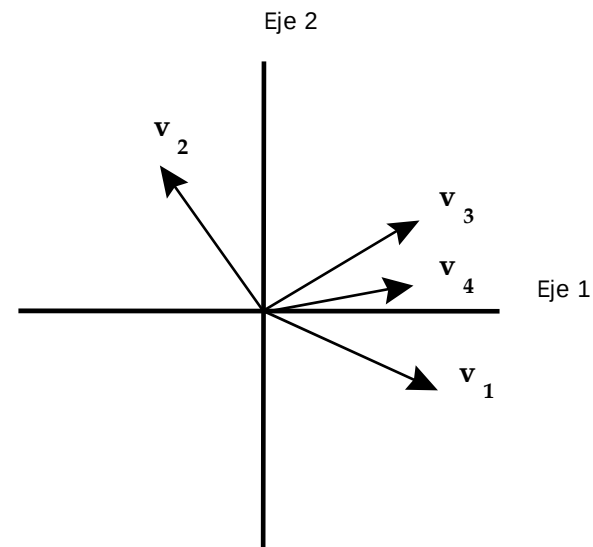
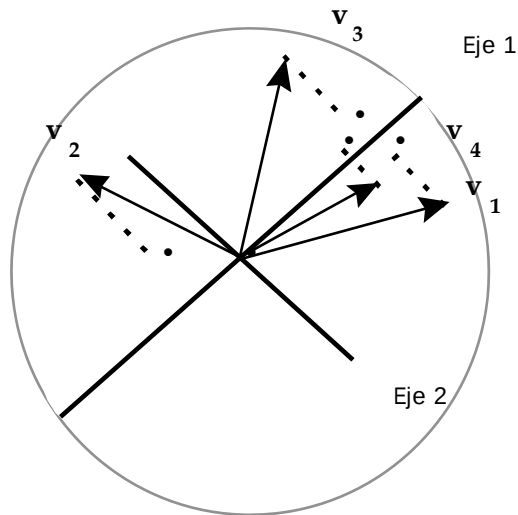


Fit in R^n (standardized data)

Original cloud



First factorial plan



Optimal joint visualisation of the correlations
between variables

Fit in R^n

Data matrix X : **centered** o **standardized**

Let $v \in R^p$ be the unit vector defining the direction maximizing the inertia:

$$\varphi = X N^{1/2} v$$

$$\left. \begin{array}{l} \underset{v}{\text{Max}} \sum_{j=1}^p \varphi_j^2 = \varphi' \varphi = v' N^{1/2} X X' N^{1/2} v \\ v' v = 1 \end{array} \right\} N^{1/2} X X' N^{1/2} v = \lambda v$$

Transition relationships between both fits:

$$\varphi_\alpha = X N^{1/2} v_\alpha$$

$$\varphi'_\alpha \varphi_\alpha = \lambda_\alpha$$

$$u_\alpha = \lambda^{-1/2} X N^{1/2} v_\alpha$$

$$v_\alpha = \lambda^{-1/2} N^{1/2} X u_\alpha$$

Indirect projection formulas

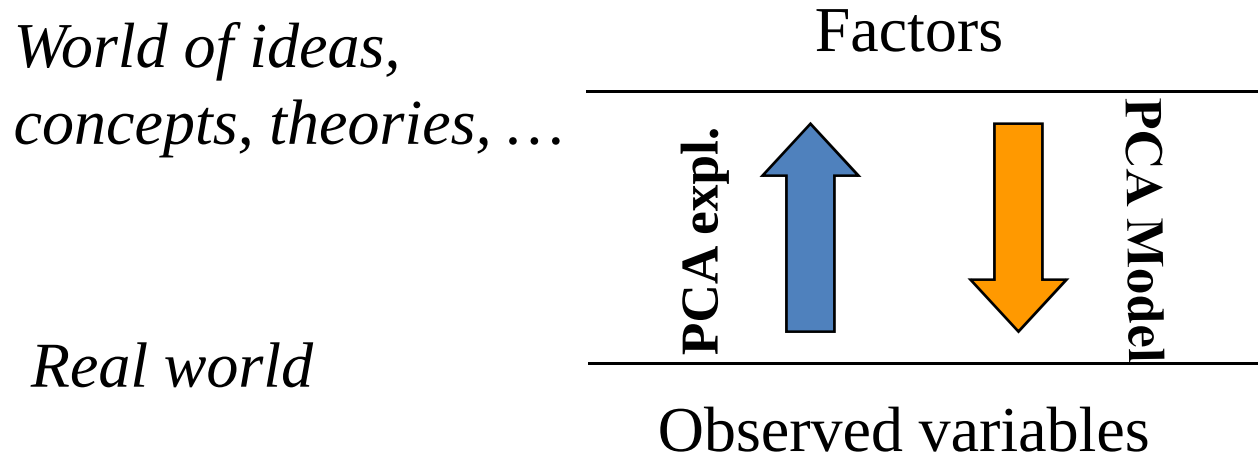
$$\varphi_\alpha = \lambda^{1/2} u_\alpha$$

$$\psi_\alpha = \lambda^{1/2} N^{-1/2} v_\alpha$$

Interpeting the projections

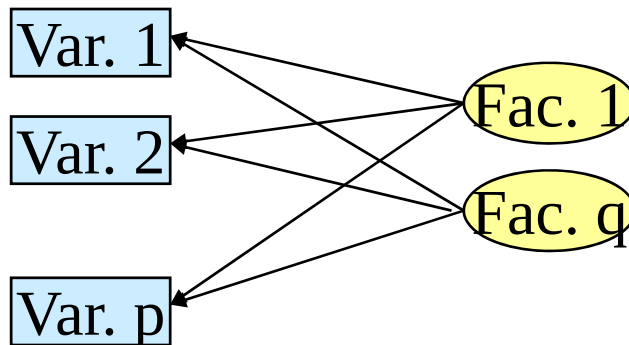
$$\varphi_\alpha = \lambda^{-1/2} X N \psi_\alpha = \begin{cases} \text{cor}(x_j, \psi_\alpha) \\ s_j \text{cor}(x_j, \psi_\alpha) \end{cases}$$

The PCA is a device to find artificial latent variables, from observed ones.



ACP: $\Psi_{\alpha} = u_{1\alpha} \mathbf{x}_1 + u_{2\alpha} \mathbf{x}_2 + \dots + u_{p\alpha} \mathbf{x}_p$

$$\Psi_{(n,p)} = \mathbf{X}_{(n,p)} \mathbf{U}_{(p,p)}$$



But only the first q Factors convey structural information, the remaining are noise

PCA in practice

- Role of the variables: Normed or non normed analysis
 - Normed PCA means to give all variables the same importance, we achieve this by standardization of data (diagonalization of the correlation matrix)
 - Non normed PCA means to give to each variable an importance proportional to its standard deviation. We achieve this working with the just centered data matrix (diagonalization of the covariance matrix)
- What variables to analyze?
 - This is the most crucial decision. Often the information contained is obvious, then try to perform partial analysis. PCA is a device of exploration.

PCA in practice

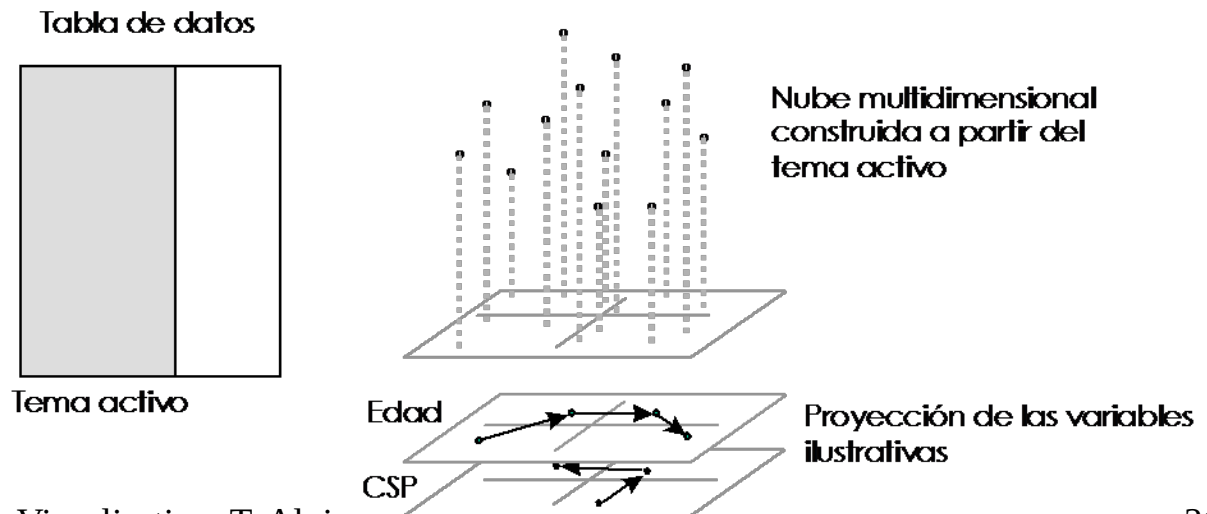
- How many factorial directions are significative?
 - Difficult to assess. How many axes remain stable with independent data?
 - Use the screeplot.
 - Perform random perturbation of data to assess stability.
- How to interpret the axes
 - The significative axes convey structural (deterministic) information of the phenomenon under study and they can be interpreted and given a name (this is the most appealing outcome).
 - Interpretation is done in the basis of the correlations between the principal component (the new artificial latent variables and the original ones, the pc is a mean variable of the most correlated).

Projection of illustrative variables

Very useful ...

It allows to relate each illustrative variable to the active topic altogether

- Continuous
 - We depict their correlations with the factorial axes.
- Categorical
 - We represent a categorical variable by the set of the centre of gravity of the different subclouds of individuals corresponding to each level of the categorical variable.



Finding the PCA solution iteratively (NIPALS)

Initialize $\mathbf{X}_1 \leftarrow \mathbf{X}$

For $h=1, \dots, r=\text{rang}(\mathbf{X})$

Ψ_h = mean column of $\mathbf{X}_h$

Repeat till convergence of $\mathbf{u}_h$

$$\mathbf{u}_h = \mathbf{X}_h' \Psi_h$$

$$\mathbf{u}_h = \mathbf{u}_h / |\mathbf{u}_h|$$

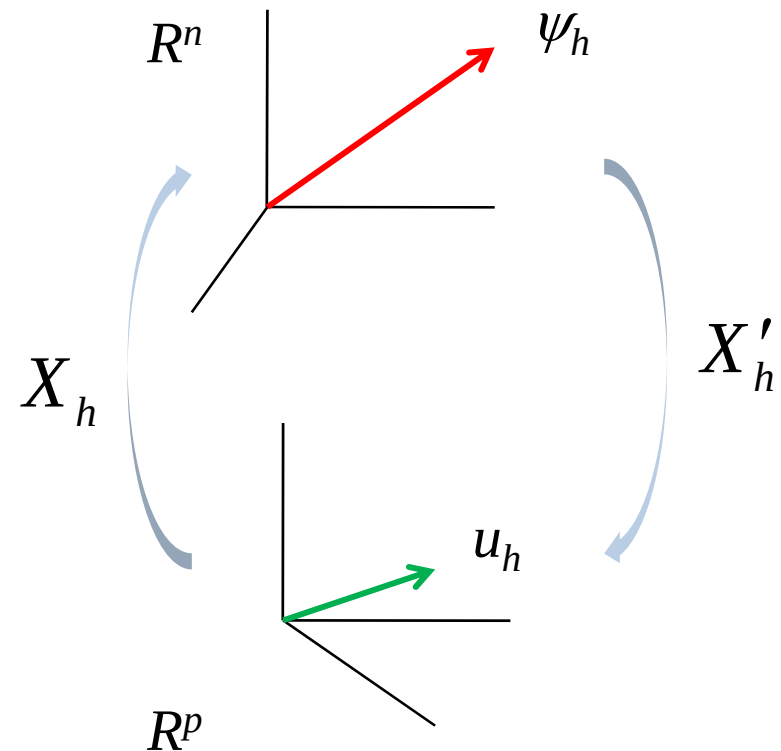
$$\Psi_h = \mathbf{X}_h \mathbf{u}_h$$

$$\mathbf{X}_h = \mathbf{X}_{h-1} - \Psi_h \mathbf{u}_h'$$

In the convergence:

$$\mathbf{X}_h' \mathbf{X}_h \mathbf{u}_h \simeq \mathbf{u}_h$$

$$\mathbf{X}_h \mathbf{X}_h' \psi_h \simeq \psi_h$$



A relevant application: Google

- GoogleTM uses SVD to accelerate finding relevant web pages. Define a web site as an *authority* if many sites link to it. Define a web site as a *hub* if it links to many sites. We want to compute a ranking $x_1; \dots; x_N$ of authorities and $y_1; \dots; y_M$ of hubs.

As a first pass, we can compute the ranking scores as follows: x_i^0 is the number of links pointing to i and y_i^0 is the number of links going out of i . But, not all links should be weighted equally. For example, links from authorities (or hubs) should count more. So, we can revise the rankings as follows

$$x_i^1 = A'y^0$$

$$y_j^1 = Ax^0$$

Where A is the adjacency matrix with $a_{ij} = 1$ if i links to j . (of 10^9 order)

But an authority depends also from the pages linking to the linking pages of the authority. Hence iterating ...

$$x_i^k = A'Ax^{k-1} \quad y_j^k = AA'y^{k-1}$$

prcomp(x, retx=T, center=T, scale.=F, tol = NULL, ...)

Arguments:

x: a numeric (or data frame) which provides the data.

retx: a logical value indicating whether the rotated variables should be returned.

center: a logical value indicating whether the variables should be shifted to be zero centered.

scale.: a logical value indicating whether the variables should be scaled to have unit variance before the analysis takes place.

tol: a value indicating the magnitude below which components should be omitted.

Attributes

sdev: the standard deviations of the principal components (i.e., the square roots of the eigenvalues).

rotation: the matrix of variable loadings (i.e., a matrix whose columns contain the eigenvectors).

x: if 'retx' is true the value of the rotated data (the centred (and scaled if requested) data multiplied by the 'rotation' matrix) is returned.

biplot(x, y, var.axes = TRUE, main = NULL, ...)

Arguments:

x: The first set of points (a two-column matrix), usually associated with observations.

y: The second set of points (a two-column matrix), usually associated with variables.

var.axes: If 'TRUE' the second set of points have arrows representing them as (unscaled) axes.

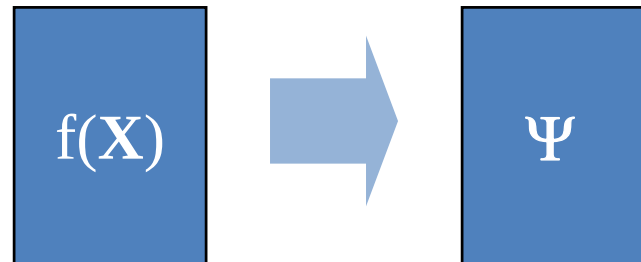


Beyond PCA $\Rightarrow$ MCA

Ludovic Lebart

French statistician, promoter of MCA

- PCA just analyzes continuous variables through their correlations, hence it just can reveal linear relationships between variables



- Thus, transform the original variables
Recode them to ordinal to take into account non linearities

$$\text{var } j \quad a_{j1} \cdots a_{jk}$$

$$x_{ij} \rightarrow 001000$$

MCA of hypercubes

- *Dimensions (= categorical variables)*
- *Measured variables in cells (=responses, they may be continuous or categorical)*
- (Hypercube can be explicit or implicit in a relational DB.)

Hypercube dimensions

$A_1 A_2 A_3 A_4$
 $B_1 B_2$
 $C_1 C_2 C_3$

A_1	B_1	C_1
A_1	B_2	C_2
A_3	B_1	C_3
A_2	B_2	C_1
...		



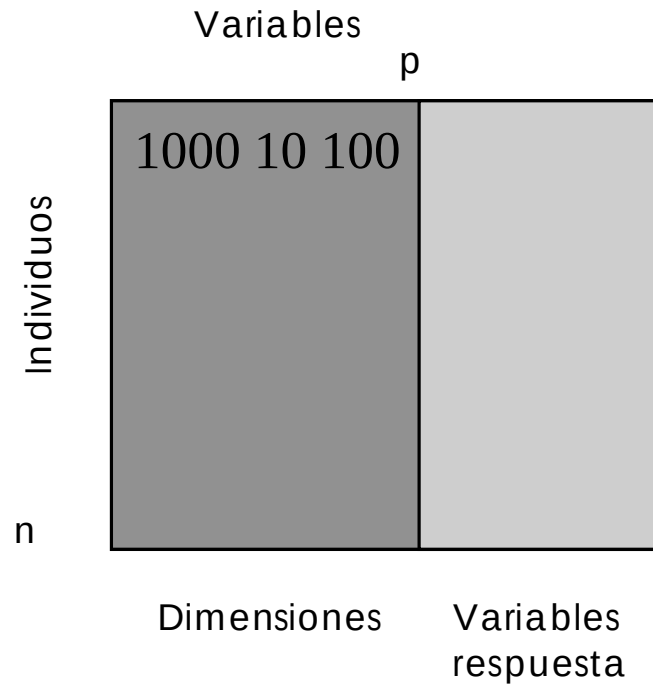
Numerical coding (binning)

1000	10	100
1000	01	010
0010	10	001
0100	01	100
...		

(=Z)

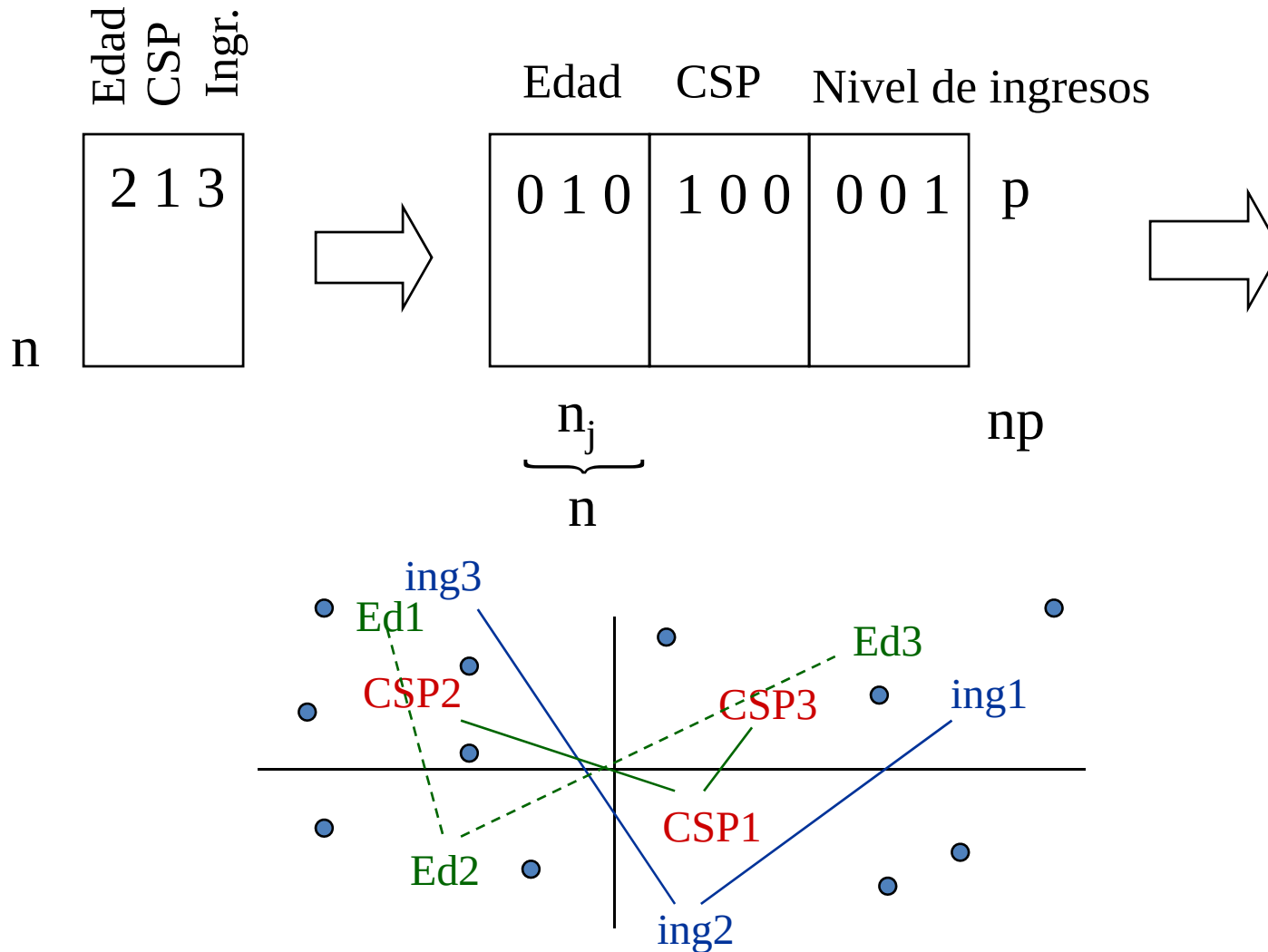
- Active Variables : Dimensions
- Illustrative variables : Responses

We will visualize the responses upon the grid provided by the dimensions

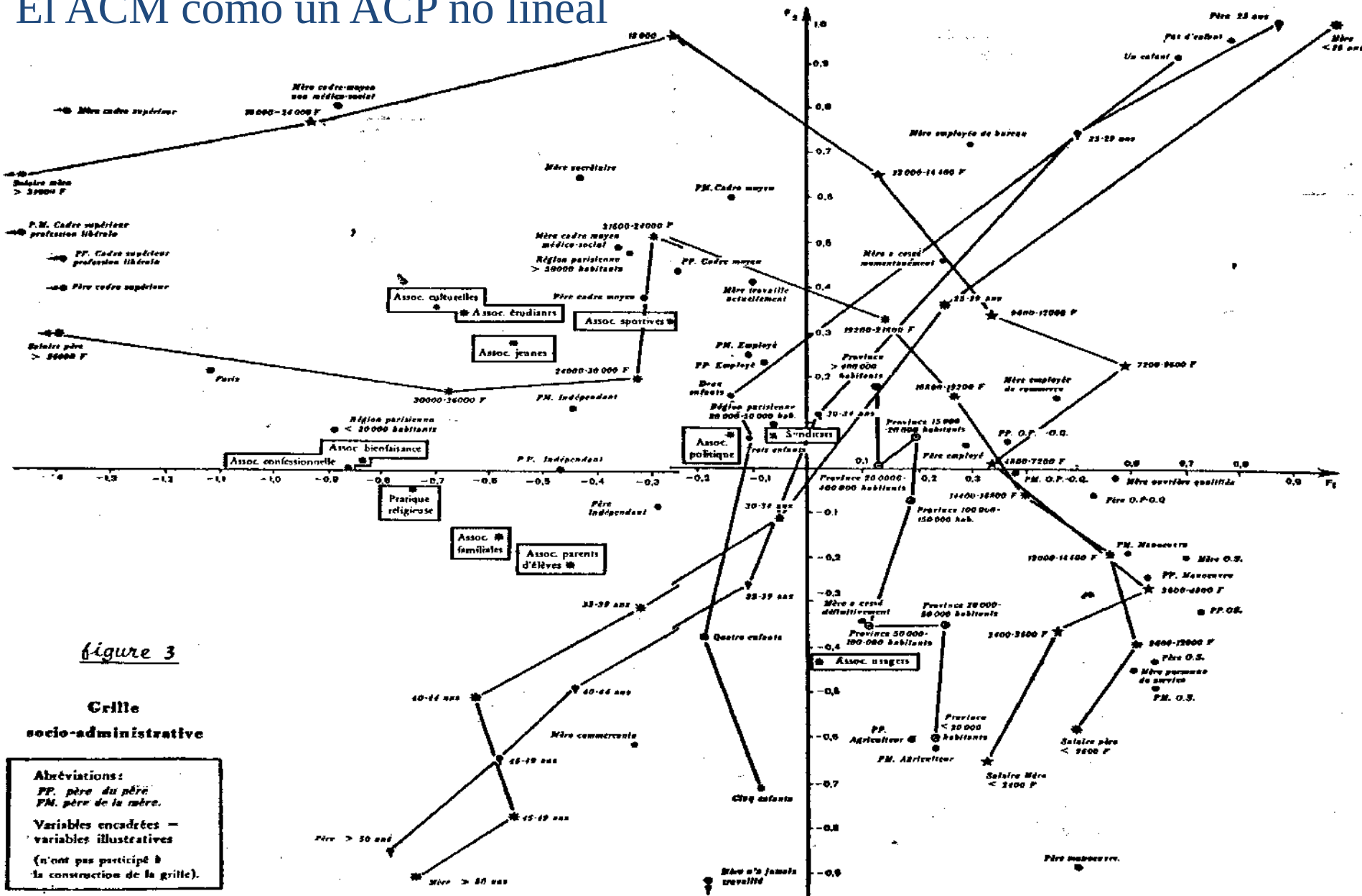


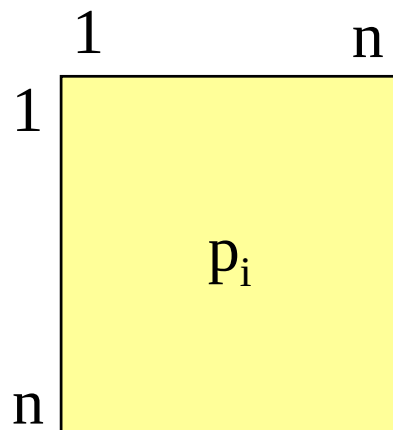
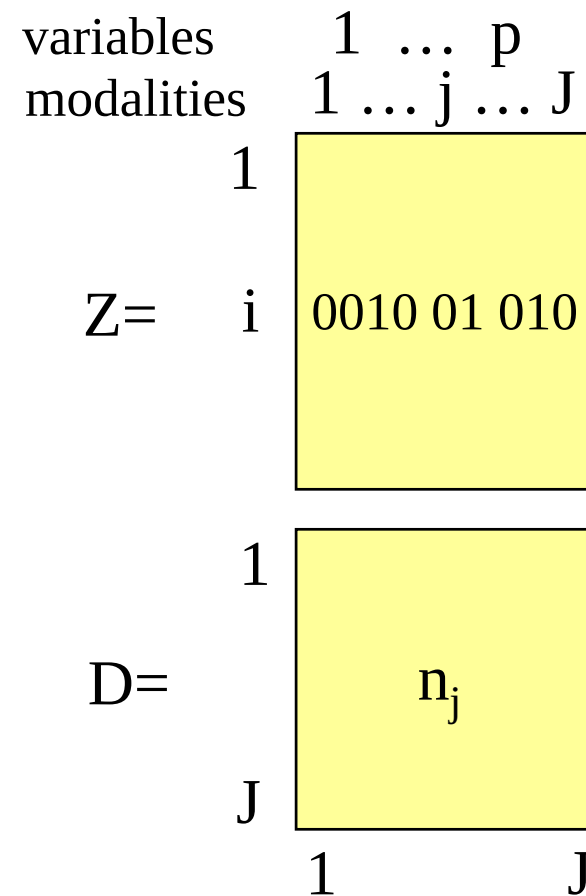
MCA

Active grid



El ACM como un ACP no lineal





$$p_i = \frac{p}{np} = \frac{1}{n}$$

$$n_j = \sum_{i=1}^n z_{ij}$$

Row profile: $\frac{1}{p} Z$

Chi-square Metric:

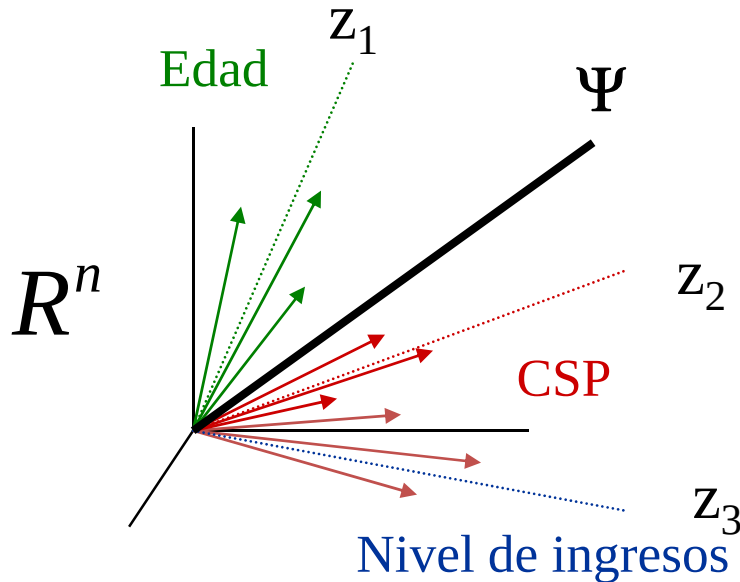
$$\left(\frac{D}{np} \right)^{-1}$$

$$\underset{u}{Max} \sum_{i=1}^n p_i \psi_i^2 = \frac{1}{n} \psi' \psi = n u' D^{-1} Z' Z D^{-1} u$$

$$np u' D^{-1} u = 1$$

$$\Rightarrow \text{eig } \frac{1}{p} Z' Z D^{-1} u = \lambda u$$

What are the factors in MCA?



$$\text{Max} \sum_{j=1}^p \text{cor}^2(\Psi, \mathbf{z}_j)$$

$$\mathbf{z}_j = \sum_k u_{jk} \mathbf{a}_{jk}$$

$\Rightarrow$ Optimal quantification of the categorical variables

Original categorical data



Equivalent continuous factors

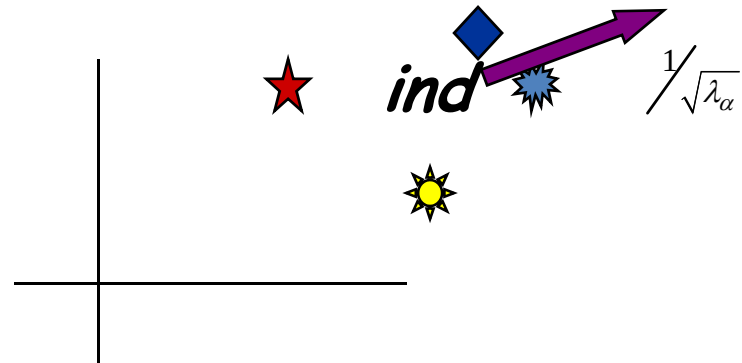


But we will work with more dimensions than in PCA

Interesting properties of the MCA displays

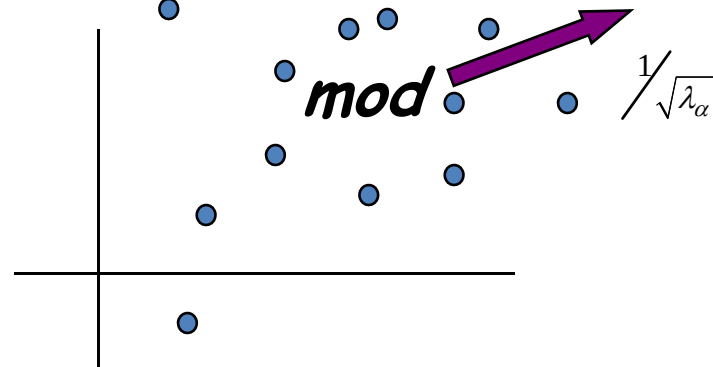
- Every individual is the cdg of their chosen modalities
(apart from a multiplicative factor)

$$\psi_{i\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \frac{\sum_j^J \varphi_{j\alpha} z_{ij}}{p}$$



- One modality (=level) is the cdg of individuals having chosen it
(apart from a multiplicative factor)

$$\varphi_{j\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \frac{\sum_i^n \psi_{i\alpha} z_{ij}}{n}$$



MCA iterative algorithm

Initialize $\mathbf{Y}_0 \leftarrow \mathbf{Z}; \mathbf{Z} \leftarrow [\mathbf{Z}_1, \dots, \mathbf{Z}_p]; \mathbf{D} = \mathbf{Z}'\mathbf{Z}; \mathbf{D}_k = \mathbf{Z}_k'\mathbf{Z}_k$

For $h=1, \dots, \text{rang}(\mathbf{Z})$

$\Psi_h = \text{rowmean of } \mathbf{Y}$

Repeat till convergence of $\mathbf{u}_h$

$$\mathbf{u}_h = \mathbf{D}^{-1} \mathbf{Y}_h' \Psi_h$$

$$\mathbf{u}_h = \mathbf{u}_h / |\mathbf{u}_h|$$

$$\Psi_h = (1/p) \mathbf{Y}_h \mathbf{u}_h$$

$$\mathbf{Y}_h = \mathbf{Y}_{h-1} - \Psi_h \mathbf{u}_h'$$

$$\mathbf{z}_k = \mathbf{Z}_k \mathbf{u}_k; \mathbf{u}_k = \mathbf{D}_k^{-1} \mathbf{Z}_k' \Psi_h \quad k=1 \dots p$$

Projection of the illustrative variables

- Continuous
 - From their correlations with the factorial axes.
- Categorical
 - As the set of cdgs of the individuals having chosen each level of the categorical variable.

library(MASS)
mca(df, nf = 2, abbrev = FALSE)

Arguments:

df: A data frame containing only factors

nf: The number of dimensions for the MCA.

Attributes:

rs: The coordinates of the rows, in 'nf' dimensions.

cs: The coordinates of the column vertices, one for each level of each factor.

fs: Weights for each row, used to interpolate additional factors in 'predict.mca'.

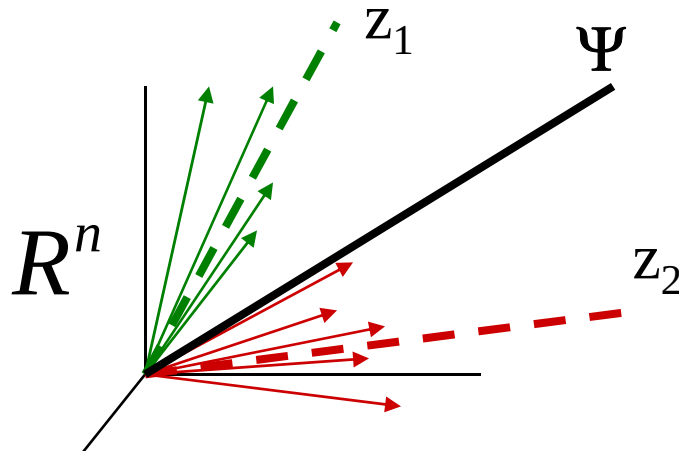
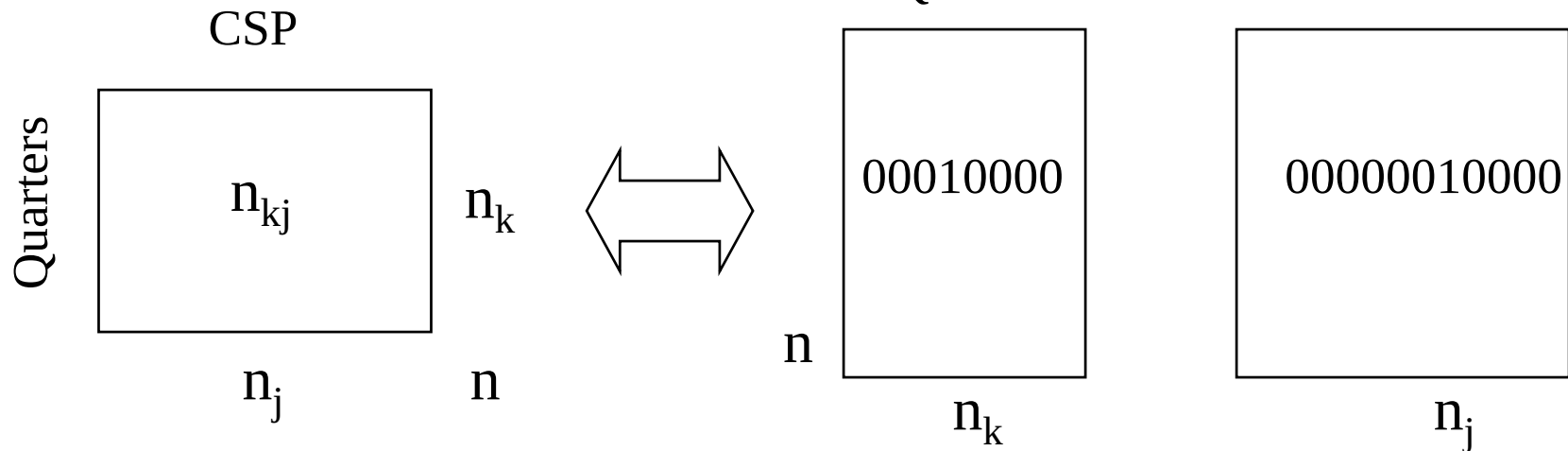
d: The singular values for the 'nf' dimensions.



Simple Correspondences Analysis

Analysis of crosstables

Jean Paul Benzecri,
Analyse des Données father



$$\text{Max cor}(\mathbf{z}_1, \mathbf{z}_2)$$

$$\mathbf{z}_1 = \sum_k u_k a_k$$

$$\mathbf{z}_2 = \sum_j v_j b_j$$

library(MASS)
corresp(x, data, ...)

Arguments

x : A two-way frequency table. Currently accepted forms are matrices, data frames ...

nf: The number of factors to be computed.
(max. value = min (nrow-1, ncol-1)).