Extensions of Principal Components Methods

D. Conti [(1)]***, K. Gibert*** [(1, 2)]

karina.gibert@upc.edu, *xavier.angerri@upc.edu*

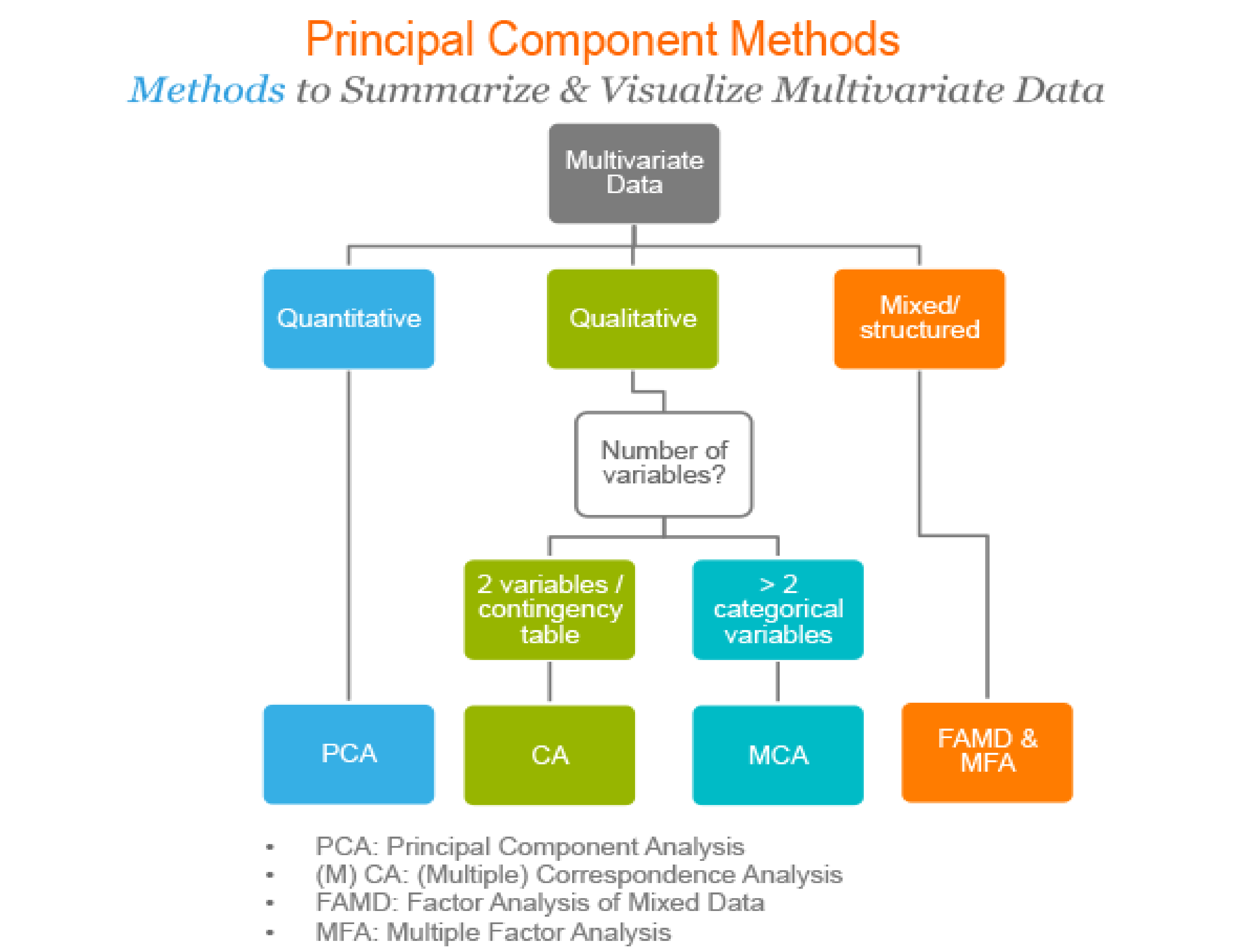
http://www.eio.upc.edu/homepages/karina

[(1)]Department d'Estadística I Investigació Operativa

(2) Knowledge Engineering and Machine Learning group @

Intelligent Data Science and Artificial Intelligence Research Center

Universitat Politècnica de Catalunya, Barcelona



Principal Component Methods

Methods to Summarize & Visualize Multivariate Data

Multivariate Data

Quantitative

Qualitative

Mixed/ structured

Number of variables?

2 variables / contingency table

> 2 categorical variables

PCA

CA

MCA

FAMD & MFA

- PCA: Principal Component Analysis
- (M) CA: (Multiple) Correspondence Analysis
- FAMD: Factor Analysis of Mixed Data
- MFA: Multiple Factor Analysis

http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/

K. Gibert, X.Angerri

Factor Analysis of Mixed Data (FAMD)

✓ Factor analysis of mixed data (FAMD), developed by (Pagès, 2004) is a principal components method that aims to analyze a data set that contains both quantitative and qualitative information. This analysis allows analyzing the similarity between individuals, taking into account a mixed set of variables. Additionally, it allows us to explore the association between all these variables, both quantitative and qualitative.

✓ In essence, the FAMD algorithm can be viewed as a combination of the PCA and the MCA. That is, it uses PCA on the quantitative variables and MCA on the qualitative variables. In the process, these variables are normalized to balance the influence of each data set.

A quick view of FAMD

Formally, the criterion maximized by the technique for a factor S can be written as:

$$\lambda_s = \sum_{k\in K} r^2\left(z_s, \nu_k\right) + \sum_{q\in Q} \eta^2\left(z_s, V_q\right)$$

Where:

- K represents the set of quantitative variables.
- Q represents the set of qualitative variables.
- r2 is the square of the correlation coefficient between vk and the factor zs of rank s.
- η2 is the square of the correlation ratio between Vq and the factor zs of rank s.
- λs is the eigenvalue of rank s.

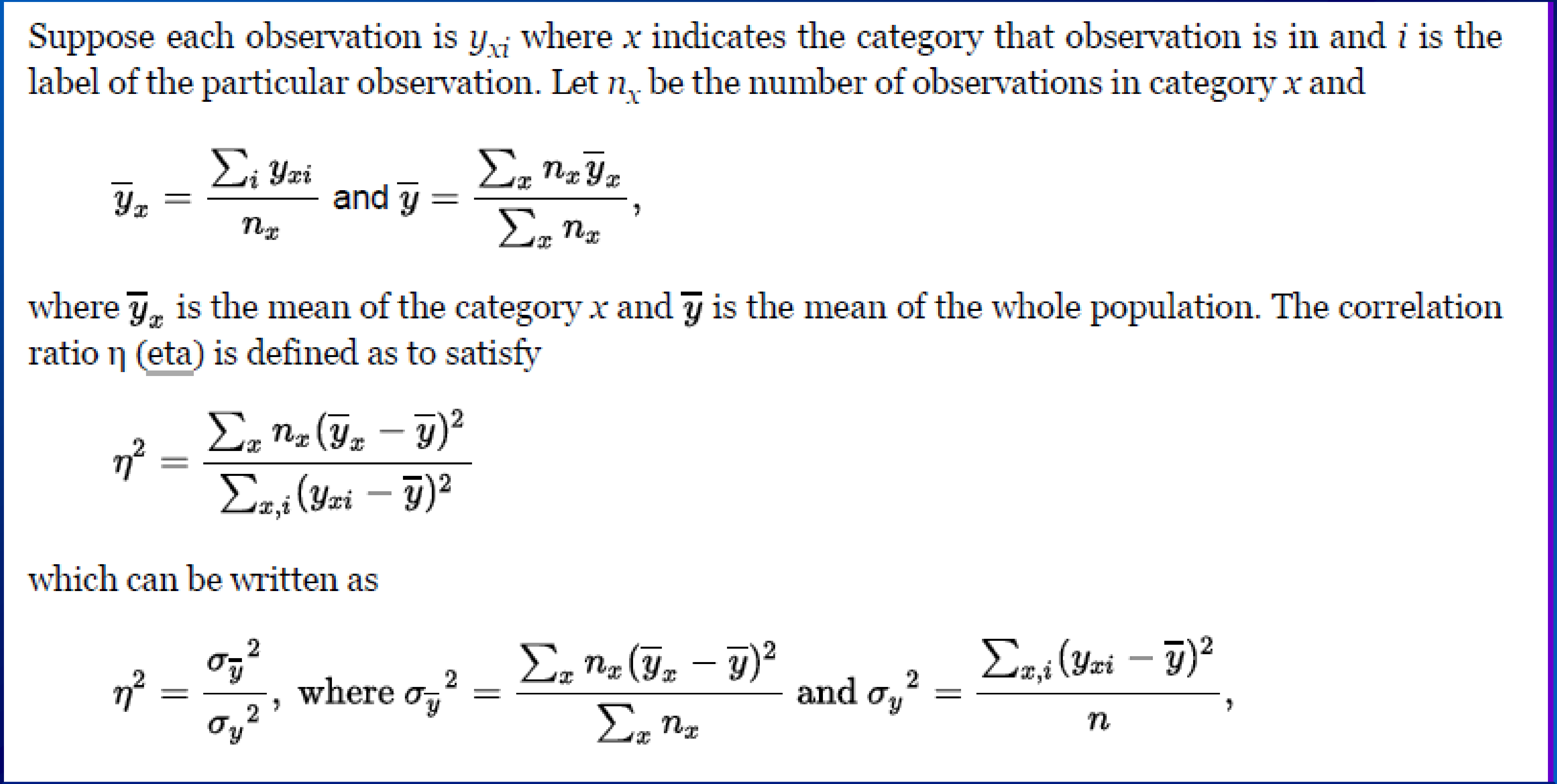
FROM PCA

$$\lambda_i = r_{1i}^2 + \cdots + r_{pi}^2 = Var(CP_i)$$

$$r_{ij} = u_{ij}\sqrt{\lambda_i}$$

$$\frac{r_{ji}^2}{r_{1i}^2 + \cdots + r_{pi}^2} = \frac{r_{ji}^2}{\lambda_i}$$

Just to remind → Correlation Ratio



Suppose each observation is y_{xi} where x indicates the category that observation is in and i is the label of the particular observation. Let n_x be the number of observations in category x and

$$\overline{y}_x = \frac{\sum_i y_{xi}}{n_x} \text{ and } \overline{y} = \frac{\sum_x n_x \overline{y}_x}{\sum_x n_x},$$

where $\overline{y}_x$ is the mean of the category x and $\overline{y}$ is the mean of the whole population. The correlation ratio η (eta) is defined as to satisfy

$$\eta^2 = \frac{\sum_x n_x (\overline{y}_x - \overline{y})^2}{\sum_{x,i} (y_{xi} - \overline{y})^2}$$

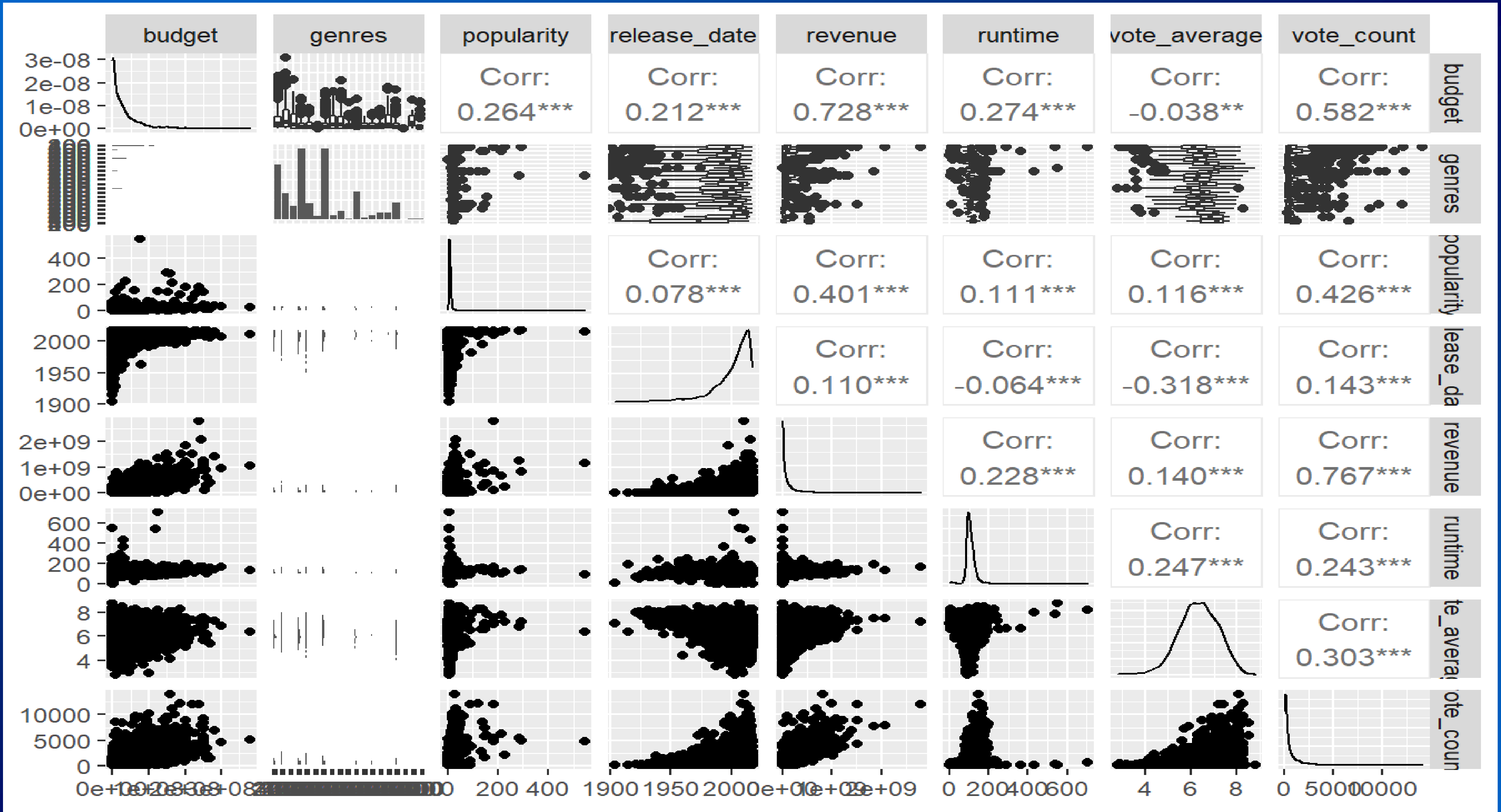
which can be written as

$$\eta^2 = \frac{{\sigma_{\overline{y}}}^2}{{\sigma_y}^2}, \text{ where } {\sigma_{\overline{y}}}^2 = \frac{\sum_x n_x (\overline{y}_x - \overline{y})^2}{\sum_x n_x} \text{ and } {\sigma_y}^2 = \frac{\sum_{x,i} (y_{xi} - \overline{y})^2}{n},$$

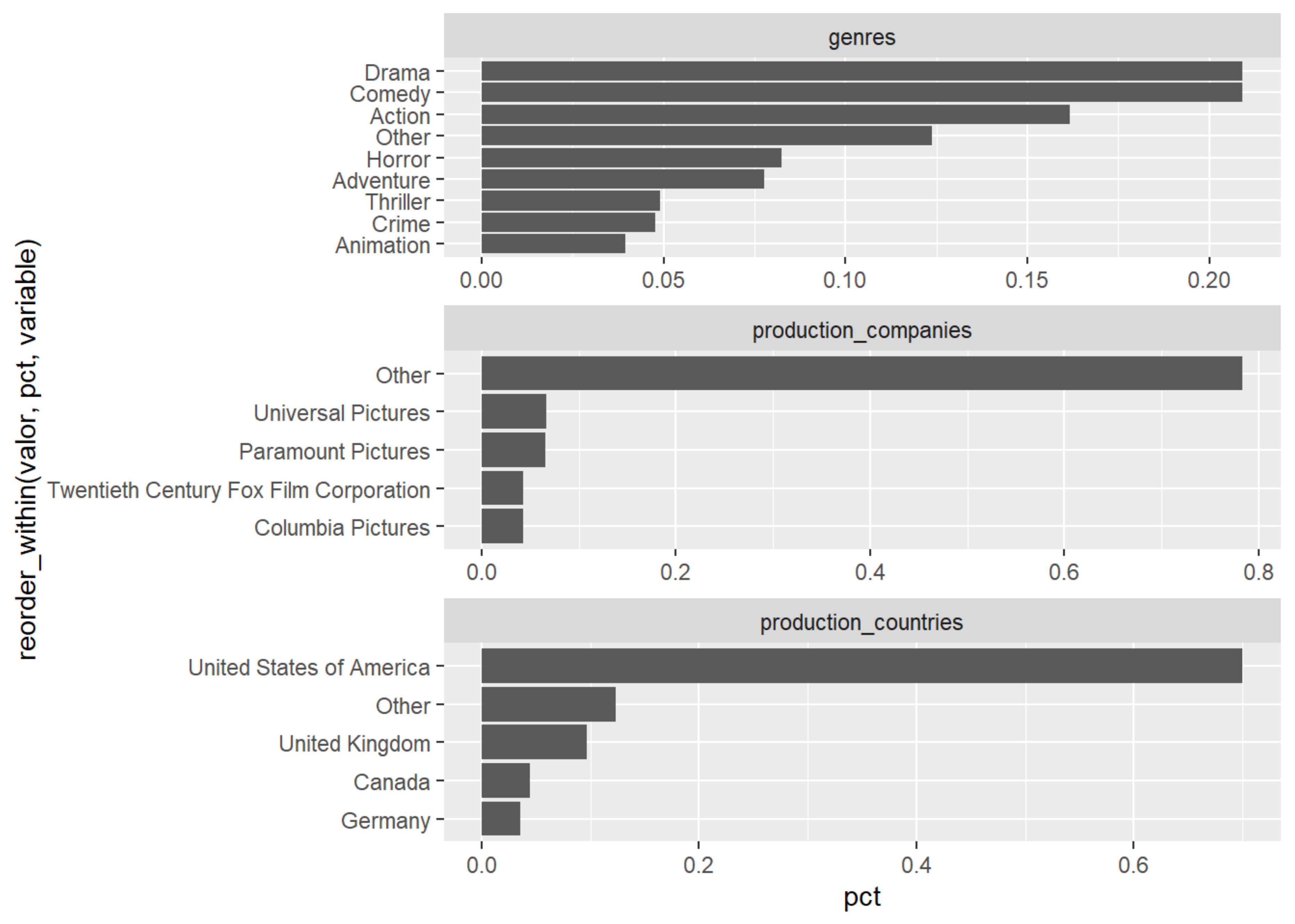
FAMD Core task

✓ Why maximize the eigenvalue of each of the factors (or new projected dimensions?

Well, because these are what determine the amount of variance explained by each of these factors



https://rpubs.com/hugoporras/desc_patrones_1


genres
Drama
Comedy
Action
Other
Horror
Adventure
Thriller
Crime
Animation
0.00
0.05
0.10
0.15
0.20
production_companies
Other
Universal Pictures
Paramount Pictures
Twentieth Century Fox Film Corporation
Columbia Pictures
0.0
0.2
0.4
0.6
0.8
production_countries
United States of America
Other
United Kingdom
Canada
Germany
0.0
0.2
0.4
0.6
pct
reorder_within(valor, pct, variable)

Pseudo-Algorithm FAMD

GOAL → Determination of the matrix X of dimensions n × k, which contains the standardized values of the components, the variance of each component and a matrix of dimensions n × k of the squared loads. These loadings are defined as the squared correlations of the quantitative variables with the PCAMIX components and as the correlation ratio for the qualitative variables.

This procedure is carried out in the following steps:

1) The real matrix Z = [Z1, Z2] of dimension n×(p1+m) is constructed where: * Z1 is the standardized version of X1 (PCA). * Z2 is the centered version of the X2 G Indicator Matrix (MCA).

The real matrix Z = [Z1, Z2] of dimension n×(p1+m) is constructed where: * Z1 is the standardized version of X1 (ACP). * Z2 is the centered version of the X2 G Indicator Matrix (ACM).

The diagonal matrix N is constructed from the weights of the rows of Z. The n rows are often weighted by 1/n.

2) The diagonal matrix M of the weights of the columns is constructed, so that the first columns p1 (corresponding to the numerical variables) are weighted by 1 (as in standard PCA) and the last m columns (corresponding to the levels of the categorical variables) are weighted by “n/ns” (as in standard ACM), where ns, s = 1,. . . , m denotes the number of observations belonging to level s. Decomposition is applied.

3) After this, the total inertia of Z with this distance and the weights 1/n is equal to p1+m−p2

K. Gibert, X.Angerri

FAMD in R

```
## ## Call:
## PCAmix(X.quanti = split$X.quanti, X.quali = split$X.quali, rename.level = TRUE,     graph = FALSE)
## Method = Principal Component of mixed data (PCAmix)
## "name" "description"
## "$eig" "eigenvalues of the principal components (PC) "
## "$ind" "results for the individuals (coord,contrib,cos2)"
## "$quanti" "results for the quantitative variables (coord,contrib,cos2)"
## "$levels" "results for the levels of the qualitative variables (coord,contrib,cos2)"
## "$quali" "results for the qualitative variables (contrib,relative contrib)"
## "$sqload" "squared loadings"
## "$coef" "coef of the linear combinations defining the PC"
```

Multiple Factor Analysis (MFA)

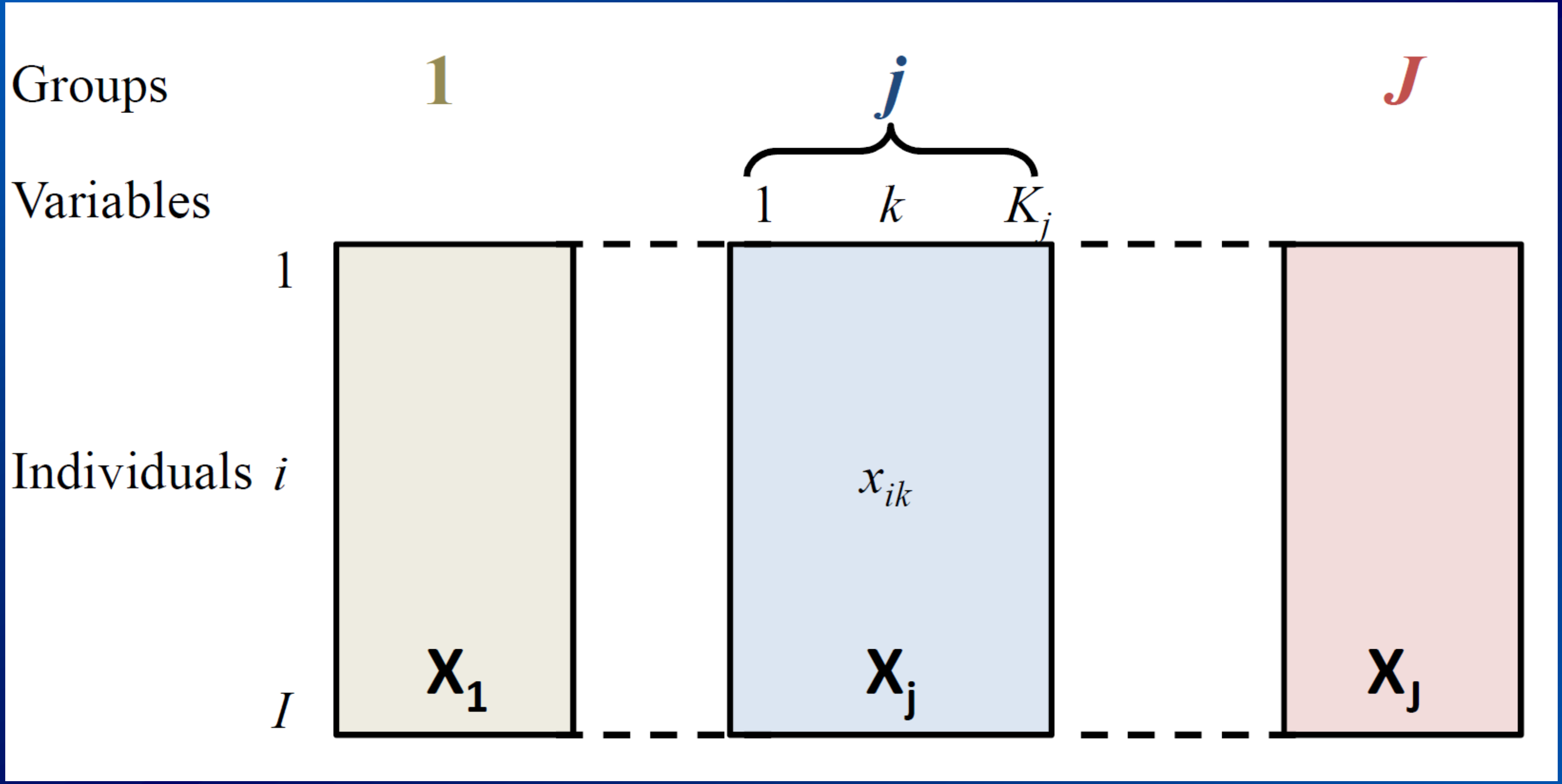
Multiple factor analysis (**MFA**) is a multivariate data analysis method for summarizing and visualizing a complex data table in which individuals are described by several sets of variables (quantitative and /or qualitative) structured into groups. It takes into account the contribution of all active groups of variables to define the distance between individuals. The number of variables in each group may differ and the nature of the variables (qualitative or quantitative) can vary from one group to the other, but the variables should be of the same nature in a given group.

MFA may be considered as a general factor analysis. Roughly, the core of MFA is based on:

- Principal component analysis (PCA) when variables are quantitative,
- Multiple correspondence analysis (MCA) when variables are qualitative.

This global analysis, where multiple sets of variables are simultaneously considered, requires to balance the influences of each set of variables. Therefore, in MFA, the variables are weighted during the analysis. Variables in the same group are normalized using the same weighting value, which can vary from one group to another. Technically, MFA assigns to each variable of group j, a weight equal to the inverse of the first eigenvalue of the analysis (PCA or MCA according to the type of variable) of the group j.

Multiple Factor Analysis (MFA)



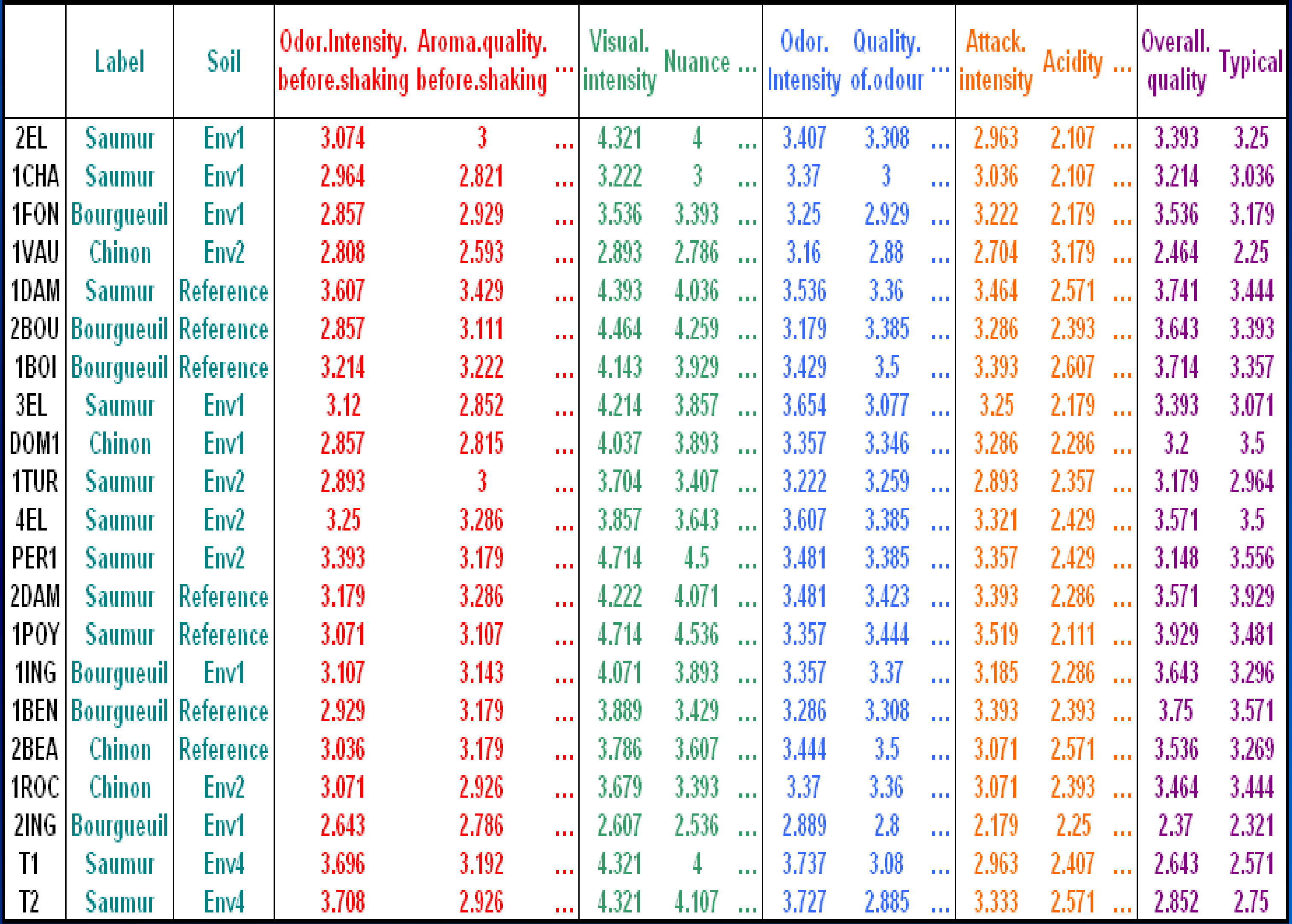
Groups of variables are quantitative and/or qualitative

Objectives :

- study the link between the sets of variables
- balance the influence of each group of variables
- give the classical graphs but also specific graphs (partial graphs)

Multiple Factor Analysis (MFA)

## [1] "Label" "Soil" ## [3] "Odor.Intensity.before.shaking" "Aroma.quality.before.shaking" ## [5] "Fruity.before.shaking" "Flower.before.shaking" ## [7] "Spice.before.shaking" "Visual.intensity" ## [9] "Nuance" "Surface.feeling" ## [11] "Odor.Intensity" "Quality.of.odour" ## [13] "Fruity" "Flower" ## [15] "Spice" "Plante" ## [17] "Phenolic" "Aroma.intensity" ## [19] "Aroma.persistency" "Aroma.quality" ## [21] "Attack.intensity" "Acidity" ## [23] "Astringency" "Alcohol" ## [25] "Balance" "Smooth" ## [27] "Bitterness" "Intensity" ## [29] "Harmony" "Overall.quality" ## [31] "Typical"



	Label	Soil	Odor.Intensity. before.shaking	Aroma.quality. before.shaking	...	Visual. intensity	Nuance	...	Odor. Intensity	Quality. of.odour	...	Attack. intensity	Acidity	...	Overall. quality	Typical
2EL	Saumur	Env1	3.074	3	...	4.321	4	...	3.407	3.308	...	2.963	2.107	...	3.393	3.25
1CHA	Saumur	Env1	2.964	2.821	...	3.222	3	...	3.37	3	...	3.036	2.107	...	3.214	3.036
1FON	Bourgueuil	Env1	2.857	2.929	...	3.536	3.393	...	3.25	2.929	...	3.222	2.179	...	3.536	3.179
1VAU	Chinon	Env2	2.808	2.593	...	2.893	2.786	...	3.16	2.88	...	2.704	3.179	...	2.464	2.25
1DAM	Saumur	Reference	3.607	3.429	...	4.393	4.036	...	3.536	3.36	...	3.464	2.571	...	3.741	3.444
2BOU	Bourgueuil	Reference	2.857	3.111	...	4.464	4.259	...	3.179	3.385	...	3.286	2.393	...	3.643	3.393
1BOI	Bourgueuil	Reference	3.214	3.222	...	4.143	3.929	...	3.429	3.5	...	3.393	2.607	...	3.714	3.357
3EL	Saumur	Env1	3.12	2.852	...	4.214	3.857	...	3.654	3.077	...	3.25	2.179	...	3.393	3.071
DOM1	Chinon	Env1	2.857	2.815	...	4.037	3.893	...	3.357	3.346	...	3.286	2.286	...	3.2	3.5
1TUR	Saumur	Env2	2.893	3	...	3.704	3.407	...	3.222	3.259	...	2.893	2.357	...	3.179	2.964
4EL	Saumur	Env2	3.25	3.286	...	3.857	3.643	...	3.607	3.385	...	3.321	2.429	...	3.571	3.5
PER1	Saumur	Env2	3.393	3.179	...	4.714	4.5	...	3.481	3.385	...	3.357	2.429	...	3.148	3.556
2DAM	Saumur	Reference	3.179	3.286	...	4.222	4.071	...	3.481	3.423	...	3.393	2.286	...	3.571	3.929
1POY	Saumur	Reference	3.071	3.107	...	4.714	4.536	...	3.357	3.444	...	3.519	2.111	...	3.929	3.481
1ING	Bourgueuil	Env1	3.107	3.143	...	4.071	3.893	...	3.357	3.37	...	3.185	2.286	...	3.643	3.296
1BEN	Bourgueuil	Reference	2.929	3.179	...	3.889	3.429	...	3.286	3.308	...	3.393	2.393	...	3.75	3.571
2BEA	Chinon	Reference	3.036	3.179	...	3.786	3.607	...	3.444	3.5	...	3.071	2.571	...	3.536	3.269
1ROC	Chinon	Env2	3.071	2.926	...	3.679	3.393	...	3.37	3.36	...	3.071	2.393	...	3.464	3.444
2ING	Bourgueuil	Env1	2.643	2.786	...	2.607	2.536	...	2.889	2.8	...	2.179	2.25	...	2.37	2.321
T1	Saumur	Env4	3.696	3.192	...	4.321	4	...	3.737	3.08	...	2.963	2.407	...	2.643	2.571
T2	Saumur	Env4	3.708	2.926	...	4.321	4.107	...	3.727	2.885	...	3.333	2.571	...	2.852	2.75

Multiple Factor Analysis (MFA)

1.First group - A group of categorical variables specifying the origin of the wines, including the variables label and soil

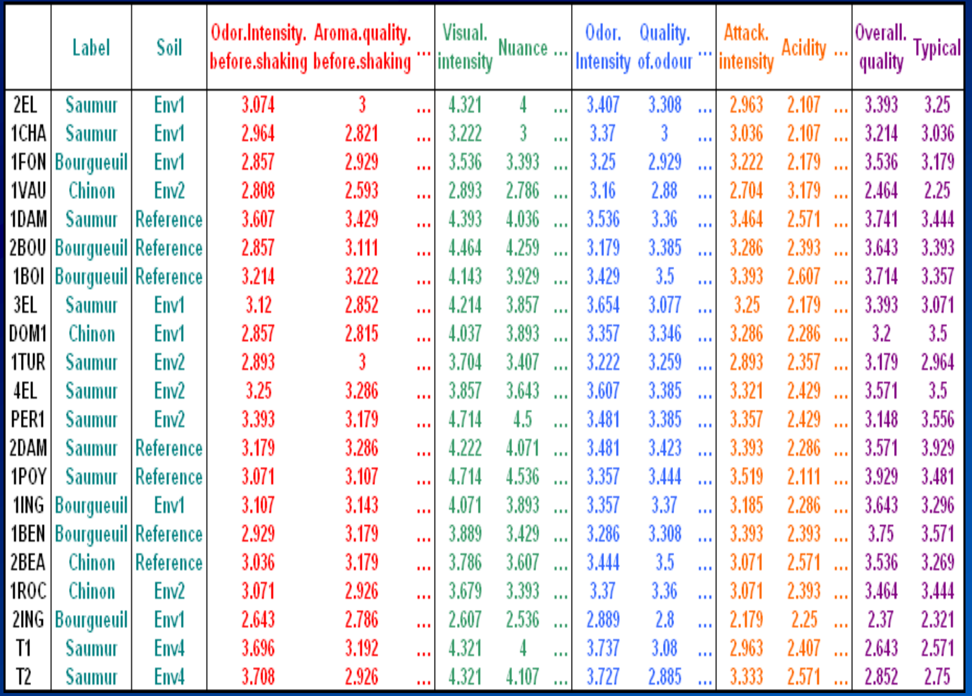
2.Second group - A group of continuous variables, describing the odor of the wines before shaking, including the variables: Odor.Intensity.before.shaking, Aroma.quality.before.shaking, Fruity.before.shaking, Flower.before.shaking and Spice.before.shaking.

3.Third group - A group of continuous variables quantifying the visual inspection of the wines, including the variables: Visual.intensity, Nuance and Surface.feeling.

4.Fourth group - A group of continuous variables concerning the odor of the wines after shaking, including the variables: Odor.Intensity, Quality.of.odour, Fruity, Flower, Spice, Plante, Phenolic, Aroma.intensity, Aroma.persistency and Aroma.quality.

5.Fith group - A group of continuous variables evaluating the taste of the wines, including the variables Attack.intensity, Acidity, Astringency, Alcohol, Balance, Smooth, Bitterness, Intensity and Harmony.

6.Sixth group - A group of continuous variables concerning the overall judgement of the wines, including the variables Overall.quality and Typical.



	Label	Soil	Odor.Intensity. before.shaking	Aroma.quality. before.shaking	...	Visual. intensity	Nuance	...	Odor. Intensity	Quality. of.odour	...	Attack. intensity	Acidity	...	Overall. quality	Typical
2EL	Saumur	Env1	3.074	3	...	4.321	4	...	3.407	3.308	...	2.963	2.107	...	3.393	3.25
1CHA	Saumur	Env1	2.964	2.821	...	3.222	3	...	3.37	3	...	3.036	2.107	...	3.214	3.036
1FON	Bourgueuil	Env1	2.857	2.929	...	3.536	3.393	...	3.25	2.929	...	3.222	2.179	...	3.536	3.179
1VAU	Chinon	Env2	2.808	2.593	...	2.893	2.786	...	3.16	2.88	...	2.704	3.179	...	2.464	2.25
1DAM	Saumur	Reference	3.607	3.429	...	4.393	4.036	...	3.536	3.36	...	3.464	2.571	...	3.741	3.444
2BOU	Bourgueuil	Reference	2.857	3.111	...	4.464	4.259	...	3.179	3.385	...	3.286	2.393	...	3.643	3.393
1BOI	Bourgueuil	Reference	3.214	3.222	...	4.143	3.929	...	3.429	3.5	...	3.393	2.607	...	3.714	3.357
3EL	Saumur	Env1	3.12	2.852	...	4.214	3.857	...	3.654	3.077	...	3.25	2.179	...	3.393	3.071
DOM1	Chinon	Env1	2.857	2.815	...	4.037	3.893	...	3.357	3.346	...	3.286	2.286	...	3.2	3.5
1TUR	Saumur	Env2	2.893	3	...	3.704	3.407	...	3.222	3.259	...	2.893	2.357	...	3.179	2.964
4EL	Saumur	Env2	3.25	3.286	...	3.857	3.643	...	3.607	3.385	...	3.321	2.429	...	3.571	3.5
PER1	Saumur	Env2	3.393	3.179	...	4.714	4.5	...	3.481	3.385	...	3.357	2.429	...	3.148	3.556
2DAM	Saumur	Reference	3.179	3.286	...	4.222	4.071	...	3.481	3.423	...	3.393	2.286	...	3.571	3.929
1POY	Saumur	Reference	3.071	3.107	...	4.714	4.536	...	3.357	3.444	...	3.519	2.111	...	3.929	3.481
1ING	Bourgueuil	Env1	3.107	3.143	...	4.071	3.893	...	3.357	3.37	...	3.185	2.286	...	3.643	3.296
1BEN	Bourgueuil	Reference	2.929	3.179	...	3.889	3.429	...	3.286	3.308	...	3.393	2.393	...	3.75	3.571
2BEA	Chinon	Reference	3.036	3.179	...	3.786	3.607	...	3.444	3.5	...	3.071	2.571	...	3.536	3.269
1ROC	Chinon	Env2	3.071	2.926	...	3.679	3.393	...	3.37	3.36	...	3.071	2.393	...	3.464	3.444
2ING	Bourgueuil	Env1	2.643	2.786	...	2.607	2.536	...	2.889	2.8	...	2.179	2.25	...	2.37	2.321
T1	Saumur	Env4	3.696	3.192	...	4.321	4	...	3.737	3.08	...	2.963	2.407	...	2.643	2.571
T2	Saumur	Env4	3.708	2.926	...	4.321	4.107	...	3.727	2.885	...	3.333	2.571	...	2.852	2.75

Multiple Factor Analysis (MFA)

MFA is a weighted PCA:

- calculate the1st eigenvalue λ_1 of the jth group of variables($j=1,...,J$)
- do an over all PCA on the weighted table:

X_j corresponds to the jth normalized or standardized table

$$\left[\frac{X_1}{\sqrt{\lambda_1^1}};\frac{X_2}{\sqrt{\lambda_1^2}};\ldots;\frac{X_J}{\sqrt{\lambda_1^J}}\right]$$

Multiple Factor Analysis (MFA)

In PCA (reminder) : $\arg\max\limits_{v_1 \in \mathbb{R}^I} \sum_{k=1}^{K} cov^2(x_{.k}, v_1)$

In MFA :

$$\arg\max_{v_1 \in \mathbb{R}^I} \sum_{j=1}^{J} \sum_{k \in K_j} cov^2\left(\frac{x_{.k}}{\sqrt{\lambda_1^j}}, v_1\right) = \arg\max_{v_1 \in \mathbb{R}^I} \sum_{j=1}^{J} \underbrace{\frac{1}{\lambda_1^j} \sum_{k \in K_j} cov^2(x_{.k}, v_1)}_{\mathcal{L}_g(K_j, v_1)}$$

Multiple Factor Analysis (MFA)

Study the similarity between individuals with respect to the whole set of variables AND the relationships between variables

Take the group structure into account

- Study the overall similarities and differences between groups (and the specific features of each group)
- Study the similarities and differences between groups from an individual’s point of view
- Compare the characteristics of individuals from the separate analyses

⇒ Balance the influence of all of the groups in the analysis and explore more information with Partial Analysis

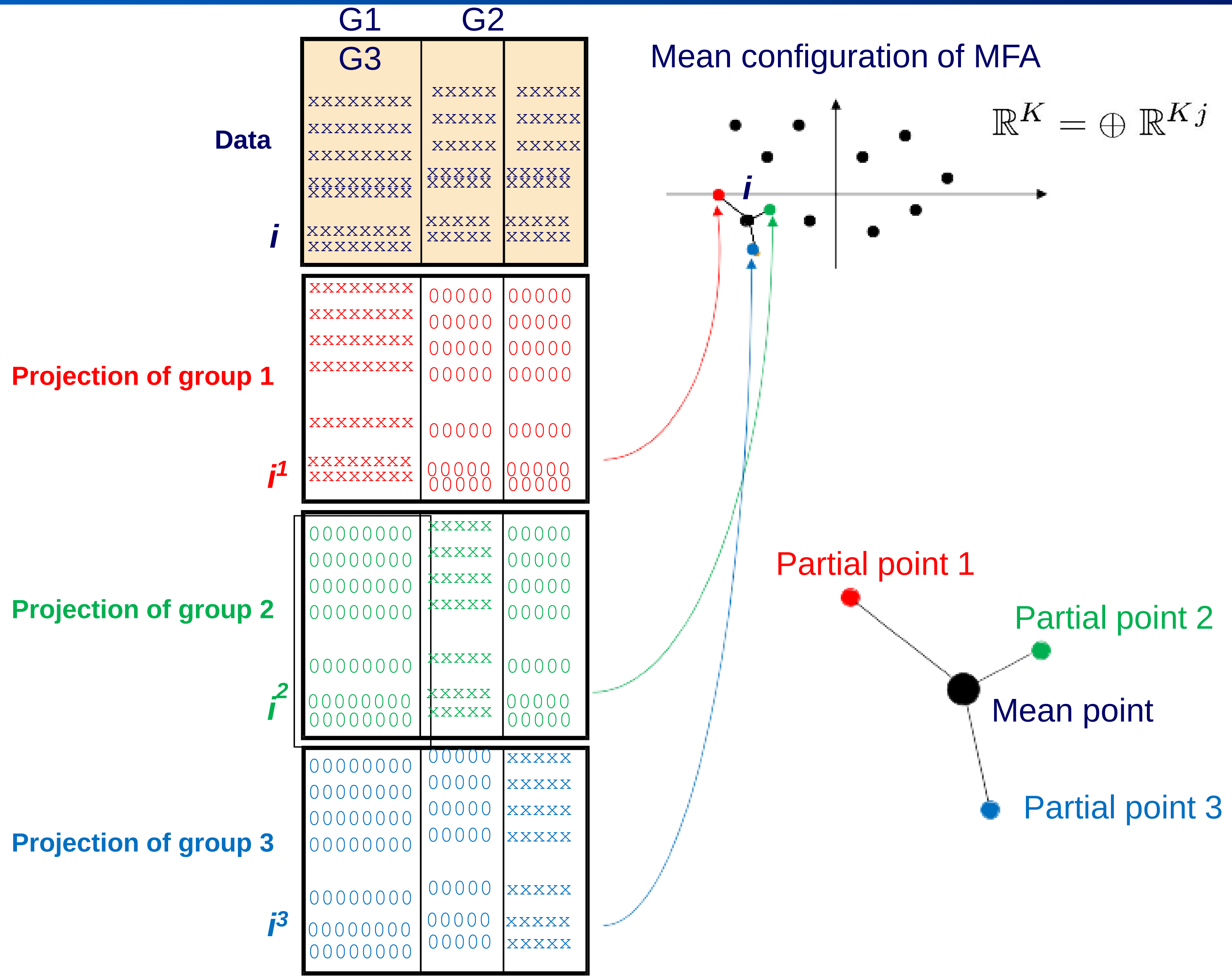
Multiple Factor Analysis (MFA)

⇒ Same plots as in PCA

- Study similarities between individuals in terms of the set of variables
- Study relationships between variables
- Characterize individuals in terms of variables

⇒ Same outputs (coordinates, cosine, contributions)
⇒ Add individuals and variables (quantitative, qualitative) as supplementary information

Projections of partial points



Data

Projection of group 1

Projection of group 2

Projection of group 3

Mean configuration of MFA

$$\mathbb{R}^K = \oplus \mathbb{R}^{K j}$$

Partial point 1

Partial point 2

Mean point

Partial point 3

Multiple Factor Analysis (MFA)

R code

The function *MFA*()[*FactoMiner* package] can be used. A simplified format is :

MFA (database, group, type = rep("s",length(group)), ind.sup = NULL, name.group = NULL, num.group.sup = NULL, graph = TRUE)

- base : a data frame with n rows (individuals) and p columns (variables)
- group: a vector with the number of variables in each group.
- type: the type of variables in each group. By default, all variables are quantitative and scaled to unit variance. Allowed values include:
    - “c” or “s” for quantitative variables. If “s”, the variables are scaled to unit variance.
    - “n” for categorical variables.
    - “f” for frequencies (from a contingency tables).
- ind.sup: a vector indicating the indexes of the supplementary individuals.
- name.group: a vector containing the name of the groups (by default, NULL and the group are named group.1, group.2 and so on).
- num.group.sup: the indexes of the illustrative groups (by default, NULL and no group are illustrative).
- graph : a logical value. If TRUE a graph is displayed.

Multiple Factor Analysis (MFA)

http://www.sthda.com/english/articles/22-principal-component-methods/73-mfa-in-r-using-factominer-quick-scripts-and-videos/

Homework