Factorial Methods

K. Gibert[1,2]

[1]*Department of Statistics and Operation Research*

[2] *Knowledge Engineering and Machine Learning group*
Universitat Politècnica de Catalunya, Barcelona

Master Oficial en Enginyeria Informàtica
Universitat Politècnica de Catalunya

Factorial Methods

- Find the isomorph transformation from original space

  keeps the adjacency relationships among variables

- Results expressed in a ficticious space

- Might produce interpretation problems

- Methods
  - PCA (Principal components analysis)
  - Simple correspondence analysis
  - Multiple correspondence analysis

Factorial Methods

- Output: K factors rotating original X variables
- Factors: Linear combinations of original variables

Several uses:

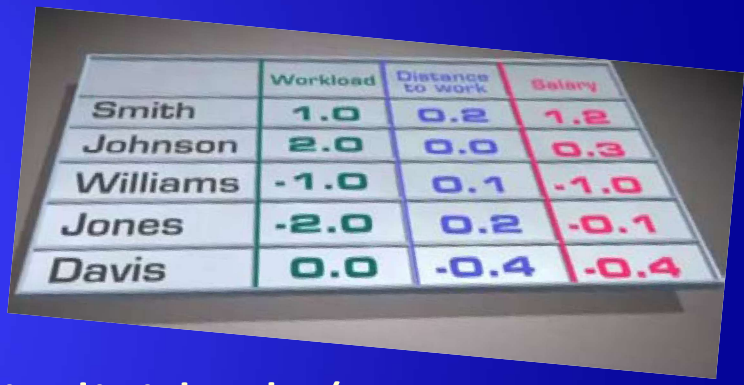
- As an associative data mining method to analyze relationships among variables
  Project variables and modalities and find associations

- As a preprocessing method for elicitation of latent variables
  Project active and illustrative variables/individuals on first/second factorial plan and interpret factors (find latent variables)

- As a preprocessing method for multidimensionality reduction

Factorial Methods

- Principal Components Analysis
  - Only numerical variables
  - Find the most informative projection planes (factorial planes, maximize projected inertia)



	Workload	Distance to work	Salary
Smith	1.0	0.2	1.2
Johnson	2.0	0.0	0.3
Williams	-1.0	0.1	-1.0
Jones	-2.0	0.2	-0.1
Davis	0.0	-0.4	-0.4

Given <X,M,D>

- A data matrix X (nxp) centered
- A matrix of individuals weights D (nxn)
- Assume euclidean metrics to compare individuals (M= $\mathbb{I}_p$)

Si les dades estan centrades l'angle entre dues variables projectades coincideix amb la correlació entre elles

Matrix $M^{1/2}$ X'DX$M^{1/2}$

- Product of data with the two metrics
- Simetric,
- Semidefinite
- Catches relationships and opositions of data

Factorial Methods

Given triplet <X,M,D>, diagonalize $M^{1/2}\,X'DXM^{1/2}$

Data	Factorial Method	X	M	D
Continuous variables	PCA	Centered data matrix	$\mathbb{I}_p$	$\mathbb{I}_n$
Contingency table (n_{ij})	CA	$F=(n_{ij}/n_{i})$	$\mathrm{diag}(1/f_j)$	$\mathrm{diag}\,(f_i)$
		$G=(n_{ij}/n_{j})$	$\mathrm{diag}(1/f_i)$	$\mathrm{diag}\,(f_j)$
Categorical variables	MCA	$F=(f_{ij}/(f_i/\sqrt{f_j}))$	$\mathbb{I}_p$	$\mathrm{diag}\,(f_i)$
		Burt table	$\mathbb{I}_{n+p}$	$\mathrm{diag}(n_{ij})$

Factorial Methods

- Principal Components Analysis

$M^{1/2}X'DXM^{1/2}$ catches well the data structures

$Rang(M^{1/2}X'DXM^{1/2}) = r$, $r = rang(X)$ r positive vaps and p-r null vaps

$Trace(M^{1/2}X'DXM^{1/2}) = \sum_{\alpha=1}^{r} \lambda_\alpha$ (λ_α, the r non null vaps)

$M = \mathbb{I}_p$: $M^{1/2}X'DXM^{1/2} = X'DX$

X centered and D diagonal : $X'DX = Cov(X)$

X standardized and D diagonal : $X'DX = Corr(X)$

(preferred, big variabilities do not dominate analysis)

Build variances and covariances matrix: $X'DX$

Diagonalize $X'DX$ (i.e. solving the equation) $X'DXu = \lambda u$

provides eigen values λ_α and

eigenvectors $u_\alpha = (u_{\alpha 1} \ldots u_{\alpha p})$

Factorial Methods

- Principal Components Analysis

Diagonalize X'DX (i.e. solving the equation) $X'DXu = \lambda u$ (1)

$\det(X'DX-\lambda)=0$ (find roots of characteristic polynomial)

provides eigen values λ_α ($\alpha = 1:r$, $r=rang(X)$)

substituting in (1) provides eigenvectors $u_\alpha = (u_{\alpha 1} \ldots u_{\alpha p})$

$u^{-1}X'DXu = \lambda$ *is a diagonal matrix*

(X'DX becomes diagonal when pre/post multiplied by u)

$u^{-1}=u'$ *in orthonormal basis:* $u'X'DXu = \lambda$

X'DX decompose in a product by a diagonal matrix $X'DX = u\lambda u'$

$X'DX = u\lambda u' = u\lambda^{1/2}\lambda^{1/2} u' = u\lambda^{1/2} I \lambda^{1/2} u' = u\lambda^{1/2} u'u \lambda^{1/2} u' = A^{1/2} A^{1/2}$

X'DX decompose in a product of something by itself (A square root)

Trace(X'DX)=Trace(λ*)* (property of diagonalization)

Factorial Methods

- Given <X,M,D>

Diagonalize correlations matrix (with normalized data X'DX)

Get r eigen values λ_α *and sort decreasingly*

$$\{\lambda_\alpha\}_{\alpha=1:r} \quad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r$$

Corresponding eigenvectors $u_\alpha = (u_{\alpha 1} \ldots u_{\alpha p})$

$$|u_\alpha| = 1$$

$$u_\alpha u_{\alpha'} = 0$$

$\{u_\alpha\}_{\alpha=1:r}$ *orthonormal base for individuals*

The subspace generated by $\{u_\alpha\}_{\alpha=1:r}$ *is the same as the subspace generated by the rows of X*

Factorial Methods

- Given <X,M,D>

Diagonalize Correlations matrix X'DX

Get r eigen values λ_α *and sort decreasingly*

$$\{\lambda_\alpha\}_{\alpha=1:r} \quad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r$$

Corresponding eigenvectors $u_\alpha = (u_{\alpha 1} \ldots u_{\alpha p})$

for $M = \mathbb{I}_p$: $u^*_\alpha = u_\alpha$; *for* $M \neq \mathbb{I}_p$: $u^*_\alpha = M^{-1/2} u_\alpha$

$\{u^*_\alpha\}_{\alpha=1:r}$ *orthonormal base for individuals*

u^*_α *are the principal factors of X : good rotation directions*

$U^* = ([u^*_1] [u^*_2] \ldots [u^*_r])$ *is the basis for the projection space*

Factorial Methods

- Given <X,M,D>

In general *Diagonalize* $M^{1/2}X'DXM^{1/2}$

Get r eigen values λ_α *and sort decreasingly (vaps are conserved!!!!)*

$$\{\lambda_\alpha\}_{\alpha=1:r} \quad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r$$

Corresponding eigenvectors $u^*_\alpha = (u^*_{\alpha 1} \ldots u^*_{\alpha p})$

by algebraic properties, u^*_α *can be found from* u^*

$$u^*_\alpha = M^{-1/2} u_\alpha$$

$\{u^*_\alpha\}_{\alpha=1:r}$ *orthonormal base for individuals*

$$|u^*_\alpha|_M = 1 : \quad u^{*\prime}_\alpha M u^*_\alpha = u'_\alpha M^{-1/2} M M^{-1/2} u_\alpha = 1$$

$$u^*_\alpha M u^*_{\alpha'} = 0: \quad u^*_\alpha M u^*_{\alpha'} = u'_\alpha M^{-1/2} M M^{-1/2} u_{\alpha'} = 0$$

Subspace generated by $\{u^*_\alpha\}_{\alpha=1:r}$ = *Subspace generated by X rows*

Factorial Methods

- Given <X,M,D>

Diagonalize Correlations matrix X'DX

Get r eigen values λ_α *and sort decreasingly*

$$\{\lambda_\alpha\}_{\alpha=1:r} \quad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r$$

Corresponding eigenvectors $u_\alpha = (u_{\alpha 1} \ldots u_{\alpha p})$

for $M = \mathbb{I}_p$: $u^*_\alpha = u_\alpha$; *for* $M \neq \mathbb{I}_p$: $u^*_\alpha = M^{-1/2} u_\alpha$

$\{u^*_\alpha\}_{\alpha=1:r}$ *orthonormal base for individuals*

u^*_α *are the principal factors of X : good rotation directions*

$U^* = ([u^*_1] [u^*_2] \ldots [u^*_r])$ *is the basis for the projection space*

How is i expressed in rotated space?

Factorial Methods

- Given <X,M,D>

Can we find coordinates in rotated space from original ones?

The projection matrix $P = U^*_k U^{*\prime}_k M$

Projection of a single individual: $Pr(i) = U^*_k U^{*\prime}_k M x_i$

Projection of all individuals: $Pr(X) = U^*_k U^{*\prime}_k M X'$

Get a matrix with projections in ROWS: $Pr(X)' = XMU^*_k U^{*\prime}_k$

Projections expressed in original vectorial space

The best possible projection over k dimensions

Factorial Methods

- Given <X,M,D>

Matrix $XMU^*_k U^{*\prime}_k$ *provides the best possible k-projection of X*

Silver-Smidth norm: $||X||^2_{MD} = \sum_{\alpha=1}^{r} \lambda_\alpha$

Measures variability, information contained in X

Property: $||XMU^*_k U^{*\prime}_k||^2_{MD} = ||X||^2_{MD}$

Any other k-projection of X

- Provides smallest values of Silver-Smidth norm
- Has less variability
- Keeps smallest information from X

Factorial Methods

- Given <X,M,D>

Diagonalize correlations matrix (with normalized data)

eigenvectors $u_\alpha = (u_{\alpha 1} \ldots u_{\alpha p})$ *(direction of factor α, α =1:p)*

$u_{\alpha p}$: contribution of variable p to the factor α

$(u_1 \ldots\ldots\ldots u_k)$ *ortonormal*

eigen values λ_k *(quantity of information converved by factor k)*

(Projected inertia)

$\{\lambda_\alpha\}_{\alpha=1:r} \quad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots\ldots \geq \lambda_r$

$\Sigma_{\forall \alpha} \lambda_\alpha$ = Total inertia of X (information in data)

Close objects project close
proximity linked with association

Factorial Methods

- Given <X,M,D>

eigenvectors $u_\alpha = (u_{\alpha 1} \ldots u_{\alpha p})$ *(direction of factor k)*

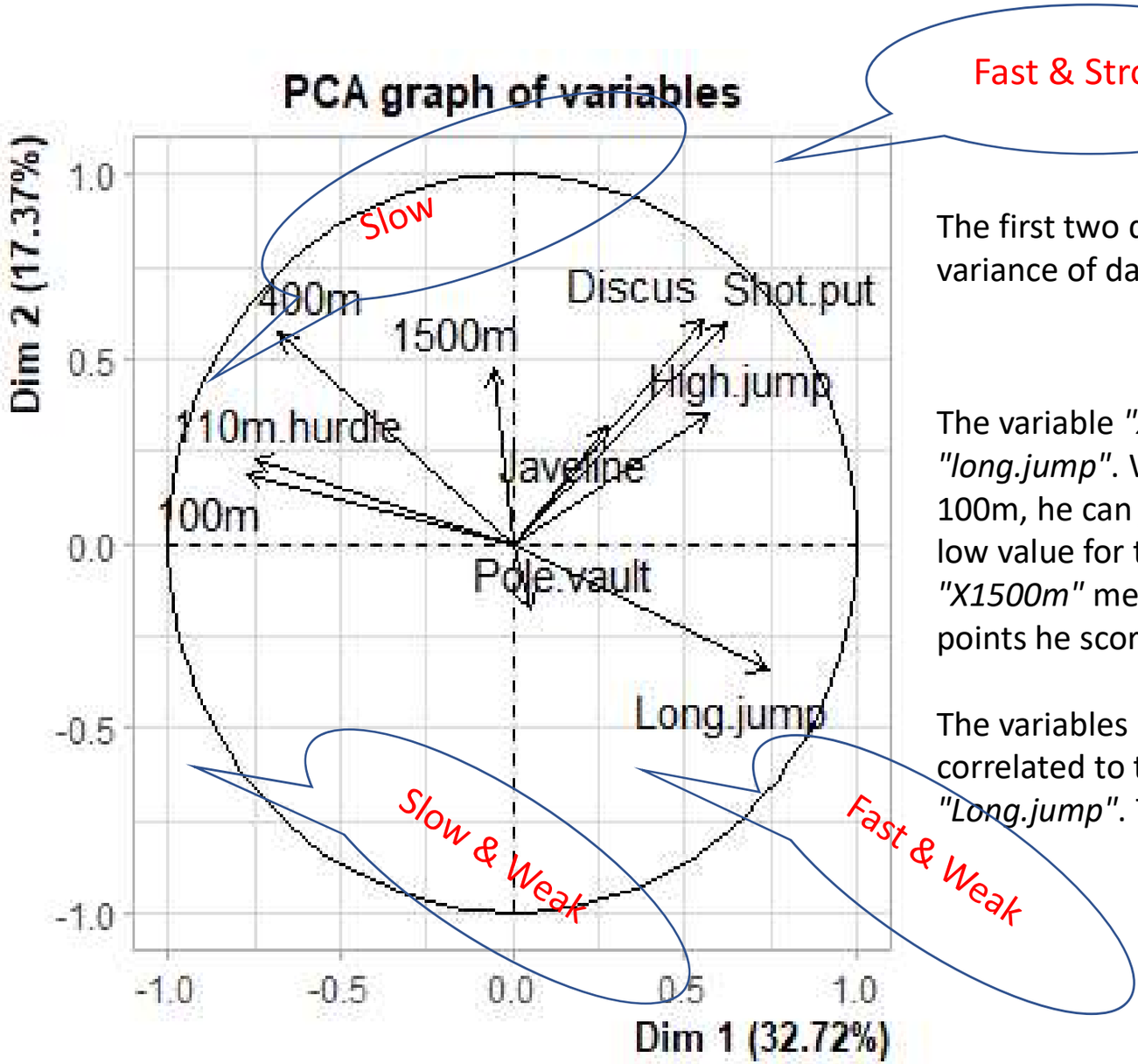
$u_{\alpha p}$: contribution of variable p to the factor α

eigen values λ_α *(quantity of information conserved by factor* α*)*

(Projected inertia)

$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r$

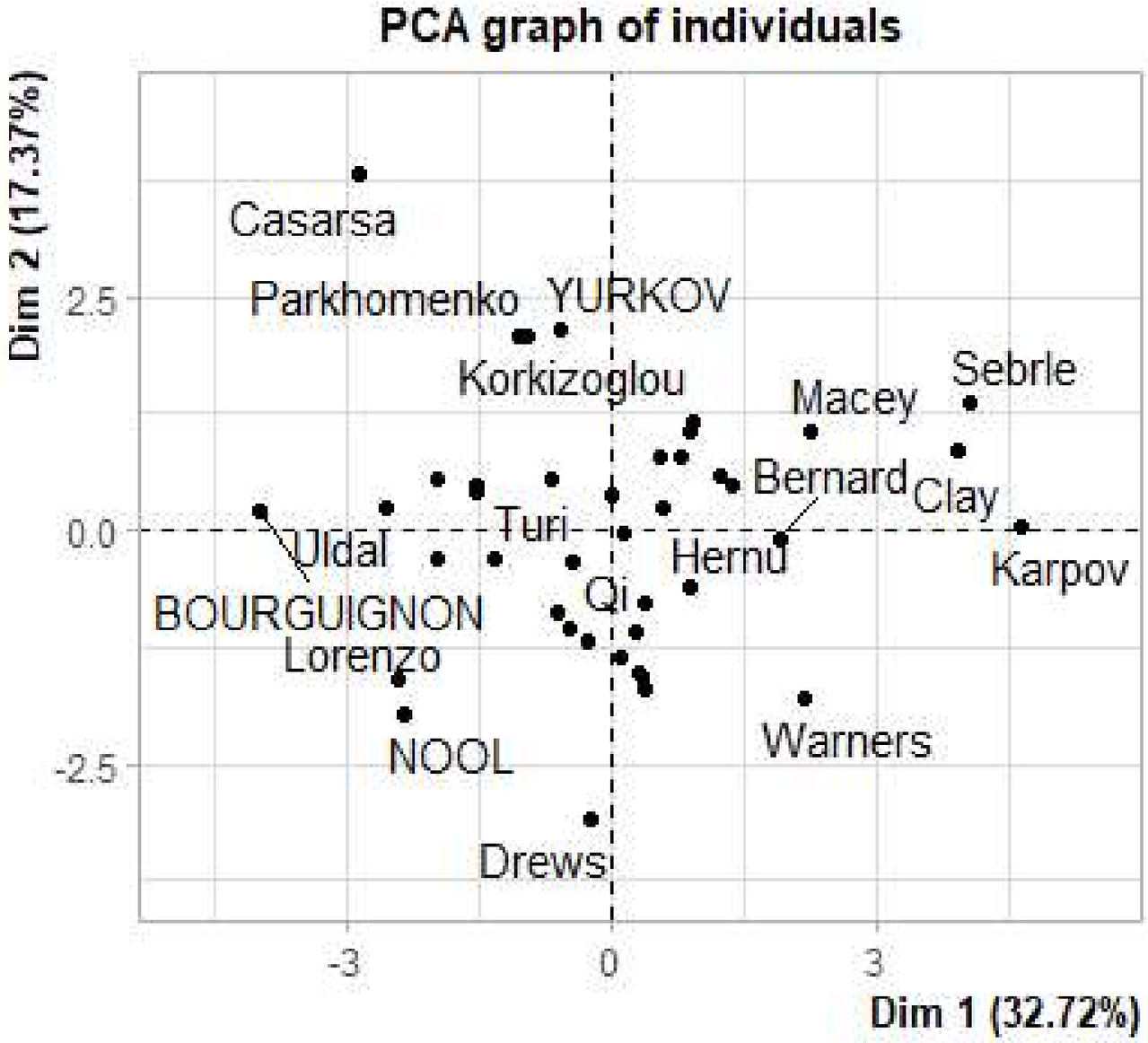
$\Sigma_{\forall \alpha} \lambda_\alpha$ = Total inertia of X (information of data)



The first two dimensions resume 50% of the total inertia (the inertia is the total variance of dataset *i.e.* the trace of the correlation matrix).

The variable *"X100m"* is correlated negatively to the variable *"long.jump"*. When an ahtlete performs a short time when running 100m, he can jump a big distance. Here one has to be careful because a low value for the variables *"X100m"*, *"X400m"*, *"X110m.hurdle"* and *"X1500m"* means a high score: the shorter an athlete runs, the more points he scores.

The variables *"Discus"*, *"Shot.put"* and *"High.jump"* are not much correlated to the variables *"X100m"*, *"X400m"*, *"X110m.hurdle"* and *"Long.jump"*. This means that strength is not much correlated to speed.



Sebrle???

Casarsa??

Nool???

Warners???

KARPOV ?? BOURGUIGNON?

Cheat Sheet – PCA Dimensionality Reduction

What is PCA?

- Based on the dataset find a new set of orthogonal feature vectors in such a way that the data spread is maximum in the direction of the feature vector (or dimension)
- Rates the feature vector in the decreasing order of data spread (or variance)
- The datapoints have maximum variance in the first feature vector, and minimum variance in the last feature vector
- The variance of the datapoints in the direction of feature vector can be termed as a measure of information in that direction.

Steps

1. Standardize the datapoints
2. Find the covariance matrix from the given datapoints
3. Carry out eigen-value decomposition of the covariance matrix
4. Sort the eigenvalues and eigenvectors

$$X_{new} = \frac{X - mean(X)}{std(X)}$$

$$C[i,j] = cov(x_i, x_j)$$

$$C = V\Sigma V^{-1}$$

$$\Sigma_{sort} = sort(\Sigma) \quad V_{sort} = sort(V, \Sigma_{sort})$$

Dimensionality Reduction with PCA

- Keep the first m out of n feature vectors rated by PCA. These m vectors will be the best m vectors preserving the maximum information that could have been preserved with m vectors on the given dataset

Steps:

1. Carry out steps 1-4 from above
2. Keep first m feature vectors from the sorted eigenvector matrix
3. Transform the data for the new basis (feature vectors)
4. The importance of the feature vector is proportional to the magnitude of the eigen value

$$V_{reduced} = V[:, 0:m]$$

$$X_{reduced} = X_{new} \times V_{reduced}$$

Figure 1

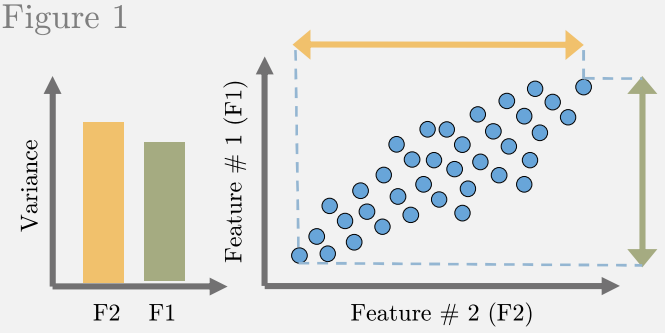


Figure 2

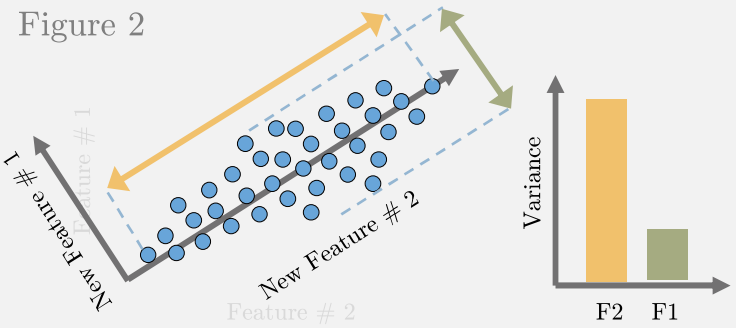


Figure 3

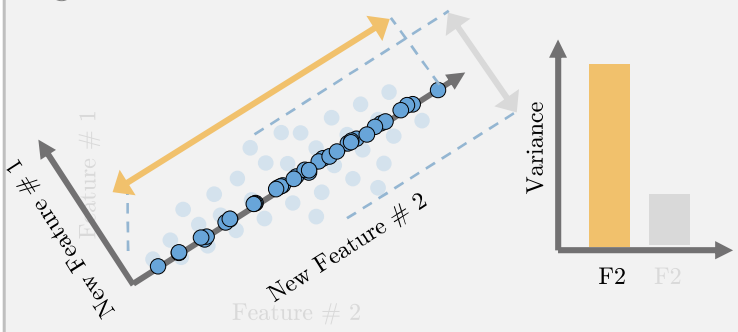


Figure 1: Datapoints with feature vectors as x and y-axis

Figure 2: The cartesian coordinate system is rotated to maximize the standard deviation along any one axis (new feature # 2)

Figure 3: Remove the feature vector with minimum standard deviation of datapoints (new feature # 1) and project the data on new feature # 2

MCA

- Multiple Correspondence Analysis
  - Only qualitative variables
  - Find the most informative projection planes (factorial planes, maximize projected inertia)
- Given <X,M,D>
  - A data matrix X (nxp) the logic table
  - A matrix of individuals weights D (nxn)
  - Assume chi2 metrics to compare individuals (M)

MCA

- Given <X,M,D>

Same algorithm as PCA

  - Matrix $M^{1/2} X'DXM^{1/2}$
  - Catches relationships and oppositions of data

Diagonalize matrix $M^{1/2} X'DXM^{1/2}$

  - Get r eigen values λ_α and sort decreasingly

$$\{\lambda_1\}_{\alpha=1:r} \qquad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots\ldots \geq \lambda_r$$

  - Corresponding eigenvectors $u_\alpha = (u_{\alpha 1} \ldots\ldots u_{\alpha p})$

$$|u_\alpha| = 1$$

$$u_\alpha u_{\alpha'} = 0$$

$\{u_\alpha\}_{\alpha=1:r}$ orthonormal base for individuals

5

MCA

- Given <X,M,D>

  Diagonalize matrix $M^{1/2} X'DXM^{1/2}$

  - The subspace generated by $\{u_\alpha\}_{\alpha=1:r}$ is the same as the subspace generated by the rows of X

  $u^*_\alpha = M^{-1/2} u_\alpha$ are the principal factors of X :

  - good rotation directions

  $U^* = ([u^*_1] [u^*_2] \ldots [u^*_r])$ is the basis for the projection space

6

MCA

- Two possibilities:
  - Analysis of Logic Table (X logic table)
  - Analysis of Burt Table (X Burt table)

7

MCA
Analysis of Logic Table

- Given <X,M,D>
- *X* =Z

  Z: logic matrix (complete disjunctive form)

- $M^{1/2} X'DXM^{1/2}$ *is sparse (high useless dimensions)*
- *Centroids of modalities ARE NOT CDG of their individuals (systematic bias of 1/sqrt(λ))*
- *Traditional dual analysis*

Systematic bias

```
Mca0<- MCA(X, method="Indicator")    ;  FactoMineR
```


8

MCA

- Original data set (only qualitative variables)

χ_{nxQ} =

n	χ_1 factor A	χ_2 factor B	χ_3 factor C	χ_4 factor D	χ_5 factor E
1	A2	B4	C3	D3	E4
2	A2	B4	C3	D2	E4
3	A3	B2	C2	D3	E4
4	A3	B9	C2	D2	E2
5	A5	B1	C2	D2	E3
6	A2	B4	C3	D4	E2
7	A1	B3	C2	D3	E2
8	A2	B5	C3	D4	E1
9	A2	B4	C2	D3	E4
1	A2	B4	C1	D2	E4
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
2000	A5	B2	C2	D2	E3

n=2000
Q=5

9

MCA

Logic Table - Complete Disjunctive Form CDF

$Z = [Z_1, Z_2, \ldots, Z_Q]$, $Z_q, \forall q \in [1:Q]$, qualitative var

J_q = number of categories for Z_q

Each χ_q coded into the binary variables Z_1, Z_q

	χ_1
	tor A
	A2
	A2
3	A3
4	A3
5	A5
6	A2
7	A1
8	A2
9	A2
1	A2
...	...
...	...
...	...
...	...
2000	A5

$$Z_{n \times \sum_{q=1}^{Q} J_q} =$$

Z1				
A1	A2	A3	A4	A5
0	1	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	1	0	0
0	0	0	0	1
0	1	0	0	0
1	0	0	0	0
0	1	0	0	0
0	1	0	0	0
0	1	0	0	0
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

...

Z_1
n=2000
J_1=5

n=2000
Q=8

MCA

- Methodology

1. Project centroids of all qualitative variables
2. Use one color for each variable
3. Add legend to the map
4. Link centroids of same variable with lines for ordinal variables

5. Evenctually project individuals

12

MCA

- Analysis of the categories

1. Two categories close if
   - Tend to be simultaneously present (absent) for large number of individuals
   - We say that categories are associated

13

MCA

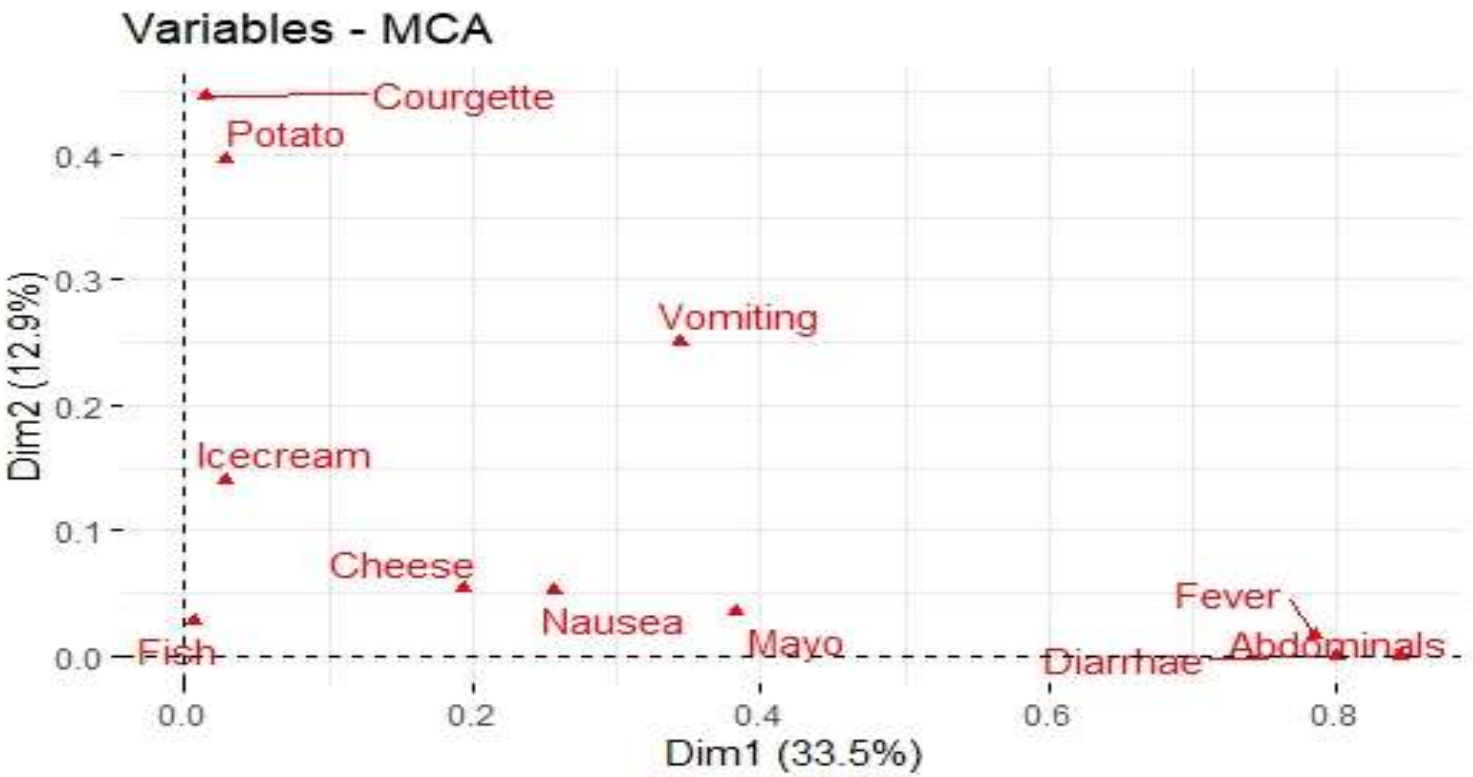
- Analysis relationships on individuals

1. Typology of the individuals
2. Close individuals: those who share many categories

- Analysis of interactions among variables

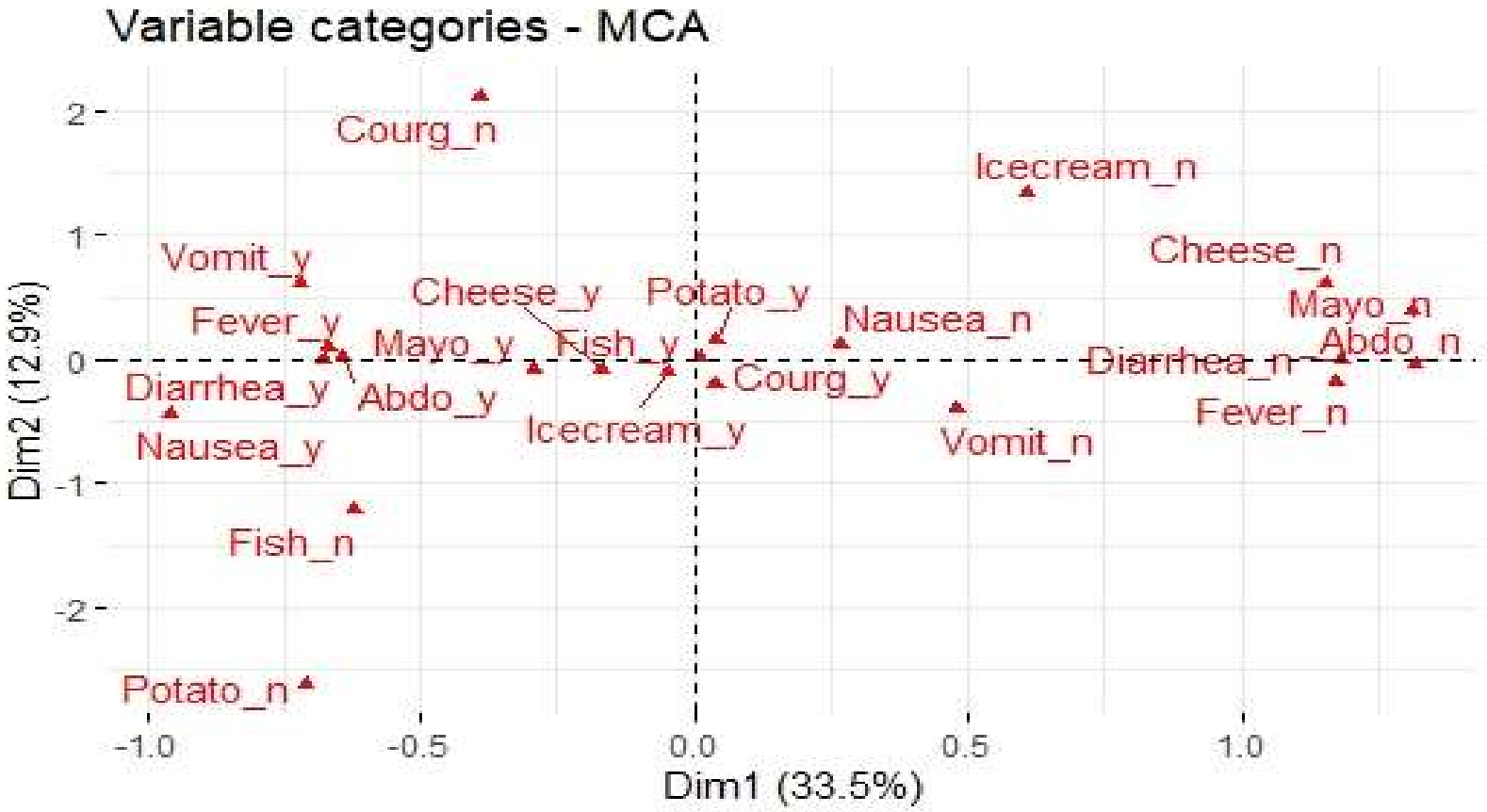
1. Relationship among the variables through the relationships among their categories

14



•The plot identifies variables that are the most correlated with each dimension. The squared correlations between variables and the dimensions are used as coordinates.

•It can be seen that, the variables Diarrhae, Abdominals and Fever are the most correlated with dimension 1. Similarly, the variables Courgette and Potato are the most correlated with dimension 2.



The plot above shows the relationships between variable categories. It can be interpreted as follow:

•Variable categories with a similar profile are grouped together.

•Negatively correlated variable categories are positioned on opposite sides of the plot origin (opposed quadrants).

•The distance between category points and the origin measures the quality of the variable category on the factor map. Category points that are away from the origin are well represented on the factor map.

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*

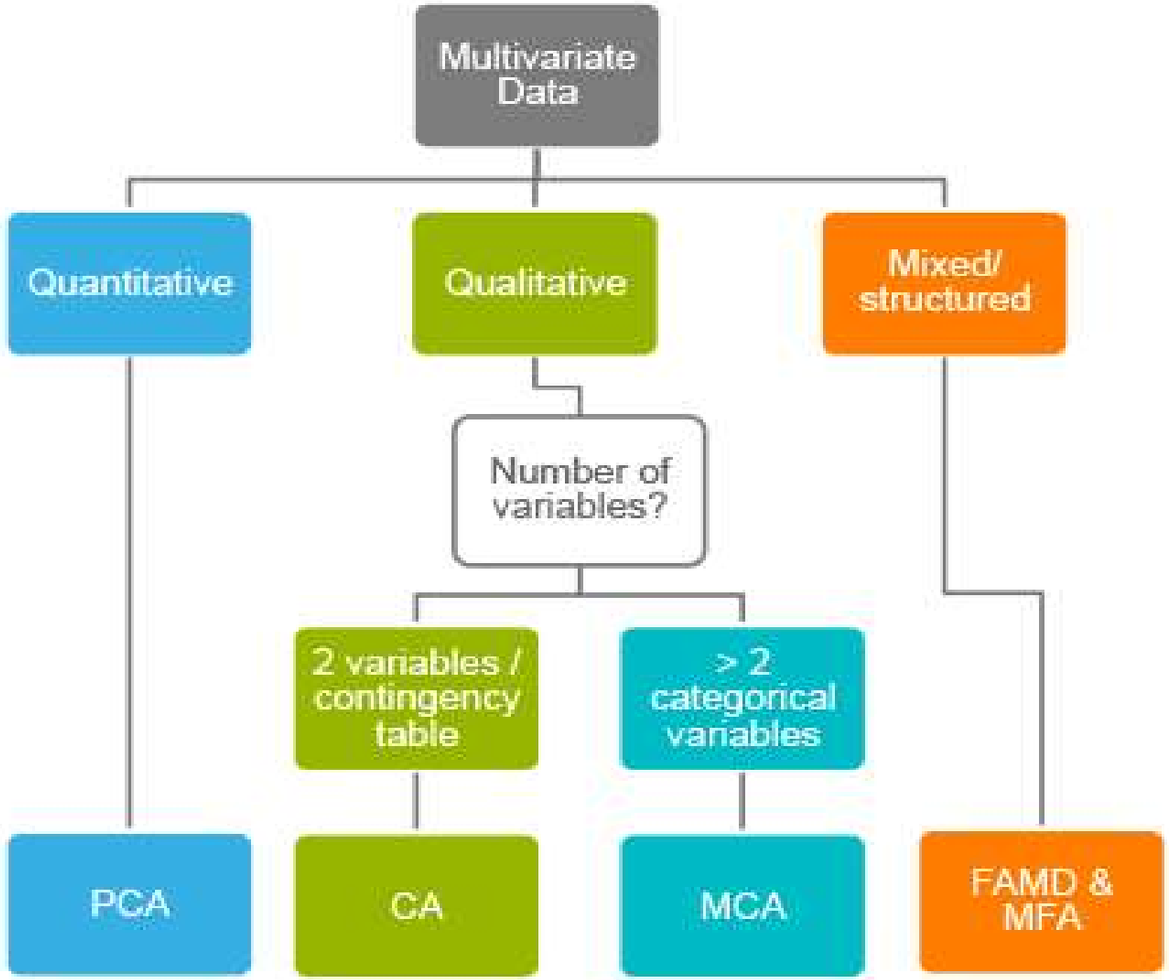
(2) Knowledge Engineering and Machine Learning group @

Intelligent Data Science and Artificial Intelligence Research Center

Universitat Politècnica de Catalunya, Barcelona

Principal Component Methods

Methods to Summarize & Visualize Multivariate Data



- 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.

K. Gibert, X.Angerri

A quick view of FAMD

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

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

FROM PCA

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

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.

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

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

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

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

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

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

which can be written as

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

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:

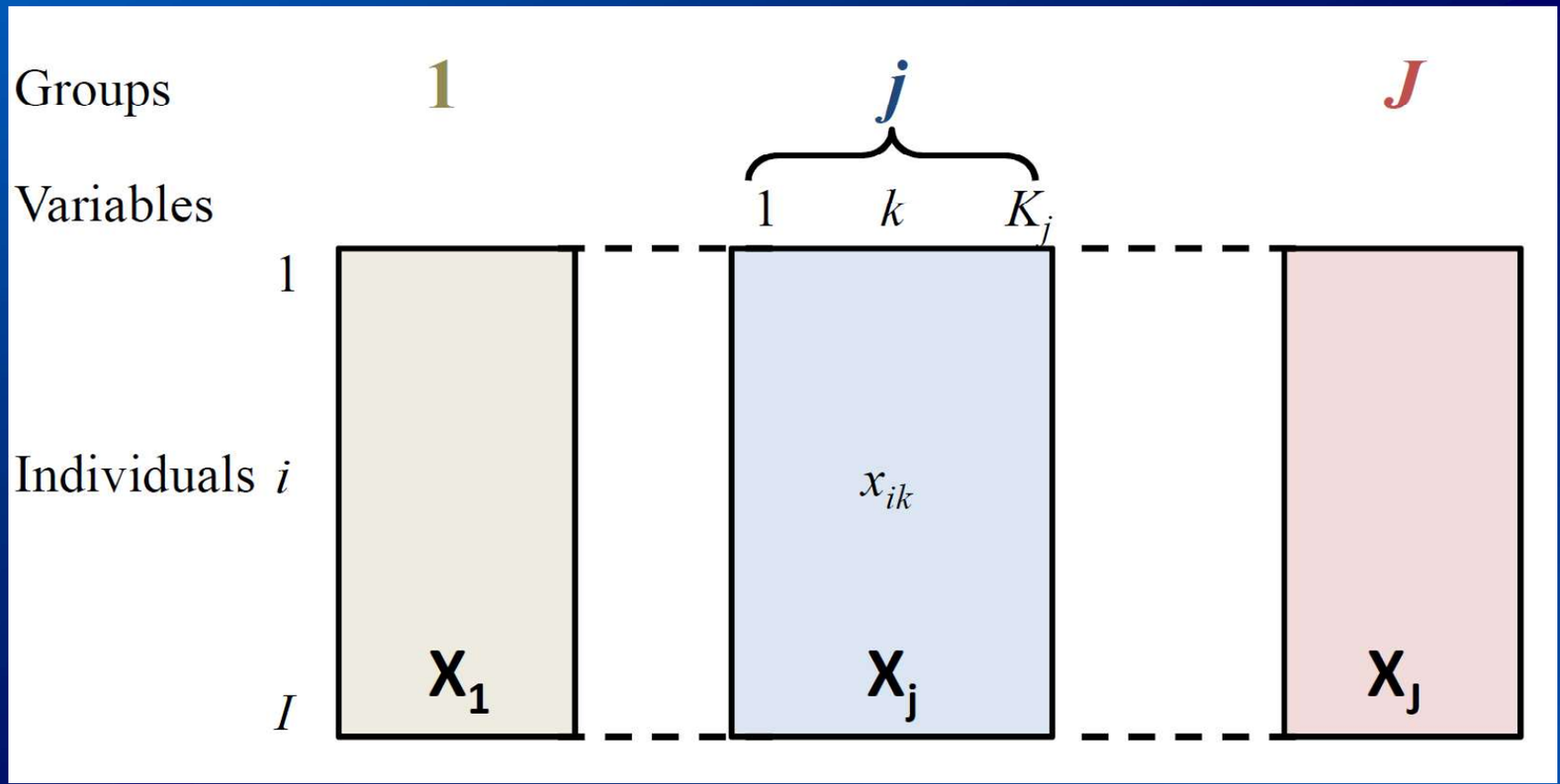
•Principal component analysis (PCA) when variables are quantitative,

•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.

K. Gibert, X.Angerri

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)

K. Gibert, X.Angerri

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

In MFA :

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

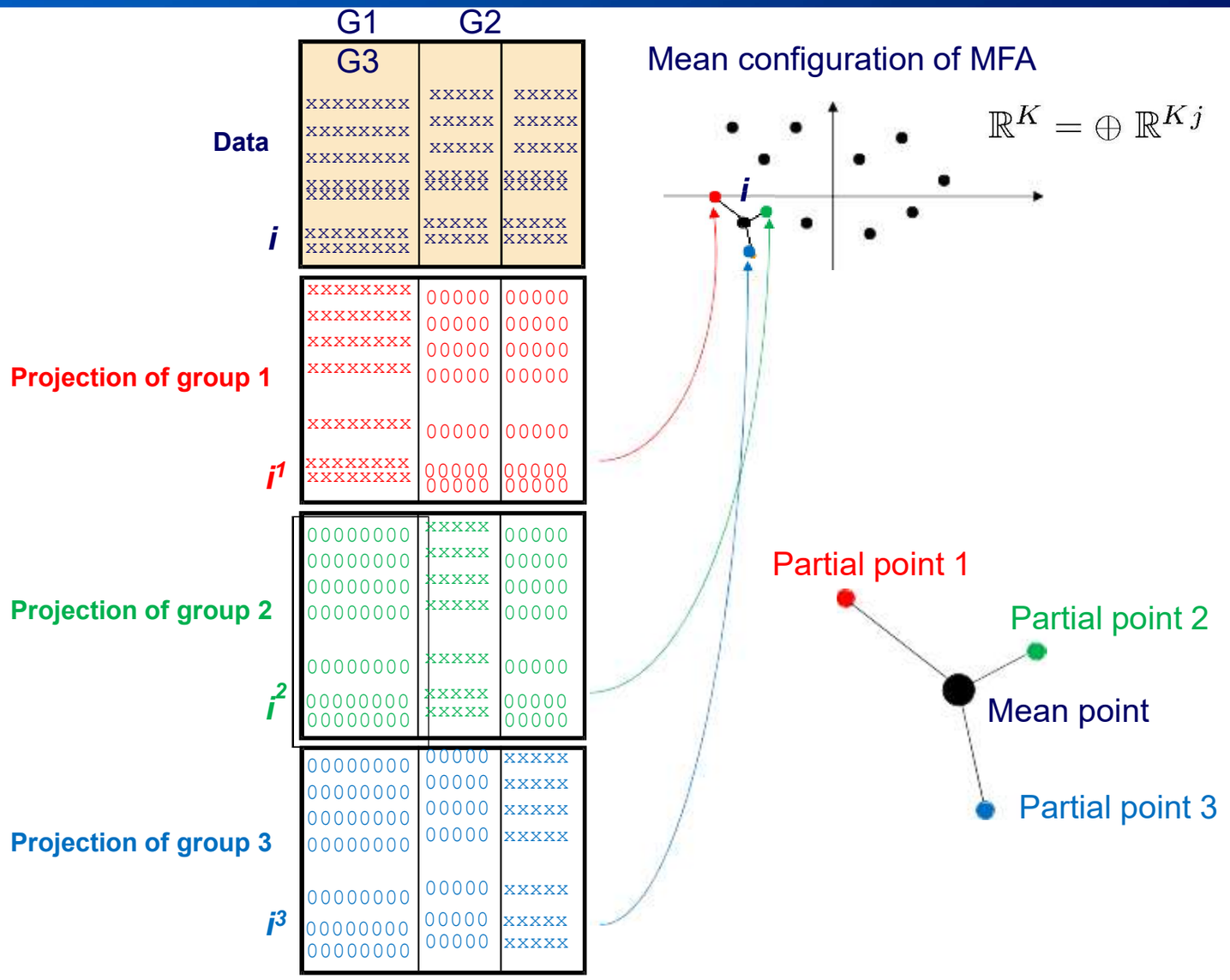
Multiple Factor Analysis (MFA)

⇒ 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



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.

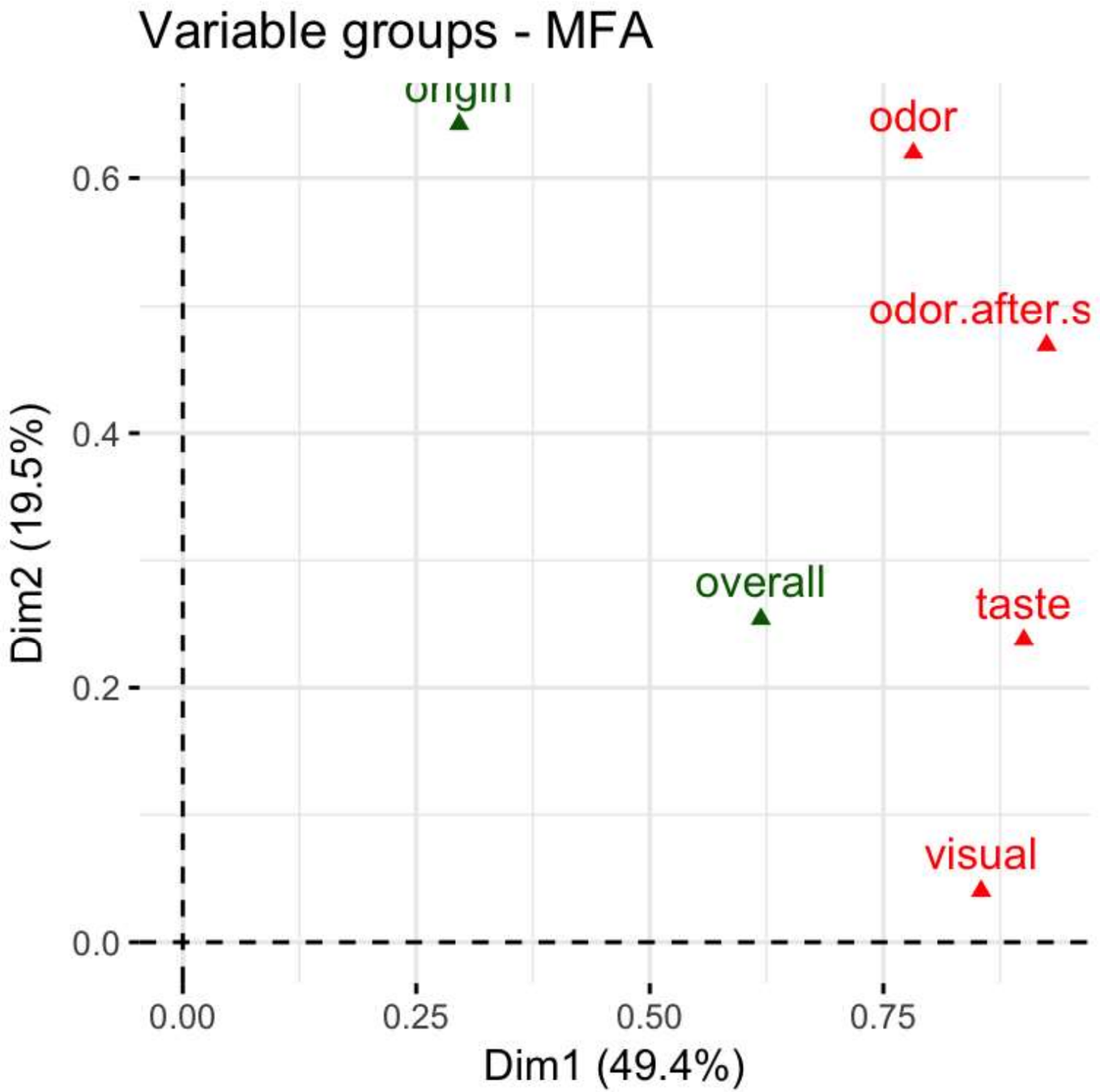
MFA

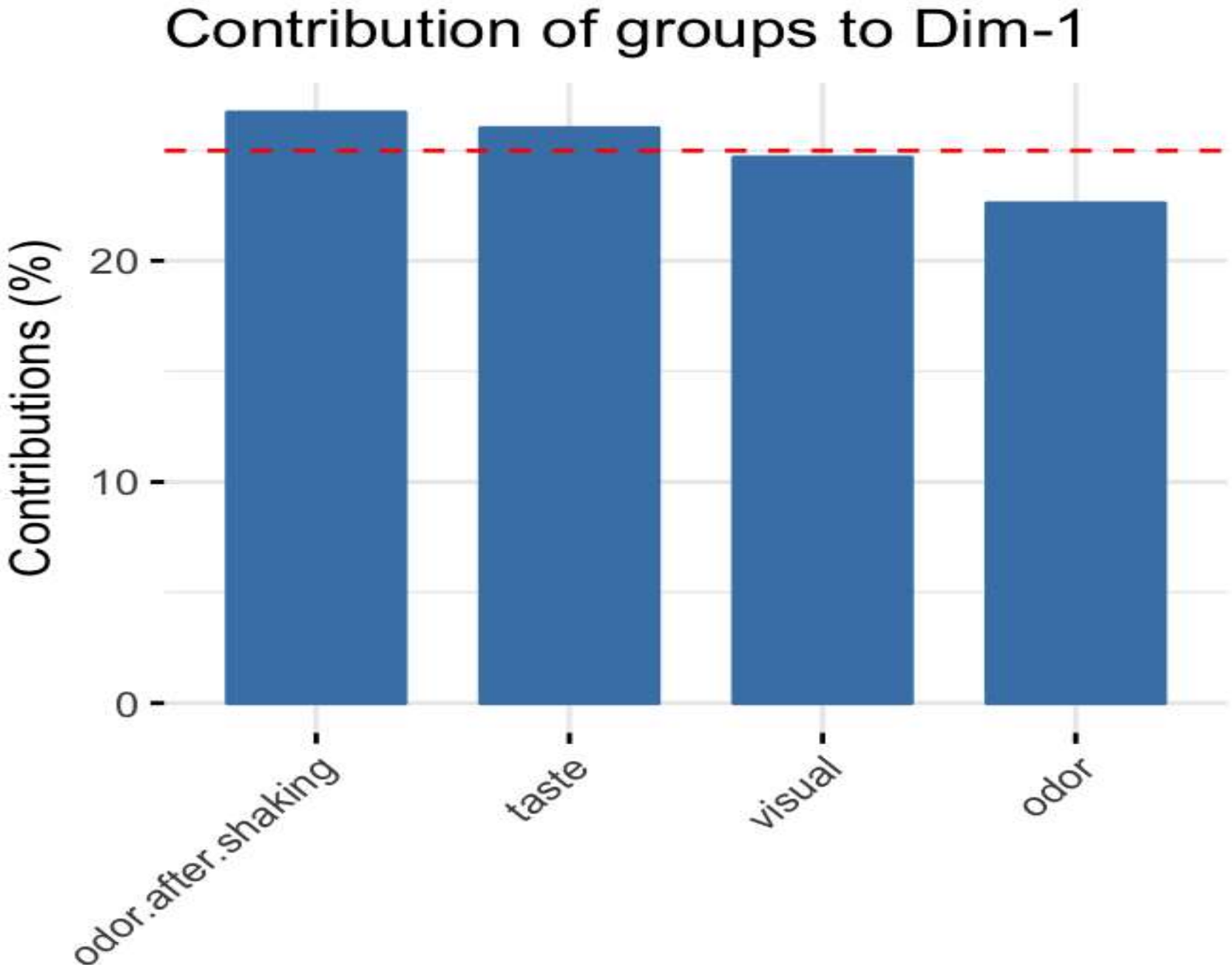
Multiple factor analysis (**MFA**) is a multivariate data analysis method for summarizing and visualizing a complex data table in which individuals are described by several sets of variables (quantitative and /or qualitative) structured into groups. It takes into account the contribution of all active 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.

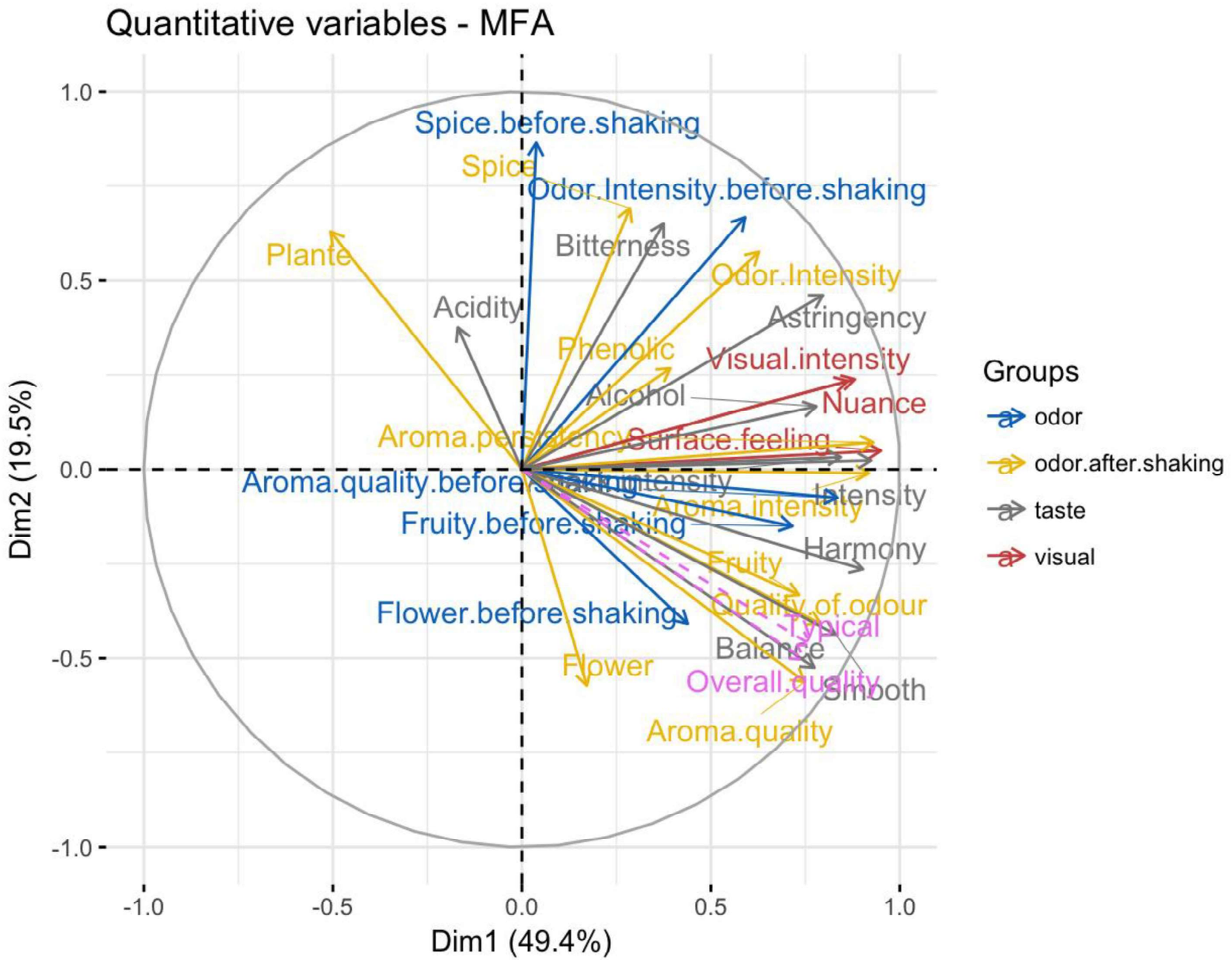
MFA may be considered as a general factor analysis. Roughly, the core of MFA is based on:

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

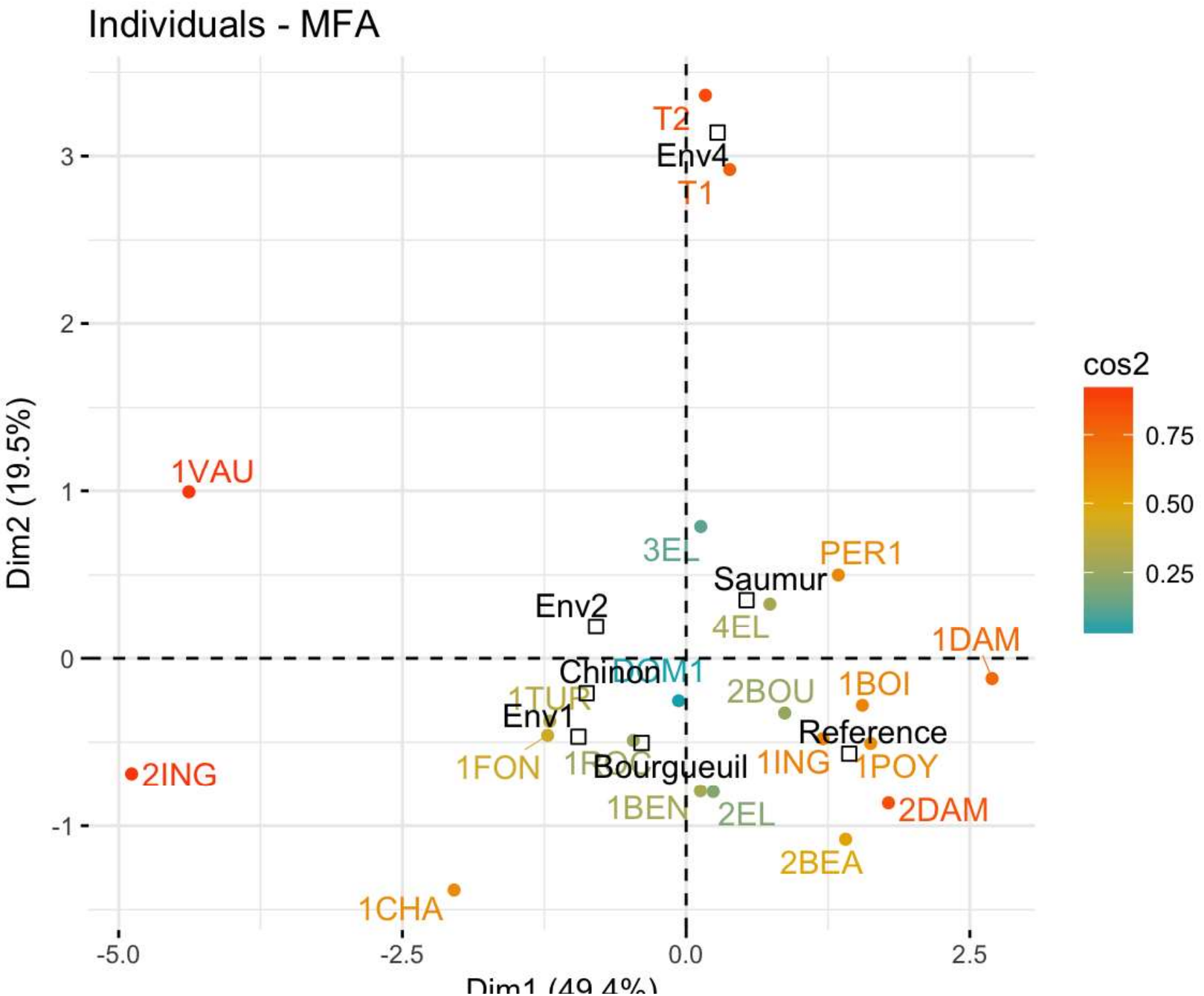
This global analysis, where multiple sets of variables are simultaneously considered, requires to balance the influences of each set of variables. Therefore, in MFA, the variables are weighted during the analysis. Variables in the same group are normalized using the same weighting value, which can vary 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.


Variable groups - MFA
origin
odor
odor.after.s
overall
taste
visual
0.6
0.4
0.2
0.0
0.00
0.25
0.50
0.75
Dim2 (19.5%)
Dim1 (49.4%)


Contribution of groups to Dim-1
Contributions (%)
20
10
0
odor.after.shaking
taste
visual
odor



For example, the first dimension represents the positive sentiments about wines: “intensity” and “harmony”. The most correlated variables to the second dimension are: i) Spice before shaking and Odor intensity before shaking for the odor group; ii) Spice, Plant and Odor intensity for the odor after shaking group and iii) Bitterness for the taste group. This dimension represents essentially the “spicyness” and the vegetal characteristic due to olfaction.



Individuals with similar profiles are close to each other on the factor map. The first axis, mainly opposes the wine 1DAM and, the wines 1VAU and 2ING. As described in the previous section, the first dimension represents the harmony and the intensity of wines. Thus, the wine 1DAM (positive coordinates) was evaluated as the most "intense" and "harmonious" contrary to wines 1VAU and 2ING (negative coordinates) which are the least "intense" and "harmonious". The second axis is essentially associated with the two wines T1 and T2 characterized by a strong value of the variables Spice.before.shaking and Odor.intensity.before.shaking.

The category Env4 has high coordinates on the second axis related to T1 and T2.
The category "Reference" is known to be related to an excellent wine-producing soil. As expected, our analysis demonstrates that the category "Reference" has high coordinates on the first axis, which is positively correlated with wines "intensity" and "harmony".