Master in Data Science (MDS)

Multivariate Analysis (MVA) Multiple Factor Analysis

Arturo Palomino

Dante Conti

Credits:Prof. Belchin Kostov

belchin.adriyanov.kostov@upc.edu

Slides modified from François Husson, with permission

1 Data - Introduction

2 Equilibrium and global PCA

3 Studying groups
　　　Group representation
　　　Partial points representation
　　　Separate analyses

4 Further topics
　　　Qualitative data
　　　Contingency tables
　　　Interpretation aids

FIB

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

- 10 white wines from the Loire valley : 5 Vouvray - 5 Sauvignon

- sensory descriptors : acidity, bitterness, citrus odor, etc.

- 10 white wines from the Loire valley : 5 Vouvray - 5 Sauvignon
- sensory descriptors : acidity, bitterness, citrus odor, etc.

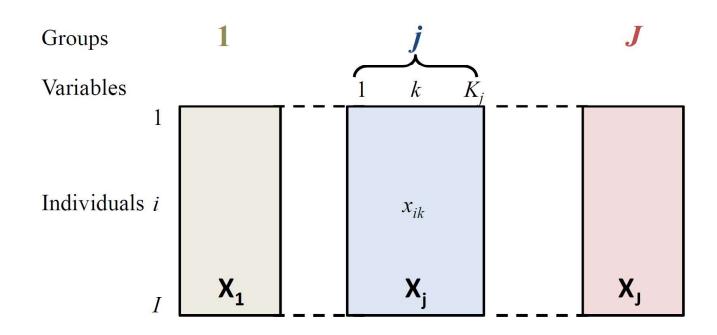|	O.fruity	O.passion	O.citrus		Sweetness	Acidity	Bitterness	Astringency	Aroma.intensity	Aroma.persistency	Visual.intensity	Grape variety
S Michaud	4.3	2.4	5.7	…	3.5	5.9	4.1	1.4	7.1	6.7	5.0	Sauvignon
S Renaudie	4.4	3.1	5.3	…	3.3	6.8	3.8	2.3	7.2	6.6	3.4	Sauvignon
S Trotignon	5.1	4.0	5.3	…	3.0	6.1	4.1	2.4	6.1	6.1	3.0	Sauvignon
S Buisse Domaine	4.3	2.4	3.6	…	3.9	5.6	2.5	3.0	4.9	5.1	4.1	Sauvignon
S Buisse Cristal	5.6	3.1	3.5	…	3.4	6.6	5.0	3.1	6.1	5.1	3.6	Sauvignon
V Aub Silex	3.9	0.7	3.3	…	7.9	4.4	3.0	2.4	5.9	5.6	4.0	Vouvray
V Aub Marigny	2.1	0.7	1.0	…	3.5	6.4	5.0	4.0	6.3	6.7	6.0	Vouvray
V Font Domaine	5.1	0.5	2.5	…	3.0	5.7	4.0	2.5	6.7	6.3	6.4	Vouvray
V Font Brûlés	5.1	0.8	3.8	…	3.9	5.4	4.0	3.1	7.0	6.1	7.4	Vouvray
V Font Coteaux	4.1	0.9	2.7	…	3.8	5.1	4.3	4.3	7.3	6.6	6.3	Vouvray

- 10 white wines from the Loire valley : 5 Vouvray - 5 Sauvignon
- sensory descriptors : acidity, bitterness, citrus odor, etc.
- tasting note of 60 consumers: overall appreciation

	Expert (27)	Student (15)	Consumer (15)	*Appreciation (60)*	*Grape variety (1)*
Wine 1					
Wine 2					
…					
Wine 10					

- How to characterize the wines?
- Are wines described in the same way by the different juries?
- Are there specific responses from certain juries?

Groups   **1**   **j**   **J**

Variables 1 k K_j

Individuals i x_{ik}

$\mathbf{X_1}$ $\mathbf{X_j}$ $\mathbf{X_J}$

Examples with **quantitative and/or qualitative** variables :

- genomics : DNA, expression, proteins
- questionnaires : student health (product consumption, psychological state, sleep, age, sex, etc.)
- Economics : annual economic indices

- 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

"Doing data analysis, in good mathematics, is simply searching for eigenvectors; all the science of it (the art) is to find the right matrix to diagonalize"
Benzécri

MFA is a weighted PCA :

- calculate the 1st eigenvalue λ_1^j of the jth group of variables ($j = 1, ..., J$)

- do an overall PCA on the weighted table :

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

X_j corresponds to the jth normalized or standardized table

	Before weighting			After weighting		
	Expert	Student	Consumer	Expert	Student	Consumer
λ_1	11.74	7.89	7.17	1.00	1.00	1.00
λ_2	6.78	3.83	2.59	0.58	0.49	0.36
λ_3	2.74	1.70	1.63	0.23	0.22	0.23

- Same weights for all variables from the same group : group structure is preserved

- For each group, the variance of the principal dimension (first eigenvalue) is equal to 1

- No group can generate the first axis on its own

- A multi-dimensional group will contribute to more axes than a one-dimensional group
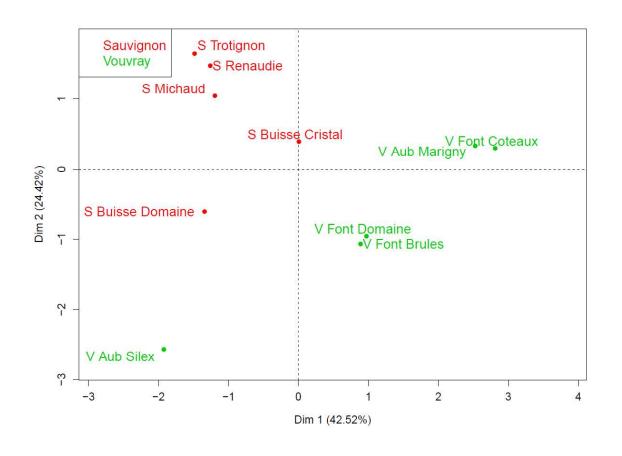
⇒ Same plots as in PCA

- 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

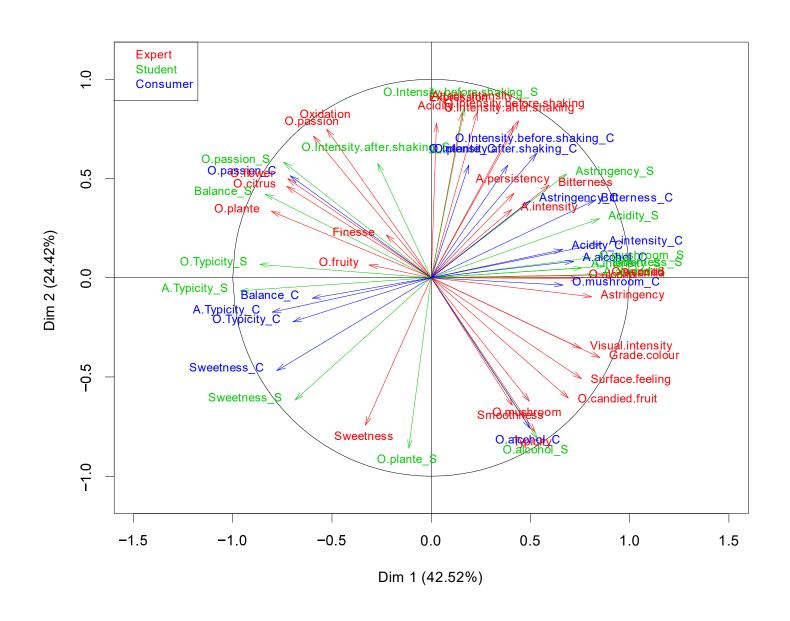- The 2 grape varieties are well-separated
- The Vouvray are more varied in terms of sensory perception
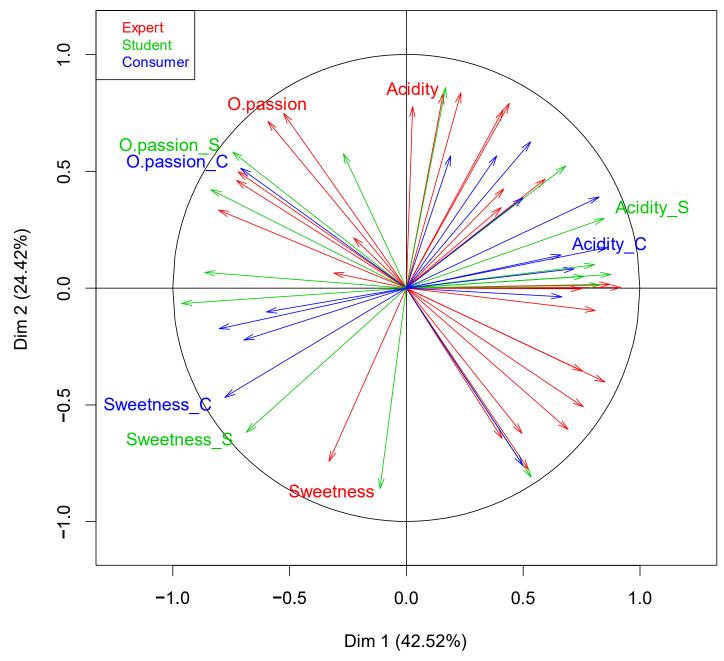- Several groups of wines …

Variables plot

① Data - Introduction

② Equilibrium and global PCA

③ **Studying groups**
   Group representation
   Partial points representation
   Separate analyses

④ Further topics
   Qualitative data
   Contingency tables
   Interpretation aids

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

In MFA :

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

$L_g(K_j, v_1)$ = projected inertia of all the variables of K_j on v_1 $\Rightarrow$ The first principal component of the MFA is the variable which maximizes the link with all groups, in the L_g sense.

$$0 \le L_g(K_j, v_1) \le 1$$

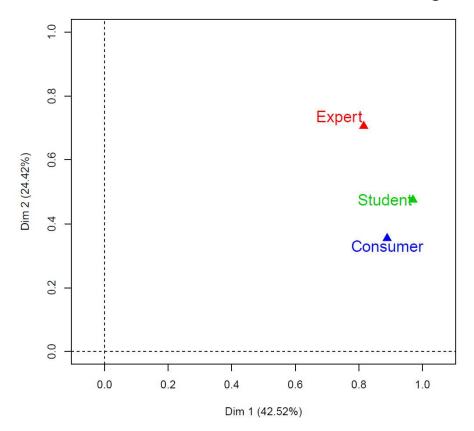$L_g = 0$: all variables in the jth group are uncorrelated with v_1
$L_g = 1$: v_1 the same as the 1st principal component of K_j

$\Rightarrow$ Using L_g to plot groups

The jth group has coordinates $L_g(K_j, v_1)$ and $L_g(K_j, v_2)$

- 1st axis is the same for all groups
- 2nd axis is due to the Experts group
- 2 groups are close to each other when they induce the same structure

$\Rightarrow$ This plot provides a synthetic comparison of the groups

$\Rightarrow$ Are the relative positions of individuals similar from one group to the next?

- The L_g coefficient measures the connection between groups of variables :

$$\mathcal{L}_g(K_j, K_m) = \sum_{k \in K_j} \sum_{l \in K_m} cov^2 \left(\frac{x_{.k}}{\sqrt{\lambda_1^j}}, \frac{x_{.l}}{\sqrt{\lambda_1^m}} \right)$$

- The L_g coefficient as an indicator of a group's dimensionality

$$\mathcal{L}_g(K_j, K_j) = \frac{\sum_{k=1}^{K_j} (\lambda_k^j)^2}{(\lambda_1^j)^2} = 1 + \frac{\sum_{k=2}^{K_j} (\lambda_k^j)^2}{(\lambda_1^j)^2}$$

$$RV(K_j, K_m) = \frac{\mathcal{L}_g(K_j, K_m)}{\sqrt{\mathcal{L}_g(K_j, K_j)} \sqrt{\mathcal{L}_g(K_m, K_m)}} \qquad 0 \le RV \le 1$$

$RV = 0$: all variable in K_j and K_m are uncorrelated
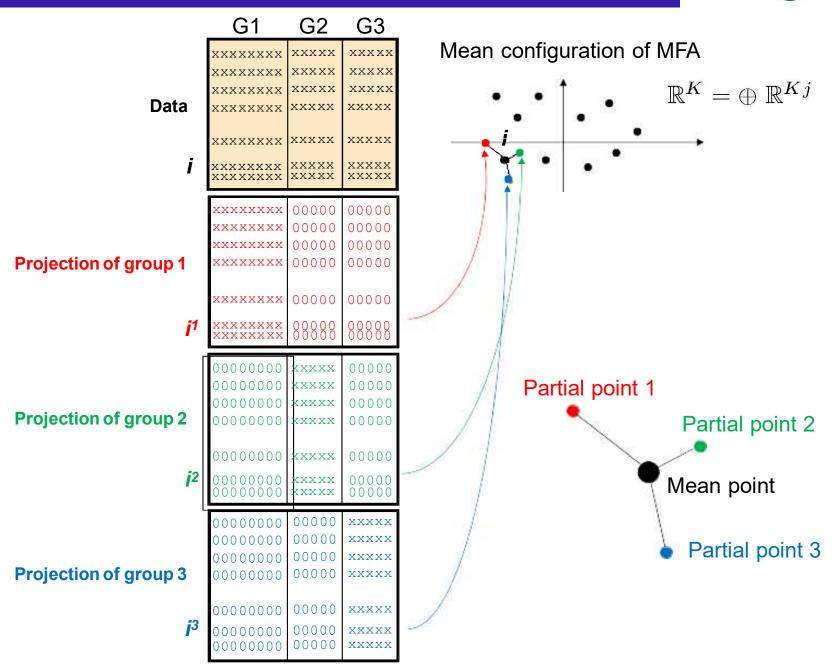$RV = 1$: the two point clouds are homothetic

⇒ Comparing groups in terms of individuals

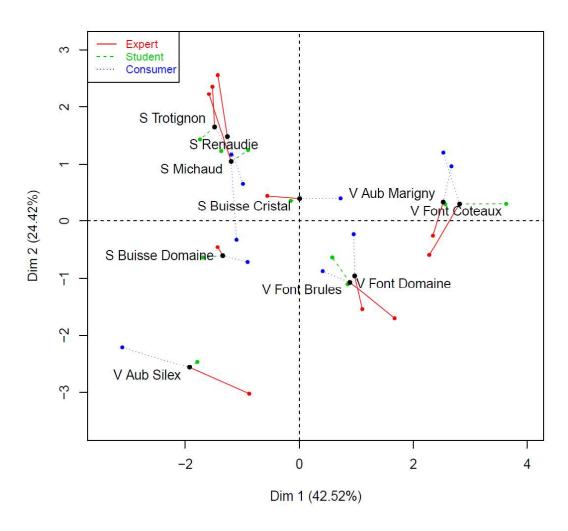⇒ Comparing descriptions provided by each group in a shared space

⇒ Are there specific individuals related to certain groups of variables?

Mean configuration of MFA

$$\mathbb{R}^K = \oplus \, \mathbb{R}^{Kj}$$

- Partial point = representing an individual as seen by a group
- An individual is at the barycenter of its partial points

1. Define the structure of the dataset (group composition)
2. Define the active groups and supplementary elements
3. Standardize the variables or not?
4. Run the MFA
5. Choose the number of dimensions to interpret
6. Simultaneous analysis of the individuals and variables plots
7. Group study
8. Partial analyses
9. Use indices to enrich the interpretation

The `MFA` function of the `FactoMineR` package

- MFA : a multi-table method for quantitative variables, qualitative variables, and frequency tables

- MFA balances the influence of each table

- Represents the information brought by each table in a shared setting

- Classical outputs (individuals, variables)

- Specific outputs (groups, separate analyses, partial points)

For more information about MFA, have a look at these videos:

https://www.youtube.com/watch?v=MOl0Aw1TTFE

https://www.youtube.com/watch?v=kGSjHD3yG84

https://www.youtube.com/watch?v=OVtNX6Or1FI

https://www.youtube.com/watch?v=Vumu7OoFHdA

https://www.youtube.com/watch?v=7HHuKsl2_fk

Husson F., Lê S. & Pagès J. (2017) *Exploratory Multivariate Analysis by Example Using R* 2nd edition, 230 p., CRC/Press.
Pagès, J. (2014). *Multiple Factor Analysis by Example Using R*. CRC Press.