Multiple Factor Analysis

François Husson

Applied Mathematics Department - Rennes Agrocampus

husson@agrocampus-ouest.fr

Outline

Sensory description of Loire wines

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

Sensory description of Loire wines

- 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

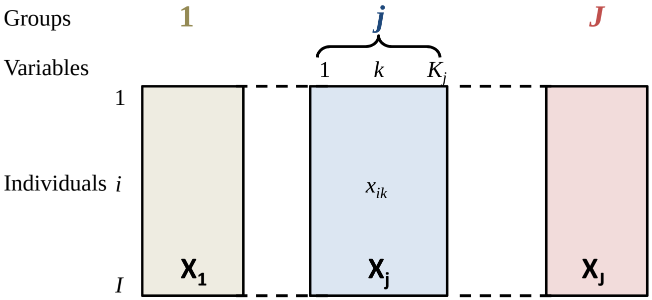
Sensory description of wines : comparing juries

- 10 white wines from the Loire valley : 5 Vouvray - 5 Sauvignon
- sensory descriptions from 3 juries : experts, consumers, students
- 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 ?

Multi-tables



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

Aims

- 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

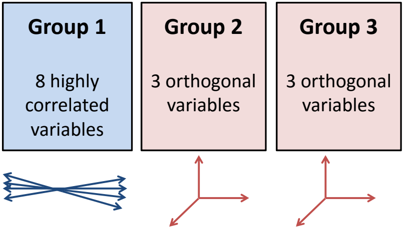
$\Rightarrow$ Balance the influence of all of the groups in the analysis

Outline

Balancing the influence of each group of variables

In PCA : normalizing balances each variable's influence (when calculating distances between individuals i and i')
In MFA, we balance in terms of groups

1st idea : divide each variable by the total inertia of the group it belongs to



2nd idea : divide each variable by the (square root of the) 1st eigenvalue of the group it belongs to

Balancing the influence of each group of variables

"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

Balancing the influence of each group of variables

	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

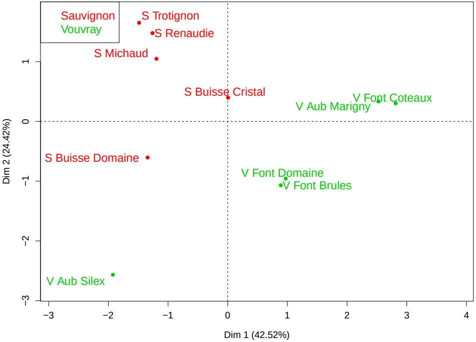
MFA - a weighted PCA

⇒ 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

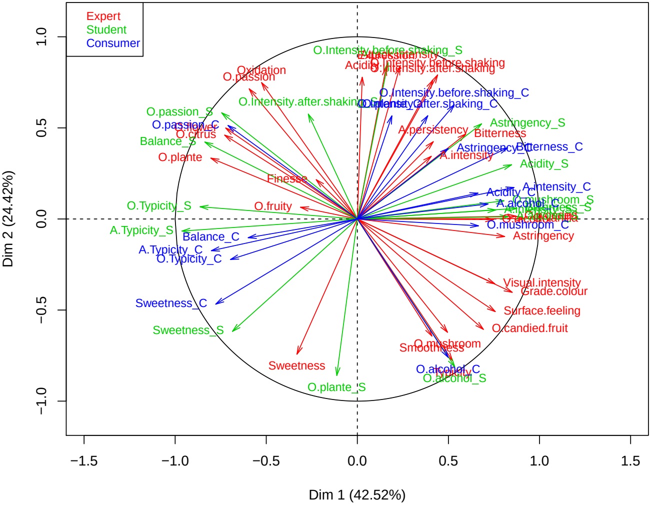
⇒ Same outputs (coordinates, cosine, contributions)
⇒ Add individuals and variables (quantitative, qualitative) as supplementary information

Individuals plot

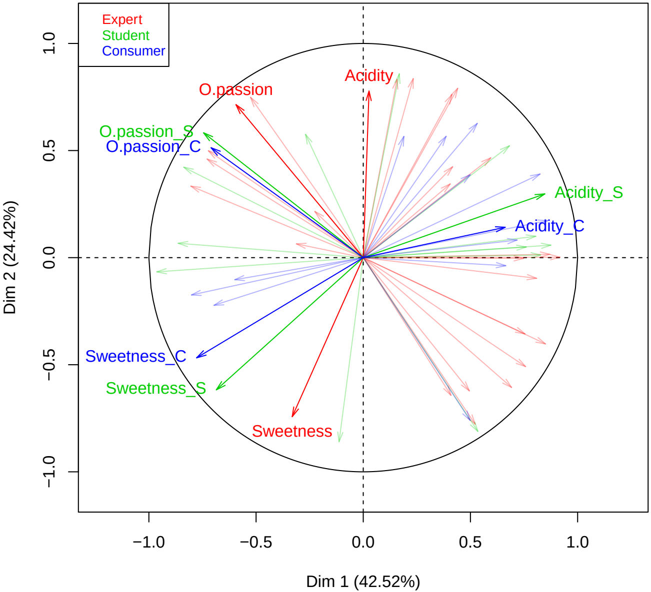


- The 2 grape varieties are well-separated
- The Vouvray are more varied in terms of sensory perception
- Several groups of wines . . .

Variables plot



Variables plot



Outline

First MFA component

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

In MFA :

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

$\mathcal{L}_g(K_j, v_1)$ = 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 $\mathcal{L}_g$ sense.

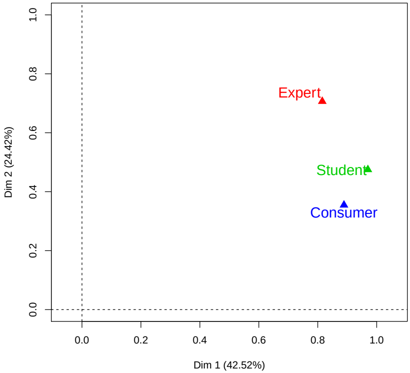
$$0 \leq \mathcal{L}_g(K_j, v_1) \leq 1$$

$\mathcal{L}_g = 0$: all variables in the jth group are uncorrelated with v_1
$\mathcal{L}_g = 1$: v_1 the same as the 1st principal component of K_j

Group plot

⇒ Using $\mathcal{L}_g$ to plot groups

The jth group has coordinates $\mathcal{L}_g(K_j, v_1)$ and $\mathcal{L}_g(K_j, v_2)$



- 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

⇒ This plot provides a synthetic comparison of the groups

⇒ Are the relative positions of individuals similar from one group to the next ?

Measuring how similar groups are

- The $\mathcal{L}_g$ 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 $\mathcal{L}_g$ coefficient as an indicator of a group's dimensionality

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

- $RV(K_j, K_m) = \frac{\mathcal{L}_g(K_j, K_m)}{\sqrt{\mathcal{L}_g(K_j, K_j)}\sqrt{\mathcal{L}_g(K_m, K_m)}}$ $\quad 0 \leq RV \leq 1$
  $RV = 0$: all variable in K_j and K_m are uncorrelated
  $RV = 1$: the two point clouds are homothetic

Measuring how similar groups are

```
> res$group$Lg
          Expert   Student   Consumer    MFA
Expert      1.45
Student     1.17      1.29
Consumer    0.94      1.04       1.25
MFA         1.33      1.31       1.21   1.44

> res$group$RV
          Expert    Student   Consumer    MFA
Expert      1.00
Student     0.85       1.00
Consumer    0.70       0.82       1.00
MFA         0.92       0.96       0.90   1.00
```

- The experts give more sophisticated descriptions (larger $\mathcal{L}_g$)
- The students and experts are quite related : $RV = 0.85$
- The students are closest to the shared configuration : $RV = 0.96$

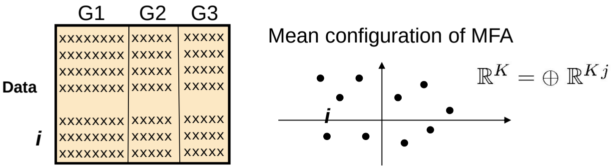
Partial points representation

⇒ Comparing groups in terms of individuals

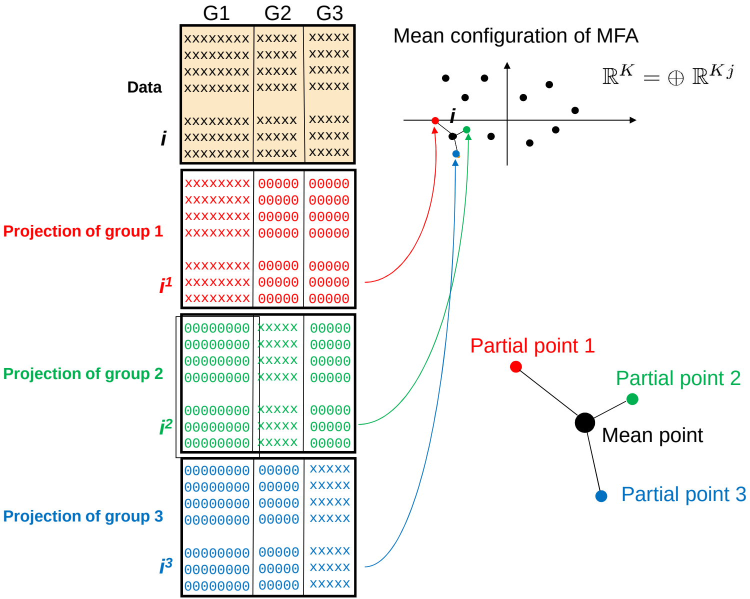
⇒ Comparing descriptions provided by each group in a shared space

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

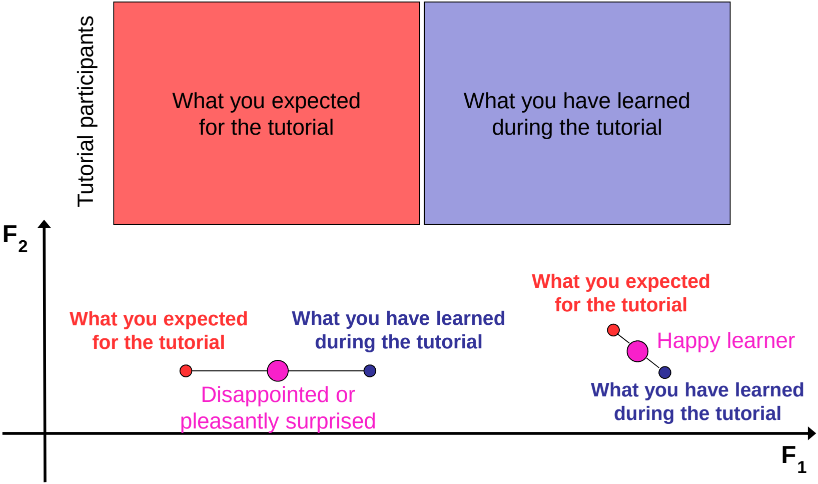
Projections of partial points



Projections of partial points



Partial points



Transition formulas

The transition formulas apply for the mean points

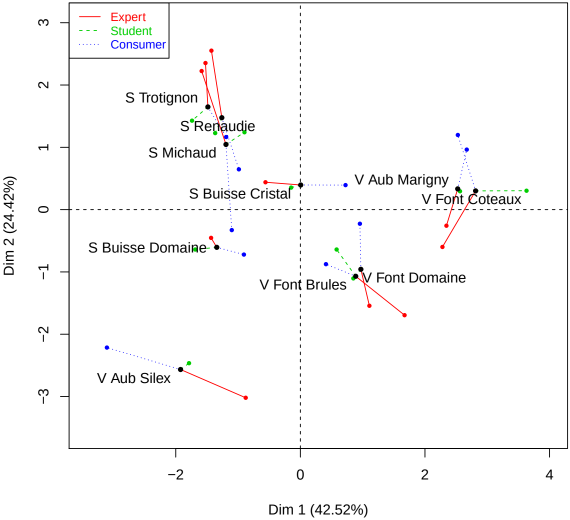
$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{J} \left(\frac{1}{\lambda_1^j} \sum_{k=1}^{K_j} x_{ik} G_s(k) \right)$$

and the partial points

$$F_s(i^j) = J \times \frac{1}{\sqrt{\lambda_s}} \frac{1}{\lambda_1^j} \sum_{k=1}^{K_j} x_{ik} G_s(k)$$

$\Rightarrow$ The superimposed plot with mean points and partial points can be analyzed in the same space

Partial points plot



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

Inertia ratios

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(F_{i^j s})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(F_{is})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(F_{i^j s} - F_{is})^2$$

total inertia = between-individual inertia + within-individual inertia

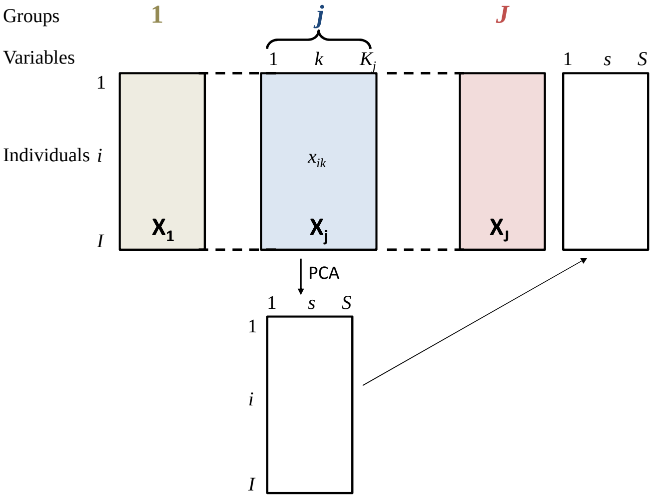
$$\frac{\text{“Between” inertia on axis } s}{\text{Total inertia on axis } s} = \frac{J \ \sum_{i=1}^{I}(F_{is})^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}(F_{i^j s})^2}$$

```
> res$inertia.ratio
Dim.1   Dim.2   Dim.3   Dim.4   Dim.5
 0.93    0.82    0.78    0.54    0.53
```

- On the first axis, the coordinates of the partial points are close to each other (0.93 close to 1)
- The within-inertia on an axis can be broken down by individual

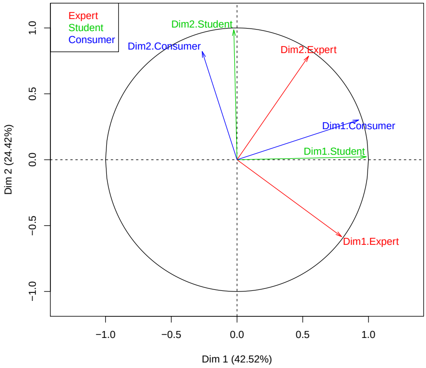
Connection with components obtained from separate PCA

Do separate analyses give comparable results to the global MFA ?



Connection with components obtained from separate PCA

⇒ Principal components of separate PCA are projected as supplementary information



- The PCA dimensions for the students are like those of the MFA
- The first two dimensions of each group are well-projected

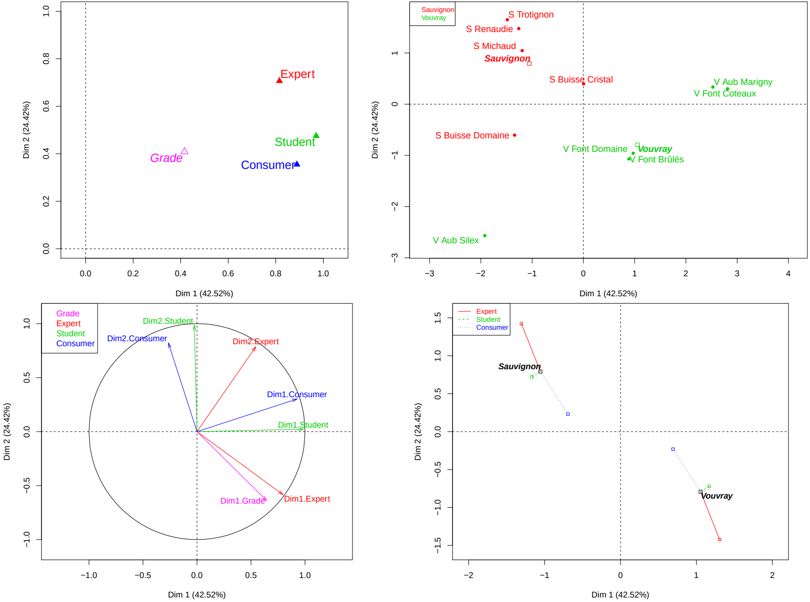
Outline

Qualitative data

- Balance the effect of each group of variables in the global analysis
- The usual plots for treating qualitative data (individuals and categories)
- Specific plots (groups plot, superimposed plot, partial axes plots, separate analyses plots)

$\Rightarrow$ Same methodological approach, just replacing PCA with MCA

Qualitative data



Mixed data

$\Rightarrow$ Some groups with quantitative variables and others with qualitative variables

"Locally", MFA behaves like :

- a PCA for the quantitative variables
- an MCA for the qualitative variables

The MFA weighting allows us to analyze the two variable types together

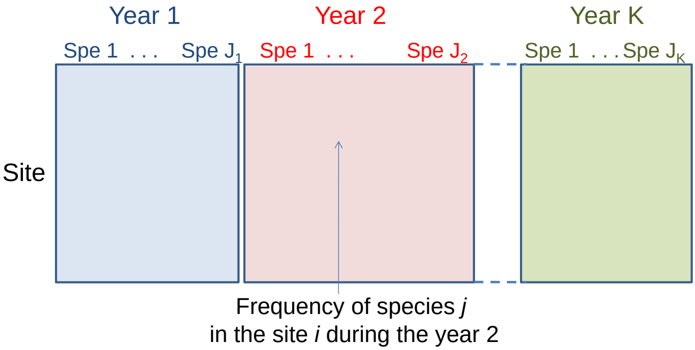
Special case : if each group has just one variable $\Longrightarrow$ **Factor Analysis of Mixed Data** (FAMD)

MFA for contingency tables

MFA can be extended to contingency tables : MFACT
The tables must have the same rows (or the same columns)

Examples

- survey in several countries (Profession × Questions / country)
- ecology : Sites × Species / Year



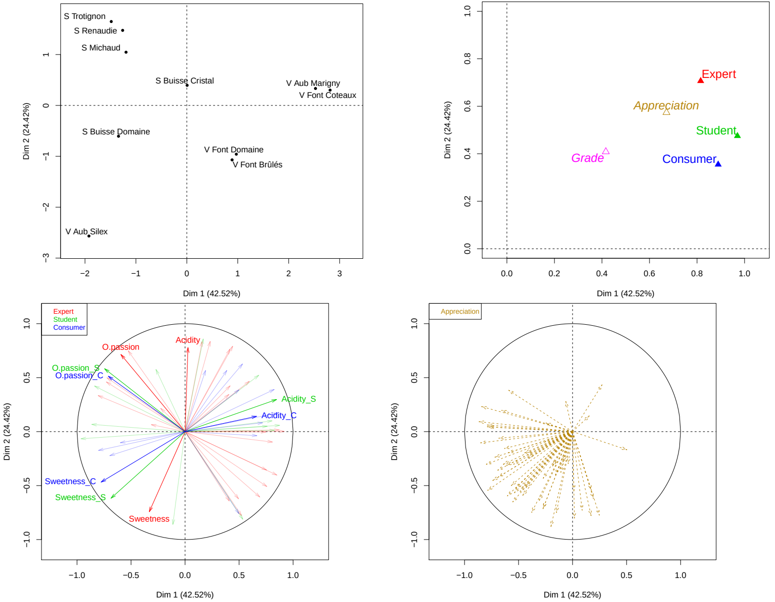
Plotting supplementary information

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

Questions :

- Are *preferences* linked with sensory characteristics ?
- Does the *grape variety* explain the sensory characteristics ?

Visualizing quantitative supplementary groups



Indices : contributions and representation quality

- Individuals and variables : same as the PCA calculations
- Contribution of the kth group to construction of the sth axis :

$$Ctr_s(k) = \frac{F_{ks}}{\sum_{k=1}^{K} F_{ks}} \ (\times 100)$$

```
> res$group$contrib
          Dim.1  Dim.2  Dim.3  Dim.4  Dim.5
Expert    30.49  45.99  33.68  44.59  40.60
Student   36.27  30.92  35.07   9.20  14.72
Consumer  33.24  23.09  31.25  46.20  44.68
```

- Representation quality of the kth group in a subspace : $\cos^2$ between the kth point and its projection

```
> res$group$cos2
         Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
Expert    0.46  0.34  0.03  0.03  0.01
Student   0.73  0.17  0.03  0.00  0.00
Consumer  0.63  0.10  0.03  0.03  0.02
```

Characterizing the axes

Using quantitative variables :

- correlation between each variable and the *s*th principal component is calculated
- the correlation coefficients are sorted and the significant ones retained

```
> dimdesc(res)
           $Dim.1$quanti                                   $Dim.2$quanti
                corr p.value                                    corr p.value
O.vanilla       0.92 1.8e-04    O.Int.bef.shaking_S   0.86   0.0015
Bitterness_S    0.88 9.0e-04    Attack.intensity      0.84   0.0026
O.wooded        0.87 1.0e-03    Expression            0.83   0.0028
A.intensity_C   0.86 1.4e-03    O.Int.bef.shaking     0.79   0.0064
Grade.colour    0.85 1.8e-03    Acidity               0.78   0.0081
Acidity_S       0.85 2.0e-03    O.Int.after.shaking   0.76   0.0110
   ...           ...     ...        ...                ...     ...
Balance_S      -0.84 2.5e-03    Typicity             -0.78   0.0081
O.Typicity_S   -0.86 1.3e-03    O.alcohol_S          -0.81   0.0044
A.Typicity_S   -0.96 7.7e-06    O.plante_S           -0.86   0.0014
```

Characterizing the axes

Using qualitative variables :

- do analysis of variance with an individual's coordinates ($F_{.s}$) described in terms of the given qualitative variable
  - one F-test per variable
  - for each category, a Student's t-test

```
> dimdesc(res)
$Dim.1$quali
                 R2      p.value
grape variety 0.416   0.04396733

$Dim.1$category
           Estimate     p.value
Vouvray       1.055   0.04396733
Sauvignon    -1.055   0.04396733
```

```
$Dim.2$quali
                 R2     p.value
grape variety 0.408 0.04667455

$Dim.2$category
            Estimate      p.value
Sauvignon      0.792   0.04667455
Vouvray       -0.792   0.04667455
```

Putting MFA into practice

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

Conclusion

- 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)

Bibliography

- Pagès, J. (2014). *Multiple Factor Analysis by Example Using R*. CRC Press.