Multivariate Data Analysis
Special focus on Clustering and Multiway Methods

François Husson & Julie Josse

Applied mathematics department, Agrocampus Rennes

useR! 2010, July 20, 2010
Why a tutorial on Multivariate Data Analysis?

- Our research focus is principal component methods
- We teach multivariate data analysis
- We have developed R packages:
  - **FactoMineR** to perform principal component methods
    - PCA, correspondence analysis (CA), multiple correspondence analysis (MCA), multiple factor analysis (MFA)
    - complementarity between clustering and principal component methods
  - **missMDA** to handle missing values in and with multivariate data analysis
    - perform principal component methods (PCA, MCA) with missing values
    - simple and multiple imputation based on principal component models for continuous and categorical data
Outline

Multivariate data analysis with a special focus on clustering and multiway methods

1. Principal Component Analysis (PCA)
2. Multiple Factor Analysis (MFA)
3. Complementarity between Clustering and Principal Component methods

⇒ Multidimensional descriptive methods
⇒ Graphical representations
Principal Component Analysis

1. Data - Issues - Preprocessing
2. Individuals Study
3. Variables Study
4. Helps to Interpret
Principal Component Analysis

Dimensionality reduction ⇒ describes the dataset with a smaller number of variables

Technique widely used for applications such as: data compression, data reconstruction, preprocessing before clustering, and ...

Descriptive methods
PCA deals with which kind of data?

PCA deals with continuous variables, but categorical variables can also be included in the analysis.

Many examples:

- Sensory analysis: products - descriptors
- Ecology: plants - measurements; waters - physico-chemical analyses
- Economy: countries - economic indicators
- Microbiology: cheeses - microbiological analyses
- etc.

**Figure:** Data table in PCA
Wine data

- **10 individuals (rows):** white wines from Val de Loire
- **30 variables (columns):**
  - 27 continuous variables: sensory descriptors
  - 2 continuous variables: odour and overall preferences
  - 1 categorical variable: label of the wines (Vouvray - Sauvignon)

|       | O.fruity | O.passion | O.citrus | ... | Sweetness | Acidity | Bitterness | Astringency | Aroma.intensity | Aroma.persistency | Visual.intensity | Odor.preference | Overall.preference | Label     |
|-------|----------|-----------|----------|-----|-----------|---------|------------|-------------|-----------------|------------------|-----------------|-----------------|-----------------|------------------|----------|
| S Michaud | 4.3      | 2.4       | 5.7      | ... | 3.5       | 5.9     | 4.1        | 1.4         | 7.1             | 6.7              | 5.0             | 6.0             | 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             | 5.4             | 5.5             | Sauvignon       |
| S Trotignon | 5.1      | 4.0       | 5.3      | ... | 3.0       | 6.1     | 4.1        | 2.4         | 6.1             | 6.1              | 3.0             | 5.0             | 5.5             | Sauvignon       |
| S Buisse Domaine | 4.3      | 2.4       | 3.6      | ... | 3.9       | 5.6     | 2.5        | 3.0         | 4.9             | 5.1              | 4.1             | 6.1             | 5.1             | Sauvignon       |
| S Buisse Cristal | 5.6      | 3.1       | 3.5      | ... | 3.4       | 6.6     | 5.0        | 3.1         | 6.1             | 6.1              | 3.6             | 6.1             | 5.0             | Sauvignon       |
| V Aub Silex | 3.9      | 0.7       | 3.3      | ... | 7.9       | 4.4     | 3.0        | 2.4         | 5.9             | 5.6              | 4.0             | 6.1             | 5.0             | Vouvray         |
| V Aub Marigny | 2.1      | 0.7       | 1.0      | ... | 3.5       | 6.4     | 5.0        | 4.0         | 6.3             | 6.7              | 6.0             | 5.1             | 4.1             | Vouvray         |
| V Font Domaine | 5.1      | 0.5       | 2.5      | ... | 3.0       | 5.7     | 4.0        | 2.5         | 6.7             | 6.3              | 6.4             | 4.4             | 5.1             | Vouvray         |
| V Font Brûlés | 5.1      | 0.8       | 3.8      | ... | 3.9       | 5.4     | 4.0        | 3.1         | 7.0             | 6.1              | 7.4             | 4.4             | 6.4             | Vouvray         |
| V Font Coteaux | 4.1      | 0.9       | 2.7      | ... | 3.8       | 5.1     | 4.3        | 4.3         | 7.3             | 6.6              | 6.3             | 6.0             | 5.7             | Vouvray         |
Problems - objectives

- **Individuals study:**
  similarity between individuals with respect to all the variables
  ⇒ partition between individuals

- **Variables study:**
  linear relationships between variables
  ⇒ visualization of the correlation matrix (denoted $S$); find synthetic variables

- **Link between the two studies:**
  characterization of the groups of individuals by the variables;
  specific individuals to better understand links between variables
Two clouds of points

Individuals study

Variables study

Figure: Two clouds of points
Preprocessing

⇒ Similarity between individuals: Euclidean distance

- Choosing active variables

\[ d^2(i, i') = \sum_{k=1}^{K} (x_{ik} - x_{i'k})^2 \]

- Variables are always centred

\[ d^2(i, i') = \sum_{k=1}^{K} ((x_{ik} - \bar{x}_k) - (x_{i'k} - \bar{x}_k))^2 \]

- Standardizing variables or not?

\[ d^2(i, i') = \sum_{k=1}^{K} \frac{1}{s^2_k} (x_{ik} - x_{i'k})^2 \]
- Study the structure, i.e. the shape of the cloud of individuals
- Individuals are in $\mathbb{R}^K$
Fit the individuals cloud

Find the subspace which better sums up the data

⇒ Closest representation by projection
⇒ Best representation of the diversity, variability

Figure: Camel vs dromedary?
Fit the individuals cloud

\[ P_{u_1}(x_i.) = u_1(u_1' u_1)^{-1} u_1' x_i. \]
\[ = < x_{i.}, u_1 > u_1 \]
\[ F_{i1} = < x_{i.}, u_1 > \]

- Minimize the distance between individuals and their projections
- Maximize the variance of the projected data

\[ u_1 = \arg \max_{u_1 \in \mathbb{R}^K} \text{var}(F_1) = \arg \max_{u_1 \in \mathbb{R}^K} \text{var}(Xu_1) \text{ with } u_1' u_1 = 1 \]

\[ \Rightarrow u_1 \text{ first eigenvector of the correlation matrix associated with the largest eigenvalue } \lambda_1: Su_1 = \lambda_1 u_1 \]

\[ \text{Var}(F_1) = \text{var}(Xu_1) = 1/I \ u_1' X' Xu_1 = u_1' Su_1 = \lambda_1 u_1' u_1 = \lambda_1 \]
Fit the individuals cloud

Additional axes are sequentially defined: each new direction maximizes the projected variance among all orthogonal directions
⇒ $Q$ eigenvectors $u_1, ..., u_Q$ associated to $\lambda_1, ..., \lambda_Q$

Representation quality: dimensionality reduction ⇒ loosing information

• Total variance of the initial individuals cloud (total inertia):

$$\frac{1}{I} \| \mathbf{x}_i - \mathbf{g} \|^2 = tr(S) = \sum_{k=1}^{K} \lambda_k \ (= K)$$

• Variance of the projected individuals cloud (Q-dimensional representation): $\text{var}(F_1) + \text{var}(F_2) + ... + \text{var}(F_Q)$

⇒ Percentage of variance explained: $\frac{\sum_{k=1}^{Q} \lambda_k}{\sum_{k=1}^{K} \lambda_k}$
Example: wine data

- Sensory descriptors are used as active variables: only these variables are used to construct the axes
- Variables are (centred and) standardized

| Label                  | O.fruity | O.passion | O.citrus | ... | Sweetness | Acidity | Bitterness | Astringency | Aroma.intensity | Aroma.persistency | Visual.intensity | Odor.preference | Overall.preference | Label       |
|------------------------|----------|-----------|----------|-----|-----------|---------|------------|-------------|------------------|-------------------|-----------------|-----------------|-----------------|---------------|-------------|
| S Michaud              | 4.3      | 2.4       | 5.7      | ... | 3.5       | 5.9     | 4.1        | 1.4         | 7.1              | 6.7               | 5.0             | 6.0             | 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             | 5.4             | 5.5             | Sauvignon     |
| S Trotignon            | 5.1      | 4.0       | 5.3      | ... | 3.0       | 6.1     | 4.1        | 2.4         | 6.1              | 6.1               | 3.0             | 5.0             | 5.5             | Sauvignon     |
| S Buisse Domaine       | 4.3      | 2.4       | 3.6      | ... | 3.9       | 5.6     | 2.5        | 3.0         | 4.9              | 5.1               | 4.1             | 5.3             | 4.6             | Sauvignon     |
| S Buisse Cristal       | 5.6      | 3.1       | 3.5      | ... | 3.4       | 6.6     | 5.0        | 3.1         | 6.1              | 5.1               | 3.6             | 6.1             | 5.0             | Sauvignon     |
| V Aub Silex            | 3.9      | 0.7       | 3.3      | ... | 7.9       | 4.4     | 3.0        | 2.4         | 5.9              | 5.6               | 4.0             | 5.0             | 5.5             | Vouvray       |
| V Aub Marigny          | 2.1      | 0.7       | 1.0      | ... | 3.5       | 6.4     | 5.0        | 4.0         | 6.3              | 6.7               | 6.0             | 5.1             | 4.1             | Vouvray       |
| V Font Domaine         | 5.1      | 0.5       | 2.5      | ... | 3.0       | 5.7     | 4.0        | 2.5         | 6.7              | 6.3               | 6.4             | 4.4             | 5.1             | Vouvray       |
| V Font Brûlés          | 5.1      | 0.8       | 3.8      | ... | 3.9       | 5.4     | 4.0        | 3.1         | 7.0              | 6.1               | 7.4             | 4.4             | 6.4             | Vouvray       |
| V Font Coteaux         | 4.1      | 0.9       | 2.7      | ... | 3.8       | 5.1     | 4.3        | 4.3         | 7.3              | 6.6               | 6.3             | 6.0             | 5.7             | Vouvray       |
Example: graph of the individuals

⇒ Need variables to interpret the dimensions of variability
Individuals coordinates considered as variables
Interpretation of the individuals graph with the variables

- Correlation between variable $x.k$ and $F.1$ (and $F.2$)

$\Rightarrow$ Correlation circle
Interpretation of the individuals graph with the variables
Cloud of variables

Since variables are centred:

$$\cos(\theta_{kl}) = \frac{<x_k, x_l>}{\|x_k\| \|x_l\|} = \frac{\sum_{i=1}^{I} x_{ik}x_{il}}{\sqrt{(\sum_{i=1}^{I} x_{ik}^2)(\sum_{i=1}^{I} x_{il}^2)}} = r(x_k, x_l)$$
Fit the variables cloud

Find $v_1$ (in $\mathbb{R}^I$, with $v_1'v_1 = 1$) which best fits the cloud

$$P_{v_1}(x.k) = v_1(v_1'v_1)^{-1}v_1'x.k$$

$$G_{k1} = 1/I < v_1, x.k >$$

$$G_{k1} = 1/I \frac{< v_1, x.k >}{\|v_1\|\|x.k\|}$$

$$\arg \max_{v_1 \in \mathbb{R}^I} \sum_{i=k}^{K} G_{k1}^2 = \arg \max_{v_1 \in \mathbb{R}^I} \sum_{i=k}^{K} r(v_1, x.k)^2$$

$\Rightarrow v_1$ is the best synthetic variable

$\Rightarrow v_1, ..., v_Q$ are the eigenvectors of $W = XX'$ the inner product matrix associated with the largest eigenvalues: $Wv_q = \lambda_q v_q$
Fit the variables cloud

⇒ Same representation! What a wonderful result!
Projections...

\[ r(A, B) = \cos(\theta_{A,B}) \]
\[ \cos(\theta_{A,B}) \approx \cos(\theta_{H_A,H_B}) \text{ if variables are well projected} \]

Only well projected variables can be interpreted!
Link between the two representations: transition formulae
Link between the two representations: transition formulae

- \( Su = X'Xu = \lambda u \)
- \( XX'Xu = X\lambda u \rightarrow W(Xu) = \lambda(Xu) \)
- \( WF = \lambda F \) and since \( Wv = \lambda v \) then \( F \) and \( v \) are collinear
- Since, \( ||F|| = \lambda \) and \( ||v|| = 1 \) we have:

\[
\begin{align*}
v &= \frac{1}{\sqrt{\lambda}} F \\
u &= \frac{1}{\sqrt{\lambda}} G \\
\Rightarrow G &= X'v = \frac{1}{\sqrt{\lambda}} X'F \\
\Rightarrow F &= Xu = \frac{1}{\sqrt{\lambda}} XG
\end{align*}
\]

\[
F_{iq} = \frac{1}{\sqrt{\lambda_q}} \sum_{k=1}^{K} x_{ik} G_{kq} \quad G_{kq} = \frac{1}{\sqrt{\lambda_q}} \sum_{i=1}^{l} x_{ik} F_{iq}
\]

\(F,q\): principal components, scores
\(G,q\): correlations between variables and principal components
Link between the two representations: transition formulae

\[ F_{iq} = \frac{1}{\sqrt{\lambda_q}} \sum_{k=1}^{K} x_{ik} G_{kq} \]

\[ G_{kq} = \frac{1}{\sqrt{\lambda_q}} \sum_{i=1}^{l} x_{ik} F_{iq} \]

What does it mean? An individual is at the same side as the variables for which it takes high values.
Supplementary information

- For the continuous variables: projection of supplementary variables on the dimensions
- For the individuals: projection
- For the categories: projection at the barycentre of the individuals who take the categories

⇒ Supplementary information do not create the dimensions
Choosing the number of components

Bar plot, test on eigenvalues, confidence interval, cross-validation (functions `estim_ncpPCA` and `estim_ncp`), etc.

Two objectives:
⇒ Interpretation
⇒ Separate structure and noise
Percentage of variance obtained under independence

⇒ Is there a structure on my data?

<table>
<thead>
<tr>
<th>nbind</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>96.5</td>
<td>93.1</td>
<td>90.2</td>
<td>87.6</td>
<td>85.5</td>
<td>83.4</td>
<td>81.9</td>
<td>80.7</td>
<td>79.4</td>
<td>78.1</td>
<td>77.4</td>
<td>76.6</td>
<td>75.5</td>
</tr>
<tr>
<td>6</td>
<td>93.3</td>
<td>88.6</td>
<td>84.8</td>
<td>81.5</td>
<td>79.1</td>
<td>76.9</td>
<td>75.1</td>
<td>73.2</td>
<td>72.2</td>
<td>70.8</td>
<td>69.8</td>
<td>68.7</td>
<td>68.0</td>
</tr>
<tr>
<td>7</td>
<td>90.5</td>
<td>84.9</td>
<td>80.9</td>
<td>77.4</td>
<td>74.4</td>
<td>72.0</td>
<td>70.1</td>
<td>68.3</td>
<td>67.0</td>
<td>65.3</td>
<td>64.3</td>
<td>63.2</td>
<td>62.2</td>
</tr>
<tr>
<td>8</td>
<td>88.1</td>
<td>82.3</td>
<td>77.2</td>
<td>73.8</td>
<td>70.7</td>
<td>68.2</td>
<td>66.1</td>
<td>64.0</td>
<td>62.8</td>
<td>61.2</td>
<td>60.0</td>
<td>59.0</td>
<td>58.0</td>
</tr>
<tr>
<td>9</td>
<td>86.1</td>
<td>79.5</td>
<td>74.8</td>
<td>70.7</td>
<td>67.4</td>
<td>65.1</td>
<td>62.9</td>
<td>61.1</td>
<td>59.4</td>
<td>57.9</td>
<td>56.5</td>
<td>55.4</td>
<td>54.3</td>
</tr>
<tr>
<td>10</td>
<td>84.5</td>
<td>77.5</td>
<td>72.3</td>
<td>68.2</td>
<td>65.0</td>
<td>62.4</td>
<td>60.1</td>
<td>58.3</td>
<td>56.5</td>
<td>55.1</td>
<td>53.7</td>
<td>52.5</td>
<td>51.5</td>
</tr>
<tr>
<td>11</td>
<td>82.8</td>
<td>75.7</td>
<td>70.3</td>
<td>66.3</td>
<td>62.9</td>
<td>60.1</td>
<td>58.0</td>
<td>56.0</td>
<td>54.4</td>
<td>52.7</td>
<td>51.3</td>
<td>50.1</td>
<td>49.2</td>
</tr>
<tr>
<td>12</td>
<td>81.5</td>
<td>74.0</td>
<td>68.6</td>
<td>64.4</td>
<td>61.2</td>
<td>58.3</td>
<td>55.8</td>
<td>54.0</td>
<td>52.4</td>
<td>50.9</td>
<td>49.3</td>
<td>48.2</td>
<td>47.2</td>
</tr>
<tr>
<td>13</td>
<td>80.0</td>
<td>72.5</td>
<td>67.2</td>
<td>62.9</td>
<td>59.4</td>
<td>56.7</td>
<td>54.4</td>
<td>52.2</td>
<td>50.5</td>
<td>48.9</td>
<td>47.7</td>
<td>46.6</td>
<td>45.4</td>
</tr>
<tr>
<td>14</td>
<td>79.0</td>
<td>71.5</td>
<td>65.7</td>
<td>61.5</td>
<td>58.1</td>
<td>55.1</td>
<td>52.8</td>
<td>50.8</td>
<td>49.0</td>
<td>47.5</td>
<td>46.2</td>
<td>45.0</td>
<td>44.0</td>
</tr>
<tr>
<td>15</td>
<td>78.1</td>
<td>70.3</td>
<td>64.6</td>
<td>60.3</td>
<td>57.0</td>
<td>53.9</td>
<td>51.5</td>
<td>49.4</td>
<td>47.8</td>
<td>46.1</td>
<td>44.9</td>
<td>43.6</td>
<td>42.5</td>
</tr>
<tr>
<td>16</td>
<td>77.3</td>
<td>69.4</td>
<td>63.5</td>
<td>59.2</td>
<td>55.6</td>
<td>52.9</td>
<td>50.3</td>
<td>48.3</td>
<td>46.6</td>
<td>45.2</td>
<td>43.6</td>
<td>42.4</td>
<td>41.4</td>
</tr>
<tr>
<td>17</td>
<td>76.5</td>
<td>68.4</td>
<td>62.6</td>
<td>58.2</td>
<td>54.7</td>
<td>51.8</td>
<td>49.3</td>
<td>47.1</td>
<td>45.5</td>
<td>44.0</td>
<td>42.6</td>
<td>41.4</td>
<td>40.3</td>
</tr>
<tr>
<td>18</td>
<td>75.5</td>
<td>67.6</td>
<td>61.8</td>
<td>57.1</td>
<td>53.7</td>
<td>50.8</td>
<td>48.4</td>
<td>46.3</td>
<td>44.6</td>
<td>43.0</td>
<td>41.6</td>
<td>40.4</td>
<td>39.3</td>
</tr>
<tr>
<td>19</td>
<td>75.1</td>
<td>67.0</td>
<td>60.9</td>
<td>56.5</td>
<td>52.8</td>
<td>49.9</td>
<td>47.4</td>
<td>45.5</td>
<td>43.7</td>
<td>42.1</td>
<td>40.7</td>
<td>39.6</td>
<td>38.4</td>
</tr>
<tr>
<td>20</td>
<td>74.1</td>
<td>66.1</td>
<td>60.1</td>
<td>55.6</td>
<td>52.1</td>
<td>49.1</td>
<td>46.6</td>
<td>44.7</td>
<td>42.9</td>
<td>41.3</td>
<td>39.8</td>
<td>38.7</td>
<td>37.5</td>
</tr>
<tr>
<td>25</td>
<td>72.0</td>
<td>63.3</td>
<td>57.1</td>
<td>52.5</td>
<td>48.9</td>
<td>46.0</td>
<td>43.4</td>
<td>41.4</td>
<td>39.6</td>
<td>38.1</td>
<td>36.7</td>
<td>35.5</td>
<td>34.5</td>
</tr>
<tr>
<td>30</td>
<td>69.8</td>
<td>61.1</td>
<td>55.1</td>
<td>50.3</td>
<td>46.7</td>
<td>43.6</td>
<td>41.1</td>
<td>39.1</td>
<td>37.3</td>
<td>35.7</td>
<td>34.4</td>
<td>33.2</td>
<td>32.1</td>
</tr>
<tr>
<td>35</td>
<td>68.5</td>
<td>59.6</td>
<td>53.3</td>
<td>48.6</td>
<td>44.9</td>
<td>41.9</td>
<td>39.5</td>
<td>37.4</td>
<td>35.6</td>
<td>34.0</td>
<td>32.7</td>
<td>31.6</td>
<td>30.4</td>
</tr>
<tr>
<td>40</td>
<td>67.5</td>
<td>58.3</td>
<td>52.0</td>
<td>47.3</td>
<td>43.4</td>
<td>40.5</td>
<td>38.0</td>
<td>36.0</td>
<td>34.1</td>
<td>32.7</td>
<td>31.3</td>
<td>30.1</td>
<td>29.1</td>
</tr>
<tr>
<td>45</td>
<td>66.4</td>
<td>57.1</td>
<td>50.8</td>
<td>46.1</td>
<td>42.4</td>
<td>39.3</td>
<td>36.9</td>
<td>34.8</td>
<td>33.1</td>
<td>31.5</td>
<td>30.2</td>
<td>29.0</td>
<td>27.9</td>
</tr>
<tr>
<td>50</td>
<td>65.6</td>
<td>56.3</td>
<td>49.9</td>
<td>45.2</td>
<td>41.4</td>
<td>38.4</td>
<td>35.9</td>
<td>33.9</td>
<td>32.1</td>
<td>30.5</td>
<td>29.2</td>
<td>28.1</td>
<td>27.0</td>
</tr>
<tr>
<td>100</td>
<td>60.9</td>
<td>51.4</td>
<td>44.9</td>
<td>40.0</td>
<td>36.3</td>
<td>33.3</td>
<td>31.0</td>
<td>28.9</td>
<td>27.2</td>
<td>25.8</td>
<td>24.5</td>
<td>23.3</td>
<td>22.3</td>
</tr>
</tbody>
</table>

**Table:** 95 % quantile inertia on the two first dimensions of 10000 PCA on data with independent variables
### Percentage of variance obtained under independence

<table>
<thead>
<tr>
<th>nbinding</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>50</th>
<th>75</th>
<th>100</th>
<th>150</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>74.9</td>
<td>74.2</td>
<td>73.5</td>
<td>72.8</td>
<td>70.7</td>
<td>68.8</td>
<td>67.4</td>
<td>66.4</td>
<td>64.7</td>
<td>62.0</td>
<td>60.5</td>
<td>58.5</td>
<td>57.4</td>
</tr>
<tr>
<td>6</td>
<td>67.0</td>
<td>66.3</td>
<td>65.6</td>
<td>64.9</td>
<td>62.3</td>
<td>60.4</td>
<td>58.9</td>
<td>57.6</td>
<td>55.8</td>
<td>52.9</td>
<td>51.0</td>
<td>49.0</td>
<td>47.8</td>
</tr>
<tr>
<td>7</td>
<td>61.3</td>
<td>60.7</td>
<td>59.7</td>
<td>59.1</td>
<td>56.4</td>
<td>54.3</td>
<td>52.6</td>
<td>51.4</td>
<td>49.5</td>
<td>46.4</td>
<td>44.6</td>
<td>42.4</td>
<td>41.2</td>
</tr>
<tr>
<td>8</td>
<td>57.0</td>
<td>56.2</td>
<td>55.4</td>
<td>54.5</td>
<td>51.8</td>
<td>49.7</td>
<td>47.8</td>
<td>46.7</td>
<td>44.6</td>
<td>41.6</td>
<td>39.8</td>
<td>37.6</td>
<td>36.4</td>
</tr>
<tr>
<td>9</td>
<td>53.6</td>
<td>52.5</td>
<td>51.8</td>
<td>51.2</td>
<td>48.1</td>
<td>45.9</td>
<td>44.4</td>
<td>42.9</td>
<td>41.0</td>
<td>38.0</td>
<td>36.1</td>
<td>34.0</td>
<td>32.7</td>
</tr>
<tr>
<td>10</td>
<td>50.6</td>
<td>49.8</td>
<td>49.0</td>
<td>48.3</td>
<td>45.2</td>
<td>42.9</td>
<td>41.4</td>
<td>40.1</td>
<td>38.0</td>
<td>35.0</td>
<td>33.2</td>
<td>31.0</td>
<td>29.8</td>
</tr>
<tr>
<td>11</td>
<td>48.1</td>
<td>47.2</td>
<td>46.5</td>
<td>45.8</td>
<td>42.8</td>
<td>40.6</td>
<td>39.0</td>
<td>37.7</td>
<td>35.6</td>
<td>32.6</td>
<td>30.8</td>
<td>28.7</td>
<td>27.5</td>
</tr>
<tr>
<td>12</td>
<td>46.2</td>
<td>45.2</td>
<td>44.4</td>
<td>43.8</td>
<td>40.7</td>
<td>38.5</td>
<td>36.9</td>
<td>35.5</td>
<td>33.5</td>
<td>30.5</td>
<td>28.8</td>
<td>26.7</td>
<td>25.5</td>
</tr>
<tr>
<td>13</td>
<td>44.4</td>
<td>43.4</td>
<td>42.8</td>
<td>41.9</td>
<td>39.0</td>
<td>36.8</td>
<td>35.1</td>
<td>33.9</td>
<td>31.8</td>
<td>28.8</td>
<td>27.1</td>
<td>25.0</td>
<td>23.9</td>
</tr>
<tr>
<td>14</td>
<td>42.9</td>
<td>42.0</td>
<td>41.3</td>
<td>40.4</td>
<td>37.4</td>
<td>35.2</td>
<td>33.6</td>
<td>32.3</td>
<td>30.4</td>
<td>27.4</td>
<td>25.7</td>
<td>23.6</td>
<td>22.4</td>
</tr>
<tr>
<td>15</td>
<td>41.6</td>
<td>40.7</td>
<td>39.8</td>
<td>39.1</td>
<td>36.2</td>
<td>34.0</td>
<td>32.4</td>
<td>31.1</td>
<td>29.0</td>
<td>26.0</td>
<td>24.3</td>
<td>22.4</td>
<td>21.2</td>
</tr>
<tr>
<td>16</td>
<td>40.4</td>
<td>39.5</td>
<td>38.7</td>
<td>37.9</td>
<td>35.0</td>
<td>32.8</td>
<td>31.1</td>
<td>29.8</td>
<td>27.9</td>
<td>24.9</td>
<td>23.2</td>
<td>21.2</td>
<td>20.1</td>
</tr>
<tr>
<td>17</td>
<td>39.4</td>
<td>38.5</td>
<td>37.6</td>
<td>36.9</td>
<td>33.8</td>
<td>31.7</td>
<td>30.1</td>
<td>28.8</td>
<td>26.8</td>
<td>23.9</td>
<td>22.2</td>
<td>20.3</td>
<td>19.2</td>
</tr>
<tr>
<td>18</td>
<td>38.3</td>
<td>37.4</td>
<td>36.7</td>
<td>35.8</td>
<td>32.9</td>
<td>30.7</td>
<td>29.1</td>
<td>27.8</td>
<td>25.9</td>
<td>22.9</td>
<td>21.3</td>
<td>19.4</td>
<td>18.3</td>
</tr>
<tr>
<td>19</td>
<td>37.4</td>
<td>36.5</td>
<td>35.8</td>
<td>34.9</td>
<td>32.0</td>
<td>29.9</td>
<td>28.3</td>
<td>27.0</td>
<td>25.1</td>
<td>22.2</td>
<td>20.5</td>
<td>18.6</td>
<td>17.5</td>
</tr>
<tr>
<td>20</td>
<td>36.7</td>
<td>35.8</td>
<td>34.9</td>
<td>34.2</td>
<td>31.3</td>
<td>29.1</td>
<td>27.5</td>
<td>26.2</td>
<td>24.3</td>
<td>21.4</td>
<td>19.8</td>
<td>18.0</td>
<td>16.9</td>
</tr>
<tr>
<td>25</td>
<td>33.5</td>
<td>32.5</td>
<td>31.8</td>
<td>31.1</td>
<td>28.1</td>
<td>26.0</td>
<td>24.5</td>
<td>23.3</td>
<td>21.4</td>
<td>18.6</td>
<td>17.0</td>
<td>15.2</td>
<td>14.2</td>
</tr>
<tr>
<td>30</td>
<td>31.2</td>
<td>30.3</td>
<td>29.5</td>
<td>28.8</td>
<td>26.0</td>
<td>23.9</td>
<td>22.3</td>
<td>21.1</td>
<td>19.3</td>
<td>16.6</td>
<td>15.1</td>
<td>13.4</td>
<td>12.5</td>
</tr>
<tr>
<td>35</td>
<td>29.5</td>
<td>28.6</td>
<td>27.9</td>
<td>27.1</td>
<td>24.3</td>
<td>22.2</td>
<td>20.7</td>
<td>19.6</td>
<td>17.8</td>
<td>15.2</td>
<td>13.7</td>
<td>12.1</td>
<td>11.1</td>
</tr>
<tr>
<td>40</td>
<td>28.1</td>
<td>27.3</td>
<td>26.5</td>
<td>25.8</td>
<td>23.0</td>
<td>21.0</td>
<td>19.5</td>
<td>18.4</td>
<td>16.6</td>
<td>14.1</td>
<td>12.7</td>
<td>11.1</td>
<td>10.2</td>
</tr>
<tr>
<td>45</td>
<td>27.0</td>
<td>26.1</td>
<td>25.4</td>
<td>24.7</td>
<td>21.9</td>
<td>20.0</td>
<td>18.5</td>
<td>17.4</td>
<td>15.7</td>
<td>13.2</td>
<td>11.8</td>
<td>10.3</td>
<td>9.4</td>
</tr>
<tr>
<td>50</td>
<td>26.1</td>
<td>25.3</td>
<td>24.6</td>
<td>23.8</td>
<td>21.1</td>
<td>19.1</td>
<td>17.7</td>
<td>16.6</td>
<td>14.9</td>
<td>12.5</td>
<td>11.1</td>
<td>9.6</td>
<td>8.7</td>
</tr>
<tr>
<td>100</td>
<td>21.5</td>
<td>20.7</td>
<td>19.9</td>
<td>19.3</td>
<td>16.7</td>
<td>14.9</td>
<td>13.6</td>
<td>12.5</td>
<td>11.0</td>
<td>8.9</td>
<td>7.7</td>
<td>6.4</td>
<td>5.7</td>
</tr>
</tbody>
</table>

**Table:** 95% quantile inertia on the two first dimensions of 100000 PCA on data with independent variables.
Quality of the representation: $\cos^2$

- For the variables: only well projected variables (high $\cos^2$ between the variable and its projection) can be interpreted!

```
round(res.pca$var$cos2,2)
```

<table>
<thead>
<tr>
<th></th>
<th>Dim.1</th>
<th>Dim.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odor.Intensity.before.shaking</td>
<td>0.01</td>
<td>0.94</td>
</tr>
<tr>
<td>Odor.Intensity.after.shaking</td>
<td>0.01</td>
<td>0.89</td>
</tr>
<tr>
<td>Expression</td>
<td>0.11</td>
<td>0.71</td>
</tr>
</tbody>
</table>

- For the individuals: (same idea) distance between individuals can only be interpreted for well projected individuals

```
round(res.pca$ind$cos2,2)
```

<table>
<thead>
<tr>
<th></th>
<th>Dim.1</th>
<th>Dim.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>S Michaud</td>
<td>0.62</td>
<td>0.07</td>
</tr>
<tr>
<td>S Renaudie</td>
<td>0.73</td>
<td>0.15</td>
</tr>
<tr>
<td>S Trotignon</td>
<td>0.78</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Contribution

⇒ Contribution to the construction of the dimension (percentage of variability):

• for each individual: \( Ctr_q(i) = \frac{F_{iq}^2}{\sum_{i=1}^{I} F_{iq}^2} = \frac{F_{iq}^2}{\lambda_q} \)

⇒ Individuals with a large coordinate contribute the most

\[
\begin{array}{l|cc}
\text{Dim.1} & \text{Dim.2} \\
\hline
\text{S Michaud} & 15.49 & 3.10 \\
\text{S Renaudie} & 15.56 & 5.56 \\
\text{S Trotignon} & 15.46 & 2.43 \\
\end{array}
\]

• for each variable: \( Ctr_q(k) = \frac{G_{kq}^2}{\lambda_q} = \frac{r(x_k,v_q)^2}{\lambda_q} \)

⇒ Variables highly correlated with the principal component contribute the most
Description of the dimensions

By the continuous variables:

- correlation between each variable and the principal component of rank \( q \) is calculated
- correlation coefficients are sorted and significant ones are given

```r
> dimdesc(res.pca)

$Dim.1$quantitative

<table>
<thead>
<tr>
<th>Variable</th>
<th>corr</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>O.candied.fruit</td>
<td>0.93</td>
<td>9.5e-05</td>
</tr>
<tr>
<td>Grade</td>
<td>0.93</td>
<td>1.2e-04</td>
</tr>
<tr>
<td>Surface.feeling</td>
<td>0.89</td>
<td>5.5e-04</td>
</tr>
<tr>
<td>Typicity</td>
<td>0.86</td>
<td>1.4e-03</td>
</tr>
<tr>
<td>O.mushroom</td>
<td>0.84</td>
<td>2.3e-03</td>
</tr>
<tr>
<td>Visual.intensity</td>
<td>0.83</td>
<td>3.1e-03</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>O.plante</td>
<td>-0.87</td>
<td>1.0e-03</td>
</tr>
<tr>
<td>O.flower</td>
<td>-0.89</td>
<td>4.9e-04</td>
</tr>
<tr>
<td>O.passion</td>
<td>-0.90</td>
<td>4.5e-04</td>
</tr>
<tr>
<td>Freshness</td>
<td>-0.91</td>
<td>2.9e-04</td>
</tr>
</tbody>
</table>

$Dim.2$quantitative

<table>
<thead>
<tr>
<th>Variable</th>
<th>corr</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odor.Intensity.before.shaking</td>
<td>0.97</td>
<td>3.1e-06</td>
</tr>
<tr>
<td>Odor.Intensity.after.shaking</td>
<td>0.95</td>
<td>3.6e-05</td>
</tr>
<tr>
<td>Attack.intensity</td>
<td>0.85</td>
<td>1.7e-03</td>
</tr>
<tr>
<td>Expression</td>
<td>0.84</td>
<td>2.2e-03</td>
</tr>
<tr>
<td>Aroma.persistency</td>
<td>0.75</td>
<td>1.3e-02</td>
</tr>
<tr>
<td>Bitterness</td>
<td>0.71</td>
<td>2.3e-02</td>
</tr>
<tr>
<td>Aroma.intensity</td>
<td>0.66</td>
<td>4.0e-02</td>
</tr>
<tr>
<td>Sweetness</td>
<td>-0.78</td>
<td>8.0e-03</td>
</tr>
</tbody>
</table>
```
Description of the dimensions

By the categorical variables:

- Perform a one-way analysis of variance with the coordinates of the individuals \((F_q)\) explained by the categorical variable
  - a F-test by variable
  - for each category, a Student’s \(t\)-test to compare the average of the category with the general mean

\[
> \text{dimdesc(res.pca)}
\]

\[
\begin{array}{lll}
\text{Dim.1$\text{quali}$} & \text{R2} & \text{p.value} \\
\text{Label} & 0.874 & 7.30e-05 \\
\end{array}
\]

\[
\begin{array}{lll}
\text{Dim.1$\text{category}$} & \text{Estimate} & \text{p.value} \\
\text{Vouvray} & 3.203 & 7.30e-05 \\
\text{Sauvignon} & -3.203 & 7.30e-05 \\
\end{array}
\]
Practice with R

1. Choose active variables
2. Scale or not the variables
3. Perform PCA
4. Choose the number of dimensions to interpret
5. Simultaneously interpret the individuals and variables graphs
6. Use indicators to enrich the interpretation

library(FactoMineR)
header=TRUE, sep=";",row.names=1)
res.pca <- PCA(Expert,scale=T,quanti.sup=29:30,quali.sup=1)
res.pca
x11()
barplot(res.pca$eig[,1],main="Eigenvalues",names.arg=1:nrow(res.pca$eig))
plot.PCA(res.pca,habillage=1)
res.pca$ind$coord
res.pca$ind$cos2
res.pca$ind$contrib
plot.PCA(res.pca,axes=c(3,4),habillage=1)
dimdesc(res.pca)
write.infile(res.pca,file="my_FactoMineR_results.csv")  #to export a list
source("http://factominer.free.fr/install-facto.r")
Handling missing values: missMDA package

⇒ Obtain the principal components from observed data with an EM-type algorithm

- Impute missing values with PCA using `imputePCA` function (tuning parameter: number of components)
- Perform the usual PCA on the completed data set

```r
library(missMDA)
data(orange)
nb.dim <- estim_ncpPCA(orange,ncp.max=5)
res.comp <- imputePCA(orange,ncp=2)
res.pca <- PCA(res.comp$completeObs)
```
MCA: problems - objectives

- Individuals study: similarity between individuals (for all the variables) → partition between individuals
  Individuals are different if they don’t take the same levels

- Variables study: find some synthetic variables (continuous variables that sum up categorical variables); link between variables ⇒ levels study

- Categories study:
  - two levels of different variables are similar if individuals that take these levels are the same (ex: 65 years and retired)
  - two levels are similar if individuals taking these levels behave the same way, they take the same levels for the other variables (ex: 60 years and 65 years)

- Link between these studies: characterization of the groups of individuals by the levels (ex: executive dynamic women)
MCA: a PCA on an indicator matrix

- Binary coding of the factors: a factor with $K_j$ levels $\rightarrow K_j$ columns containing binary values, also called dummy variables

$$d^2(i, i') = \frac{1}{J} \sum_{j=1}^{J} \sum_{k=1}^{K_j} \frac{1}{I_k} (x_{ik} - x_{i'k})^2$$
MCA: the superimposed representation

\[ F_{iq} = \frac{1}{\sqrt{\lambda_q}} \sum_{k=1}^{K} K \frac{x_{ik}}{J} G_{kq} \]

\[ G_{kq} = \frac{1}{\sqrt{\lambda_q}} \sum_{i=1}^{I} \frac{x_{ik}}{I_k} F_{iq} \]

⇒ Individual \( i \) at the barycenter of its levels

⇒ Level \( k \) at the barycenter of the individuals who take this level
Multiple Factor Analysis

1. Data - Issues
2. Common Structure
3. Groups Study
4. Partial Analyses
5. Example

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

Benzécri
Multiway data set

Sets

Variables

Individuals

Examples with continuous and/or categorical sets of variables:

- genomic: DNA, protein
- sensory analysis: sensorial, physico-chemical
- survey: student health (addicted consumptions, psychological conditions, sleep, identification, etc.)
- economy: economic indicators for countries by year
Example: gliomas brain tumors

Gliomas: Brain tumors, WHO classification
- astrocytoma (A) .................. x5
- oligodendroglioma (O) ....... x8
- oligo-astrocytoma (OA) ...... x6
- glioblastoma (GBM) .......... x24

43 tumor samples (Bredel et al., 2005)

- Transcriptional modification (RNA), microarrays: 489 variables
- Damage to DNA (CGH array): 113 variables

\[
X_{1_{ij_1}} = \log_2\left(\frac{RNA_{tumor}}{RNA_{ref}}\right)
\]

\[
X_{2_{ij_2}} = \log_2\left(\frac{DNA_{tumor}}{DNA_{ref}}\right)
\]
Objectives

- Study the similarities between individuals with respect to all the variables
- Study the linear relationships between variables

⇒ taking into account the structure on the data (balancing the influence of each group)

- Find the common structure with respect to all the groups - highlight the specificities of each group
- Compare the typologies obtained from each group of variables (separate analyses)
Balancing the groups of variables

MFA is a weighted PCA:

- compute the first eigenvalue $\lambda_1^j$ of each group of variables
- perform a global PCA on the weighted data table:

$$
\begin{bmatrix}
\frac{X_1}{\sqrt{\lambda_1^1}}; & \frac{X_2}{\sqrt{\lambda_1^2}}; & \ldots; & \frac{X_J}{\sqrt{\lambda_1^J}}
\end{bmatrix}
$$

$\Rightarrow$ Same idea as in PCA when variables are standardized: variables are weighted to compute distances between individuals $i$ and $i'$
Balancing the groups of variables

This weighting allows that:

- same weight for all the variables of one group: the structure of the group is preserved
- for each group the variance of the main dimension of variability (first eigenvalue) is equal to 1
- no group can generate by itself the first global dimension
- a multidimensional group will contribute to the construction of more dimensions than a one-dimensional group
Individuals and variables representations

Same representations and same interpretation as in PCA
Groups study

⇒ Synthetic comparison of the groups

⇒ Are the relative positions of individuals globally similar from one group to another? Are the partial clouds similar?

⇒ Do the groups bring the same information?
Similarity between two groups

Measure of similarity between groups $K_j$ and $K_m$:

$$L_g(K_j, K_m) = \sum_{k \in K_j} \sum_{l \in K_m} \text{cov}^2 \left( \frac{x_k}{\sqrt{\lambda_k}}, \frac{x_l}{\sqrt{\lambda_l}} \right)$$

MFA = weighted PCA $\Rightarrow$ first principal component of MFA maximizes

$$\sum_{j=1}^{J} L_g(v_1, K_j) = \sum_{j=1}^{J} \sum_{k \in K_j} \text{cov}^2 \left( \frac{x_k}{\sqrt{\lambda_k}}, v_1 \right)$$

Inertia of $K_j$ projected on $v_1$
Representation of the groups

Group $j$ has the coordinates $(\mathcal{L}_g(v_1, K_j), \mathcal{L}_g(v_2, K_j))$

- 2 groups are all the more close that they induce the same structure
- The 1st dimension is common to all the groups
- 2nd dimension mainly due to CGH

$$0 \leq \mathcal{L}_g(v_1, K_j) = \frac{1}{\lambda_1^j} \sum_{k \in K_j} \text{cov}^2(x_k, v_1) \leq 1$$
 Numeric indicators 

\[ \mathcal{L}_g(K_j, K_j) = \frac{\sum_{k=1}^{K_j} (\lambda_{jk}^j)^2}{(\lambda_{1j}^j)^2} = 1 + \frac{\sum_{k=2}^{K_j} (\lambda_{jk}^j)^2}{(\lambda_{1j}^j)^2} \]

- CGH gives richer description (\(\mathcal{L}_g\) greater)
- RV: a standardized \(\mathcal{L}_g\)
- CGH and expr are not linked (RV=0.36)
- CGH closest to the overall (RV=0.90)

Contribution of each group to each component of the MFA

- Similar contribution of the 2 groups to the first dimension
- Second dimension only due to CGH
The RV coefficient

\( X_{j(I \times K_j)} \) and \( X_{m(I \times K_m)} \) not directly comparable
\( W_{j(I \times I)} = X_jX'_j \) and \( W_{m(I \times I)} = X_mX'_m \) can be compared
Inner product matrices = relative position of the individuals

Covariance between two groups:

\[
< W_j, W_m >= \sum_{k \in K_j} \sum_{l \in K_m} \text{cov}^2(x_k, x_l)
\]

Correlation between two groups:

\[
RV(K_j, K_m) = \frac{< W_j, W_m >}{\| W_j \| \| W_m \|} \quad 0 \leq RV \leq 1
\]

\( RV = 0 \): variables of \( K_j \) are uncorrelated with variables of \( K_m \)
\( RV = 1 \): the two clouds of points are homothetic
Partial analyses

- Comparison of the groups through the individuals

  ⇒ Comparison of the typologies provided by each group in a common space
  ⇒ Are there individuals very particular with respect to one group?

- Comparison of the separate PCA
Projection of partial points

\[ R^K = \oplus R^K_j \]

<table>
<thead>
<tr>
<th>G1</th>
<th>G2</th>
<th>G3</th>
</tr>
</thead>
<tbody>
<tr>
<td>xxxxxxx</td>
<td>xxxxx</td>
<td>xxxxx</td>
</tr>
<tr>
<td>xxxxxxx</td>
<td>xxxxx</td>
<td>xxxxx</td>
</tr>
<tr>
<td>xxxxxxx</td>
<td>xxxxx</td>
<td>xxxxx</td>
</tr>
</tbody>
</table>

Data

<table>
<thead>
<tr>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000000</td>
</tr>
<tr>
<td>00000000</td>
</tr>
<tr>
<td>00000000</td>
</tr>
</tbody>
</table>

Projection of group 1

<table>
<thead>
<tr>
<th>i'1</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000000</td>
</tr>
<tr>
<td>00000000</td>
</tr>
<tr>
<td>00000000</td>
</tr>
</tbody>
</table>

Projection of group 2

<table>
<thead>
<tr>
<th>i'2</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000000</td>
</tr>
<tr>
<td>00000000</td>
</tr>
<tr>
<td>00000000</td>
</tr>
</tbody>
</table>

Projection of group 3

<table>
<thead>
<tr>
<th>i'3</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000000</td>
</tr>
<tr>
<td>00000000</td>
</tr>
<tr>
<td>00000000</td>
</tr>
</tbody>
</table>

MFA individuals configuration

Partial point 1

Mean point

Partial point 2

Partial point 3
Partial points

opinion

attitude

individuals

What you think

behavioral conflict

What you do
Partial points

What you expected for the tutorial

What you have learned during the tutorial

Tutorial participants

What you expected for the tutorial

What you have learned during the tutorial

What you expected for the tutorial

What you have learned during the tutorial
Partial points

What you expected for the tutorial

What you have learned during the tutorial

F1

F2

Tutorial participants

Disappointed learner

Happy learner

What you expected for the tutorial

What you have learned during the tutorial
Representation of the partial points

- an individual is at the barycentre of its partial points
- an individual is all the more "homogeneous" that its superposed representations are close

(res.mfa$ind$within.inertia)
Representation of the partial components

Do the separate analyses give similar dimensions as MFA?

Sets

Variables

Individuals

1

j

1

k

K_j

x_{ik}

1

q

Q

1

q

Q

1
i

I

1

q

Q

1

q

Q

1

q

Q

PCA
Representation of the partial components

- The first dimension of each group is well projected
- CGH has same dimensions as MFA
Representation of the partial components

Separate PCA maps that can be compared to the MFA map slide 7
Use of biological knowledge

Genes can be grouped by gene ontology (GO) biological process

**GO:0006928** cell motility
- ANXA1
- CALD1
- EGFR
- ENPP2
- FN1
- FPRL2
- LSP1
- MSN
- PDPN
- PLAUR
- PRSS3
- SAA2
- SPINT2
- TNFRSF12A
- VEGF
- WASF1
- YARS

**GO:0009966** regulation of signal transduction
- CASP1
- EDG2
- F2R
- HCLS1
- HMOX1
- IGFBP3
- IQSEC1
- LYN
- MALT1
- TCF7L1
- TNFAIP3
- TRIO
- VEGF
- YWHAG
- YWHAH

**GO:0052276** chromosome organisation and biogenesis
- CBX6
- NUSAP1
- PCOLN3
- PTTG1
- SUV39H1
- TCF7L1
- TSPYL1
Use of biological knowledge

- Biological processes considered as supplementary groups of variables

Modular approach

Tumors

\[ X_{1ij_1} \]

\[ X_{2ij_2} \]

\[ \text{'-omics' data} \]

\[ \text{Modules} \]

\[ \text{M1, M2, M3, etc.} \]
Use of biological knowledge

Many biological processes induce the same structure on the individuals than MFA
Back to the wine example!

<table>
<thead>
<tr>
<th></th>
<th>Expert (27)</th>
<th>Consumer (15)</th>
<th>Student (15)</th>
<th>Preference (60)</th>
<th>Label (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>wine 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wine 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wine 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Objectives:

- How are the products described by the panels?
- Do the panels describe the products in a same way? Is there a specific description done by one panel?
Practice with R

1. Define groups of active and supplementary variables
2. Scale or not the variables
3. Perform MFA
4. Choose the number of dimensions to interpret
5. Simultaneously interpret the individuals and variables graphs
6. Study the groups of variables
7. Study the partial representations
8. Use indicators to enrich the interpretation
library(FactoMineR)
               header=TRUE, sep=";", row.names=1)
Consu <- read.table(".../Consumer_wine.csv",header=T,sep=";",row.names=1)
Stud <- read.table(".../Student_wine.csv",header=T,sep=";",row.names=1)
Pref <- read.table(".../Pref_wine.csv",header=T,sep=";",row.names=1)

palette(c("black","red","blue","orange","darkgreen","maroon","darkviolet"))
complet <- cbind.data.frame(Expert[,1:28],Consu[,2:16],Stud[,2:16],Pref)
res.mfa <- MFA(complet,group=c(1,27,15,15,60),type=c("n",rep("s",4)),
               num.group.sup=c(1,5),graph=FALSE,
               name.group=c("Label","Expert","Consumer","Student","Preference"))
plot(res.mfa,choix="group",palette=palette())
plot(res.mfa,choix="var",invisible="sup",hab="group",palette=palette())
plot(res.mfa,choix="var",invisible="actif",lab.var=FALSE,palette=palette())
plot(res.mfa,choix="ind",partial="all",habillage="group",palette=palette())
plot(res.mfa,choix="axes",habillage="group",palette=palette())
dimdesc(res.mfa)
write.infile(res.pca,file="my_FactoMineR_results.csv") #to export a list
The two labels are well separated

Vouvray are sensorially more different

Several groups of wines, ...
**Representation of the active variables**

- Principal Component Analysis
- Multiple Factor Analysis
- Clustering and Principal Component Methods

```
Dim 1 (42.52 %)
Dim 2 (24.42 %)
Expert
Consumer
Student
```

```
O.Intensity.before.shaking
O.Intensity.after.shaking
O.fruity
O.passion
O.citrus
O.candied.fruit
O.vanilla
O.wooded
O.mushroom
O.plante
O.flower
O.alcohol

Typicity
Attack.intensity
Sweetness
Acidity
Bitterness
Astringency
Freshness
Oxidation
Smoothness
A.intensity
A.persistency
Visual.intensity
Grade
Surface.feeling
O.candied.fruit
```

```
O.Intensity.before.shaking_C
O.Intensity.after.shaking_C
O.alcohol_C
O.plante_C
O.mushroom_C
O.passion_C
O.Typicity_C
A.intensity_C
Sweetness_C
Acidity_C
Bitterness_C
Astringency_C
Balance_C
Typical_C
Typicity_C
```

```
O.Intensity.before.shaking_S
O.Intensity.after.shaking_S
O.alcohol_S
O.plante_S
O.mushroom_S
O.passion_S
O.Typicity_S
A.intensity_S
Sweetness_S
Acidity_S
Bitterness_S
Astringency_S
Balance_S
Typical_S
```

```
-1.5 -1.0 -0.5 0.0 0.5 1.0 1.5
-1.0 -0.5 0.0 0.5 1.0
```

```
Dim 2 (24.42 %)
```

```
Dim 1 (42.52 %)
```

```
67 / 98
```
Representation of the active variables
Representation of the groups

- 2 groups are all the more close that they induce the same structure
- The 1st dimension is common to all the panels
- 2nd dimension mainly due to the experts
- Preference linked to sensory description
Representation of the partial dimensions

- The two first dimensions of each group are well projected
- Consumer has same dimensions as MFA
Preferences are linked to sensory description
The favourite wine is *Vouvray Aubussière Silex*
Helps to interpret

- Contribution of each group of variables to each component of the MFA

```r
> res.mfa$group$contrib
   Dim.1 Dim.2 Dim.3
Expert 30.5  46.0  33.7
Consumer 33.2  23.1  31.2
Student  36.3  30.9  35.1
```

- Similar contribution of the 3 groups to the first dimension
- Second dimension mainly due to the expert

- Correlation between the global cloud and each partial cloud

```r
> res.mfa$group$correlation
   Dim.1 Dim.2 Dim.3
Expert   0.95  0.95  0.96
Consumer 0.95  0.83  0.87
Student  0.99  0.99  0.84
```

First components are highly linked to the 3 groups: the 3 clouds of points are nearly homothetic
### Similarity measures between groups

```
> res.mfa$group$Lg
                           Expert  Consumer  Student  Preference  Label  MFA
Expert                   1.45      0.94      1.17       1.01    0.89    1.33
Consumer                 0.94      1.25      1.04       1.11    0.28    1.21
Student                  1.17      1.04      1.29       1.03    0.62    1.31
Preference               1.01      1.11      1.03       1.47    0.37    1.18
Label                    0.89      0.28      0.62       0.37    1.00    0.67
MFA                      1.33      1.21      1.31       1.18    0.67    1.44
```

```
> res.mfa$group$RV
                           Expert  Consumer  Student  Preference  Label  MFA
Expert                   1.00      0.70      0.85       0.69    0.74    0.92
Consumer                 0.70      1.00      0.82       0.82    0.25    0.90
Student                  0.85      0.82      1.00       0.75    0.55    0.96
Preference               0.69      0.82      0.75       1.00    0.31    0.81
Label                    0.74      0.25      0.55       0.31    1.00    0.56
MFA                      0.92      0.90      0.96       0.81    0.56    1.00
```

- Expert gives a richer description ($L_g$ greater)
- Groups Student and Expert are linked (RV = 0.85)
- Group Student is the closest to the overall (RV = 0.96)
To go further

- Mixed data: MFA with 1 group = 1 variable
  if there are only continuous variables, PCA is recovered; if
  there are only categorical variables, MCA is recovered
  a specific function: AFDM

- MFA used for methodological purposes:
  - comparison of coding (continuous or categorical)
  - comparison between preprocessing (standardized PCA and
    unstandardized PCA)
  - comparison of results from different analyses

- Hierarchical Multiple Factor Analysis
  Takes into account a hierarchy on the variables: variables are
  grouped and subgrouped (like in questionnaires structured in
  topics and subtopics)
Clustering and Principal Component Methods

1. Clustering Methods
2. Principal Components Methods as a Preprocessing Step
3. Graphical Complementarity
Unsupervised classification

- Data set: table individuals × variables (or a distance matrix)
- Objective: to produce homogeneous groups of individuals (or groups of variables)
- Two kinds of clustering to define two structures on individuals: hierarchy or partition
Hierarchical Clustering

Principle: sequentially agglomerate (clusters of) individuals using

- a distance between individuals: City block, Euclidean
- an agglomerative criterion: single linkage, complete linkage, average linkage, Ward’s criterion

Representation with a dendrogram

⇒ Euclidean distance is used in principal component methods
⇒ Ward’s criterion is based on multidimensional variance (inertia) which is the core of principal component methods
Ascending Hierarchical Clustering

AHC algorithm:

- Compute the Euclidean distance matrix \((I \times I)\)
- Consider each individual as a cluster
- Merge the two clusters \(A\) and \(B\) which are the closest with respect to the Ward’s criterion:

\[
\Delta_{\text{ward}}(A, B) = \frac{l_A l_B}{l_A + l_B} d^2(\mu_A, \mu_B)
\]

with \(d\) the Euclidean distance, \(\mu_A\) the barycentre and \(l_A\) the cardinality of the set \(A\)
- Repeat until the number of clusters is equal to one
Ward’s criterion

- Individuals can be represented by a cloud of points in $\mathbb{R}^K$
- Total inertia = multidimensional variance

With $Q$ groups of individuals, inertia can be decomposed as:

$$
\sum_{k=1}^{K} \sum_{q=1}^{Q} \sum_{i=1}^{l_q} (x_{iqk} - \bar{x}_k)^2 = \sum_{k=1}^{K} \sum_{q=1}^{Q} l_q (\bar{x}_{qk} - \bar{x}_k)^2 + \sum_{k=1}^{K} \sum_{q=1}^{Q} \sum_{i=1}^{l_q} (x_{iqk} - \bar{x}_{qk})^2
$$

Total inertia = Between inertia + Within inertia
Ward’s criterion

Step 1: 1 cluster = 1 individual
Within = 0
Between = Total

Step I: only 1 cluster
Within = Total
Between = 0

Step I-2: 3 clusters

Step I-1: 2 clusters to define

⇒ Ward minimizes the increasing of within inertia
K-means algorithm

1. Choose $Q$ points at random (the barycentre)
2. Affect the points to the closest barycentre
3. Compute the new barycentre
4. Iterate 2 and 3 until convergence
K-means algorithm

1. Choose $Q$ points at random (the barycentre)
2. Affect the points to the closest barycentre
3. Compute the new barycentre
4. Iterate 2 and 3 until convergence
K-means algorithm

1. Choose $Q$ points at random (the barycentre)
2. Affect the points to the closest barycentre
3. Compute the new barycentre
4. Iterate 2 and 3 until convergence
**K-means algorithm**

1. Choose \( Q \) points at random (the barycentre)
2. Affect the points to the closest barycentre
3. Compute the new barycentre
4. Iterate 2 and 3 until convergence
K-means algorithm

1. Choose \( Q \) points at random (the barycentre)
2. Affect the points to the closest barycentre
3. Compute the new barycentre
4. Iterate 2 and 3 until convergence
K-means algorithm

1. Choose $Q$ points at random (the barycentre)
2. Affect the points to the closest barycentre
3. Compute the new barycentre
4. Iterate 2 and 3 until convergence
K-means algorithm

1. Choose $Q$ points at random (the barycentre)
2. Affect the points to the closest barycentre
3. Compute the new barycentre
4. Iterate 2 and 3 until convergence
K-means algorithm

1. Choose $Q$ points at random (the barycentre)
2. Affect the points to the closest barycentre
3. Compute the new barycentre
4. Iterate 2 and 3 until convergence
K-means algorithm

1. Choose $Q$ points at random (the barycentre)
2. Affect the points to the closest barycentre
3. Compute the new barycentre
4. Iterate 2 and 3 until convergence
PCA as a preprocessing

With continuous variables:
⇒ AHC and k-means onto the raw data
⇒ AHC or k-means onto principal components

PCA transforms the raw variables into orthogonal principal components $F_1, \ldots, F_K$ with decreasing variance $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_K$

⇒ Keeping the first components makes the clustering more robust
⇒ But, how many components do you keep to denoise?
MCA as a preprocessing

Clustering on categorical variables: which distance to use?

- with two categories: Jaccard index, Dice’s coefficient, simple match, etc. Indices well-fitted for presence/absence data
- with more than 2 categories: use for example the $\chi^2$-distance

Using the $\chi^2$-distance ⇔ computing distances from all the principal components obtained from MCA

In practice, MCA is used as a preprocessing in order to
- transform categorical variables in continuous ones
- delete the last dimensions to make the clustering more robust
MFA as a preprocessing

MFA balances the influence of the groups when computing distances between individuals

\[ d^2(i, i') = \sum_{j=1}^{J} \frac{1}{\sqrt{\lambda_j}} \sum_{k=1}^{K_j} (x_{ik} - x_{i'k})^2 \]

AHC or k-means onto the first principal components \((F_1, \ldots, F_Q)\) obtained from MFA allows to

- take into account the groups structure in the clustering
- make the clustering more robust by deleting the last dimensions
Back to the wine data!

AHC onto the first 5 principal components from MFA

Individuals are sorted according to their coordinate $F_1$
Why sorting the tree?

X <- c(6,7,2,0,3,15,11,12)
names(X) <- X
library(cluster)
par(mfrow=c(1,2))
plot(as.dendrogram(agnes(X)))
plot(as.dendrogram(agnes(sort(X)))))
Partition from the tree

An empirical number of clusters is suggested \( \min_q \frac{W_q - W_{q+1}}{W_{q-1} - W_q} \)
Partition from the tree

An empirical number of clusters is suggested \( \min_q \frac{W_q - W_{q+1}}{W_{q-1} - W_q} \)
Hierarchical clustering on the principal component map

Hierarchical tree gives an idea of the other dimensions
Partition on the principal component map

Continuous view (principal components) and discontinuous (clusters)
Cluster description by variables

\[ v.\text{test} = \frac{\bar{x}_q - \bar{x}}{\sqrt{\frac{s^2}{I_q} \left( \frac{I-I_q}{I-1} \right)}} \sim \mathcal{N}(0,1) \quad H_0 : \bar{x}_q = \bar{x} \]

with \( \bar{x}_q \) the mean of variable \( x \) in cluster \( q \), \( \bar{x} \) (s) the mean (standard deviation) of the variable \( x \) in the data set, \( I_q \) the cardinal of cluster \( q \)

\( \text{desc.var} \text{quantile}'2' \)

<table>
<thead>
<tr>
<th>Category</th>
<th>v.test</th>
<th>Mean in Overall category</th>
<th>sd in category</th>
<th>Overall sd</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.passion_C</td>
<td>2.58</td>
<td>6.17</td>
<td>4.61</td>
<td>0.79</td>
<td>1.18</td>
</tr>
<tr>
<td>0.citrus</td>
<td>2.50</td>
<td>5.40</td>
<td>3.66</td>
<td>0.22</td>
<td>1.37</td>
</tr>
<tr>
<td>0.passion_S</td>
<td>2.45</td>
<td>5.69</td>
<td>4.18</td>
<td>0.54</td>
<td>1.20</td>
</tr>
<tr>
<td>Typicity</td>
<td>-2.42</td>
<td>1.36</td>
<td>3.91</td>
<td>0.72</td>
<td>2.07</td>
</tr>
<tr>
<td>0.candied.fruit</td>
<td>-2.44</td>
<td>0.78</td>
<td>2.58</td>
<td>0.16</td>
<td>1.45</td>
</tr>
<tr>
<td>0.alcohol_S</td>
<td>-2.48</td>
<td>3.98</td>
<td>4.33</td>
<td>0.13</td>
<td>0.28</td>
</tr>
<tr>
<td>Surface.feeling</td>
<td>-2.52</td>
<td>2.63</td>
<td>3.62</td>
<td>0.12</td>
<td>0.77</td>
</tr>
</tbody>
</table>
Cluster description

• by the principal components (individuals coordinates) : same description than for continuous variables

$desc.axes$quanti$‘2‘

<table>
<thead>
<tr>
<th>v.test</th>
<th>Mean in Overall</th>
<th>sd in Overall</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>category</td>
<td>mean</td>
<td>category</td>
<td>sd</td>
</tr>
<tr>
<td>Dim.2</td>
<td>2.20</td>
<td>1.39</td>
<td>7.77e-17</td>
</tr>
</tbody>
</table>

• by categorical variables : chi-square and hypergeometric test

⇒ Active and supplementary elements are used
⇒ Only significant results are presented
Cluster description by individuals

- **paragon**: the closest individuals to the barycentre of the cluster
  \[
  \min_{i \in q} d(x_i, \mu_q) \quad \text{with} \quad \mu_q \text{ the barycentre of cluster } q
  \]

- **specific individuals**: the furthest individuals to the barycentres of the other clusters (the individuals sorted according to their distance from the highest to the smallest to the closest barycentre)
  \[
  \max_{i \in q} \min_{q' \neq q} d(x_i, \mu_{q'})
  \]

---

desc.ind$para
cluster: 2
S Renaudie   S Trotignon   S Michaud
0.1002890   0.3101154   0.3640145

-----------------------------

desc.ind$dist
cluster: 2
S Trotignon   S Renaudie   S Michaud
1.934103     1.687849     1.265386

-----------------------------
Complementarity between hierarchical clustering and partitioning

- Partitioning after AHC: the k-means algorithm is initialized from the barycentres of the partition obtained from the tree
  - consolidate the partition
  - loss of the hierarchy

- AHC with many individuals: time-consuming
  ⇒ partitioning before AHC
  - compute k-means with approximately 100 clusters
  - AHC on the weighted barycentres obtained from the k-means
  ⇒ top of the tree is approximately the same
Practice with R

res.hcpc <- HCPC(res.mfa)

### Example of clustering on categorical data

data(tea)
res.mca <- MCA(tea, quanti.sup=19, quali.sup=20:36)
plot(res.mca, invisible=c("var", "quali.sup", "quanti.sup"), cex=0.7)
plot(res.mca, invisible=c("ind", "quali.sup", "quanti.sup"), cex=0.8)
plot(res.mca, invisible=c("quali.sup", "quanti.sup"), cex=0.8)
dimdesc(res.mca)

res.mca <- MCA(tea, quanti.sup=19, quali.sup=20:36, ncp=10)
res.hcpc <- HCPC(res.mca)
CARME conference

International conference on Correspondence Analysis and Related Methods
Agrocampus Rennes (France), February 8-11, 2011

R tutorials for corresp. ana. and related methods of visualization:
- S. Dray: multivariate analysis of ecological data with ade4
- O. Nenadić & M. Greenacre: correspondence analysis with ca
- S. Lê: from one to multiple data tables with FactoMineR
- J. de Leeuw & P. Mair: multidimensional scaling using majorisation with smacof

Invited speakers: Monica Bécue, Cajo ter Braak, Jan de Leeuw, Stéphane Dray, Michael Friendly, Patrick Groenen, Pieter Kroonenberg
Bibliography

Packages’ bibliography

http://cran.r-project.org/web/views/Multivariate.html
http://cran.r-project.org/web/views/Cluster.html

- ade4 package: data analysis functions to analyse Ecological and Environmental data in the framework of Euclidean Exploratory methods
  http://pbil.univ-lyon1.fr/ADE-4
- ca package (Greenacre and Nenadic) deals with simple, multiple and joint correspondence analysis
- cluster package: basic and hierarchical clustering
- dynGraph package: visualization software to explore interactively graphical outputs provided by multidimensional methods
  http://dyngraph.free.fr
- FactoMineR package
  http://factominer.free.fr
- hopach package: builds hierarchical tree of clusters
- missMDA package: imputes missing values with multivariate data analysis methods
FactoMineR

A website with documentation, examples, data sets:
http://factominer.free.fr

How to install the Rcmdr menu:
copy and paste the following line of code in a R session

source("http://factominer.free.fr/install-facto.r")

A book: