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

$\Rightarrow$ Multidimensional descriptive methods
$\Rightarrow$ Graphical representations

Principal Component Analysis

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

Principal Component Analysis

Dimensionality reduction $\Rightarrow$ 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

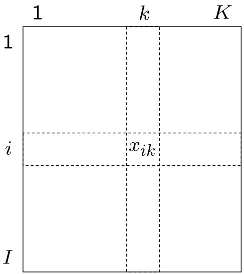


Figure: Data table in PCA

Many examples:

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

Wine data

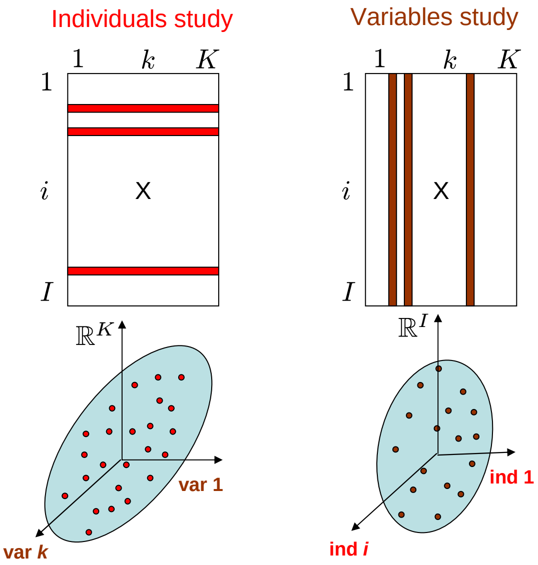
- 10 individuals (rows): white wines from Val de Loire
- 30 variables (columns):
  - 27 continuous variables: sensory descriptors
  - 2 continuous variables: odour and overall preferences
  - 1 categorical variable: label of the wines (Vouvray - Sauvignon)

	O.fruity	O.passion	O.citrus	...	Sweetness	Acidity	Bitterness	Astringency	Aroma.intensity	Aroma.persistency	Visual.intensity	Odor.preferene	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

Problems - objectives

- Individuals study:
  similarity between individuals with respect to all the variables
  $\Rightarrow$ partition between individuals
- Variables study:
  linear relationships between variables $\Rightarrow$ 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



Preprocessing

⇒ Similarity between individuals: Euclidean distance

- Choosing active variables

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

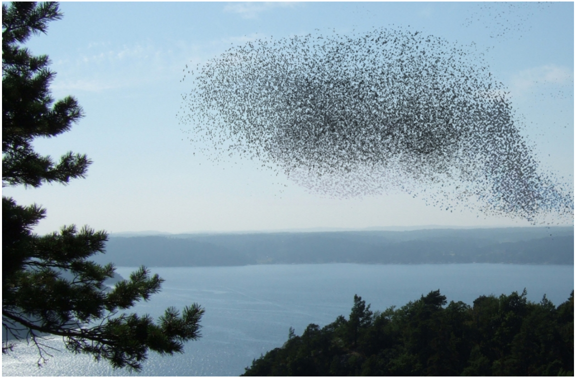
- Variables are always centred

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

- Standardizing variables or not?

$$d^2(i,i') = \sum_{k=1}^{K}\frac{1}{s_k^2}(x_{ik} - x_{i'k})^2$$

Individuals cloud



- 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

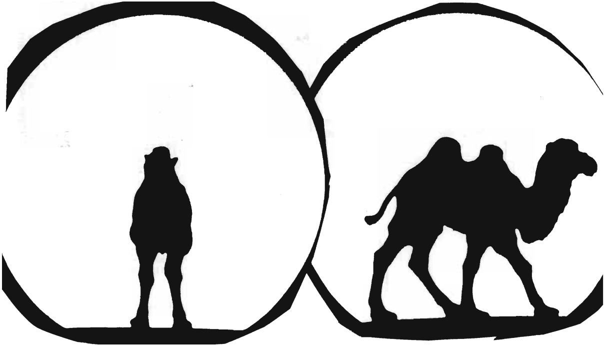
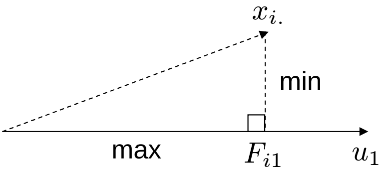


Figure: Camel vs dromedary?

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

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

$$u_1 = \arg\max_{u_1 \in \mathbb{R}^K} (var(F_{.1})) = \arg\max_{u_1 \in \mathbb{R}^K} (var(X u_1)) \text{ with } u_1' u_1 = 1$$

$\Rightarrow$ u_1 first eigenvector of the correlation matrix associated with the largest eigenvalue λ_1: $S u_1 = \lambda_1 u_1$

$$Var(F_{.1}) = var(X u_1) = 1/I \; u_1' X' X u_1 = u_1' S u_1 = \lambda_1 u_1' u_1 = \lambda_1$$

Fit the individuals cloud

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

Representation quality: dimensionality reduction $\Rightarrow$ loosing information

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

$$\frac{1}{I}\|x_{i.} - g\|^2 = tr(S) = \sum_{k=1}^{K} \lambda_k \quad (= K)$$

- Variance of the projected individuals cloud (Q-dimensional representation): $var(F_1) + var(F_2) + \dots + var(F_Q)$

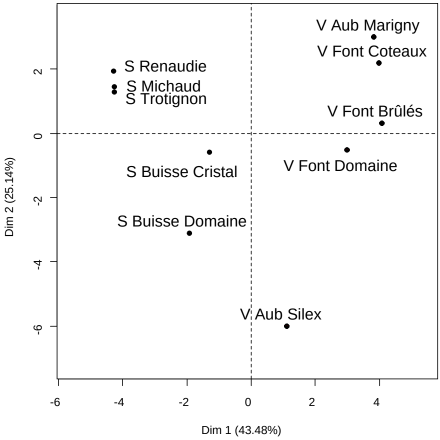
$\Rightarrow$ Percentage of variance explained: $\frac{\sum_{k=1}^{Q} \lambda_k}{\sum_{k=1}^{K} \lambda_k}$

Example: wine data

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

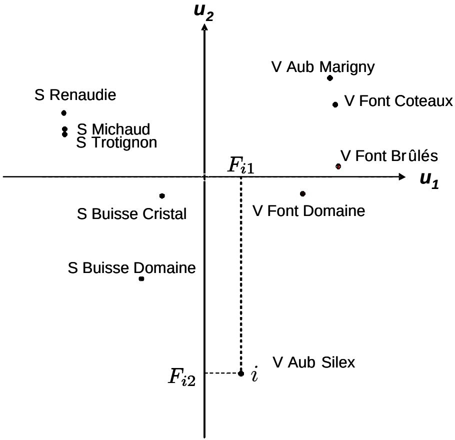
	O.fruity	O.passion	O.citrus	...	Sweetness	Acidity	Bitterness	Astringency	Aroma.intensity	Aroma.persistency	Visual.intensity	Odor.preferene	Overall.preference	Label
S Michaud	4.3	2.4	5.7	...	3.5	5.9	4.1	1.4	7.1	6.7	5.0	6.0	5.0	Sauvignon
S Renaudie	4.4	3.1	5.3	...	3.3	6.8	3.8	2.3	7.2	6.6	3.4	5.4	5.5	Sauvignon
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

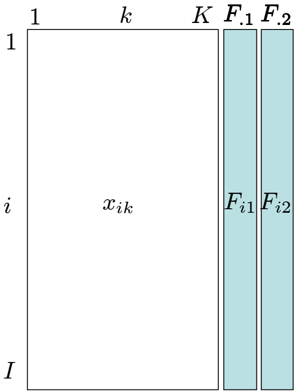
Example: graph of the individuals



⇒ Need variables to interpret the dimensions of variability

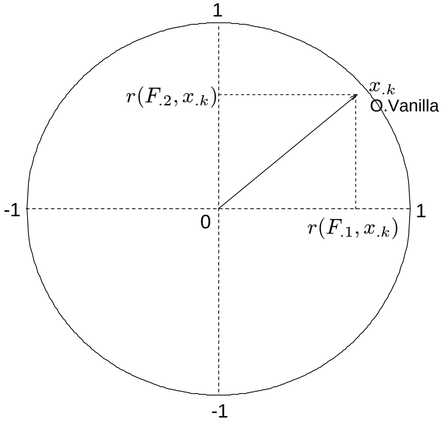
Individuals coordinates considered as variables





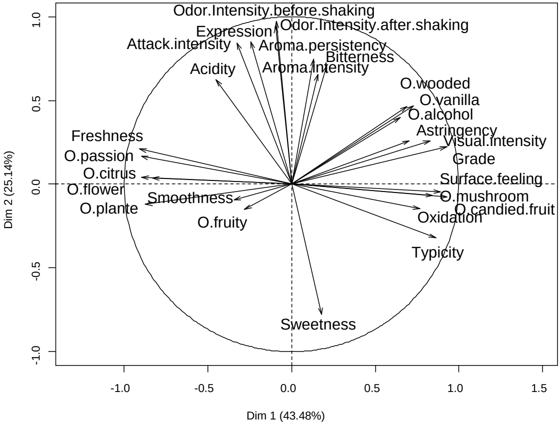
Interpretation of the individuals graph with the variables

- Correlation between variable $x_{.k}$ and $F_{.1}$ (and $F_{.2}$)

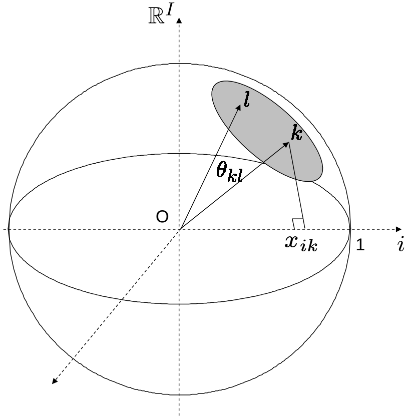


$\Rightarrow$ Correlation circle

Interpretation of the individuals graph with the variables



Cloud of variables

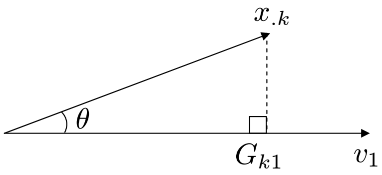


Since variables are centred:

$$cos(\theta_{kl}) = \frac{< x_{.k}, x_{.l} >}{\|x_{.k}\| \ \|x_{.l}\|} = \frac{\sum_{i=1}^{I} x_{ik} x_{il}}{\sqrt{(\sum_{i=1}^{I} x_{ik}^2)(\sum_{i=1}^{I} x_{il}^2)}} = r(x_{.k}, x_{.l})$$

Fit the variables cloud

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



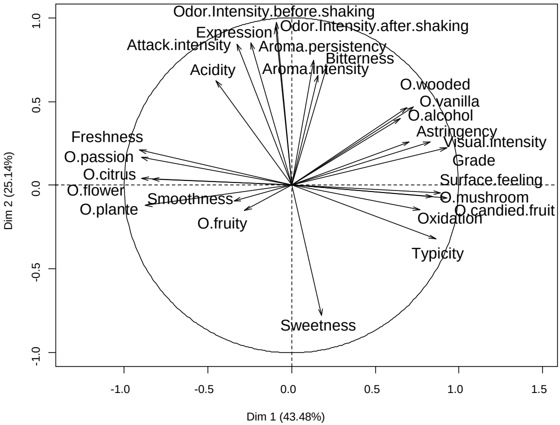
$$
\begin{aligned}
P_{v_1}(x_{.k}) &= v_1(v_1'v_1)^{-1}v_1'x_{.k} \\
G_{k1} &= 1/I \;< v_1, x.k > \\
G_{k1} &= 1/I \;\frac{< v_1, x.k >}{\|v_1\| \; \|x_{.k}\|}
\end{aligned}
$$

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

$\Rightarrow$ v_1 is the best synthetic variable

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

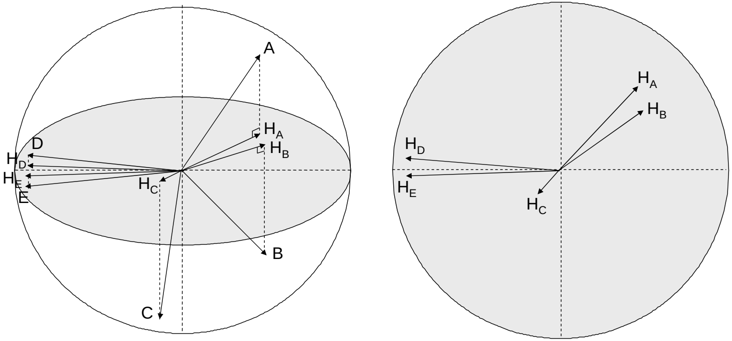
Fit the variables cloud



⇒ Same representation! What a wonderful result!

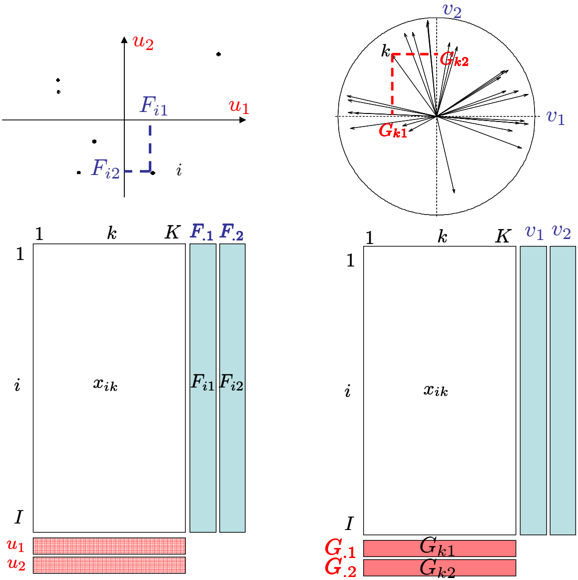
Projections...

$r(A,B) = cos(\theta_{A,B})$
$cos(\theta_{A,B}) \approx cos(\theta_{H_A,H_B})$ 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:

$$
\begin{aligned}
v &= \frac{1}{\sqrt{\lambda}} F \quad &\Rightarrow G = X'v = \frac{1}{\sqrt{\lambda}} X'F \\
u &= \frac{1}{\sqrt{\lambda}} G \quad &\Rightarrow F = Xu = \frac{1}{\sqrt{\lambda}} XG
\end{aligned}
$$

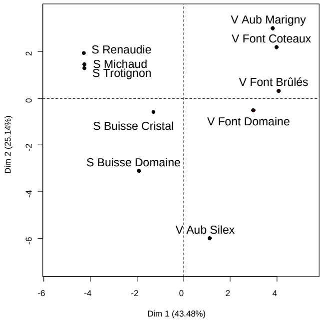
$$\boxed{F_{iq} = \frac{1}{\sqrt{\lambda_q}} \sum_{k=1}^{K} x_{ik} G_{kq}} \qquad \boxed{G_{kq} = \frac{1}{\sqrt{\lambda_q}} \sum_{i=1}^{I} x_{ik} F_{iq}}$$

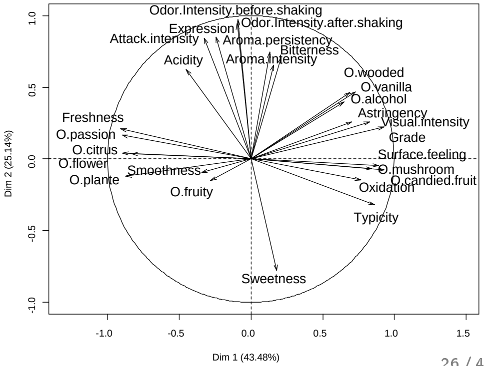
$F_{.q}$: principal components, scores
$G_{.q}$: correlations between variables and principal components

Link between the two representations: transition formulae

$$F_{iq} = \frac{1}{\sqrt{\lambda_q}} x_{ik} G_{kq} \qquad G_{kq} = \frac{1}{\sqrt{\lambda_q}} x_{ik} F_{iq}$$

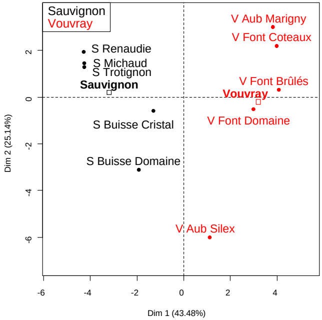
What does it mean? An individual is at the same side as the variables for which it takes high values

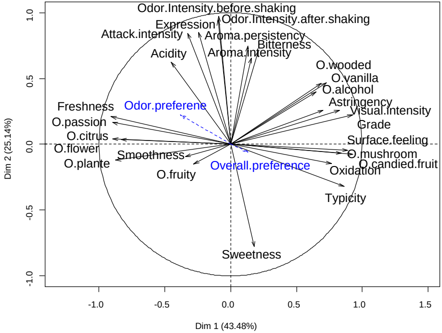




Supplementary information

- 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

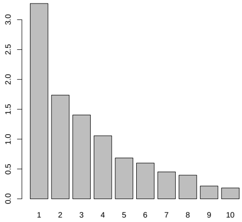




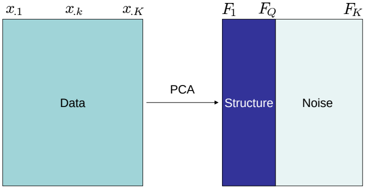
⇒ Supplementary information do not create the dimensions

Choosing the number of components

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



Two objectives:
$\Rightarrow$ Interpretation
$\Rightarrow$ Separate structure and noise



Percentage of variance obtained under independence

⇒ Is there a structure on my data?

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

Table: 95 % quantile inertia on the two first dimensions of 10000 PCA on data with independent variables

Percentage of variance obtained under independence

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

Table: 95 % quantile inertia on the two first dimensions of 10000 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)
                              Dim.1 Dim.2
Odor.Intensity.before.shaking  0.01  0.94
Odor.Intensity.after.shaking   0.01  0.89
Expression                     0.11  0.71
```

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

```
round(res.pca$ind$cos2,2)
              Dim.1 Dim.2
S Michaud      0.62  0.07
S Renaudie     0.73  0.15
S Trotignon    0.78  0.07
```

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

```
round(res.pca$ind$contrib,2)
                  Dim.1 Dim.2
S Michaud         15.49  3.10
S Renaudie        15.56  5.56
S Trotignon       15.46  2.43
```

- 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

```
> dimdesc(res.pca)
            $Dim.1$quanti                                           $Dim.2$quanti
                 corr p.value                                  corr p.value
O.candied.fruit  0.93 9.5e-05  Odor.Intensity.before.shaking 0.97 3.1e-06
Grade            0.93 1.2e-04  Odor.Intensity.after.shaking  0.95 3.6e-05
Surface.feeling  0.89 5.5e-04  Attack.intensity              0.85 1.7e-03
Typicity         0.86 1.4e-03  Expression                    0.84 2.2e-03
O.mushroom       0.84 2.3e-03  Aroma.persistency             0.75 1.3e-02
Visual.intensity 0.83 3.1e-03  Bitterness                    0.71 2.3e-02
   ...            ...     ...  Aroma.intensity               0.66 4.0e-02
O.plante        -0.87 1.0e-03
O.flower        -0.89 4.9e-04
O.passion       -0.90 4.5e-04
Freshness       -0.91 2.9e-04  Sweetness                    -0.78 8.0e-03
```

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

```
> dimdesc(res.pca)
Dim.1$quali
             R2      p.value
Label     0.874     7.30e-05

Dim.1$category
           Estimate      p.value
Vouvray       3.203     7.30e-05
Sauvignon    -3.203     7.30e-05
```

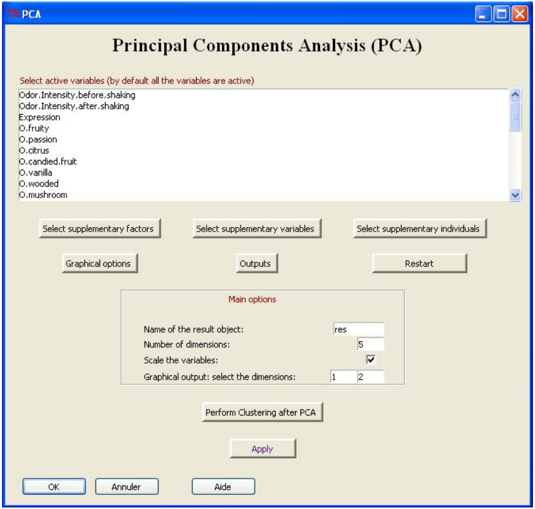
Practice with R

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)
Expert <- read.table("http://factominer.free.fr/useR2010/Expert_wine.csv",
   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
```

Practice with GUI

```
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

```
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

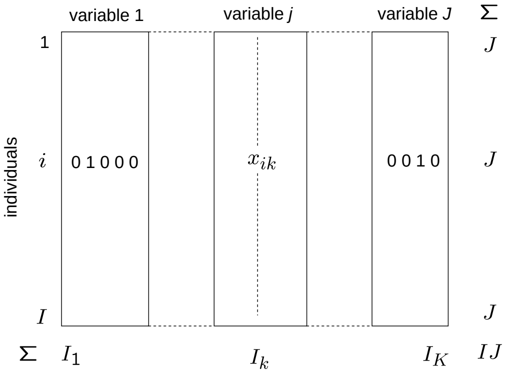
- Individuals study: similarity between individuals (for all the variables) $\rightarrow$ partition between individuals
  Individuals are different if they don't take the same levels

- Variables study: find some synthetic variables (continuous variables that sum up categorical variables); link between variables $\Rightarrow$ 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



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

MCA: the superimposed representation

$$F_{iq} = \frac{1}{\sqrt{\lambda_q}} \sum_k \frac{x_{ik}}{J} G_{kq}$$

$$G_{kq} = \frac{1}{\sqrt{\lambda_q}} \sum_i \frac{x_{ik}}{I_k} F_{iq}$$

$\Rightarrow$ Individual i at the barycenter of its levels

$\Rightarrow$ Level k at the barycenter of the individuals who take this level