Principal Component Analysis (PCA)

François Husson

Applied Mathematics Department - Rennes Agrocampus

husson@agrocampus-ouest.fr

Which kinds of data?

PCA applies to data tables where rows are considered as individuals and columns as quantitative variables

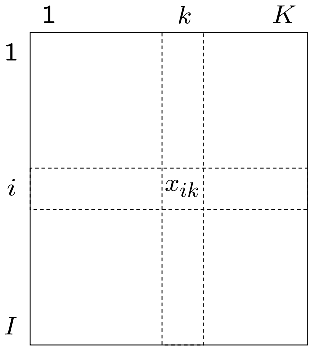


Figure: Data table in PCA

For variable k, we note:

the mean: $\bar{x}_k = \frac{1}{I} \sum_{i=1}^{I} x_{ik}$

the standard-deviation:

$$s_k = \sqrt{\frac{1}{I} \sum_{i=1}^{I} (x_{ik} - \bar{x}_k)^2}$$

Examples

- Sensory analysis: score for attribute k of product i
- Ecology: concentration of pollutant k in river i
- Economics: indicator value k for year i
- Genetics: expression of gene k for patient i
- Biology: measure k for animal i
- Marketing: value of measure k for brand i
- Sociology: time spent on activity k by individuals from social class i
- etc.

$\Rightarrow$ There exist many data tables like these

Wine data

- 10 individuals (rows): white wines from the Loire region
- 30 variables (columns):
  - 27 continuous variables: sensory descriptors
  - 2 continuous variables: odor and overall preference
  - 1 categorical variable: wine label (Vouvray or 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	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

Issues – goals

The data table can be seen as a set of rows or a set of columns

Studying individuals

- When can we say that 2 individuals are similar (or dissimilar) with respect to all the variables?
- If there are many individuals, is it possible to categorize them?

$\Rightarrow$ groups of individuals, partitions between them

Issues – goals

Studying variables

- For individuals, we interpret similarity in terms of the variables' values
- Between variables, we talk instead of "relationships"
- Linear relationships are commonplace, and a first approximation of many links $\Rightarrow$ correlation coefficient

$\Rightarrow$ visualization of the correlation matrix
$\Rightarrow$ find a small number of synthetic variables to summarize many variables (e.g. of a prior synthetic variable: the mean. But here we search for posterior synthetic variables from the data)

Issues – goals

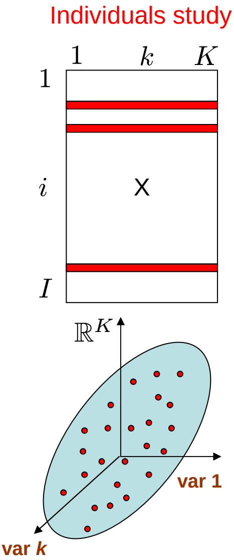
Links between the two points-of-view

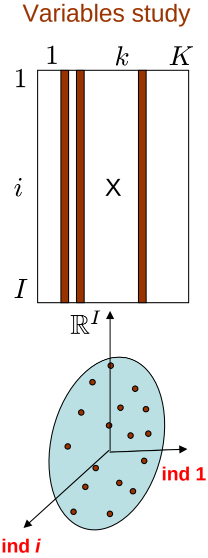
- Characterize groups of individuals using the variables
  $\Rightarrow$ need an automatic procedure
- Use specific individuals to better understand links between variables
  $\Rightarrow$ use of extreme individuals (return to individuals to understand more simply)

PCA issues:

- Descriptive method to explore data: visualization of data with simple plots
- Data compression - summarize a big data table of *individuals* $\times$ *quantitative variables*

Two point clouds





The cloud of individuals N_I

1 individual = 1 row of the data table $\Rightarrow$ 1 point in $\mathbb{R}^k$

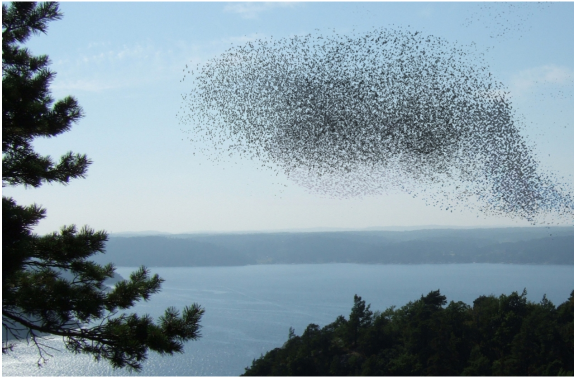
- If $K = 1$: axial representation
- If $K = 2$: scatter plot
- If $K = 3$: 3D graphical representation (more difficult)
- If $K = 4$: impossible to "see" BUT the concept is easy

Notion of similarity: (squared) distance between individuals i and i':

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

Studying the individuals $\equiv$ Studying the shape of the cloud N_I

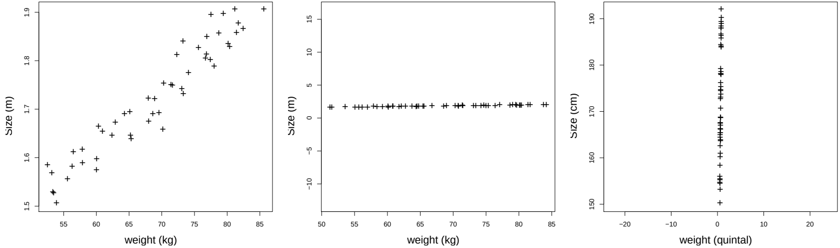
The cloud of individuals N_I



- Study the structure, *i.e.*, the shape, of the cloud of individuals
- Individuals are in $\mathbb{R}^K$

Centering – standardizing data

- Centering does not modify the shape of the cloud
  $\Rightarrow$ centering is always done



- Standardizing data is necessary if units are different between variables

$$x_{ik} \hookrightarrow \frac{x_{ik} - \bar{x}_k}{s_k}$$

Centering – standardizing data

	O.fruity	O.passion	O.citrus	...	Sweetness	Acidity	Bitterness	Astringency	Aroma.intensity	Aroma.persistency	Visual.intensity
S Michaud	-0,17	0,45	1,50	...	-0,30	0,11	0,20	-1,79	0,95	1,07	0,06
S Renaudie	0,02	1,03	1,16	...	-0,46	1,39	-0,31	-0,65	0,99	0,82	-1,08
S Trotignon	0,79	1,73	1,16	...	-0,67	0,48	0,20	-0,60	-0,44	0,07	-1,34
S Buisse Domaine	-0,17	0,45	-0,07	...	-0,02	-0,25	-2,01	0,19	-2,24	-1,66	-0,55
S Buisse Cristal	1,30	1,03	-0,12	...	-0,39	1,20	1,39	0,34	-0,44	-1,66	-0,90
V Aub Silex	-0,60	-0,97	-0,27	...	2,93	-2,07	-1,33	-0,60	-0,84	-0,92	-0,64
V Aub Marigny	-2,44	-0,97	-1,94	...	-0,30	0,84	1,39	1,45	-0,18	0,98	0,76
V Font Domaine	0,79	-1,11	-0,85	...	-0,67	-0,12	0,03	-0,44	0,29	0,41	1,03
V Font Brûlés	0,79	-0,84	0,13	...	-0,02	-0,61	0,03	0,34	0,75	0,07	1,73
V Font Coteaux	-0,29	-0,82	-0,69	...	-0,11	-0,98	0,37	1,76	1,15	0,82	0,94

PCA $\equiv$ Studying the standardized data set

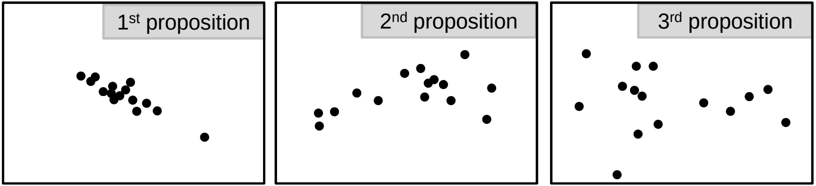
Difficult to visualize the cloud N_I $\Rightarrow$ try to get an approximate view of it

Fitting the cloud of individuals

PCA searches for the best summary space for optimal visualization of N_I
$\Longleftrightarrow$ Find a subspace that sums up the data the best

Viewpoint quality:

- faithfully reproduce the cloud's shape (*animation*)



Fitting the cloud of individuals

PCA searches for the best summary space for optimal visualization of N_I

$\Longleftrightarrow$ Find a subspace that sums up the data the best

Viewpoint quality:

- faithfully reproduce the cloud's shape (*animation*)
- best representation of diversity, variability
- doesn't distort distances between individuals

How to quantify the quality of a viewpoint?

notion of dispersion, of variability, also called **inertia**

inertia $\equiv$ variance generalized to several dimensions

Fit the individuals' cloud

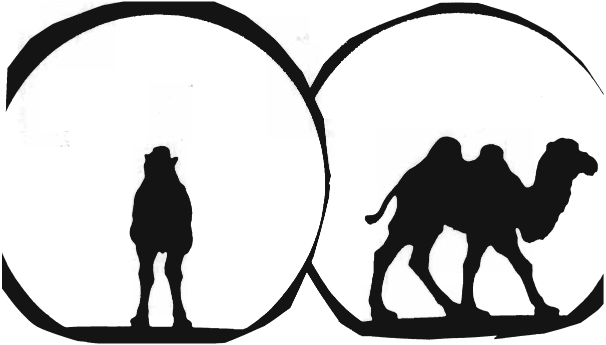
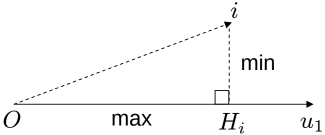


Figure: Camel or dromedary? (*illustration by J.P. Fénelon*)

Fit the individuals' cloud

How to find the best view to approximate the cloud?

1. find an axis that distorts the cloud the least



$(iH_i)^2$ small with $H_i \in$ axis $\Leftrightarrow$ $(OH_i)^2$ large (Pythagoras)
$\Rightarrow$ we want $\sum_i (OH_i)^2$ large

2. Find the best plane: maximize $\sum_i (OH_i)^2$ with $H_i \in$ plane
   The best plane contains the best axis: we search for $u_2 \perp u_1$ and maximizing $\sum_i (OH_i)^2$

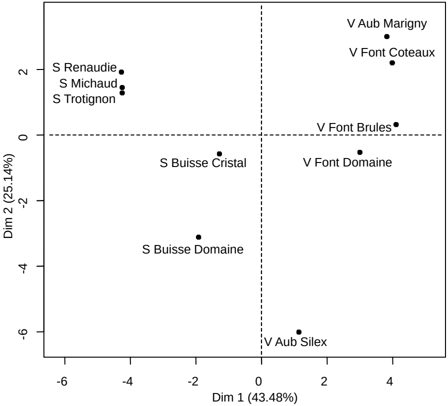
3. we can look for a third axis (etc.) with maximum inertia

Example: wine data

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

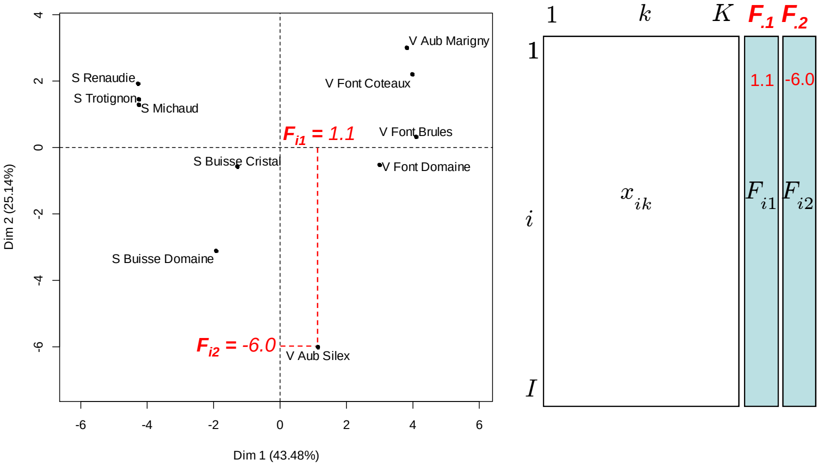
	O.fruity	O.passion	O.citrus	...	Sweetness	Acidity	Bitterness	Astringency	Aroma.intensity	Aroma.persistency	Visual.intensity	Odor.preference	Overall.preference	Label
S Michaud	4,3	2,4	5,7	...	3,5	5,9	4,1	1,4	7,1	6,7	5,0	6,0	5,0	Sauvignon
S Renaudie	4,4	3,1	5,3	...	3,3	6,8	3,8	2,3	7,2	6,6	3,4	5,4	5,5	Sauvignon
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: graphing the individuals



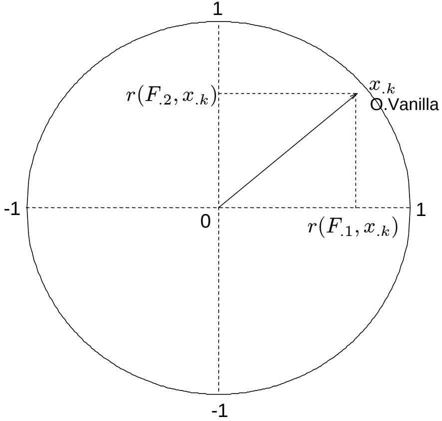
How to interpret the dimensions? Why are S. Trotignon and V. Font Brules far apart? $\Rightarrow$ Need variables to interpret the directions of variability

Individuals' coordinates considered as variables



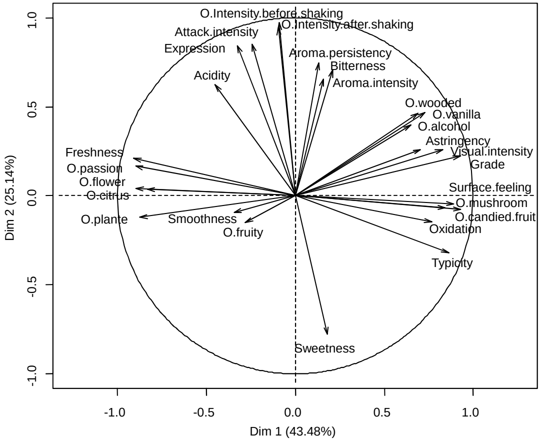
Representation of the variables as an interpretation aid for the individuals' cloud

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



$\Rightarrow$ Correlation circle

Representation of the variables as an interpretation aid for the individuals' cloud

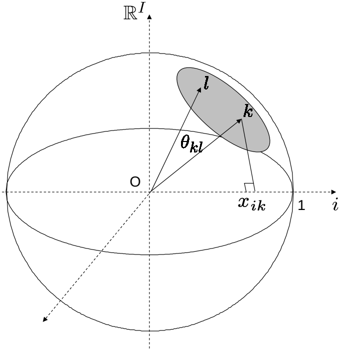


How to interpret the first dimension?

How to interpret the second dimension?

Main directions of variability: ....

Fitting the variables' cloud N_K



1 variable = 1 point in an I-dimensional space

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

Since variables are centered, $\cos(\theta_{kl}) = r(x_{.k}, x_{.l})$

If variables are standardized $\Rightarrow$ points are on an I-sphere of radius 1

Fitting the variables' cloud N_K

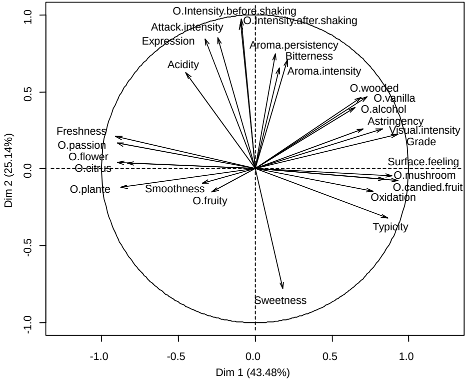
Similar strategy as for individuals: sequentially find orthogonal axes:

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

$\Rightarrow$ v_1 is the best synthetic variable for summarizing the variables

Find the 2^{nd} axis, then the 3^{rd}, etc.

Fitting the variables' cloud N_K



$\Rightarrow$ Same graph as before!!!!

- interpretation aid for the individuals' graph
- optimal representation of the variables' cloud
- visualization of the correlation matrix

Linking the two representations: transition formulas

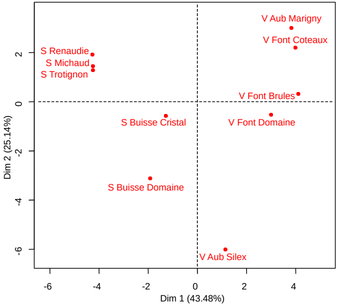
Scores: $F_{\bullet s}$

Loadings: $G_{\bullet s}/\sqrt{\lambda_s}$

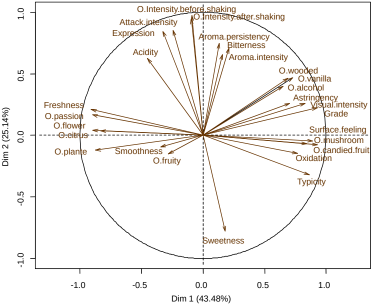
$$\boxed{F_{is} = \frac{1}{\sqrt{\lambda_s}} \sum_{k=1}^{K} x_{ik} G_{ks}}$$

$$\boxed{G_{ks} = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} x_{ik} F_{is}}$$

$\Longrightarrow$ Individuals are on the same side as their corresponding variables with high values

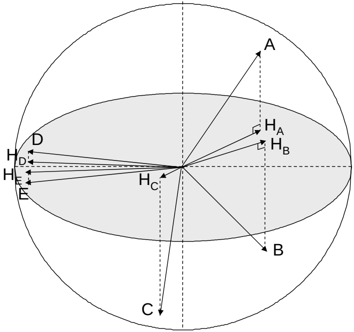


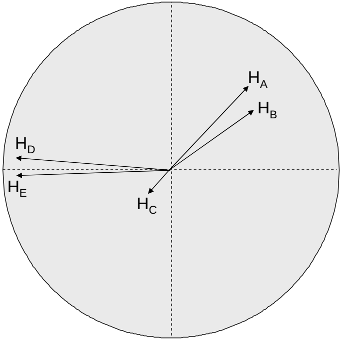
	Aub Silex
O.intensity.after.shaking	-2,54
O.intensity.before.shaking	-2,37
Expression	-2,25
Acidity	-2,07
Attack.intensity	-1,36
Bitterness	-1,33
Freshness	-1,15
...	...
Typicity	1,01
Sweetness	2,93



Projections...

$r(A, B) = cos(\theta_{A,B})$
$cos(\theta_{A,B}) \approx cos(\theta_{H_A,H_B})$ if the variables are well-projected

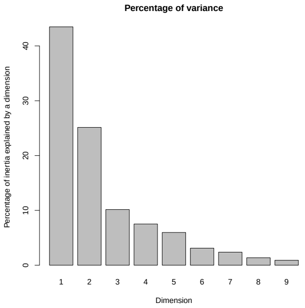




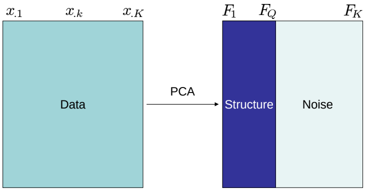
Only well-projected variables can be interpreted!

Choosing the number of dimensions

Bar chart of eigenvalues,
tests,
confidence intervals,
cross-validation (`estim_ncp` function),
etc.



Two goals:
⇒ Interpretation
⇒ Separate structure from noise



Percentage of variance obtained under independence

⇒ Is there structure in 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 for inertia in the two first axes of 10 000 PCA on data with independent variables

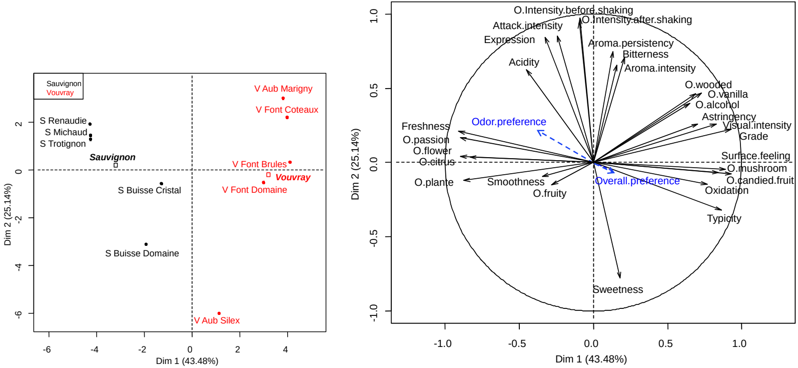
Percentage of variance obtained under independence

	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 for inertia in the two first axes of 10 000 PCA on data with independent variables

Supplementary information

- For the quantitative variables: project supplementary variables onto the axes
- For categorical variables: project the barycenter of individuals in each category



$\Rightarrow$ Supplementary information not used to build the axes

Quality of the representation: $\cos^2$

- $\cos^2(\theta_{iH_i})$ for the individuals: 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
```

- $\cos^2(\theta_{kH_k})$ for the variables: only well-projected variables (high $\cos^2$) can be interpreted!

```
> round(res.pca$var$cos2,2)
                 Dim.1 Dim.2
O.fruity          0.08  0.02
O.passion         0.80  0.03
O.citrus          0.69  0.00
```

Contributions

⇒ Contributions to components:

- for an individual: $Ctr_s(i) = \frac{F_{is}^2}{\sum_{i=1}^{I} F_{is}^2} = \frac{F_{is}^2}{\lambda_s}$

  ⇒ Individuals with a large coordinate value contribute 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 a variable: $Ctr_s(k) = \frac{r(x_{.k},v_s)^2}{\sum_{k=1}^{K} r(x_{.k},v_s)^2} = \frac{r(x_{.k},v_s)^2}{\lambda_s}$

  ⇒ Variables highly correlated with the principal component contribute the most

```
> round(res.pca$var$contrib,2)
               Dim.1 Dim.2
O.fruity        0.67  0.34
O.passion       6.84  0.40
O.citrus        5.89  0.02
```

Characterizing the axes

Using the continuous variables:

- correlation between each variable and the principal component of rank s is calculated
- correlation coefficients are sorted and significant ones are output

```
> dimdesc(res.pca)
                $Dim.1$quanti                                   $Dim.2$quanti
                 corr p.value                                    corr p.value
O.candied.fruit  0.93 9.5e-05  O.intensity.before.shaking 0.97 3.1e-06
Grade            0.93 1.2e-04  O.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
```

Characterizing the axes

Using the categorical variables:

- Do one-way analysis of variance with the coordinates of the individuals ($F_{.s}$) described by the categorical variable
  - an F-test by variable
  - for each category, a Student's t-test to compare the average of the category with the general mean

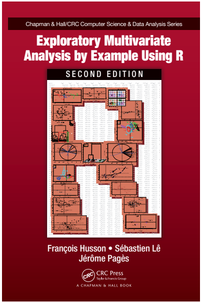
```
> dimdesc(res.pca)
Dim.1$quali
              R2      p.value
Label      0.874     7.30e-05

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

PCA in practice

1. Choose active variables
2. Rescale (or not) the variables
3. Perform PCA
4. Choose the number of dimensions to interpret
5. Joint analysis of the cloud of individuals and the cloud of variables
6. Use indicators to enrich interpretation
7. Go back to raw data for interpretation

More



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

The `FactoMineR` package for doing PCA:
`http://factominer.free.fr/`

Videos on Youtube:

- Youtube channel: `youtube.com/HussonFrancois`
- a playlist with movies in English
- a playlist with movies in French