Data Analysis

Principal component analysis

National Research University Higher School of Economics
Master's Program "Big Data Systems"

Fall 2019

Data matrix

$$\begin{array}{cc|cccc} & & \multicolumn{4}{c}{\text{Variables}} \\ & & V_1 & V_2 & \cdots & V_K \\ \hline \text{Individuals} & 1 & x_{11} & x_{11} & \cdots & x_{1K} \\ & 2 & x_{21} & \ddots & & \\ & \vdots & \vdots & & & \\ & I & x_{I1} & & & x_{IK} \end{array}$$

Many applications:

- Economics and finance: countries - economic indicators
- Marketing: brands - measures of satisfaction

- Individuals = observations
- Vector $\boldsymbol{x}_{\bullet k} = (x_{1k}, x_{2k}, \ldots, x_{Ik})^T$, $k = 1, \ldots, K$, contains the values of the variable V_k for $1, \ldots, I$ individuals.
- The $I \times K$ data matrix

$$\mathbf{X} = \{x_{ik}\} = \left(\underset{\downarrow}{\boldsymbol{x}_{\bullet 1}}, \underset{\downarrow}{\boldsymbol{x}_{\bullet 2}}, \ldots, \underset{\downarrow}{\boldsymbol{x}_{\bullet K}} \right)$$

Data matrix: example

- 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)

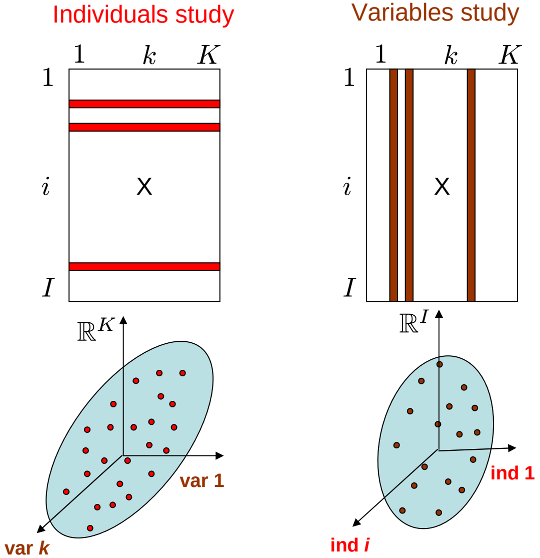
	O.fruity	O.passion	O.citrus	...	Sweetness	Acidity	Bitterness	Astringency	Aroma.intensity	Aroma.persistency	Visual.intensity	Odor.preferene	Overall.preference	Label
S Michaud	4.3	2.4	5.7	...	3.5	5.9	4.1	1.4	7.1	6.7	5.0	6.0	5.0	Sauvignon
S Renaudie	4.4	3.1	5.3	...	3.3	6.8	3.8	2.3	7.2	6.6	3.4	5.4	5.5	Sauvignon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

The PCA objectives

- Study of individuals: similarity between individuals with respect to all variables => partition between individuals
- Study of variables: linear relationships between variables => visualization of the correlation matrix
- Link between the two studies: characterization of the groups of individuals by the variables; specific individuals that help better understand links between variables

Geometrical view on data

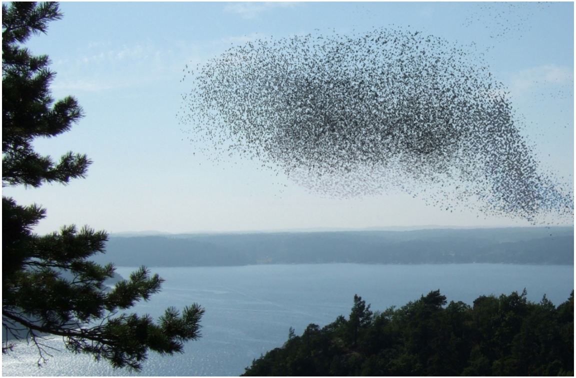
- For >3 variables, the visualization is impossible.



Individuals study

Variables study

Geometrical view on data (2)



- Study the structure, i.e., the shape of the cloud of individuals.

Basic statistics of the data

- Sample mean of the k-th variable:

$$\bar{x}_k = \frac{1}{I}\sum_{i=1}^{I} x_{ik} = \frac{1}{I}\mathbf{1}^T \boldsymbol{x}_{\bullet k}$$

where $\mathbf{1} = (\underbrace{1, ..., 1}_{I})^T$.

- Sample variance of the k-th variable:

$$\text{Var}(V_k) = \frac{1}{I}\sum_{i=1}^{I}(x_{ik} - \bar{x}_k)^2 = \|\boldsymbol{x}_{\bullet k} - \bar{x}_k\mathbf{1}\|^2$$

- Sample standard deviation of the k-th variable:

$$s(V_k) = \sqrt{\text{Var}(V_k)} = \|\boldsymbol{x}_{\bullet k} - \bar{x}_k\mathbf{1}\|$$

- Sample covariation between variables V_m and V_n:

$$\text{Cov}(V_m, V_n) = \frac{1}{I}\sum_{i=1}^{I}(x_{im} - \bar{x}_m)(x_{in} - \bar{x}_n) = \frac{1}{I}(\boldsymbol{x}_{\bullet m} - \bar{x}_m\mathbf{1})(\boldsymbol{x}_{\bullet n} - \bar{x}_n\mathbf{1})$$

Basic statistics of the data

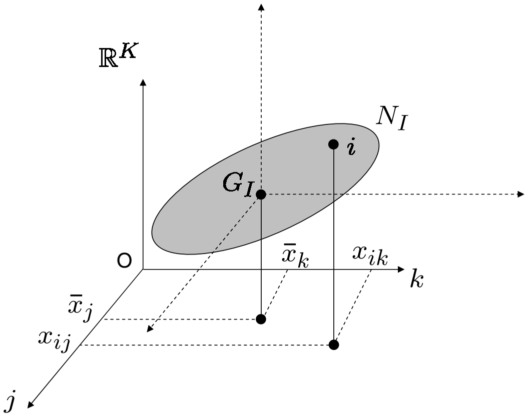
- Sample correlation between variables V_m and V_n:

$$r(V_m, V_n) = \frac{\text{Cov}(V_m, V_n)}{s(V_m)s(V_n)}$$

Centering the data

- Centering the data (always):

$$x_{ik} - \bar{x}_k, \quad \text{i.e.,} \quad \boldsymbol{x}_{\bullet k} - \bar{x}_k \mathbf{1}^T$$

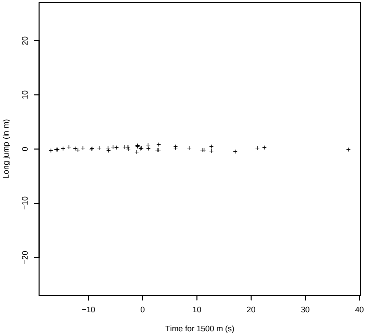


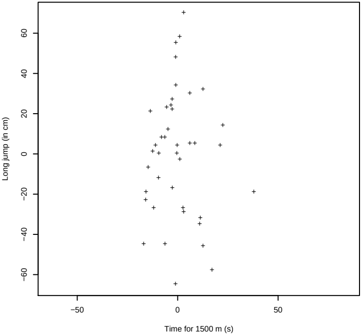
Scaling the data

- Centering AND standardizing (sometimes):

$$\frac{x_{ik}-\bar{x}_k}{s_k}, \quad \text{i.e.,} \quad \frac{\boldsymbol{x}_{\bullet k}-\bar{x}_k\mathbf{1}}{s_k}$$

- Standardizing: variables are always scaled when they are not in the same units





Centering and scaling the data

- Matrix $\mathbf{X}$ is redefined via the centered sample vectors, i.e.,

$$\mathbf{X} = \left\{x_{ik} - \bar{x}_k\right\} = \left(\underset{\downarrow}{\boldsymbol{x}_{\bullet 1} - \bar{x}_1 \mathbf{1}}, \underset{\downarrow}{\boldsymbol{x}_{\bullet 2} - \bar{x}_2 \mathbf{1}}, \ldots, \underset{\downarrow}{\boldsymbol{x}_{\bullet K} - \bar{x}_K \mathbf{1}}\right) \qquad (1)$$

or, if standardization is used,

$$\mathbf{X} = \left\{\frac{x_{ik} - \bar{x}_k}{s_k}\right\} = \left(\underset{\downarrow}{\frac{\boldsymbol{x}_{\bullet 1} - \bar{x}_1 \mathbf{1}}{s_1}}, \underset{\downarrow}{\frac{\boldsymbol{x}_{\bullet 2} - \bar{x}_2 \mathbf{1}}{s_2}}, \ldots, \underset{\downarrow}{\frac{\boldsymbol{x}_{\bullet K} - \bar{x}_K \mathbf{1}}{s_K}}\right) \qquad (2)$$

Sample covariance matrix

- Using centered $\mathbf{X}$ (1) one can compute the sample $K \times K$ covariance matrix (verify!):

$$\Sigma = \frac{1}{I}\mathbf{X}^T\mathbf{X} = \begin{pmatrix} \text{Var}(V_1) & \text{Cov}(V_1, V_2) & \cdots & \\ \text{Cov}(V_2, V_1) & \text{Var}(V_2) & & \\ \vdots & & \ddots & \\ & & & \text{Var}(V_K) \end{pmatrix} \tag{3}$$

- The standardized quantities (2) give a sample $K \times K$ correlation matrix

$$\mathbf{P} = \frac{1}{I}\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & r(V_1, V_2) & \cdots & \\ r(V_2, V_1) & 1 & & \\ \vdots & & \ddots & \\ & & & 1 \end{pmatrix}$$

Trace of Σ

- The trace of Σ (sum of the diagonal elements) is a total variance of the data:

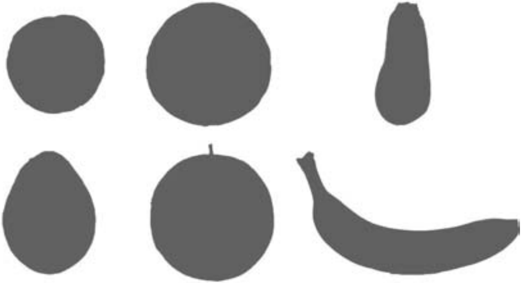
$$\boldsymbol{Tr}\boldsymbol{\Sigma} = \mathrm{Var}(V_1) + \mathrm{Var}(V_2) + \ldots + \mathrm{Var}(V_K) = \frac{1}{I}\sum_{k=1}^{K}\sum_{i=1}^{I}(x_{ik} - \bar{x}_k)^2 = \frac{1}{I}\sum_{i=1}^{I}\left\|\boldsymbol{x}_{i\bullet} - \bar{\boldsymbol{x}}\right\|^2 \quad (4)$$

  where the vector $\boldsymbol{x}_{i\bullet} = (x_{i1}, x_{i2}, \ldots, x_{iK})^T$, $i = 1, \ldots, I$, describes the i-th individual and $\bar{\boldsymbol{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K)^T$ is the mean vector of variables over individuals.

- Under all subsequent approximations, we should keep this quantity as close to its genuine value as possible.

Reduction of dimensionality: geometry

- Geometrically, the PCA seeks for a subspace (normally, the plane) projection onto which retains as much of data variability (4) as possible:



- Two-dimensional representations of fruits: from left to right an avocado, a melon and a banana, each row corresponds to a different perspective.

Reduction of dimensionality: geometry (2)

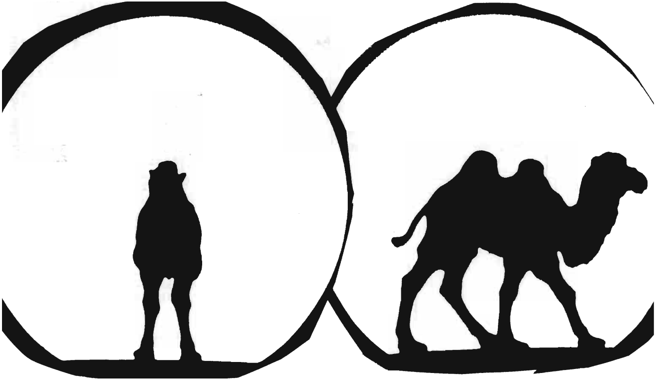
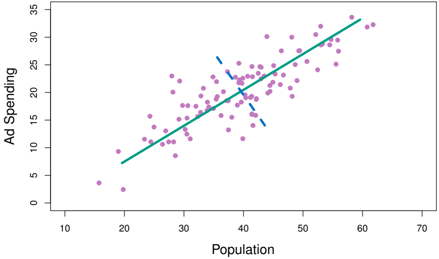


Figure: Who is that? Depends on the viewpoint.

2D example

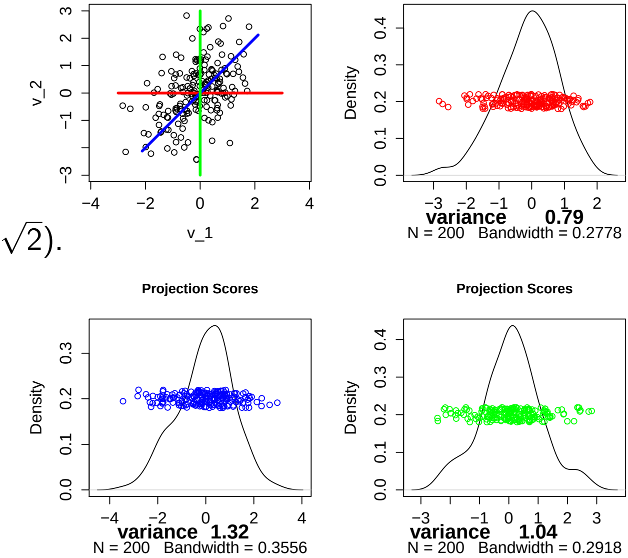
- Find the direction (1st principal component (PC) direction) along which variability of data (more precisely, their projections on the line) is the highest



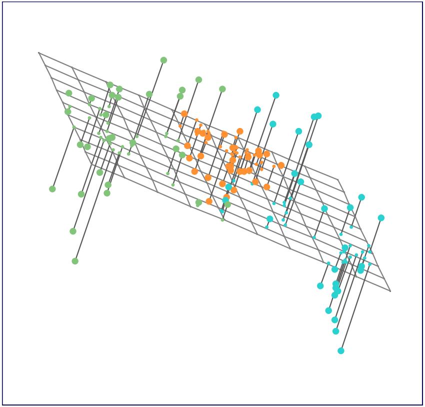
- Once the 1st PC has been found, find the direction (2nd principal component direction) such that
  - PC2 ⊥ PC1 (Why?)
  - Variability of data along which is the highest
- Repeat K (=number of variables) times

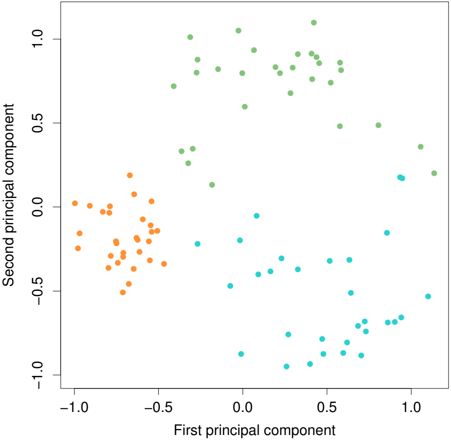
Another 2D example

Take $\boldsymbol{a} = (1,0)'$, $(0,1)', (1/\sqrt{2}, 1/\sqrt{2})$. Which one is better?



3D example




Second principal component
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
First principal component

PCA as an optimization problem

- Consider $v_1 \in \mathbb{R}^K$.
- Consider projections of individuals $x_{i\bullet}$, $i = 1, \dots, I$, onto v_1:

$$(x_{i\bullet} v_1) e_{v_1}$$

- Choose v_1 so that the (sample) variance of the projected data is maximal (verify!):

$$\text{Sample variance of projected individuals} = v_1^T \mathbf{X}^T \mathbf{X} v_1 \to \max,$$

$$v_1^T v_1 = 1.$$

- Unconstraint optimization:

$$\mathcal{L} = v_1^T \mathbf{X}^T \mathbf{X} v_1 - \lambda (v_1^T v_1 - 1) \to \max \tag{5}$$

- Solution – an eigenvalue problem

$$\mathbf{X}^T \mathbf{X} v_1 = \lambda_1 v_1 \tag{6}$$

PCA as an optimization problem (2)

- Next, find v_2 such that the sample variance projected onto v_2

$$\boldsymbol{v}_2^T \mathbf{X}^T \mathbf{X} \boldsymbol{v}_2 \to \max,$$

$$\boldsymbol{v}_1^T \boldsymbol{v}_1 = 1, \quad \boldsymbol{v}_2^T \boldsymbol{v}_1 = 0.$$

- Again, the problem is reduced to the eigenvalue problem:

$$\mathbf{X}^T \mathbf{X} \boldsymbol{v}_2 = \lambda_2 \boldsymbol{v}_2$$

- Repeat K times.
- Recall, by definition $\mathbf{X}^T \mathbf{X} = \mathbf{\Sigma}$

Diagonalizing Σ

- Matrix Σ is symmetric => there exist K real eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_K$ and the respective eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K$
- The vector of the ith individual $\boldsymbol{x}_{i\bullet} = (x_{i1}, x_{i2}, \ldots, x_{iK})^T$ is defined in some basis of $\mathbb{R}^K$.
- Rotation of the coordinate frame such that the new basis is comprised of the eigenvectors of Σ yields the diagonal covariance matrix

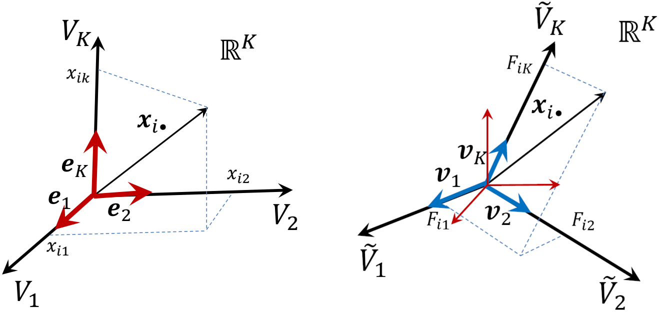
$$\boldsymbol{\Sigma}' = \boldsymbol{V}^T \boldsymbol{\Sigma} \, \boldsymbol{V} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \\ 0 & & & \lambda_K \end{pmatrix} \tag{7}$$

where $\boldsymbol{V} = \left(\underset{\downarrow}{\boldsymbol{v}_1}, \underset{\downarrow}{\boldsymbol{v}_2}, \ldots, \underset{\downarrow}{\boldsymbol{v}_K} \right)$

Rotating the frame

- Original frame:

$$\boldsymbol{x}_{i\bullet} = x_{i1}\,\boldsymbol{e}_1 + x_{i2}\,\boldsymbol{e}_2 + \ldots + x_{iK}\,\boldsymbol{e}_K$$



- Rotated frame made up of eigenvectors (always orthogonal):

$$\boldsymbol{x}_{i\bullet} = F_{i1}\,\boldsymbol{v}_1 + F_{i2}\,\boldsymbol{v}_2 + \ldots + F_{iK}\,\boldsymbol{v}_K$$

New variables

- The new coordinates are the linear combinations of the old ones

$$\begin{pmatrix} F_{i1} \\ F_{i2} \\ \vdots \\ F_{iK} \end{pmatrix} = \begin{pmatrix} \boldsymbol{v}_1 \rightarrow & \\ \boldsymbol{v}_2 \rightarrow & \\ & \\ \boldsymbol{v}_K \rightarrow & \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iK} \end{pmatrix} \tag{8}$$

- The new variables $\tilde{V}_1,...,\tilde{V}_K$ (principal components) are uncorrelated, see (7).
- The trace of Σ (recall, the total variability of the data) is invariant under change of basis:

$$\boldsymbol{Tr}\boldsymbol{\Sigma} = \boldsymbol{Tr}\boldsymbol{\Sigma}' = \sum_{k=1}^{K} \lambda_k$$

- So, Var$(\tilde{V}_k) = \lambda_k$, $k = 1, \ldots, K$, see (7).

Reduction of dimensionality

- The key idea of the PCA is to retain just a few new basis elements $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, ... (normally, 2 or 3) which provide the largest contribution to the trace $\boldsymbol{Tr}\boldsymbol{\Sigma}$ = total variability.
- The contributions of the new variables (8) to $\boldsymbol{Tr}\boldsymbol{\Sigma}$ are given by $\lambda_1 > \lambda_2 > \ldots > \lambda_K$.
- For example, if

$$(\lambda_1 + \lambda_2)\Big/\sum \lambda_i \approx 1$$

it suffices to retain only $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$:

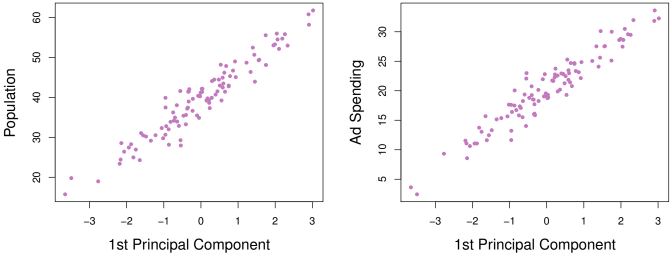
$$\begin{pmatrix} \tilde{V}_1 \\ \tilde{V}_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{v}_1 \rightarrow & \qquad\qquad \\ \boldsymbol{v}_2 \rightarrow & \\ \boldsymbol{0} \rightarrow & \\ \vdots & \\ \boldsymbol{0} \rightarrow & \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \\ \vdots \\ V_K \end{pmatrix}$$

Terminology

- F_{i1} and F_{i2} are called scores of the ith individual on the principal components directions $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ and present the coordinates of individual i in the reduced basis $\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$.
- Vector $\boldsymbol{v}_1$ is called the 1st PC loading vector and so on.
- The differences between individuals (variability) will still be captured well using only two new variables, $\tilde{V}_1$ and $\tilde{V}_2$ instead of the original set V_k, $k = 1, .., K$.

Use of PC scores

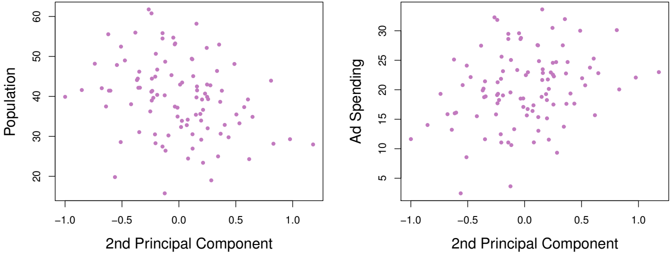
- Back to Slide 16
- Plots of the 1st PC scores F_{i1} vs the original variables



- 1st PC score is correlated with both original variables => one can judge on the orig. vars. using this PC score

Use of PC scores (2)

- Back to Slide 16
- Plots of the 2nd PC scores F_{i2} vs the original variables



- 2nd PC score is far less correlated with both original variables => poor conclusions on the orig. vars. using this PC score

PCA via SVD

- Any matrix $\mathbf{X}$ admits a singular value decomposition (SVD):

$$\underbrace{\mathbf{X}}_{I\times K} = \underbrace{\mathbf{U}}_{I\times K} \cdot \underbrace{\mathbf{D}}_{K\times K} \cdot \underbrace{\mathbf{V}^T}_{K\times K}$$

- Matrices $\boldsymbol{V} = \left(\underset{\downarrow}{\boldsymbol{v}_1}, \underset{\downarrow}{\boldsymbol{v}_2}, \ldots, \underset{\downarrow}{\boldsymbol{v}_K}\right)$, $\boldsymbol{U} = \left(\underset{\downarrow}{\boldsymbol{u}_1}, \underset{\downarrow}{\boldsymbol{u}_2}, \ldots, \underset{\downarrow}{\boldsymbol{u}_K}\right)$ are orthogonal, i.e.,

$$\boldsymbol{U}^T\boldsymbol{U} = \mathbf{E}, \ \boldsymbol{V}^T\boldsymbol{V} = \mathbf{E} \tag{9}$$

- Eq.(9) implies orthonormality of columns of $\boldsymbol{U}, \boldsymbol{V}$, e.g.,

$$\boldsymbol{u}_1\boldsymbol{u}_2 = 0, \quad \boldsymbol{u}_1\boldsymbol{u}_1 = 1$$

- $\mathbf{D}$ = diag$\{\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_K}\}$
- Singular values of $\mathbf{X}$

$$\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq, \ldots, \geq \sqrt{\lambda_K} \geq 0$$

PCA via SVD (2)

- Eigen decomposition of $\mathbf{X}^T\mathbf{X}$ (verify!) :

$$\mathbf{X}^T\mathbf{X} = \boldsymbol{V}\mathbf{D}^2\boldsymbol{V}^T$$

- Hence, $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_K$ and $\lambda_1, \lambda_2, ..., \lambda_K$ are the eigenvectors and eigenvalues of $\mathbf{X}^T\mathbf{X}$ (cf. Eq.(7))
- $\boldsymbol{v}_1, \boldsymbol{v}_2, ...$ are the 1st, 2nd ,... PC directions
- Projection of data $\mathbf{X}$ onto $\boldsymbol{v}_1$ (verify!):

$$\boldsymbol{z}_1 = \mathbf{X}\boldsymbol{v}_1 = \sqrt{\lambda_1}\boldsymbol{u}_1 \tag{10}$$

- A $I \times 1$ vector $\mathbf{X}\boldsymbol{v}_1$ is a linear combination of old vars (with coefs $\boldsymbol{v}_1$) for each individual.
- $\boldsymbol{z}_1$ is a $I \times 1$ vector with scores (coordinates) along the 1st PC direction
- So, columns of $\boldsymbol{V}$ are the orthogonal PC directions, columns of $\boldsymbol{U}$ are the respective PC scores (up to a constant $\sqrt{\lambda_i}$).

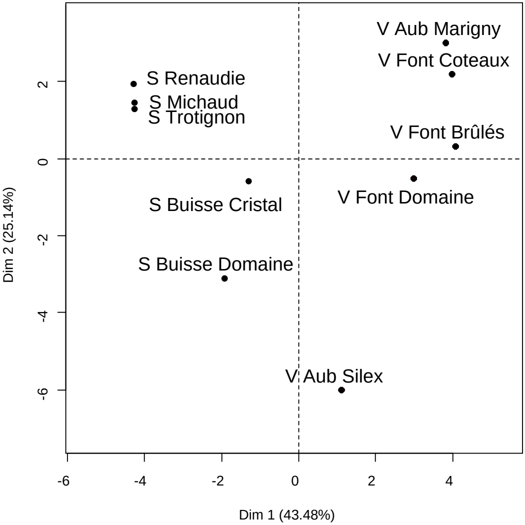
PCA via SVD (3)

- Recall, $\mathbf{X}$ is centered, i.e., sum over any column is zero.
- Hence, sum of all elements of vector $\mathbf{X}\boldsymbol{v}_1$ is also zero (verify!).
- Hence (verify!),

$$\mathrm{Var}(\mathbf{X}\boldsymbol{v}_1) = \mathrm{Var}(\boldsymbol{z}_1) = \mathrm{Var}(\sqrt{\lambda_1}\boldsymbol{u}_1) = \frac{\lambda_1}{I}$$

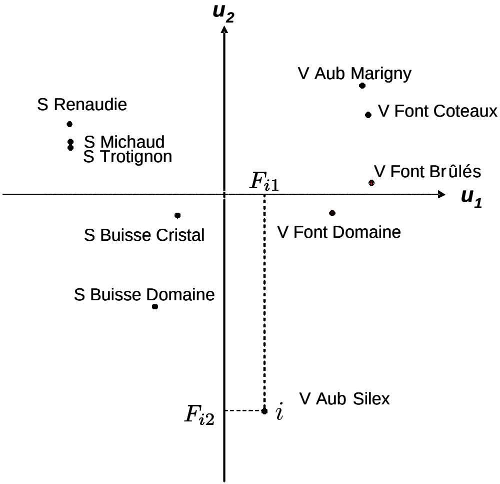
- As $\lambda_1 \geq \lambda_2 \geq, \ldots, \geq \lambda_K$, the linear combination of old vars (10) has the largest variance.

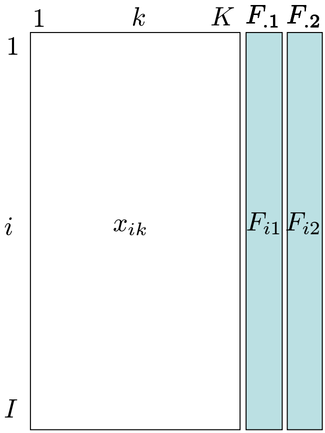
Wine example



- λ_1 and λ_2 account for 43.48% and 25.14% of the total variability => keeping only $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.
- An approximate 2D depiction of similarity among wines in terms of all the variables.

Coordinates of individuals in the reduced basis



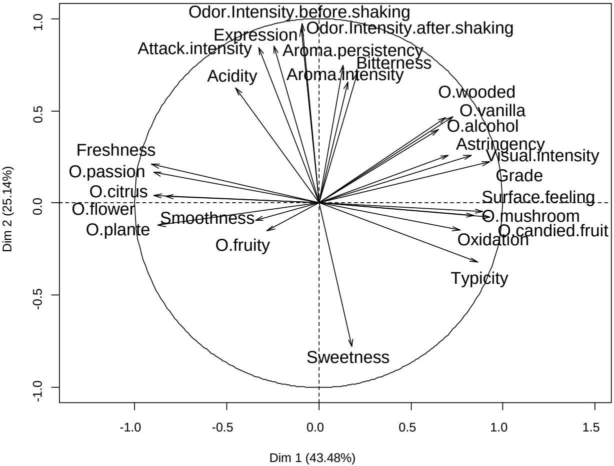


- $F_{i,1} = v_1^T x_{\cdot,i}$, $F_{i,2} = v_2^T x_{\cdot,i}$ are the coordinates of the i-th individual along first two PC directions v_1 and v_2.

Correlation between new and original variables

- Observations of the old variable V_k, $k = 1, \dots, K$ from $\mathbf{X}$ (for I individuals):
$$\boldsymbol{x}_{\bullet k} = \{x_{1k}, \dots, x_{Ik}\} \tag{11}$$
- Values of the new variable $\tilde{V}_{k'}$, $k' = 1, 2$ (for I individuals):
$$F_{\bullet k'} = \{F_{1k'}, \dots, F_{Ik'}\} \tag{12}$$
- Correlation between V_k and $\tilde{V}_{k'}$ (using (11), (12) data sets):
$$\text{cor}(V_k, \tilde{V}_{k'}) = (\boldsymbol{v}_{k'})_k \sqrt{\frac{\lambda_{k'}}{s(V_k)}} \tag{13}$$
- Interpretation: if (13) is positive, an individual with a high value of $\tilde{V}_{k'}$ is likely to have a high value of V_k (unobservable on the PC plot) and vice versa for negative correlation.

Correlation between variables, cntd.



- For instance, make “V Font Brules”, see Slide 31, tends to have a high value of `Visual.Intensity`

Study of variables

- Consider the $I \times I$ matrix built from the centered AND standardized data (2)

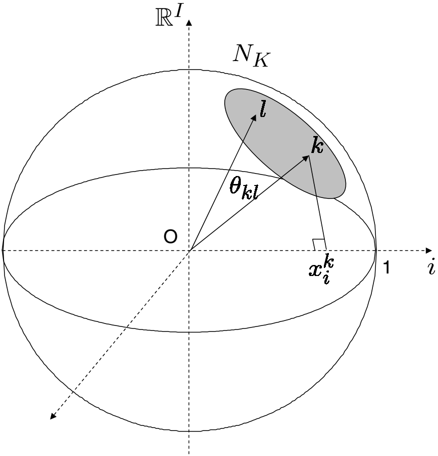
$$\frac{1}{K}\mathbf{X}\mathbf{X}^T = \begin{pmatrix} 1 & \boldsymbol{x}_{1\bullet}\boldsymbol{x}_{2\bullet} & \cdots & \\ \boldsymbol{x}_{2\bullet}\boldsymbol{x}_{1\bullet} & 1 & & \\ \vdots & & \ddots & \\ & & & 1 \end{pmatrix}$$

- Apply the PCA to the above matrix, i.e., find the eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_I$ and retain a few first that keep enough variability of the data.
- The new variables, $k = 1, \ldots, K$:

$$\begin{pmatrix} G_{1k} \\ G_{2k} \\ \vdots \\ G_{Ik} \end{pmatrix} = \begin{pmatrix} \boldsymbol{v}_1 \rightarrow & & \\ \boldsymbol{v}_2 \rightarrow & & \\ & & \\ \boldsymbol{v}_I \rightarrow & & \end{pmatrix} \begin{pmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{Ik} \end{pmatrix} \tag{14}$$

- Keep those corresponding to the truncated basis, e.g., G_1 and G_2.

Study of variables: geometry



- The inner product:

$$\boldsymbol{x}_{\bullet k}\boldsymbol{x}_{\bullet l} = \|\boldsymbol{x}_{\bullet k}\|\,\|\boldsymbol{x}_{\bullet l}\|\cos\theta_{kl}$$

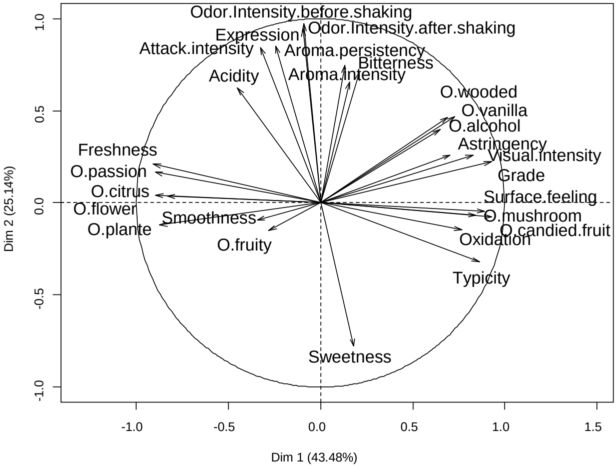
- Sample correlation between V_k and V_l (verify!):

$$\text{cor}(V_k, V_l) = \frac{\boldsymbol{x}_{\bullet k}\boldsymbol{x}_{\bullet l}}{\|\boldsymbol{x}_{\bullet k}\|\,\|\boldsymbol{x}_{\bullet l}\|} = \cos\theta_{kl}$$

- Implication: the PCA provides the 2D visual representation of correlation between the original variables.

- Small angles between variables mean high correlation.

Study of variables: the PCA

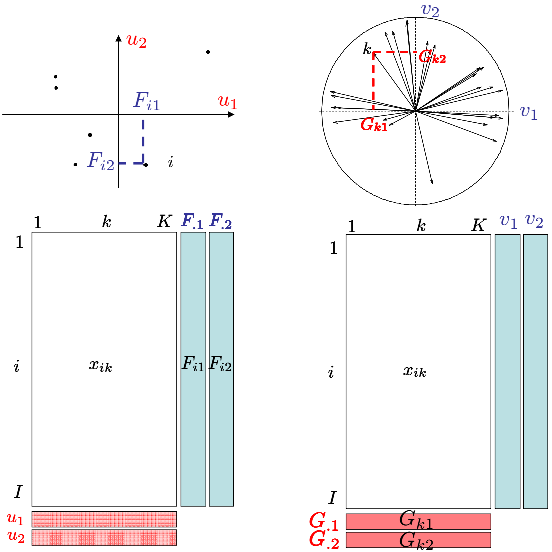


- The same plot as in Slide 34.
- Shows the extent of correlation between the original variables.

Study of variables: the PCA (2)

- The representational quality of a variable in a given plane is its distance from the circle of radius 1.

Link between the two representations



- Transition formulae

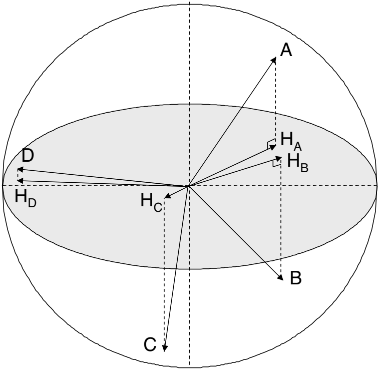
$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k=1}^{K} x_{ik} G_s(k)$$

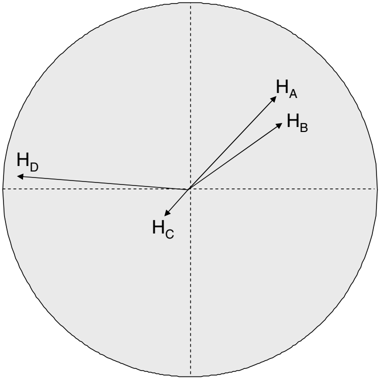
$$G_s(k) = \frac{1}{I\sqrt{\lambda_s}} \sum_{i=1}^{I} x_{ik} F_s(k)$$

Quality of representation

- The quality of representation of individual i on the component s:

$$\mathrm{qlt}_s(i) = \frac{\text{Projected inertia of } i \text{ on } \boldsymbol{v}_s}{\text{Total inertia of } i}$$





Quality of representation, $\cos^2 \theta$

- For individuals: similarity 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
```

- For the variables: only well projected variables 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
```

Contribution to the PCs

- Inertia of the sth PC explained by by the ith individual:

$$\text{Ctr}_s(i) = \frac{\text{Inertia along } v_s \text{ explained by ind. } i}{\text{Total inertia along } v_s} = \frac{F_{is}^2}{\lambda_s}$$

  => Individuals with a large PC 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
```

- Inertia of the sth PC explained by by the kth variable:

$$\text{Ctr}_s(k) = \frac{G_{ks}^2}{\lambda_s}$$

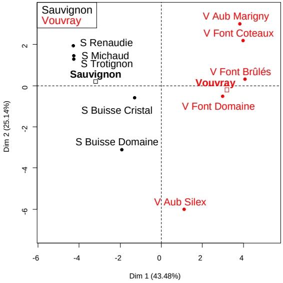
  => variables highly correlated with the sth principal component (i.e., large coordinate along v_s) contribute the most

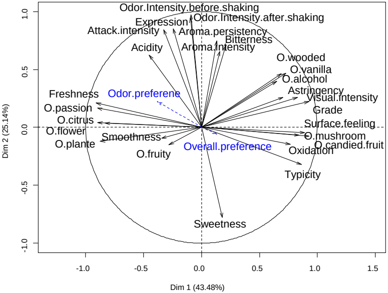
- Possible application – assessing the robustness.

Supplementary information

Supplementary data are not used for computing variability. Can be quantitative or categorical. Supplementary information do not create new dimensions.

- 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: categorical variables

How to project the supplementary categorical variables?

```
            X100m Long.jump Shot.put High.jump Competition
HERNU       11.37      7.56    14.41      1.86    Decastar
BARRAS      11.33      6.97    14.09      1.95    Decastar
NOOL        11.33      7.27    12.68      1.98    Decastar
BOURGUIGNON 11.36      6.80    13.46      1.86    Decastar
Sebrle      10.85      7.84    16.36      2.12    OlympicG
Clay        10.44      7.96    15.23      2.06    OlympicG
```

```
            X100m Long.jump Shot.put High.jump
HERNU       11.37      7.56    14.41      1.86
BARRAS      11.33      6.97    14.09      1.95
NOOL        11.33      7.27    12.68      1.98
BOURGUIGNON 11.36      6.80    13.46      1.86
Sebrle      10.85      7.84    16.36      2.12
Clay        10.44      7.96    15.23      2.06

Decastar    11.18      7.25    14.16      1.98
Olympic.G   10.92      7.27    14.62      1.98
```

- Take the average of the variables within a category

Supplementary information: categorical variables – confidence ellipses

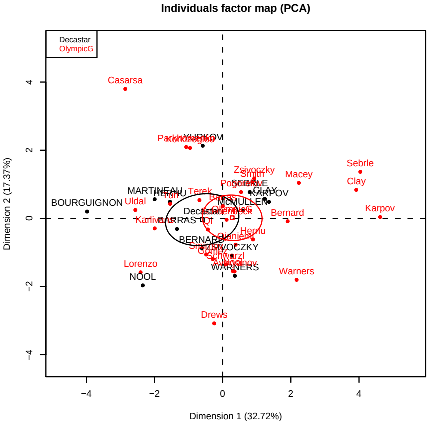


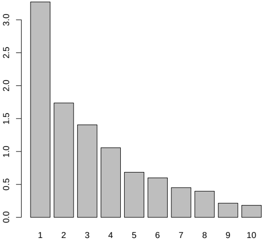
Figure: Confidence ellipses around the barycenter of each category

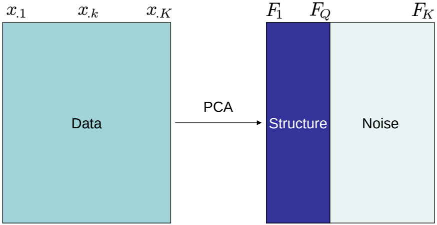
Choosing the number of components

- Percentage of variance explained by each axis: information brought by the dimension.
- Quality of the approximation:

$$\sum_{s=1}^{Q} \lambda_s \Big/ \sum_{s=1}^{K} \lambda_s$$

  - Bar plot of the eigenvalues: scree test





- Dimensionality reduction implies loss of information

Description of the dimensions

By the continuous variables:

- correlation between each variable and the principal component of rank q is calculated
- The variables are ranked w.r.t. correlation and the ones with the highest correlation (absolute values) can be retained.

```
> 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 (2)

By the categorical variables:

- Perform a one-way analysis of variance with the coordinates of the individuals $F_{i,q}$ explained by the categorical variable
- the F-test by variable

```
> 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
```

Some practice with R ...

```
library(FactoMineR)
data(decathlon)
res <- PCA(decathlon,quanti.sup=11:12,quali.sup=13)
plot(res,habillage=13)
res$eig
x11()
barplot(res$eig[,1],main="Eigenvalues",names.arg=1:nrow(res$eig))
res$ind$coord
res$ind$cos2
res$ind$contrib
dimdesc(res)
aa=cbind.data.frame(decathlon[,13],res$ind$coord)
bb=coord.ellipse(aa,bary=TRUE)
plot.PCA(res,habillage=13,ellipse=bb)
#write.infile(res,file="my_FactoMineR_results.csv") #to export a list
```

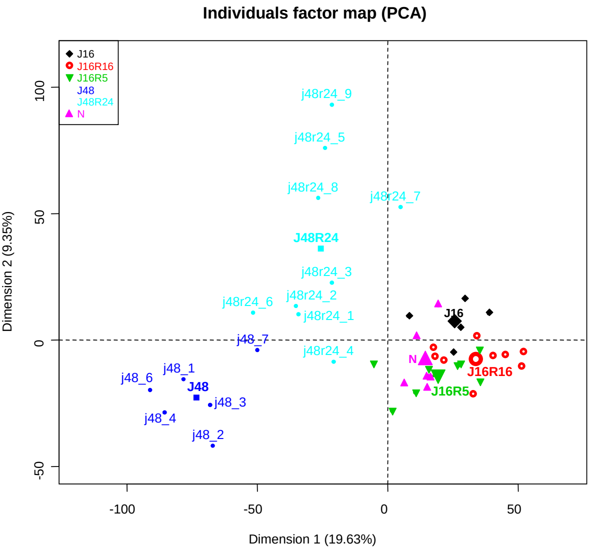
Example

Chicken data:

- 43 chickens (individuals)
- 7407 genes (variables)
- One categorical variable: 6 diets corresponding to different stresses
- Do genes differentially expressed from one stress to another?

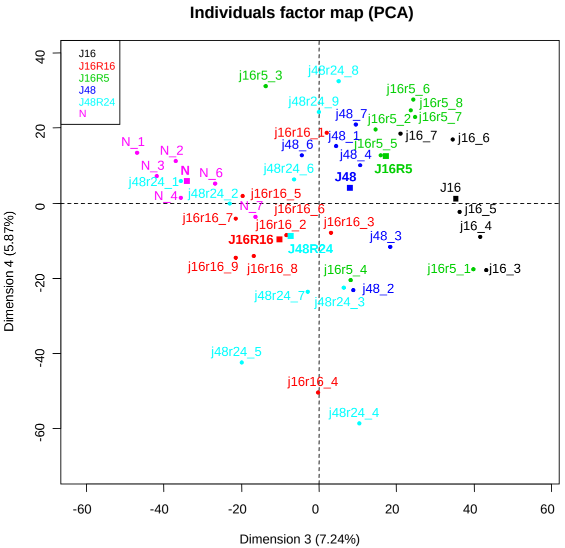
Dimensionality reduction: with a few principal components, we identify the structure of the data

Example (2)

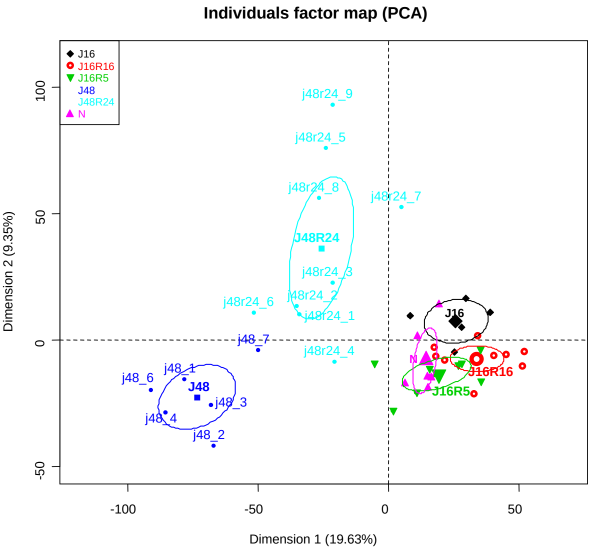


Individuals factor map (PCA)
◆ J16
○ J16R16
▼ J16R5
J48
J48R24
▲ N
j48r24_9
j48r24_5
j48r24_8
j48r24_7
J48R24
j48r24_3
j48r24_2
j48r24_6
j48r24_1
j48_7
j48r24_4
J16
N
J16R16
J16R5
j48_1
j48_6
J48
j48_3
j48_4
j48_2
100
50
0
-50
-100
-50
0
50
Dimension 2 (9.35%)
Dimension 1 (19.63%)

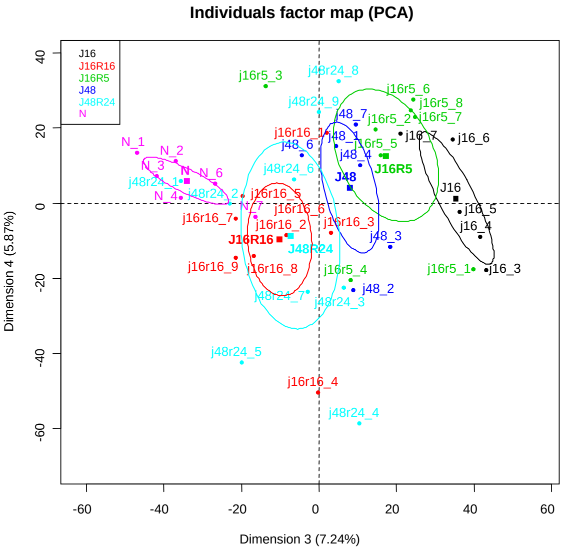
Example (3)


Individuals factor map (PCA)
J16
J16R16
J16R5
J48
J48R24
N
j16r5_3
j48r24_8
j16r5_6
j48r24_9
j16r5_8
j48_7
j16r5_2
j16r5_7
j16r16_1
j48_1
j16_7
j16_6
N_1
N_2
j48_6
j16r5_5
j48_4
J16R5
N_3
N
N_6
j48r24_6
J48
j48r24_1
J16
N_4
j48r24_2
j16r16_5
j16_5
N_7
j16r16_6
j16r16_7
j16r16_2
j16r16_3
j16_4
j48_3
J16R16
J48R24
j16r16_9
j16r16_8
j16r5_4
j16r5_1
j16_3
j48_2
j48r24_7
j48r24_3
j48r24_5
j16r16_4
j48r24_4
Dimension 4 (5.87%)
40
20
0
-20
-40
-60
Dimension 3 (7.24%)
-60
-40
-20
0
20
40
60

Example (4)


Individuals factor map (PCA)
◆ J16
○ J16R16
▼ J16R5
J48
J48R24
▲ N
j48r24_9
j48r24_5
j48r24_8
j48r24_7
J48R24
j48r24_3
j48r24_2
j48r24_6
j48r24_1
j48_7
j48r24_4
J16
N
J16R16
J16R5
j48_6
j48_1
J48
j48_3
j48_4
j48_2
Dimension 2 (9.35%)
100
50
0
-50
-100
-50
0
50
Dimension 1 (19.63%)

Example (5)


Individuals factor map (PCA)
J16
J16R16
J16R5
J48
J48R24
N
Dimension 3 (7.24%)
Dimension 4 (5.87%)
-60
-40
-20
0
20
40
60
j16r5_3
j48r24_8
j16r5_6
j16r5_8
j16r5_7
j16r5_2
j48r24_9
j48_7
j16r16_1
j48_1
j16_7
j16_6
j16r5_5
j48_6
j48_4
J16R5
j48r24_6
J48
N_1
N_2
N_3
N
N_6
j48r24_1
N_4
j48r24_2
j16r16_5
J16
j16_5
N_7
j16r16_6
j16r16_7
j16r16_2
j16r16_3
j16_4
J16R16
j48_3
J48R24
j16r16_9
j16r16_8
j16r5_4
j16r5_1
j16_3
j48_2
j48r24_7
j48r24_3
j48r24_5
j16r16_4
j48r24_4

References

See Chapter 1 of [1] for more.

[1] F. Husson, S. Le, J. Pagès, Exploratory Multivariate Analysis by Example Using R, Second Edition, Chapman & Hall/CRC Computer Science & Data Analysis, CRC Press, 2017. URL https://books.google.com/books?id=nLrODgAAQBAJ