

Data Analysis

Principal component analysis

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

Fall 2018

Data for the PCA

		Variables			
		V_1	V_2	\dots	V_K
Individuals	1	x_{11}	x_{11}	\dots	x_{1K}
	2	x_{21}	\ddots		
	\vdots	\vdots			
	I	x_{I1}			x_{IK}

Many applications:

- Sensory analysis: products - descriptors
- Ecology: plants - measurements; waters - physico-chemical analyses
- Economics and finance: countries - economic indicators
- Microbiology: cheeses - microbiological analyses
- Marketing: brands - measures of satisfaction

Important: Vector $\mathbf{x}_{\bullet k} = (x_{1k}, x_{2k}, \dots, x_{Ik})^T$, $k = 1, \dots, K$, contains the values of the variable V_k for $1, \dots, I$ individuals. The entire $I \times K$ matrix

$$\mathbf{X} = \{x_{ik}\} = \begin{pmatrix} \mathbf{x}_{\bullet 1} & \mathbf{x}_{\bullet 2} & \dots & \mathbf{x}_{\bullet K} \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}$$

Data for the PCA: 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)

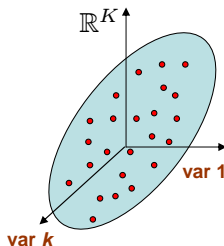
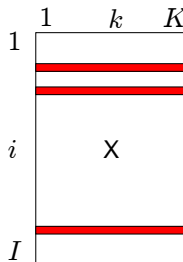
	O.fruity	O.passion	O.citrus	...	Sweetness	Acidity	Bitterness	Astringency	Aroma.intensity	Aroma.persistency	Visual.intensity	Odor.preference	Overall.preference	Label
S Michaud	4.3	2.4	5.7	...	3.5	5.9	4.1	1.4	7.1	6.7	5.0	6.0	5.0	Sauvignon
S Renaudie	4.4	3.1	5.3	...	3.3	6.8	3.8	2.3	7.2	6.6	3.4	5.4	5.5	Sauvignon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

The PCA: objectives

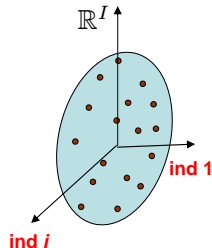
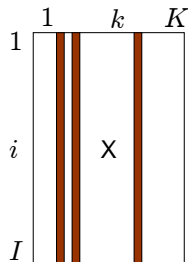
- **Study of individuals**: similarity between individuals with respect to all the variables => partition between individuals
- **Study of variables**: linear relationships between variables => visualization of the correlation matrix; seek for synthetic variables
- **Link between the two studies**: characterization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

Geometrical view on data

Individuals study

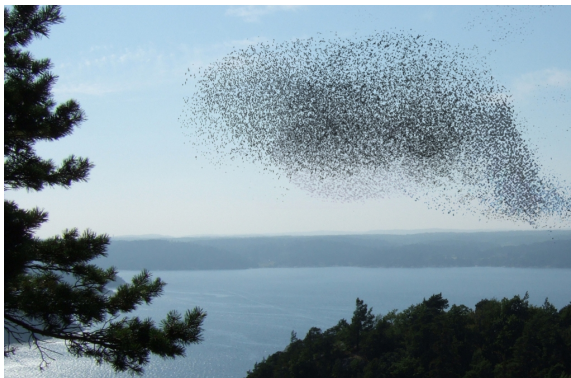


Variables study



- For >3 variables, the visualization is impossible.

Geometrical view on data (2)



- Study the structure, i.e., the shape of the cloud of individuals.
- Individuals are in \mathbb{R}^K . How alike are the individuals? If two points coincide \Rightarrow two individuals do not differ from each other in terms of the variables involved.

Basic statistics of the data

- Sample mean of the k -th variable:

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

where $\mathbf{1} = \underbrace{(1, \dots, 1)}_I^T$.

- Sample variance of the k -th variable:

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

- Sample standard deviation of the k -th variable:

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

- Sample covariation between variables $X_{\bullet m}$ and $X_{\bullet n}$:

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

Basic statistics of the data

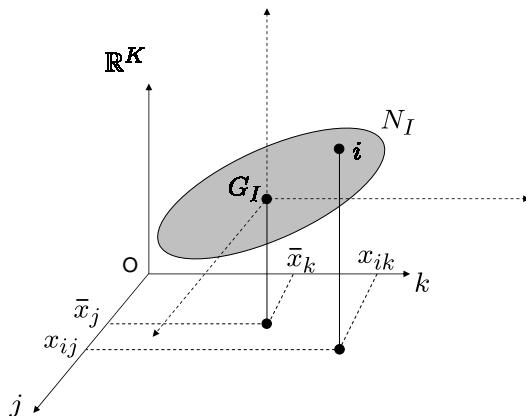
- Sample correlation between variables V_m and V_n :

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

Centering the data

- Centering the data (**always**):

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

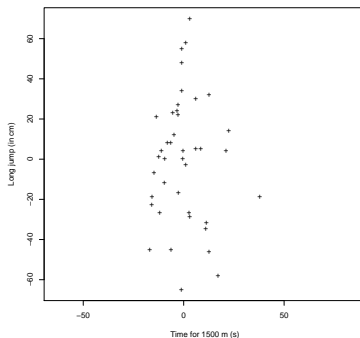
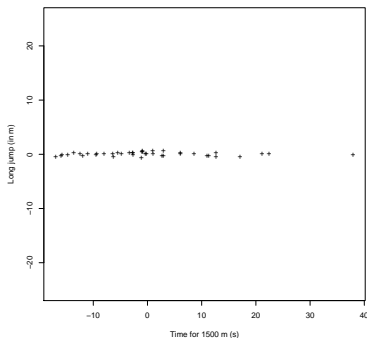


Scaling the data

- Centering AND standardizing (sometimes):

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

- 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.,

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

or, if standardization is used,

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

Sample covariance matrix

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

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

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

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

Trace of Σ

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

$$\begin{aligned} \text{Tr} \Sigma &= \text{Var}(V_1) + \text{Var}(V_2) + \dots + \text{Var}(V_K) = \\ &= \frac{1}{I} \sum_{k=1}^K \sum_{i=1}^I (x_{ik} - \bar{x}_k)^2 = \frac{1}{I} \sum_{i=1}^I \| \mathbf{x}_{i\bullet} - \bar{\mathbf{x}} \|^2 \quad (3) \end{aligned}$$

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

- From (3), the trace of Σ measures the total variability of data.
- Under all subsequent approximations, **we should keep this quantity as close to its genuine value as possible.**
- Mechanically speaking, the trace of Σ is a moment of inertia of the cloud of individuals w.r.t. the point $\bar{\mathbf{x}}$.

PCA as an optimization problem

- Consider $\mathbf{u}_1 \in \mathbb{R}^K$.
- Consider projections of individuals $\mathbf{x}_{i\bullet}$, $i = 1, \dots, I$, onto \mathbf{u}_1 :

$$(\mathbf{x}_{i\bullet} \mathbf{u}_1) \mathbf{e}_{\mathbf{u}_1}$$

- Choose \mathbf{u} so that the (sample) variance of the projected data is maximal (recall portfolio theory as for s.v.):

Sample variance of projected individuals = $\mathbf{u}_1^T \mathbf{X}^T \mathbf{X} \mathbf{u}_1 \rightarrow \max$,

$$\mathbf{u}_1^T \mathbf{u}_1 = 1.$$

- Constraint optimization:

$$\mathcal{L} = \mathbf{u}_1^T \mathbf{X}^T \mathbf{X} \mathbf{u}_1 - \lambda(\mathbf{u}_1^T \mathbf{u}_1 - 1) \rightarrow \max$$

- Solution – eigenvalue problem

$$\mathbf{X}^T \mathbf{X} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

PCA as an optimization problem (2)

- Next, find u_2 such that

Sample variance of individuals projected onto $u_2 = u_2^T \mathbf{X}^T \mathbf{X} u_2 \rightarrow \max$,

$$u_1^T u_1 = 1, \quad u_2^T u_1 = 0.$$

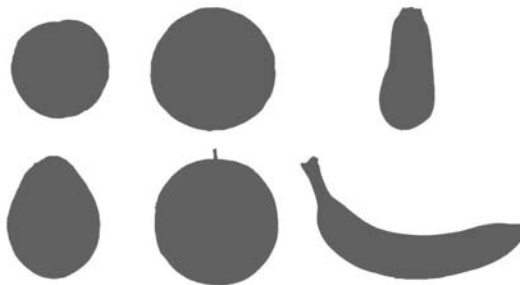
- Again, the problem is reduced to the eigenvalue problem:

$$\mathbf{X}^T \mathbf{X} u_2 = \lambda_2 u_2$$

- Repeat K times.

Reduction of dimensionality: geometrical interpretation

- Geometrically, the PCA seeks for a subspace (normally, the plane) projection onto which retains as much of the inertia (3) 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: geometrical interpretation (2)

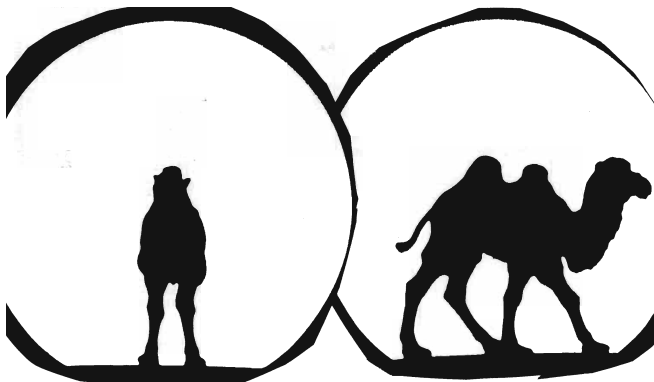


Figure: Who is it? Depends on the viewpoint.

Diagonalizing Σ

- Matrix Σ is symmetric \Rightarrow there exist K real eigenvalues $\lambda_1 > \lambda_2 > \dots > \lambda_K$ and the respective eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_K$
- The vector of the i th individual $\mathbf{x}_{i\bullet} = (x_{i1}, x_{i2}, \dots, x_{iK})^T$ 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 gives the diagonal covariance matrix

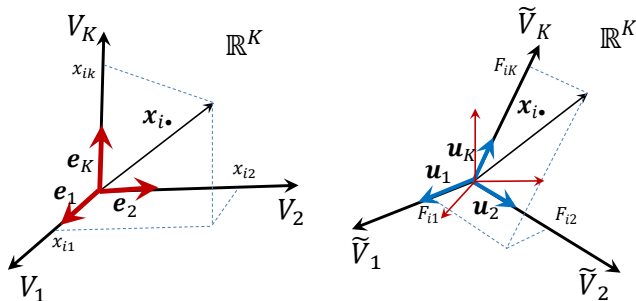
$$\Sigma' = \Lambda^T \Sigma \Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \\ 0 & & & \lambda_K \end{pmatrix} \quad (4)$$

$$\text{where } \Lambda = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_K \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}$$

Rotating the frame

- Original frame:

$$\mathbf{x}_{i\bullet} = x_{i1} \mathbf{e}_1 + x_{i2} \mathbf{e}_2 + \dots + x_{iK} \mathbf{e}_K$$



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

$$\mathbf{x}_{i\bullet} = F_{i1} \mathbf{u}_1 + F_{i2} \mathbf{u}_2 + \dots + F_{iK} \mathbf{u}_K$$

New variables

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

$$\begin{pmatrix} F_{i1} \\ F_{i2} \\ \vdots \\ F_{iK} \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1 \rightarrow \\ \mathbf{u}_2 \rightarrow \\ \\ \mathbf{u}_K \rightarrow \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iK} \end{pmatrix} \quad (5)$$

- The new variables $\tilde{V}_1, \dots, \tilde{V}_K$ (principal components) are **uncorrelated**, see (4).
- The trace of Σ (recall, the total variability of the data) is invariant under change of basis:

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

- So, $\text{Var}(\tilde{V}_k) = \lambda_k$, $k = 1, \dots, K$, see (4).

Reduction of dimensionality

- The key idea of the PCA is to retain just a few new basis elements u_1, u_2, \dots (normally, 2 or 3) which provide the largest contribution to the trace $Tr \Sigma = \text{total variability}$.
- The contributions of the new variables (5) to $Tr \Sigma$ are given by $\lambda_1 > \lambda_2 > \dots > \lambda_K$.
- For example, if

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

it suffices to retain only u_1 and u_2 :

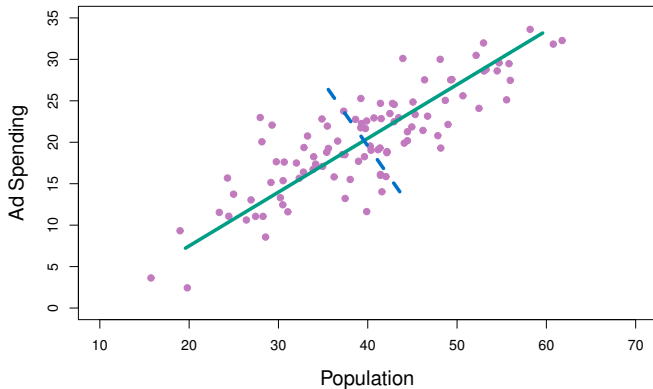
$$\begin{pmatrix} F_{i1} \\ F_{i2} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} u_1 \rightarrow \\ u_2 \rightarrow \\ 0 \rightarrow \\ \vdots \\ 0 \rightarrow \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ \vdots \\ x_{iK} \end{pmatrix}$$

Reduction of dimensionality (2)

- F_{i1} and F_{i2} are called **scores** on the **principal components directions** u_1, u_2 and present the coordinates of individual i in the reduced basis $\{u_1, u_2\}$.
- 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 $V_k, k = 1, \dots, K$.

2D example

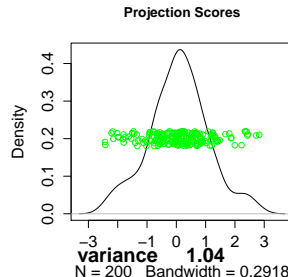
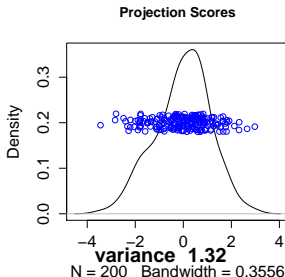
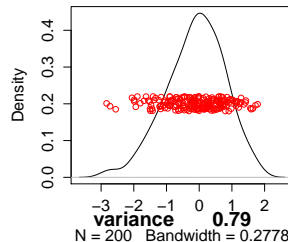
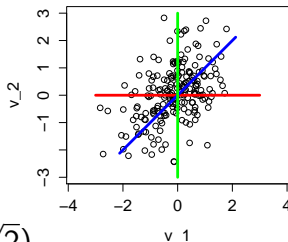
- Projecting onto two orthogonal lines:



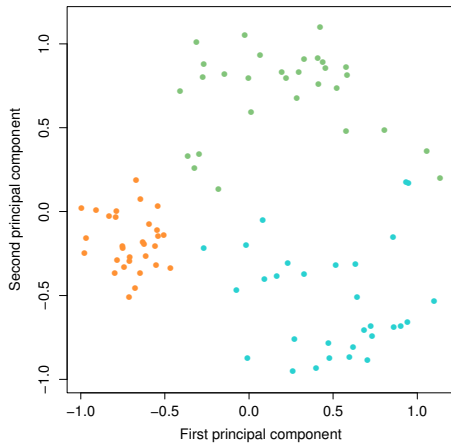
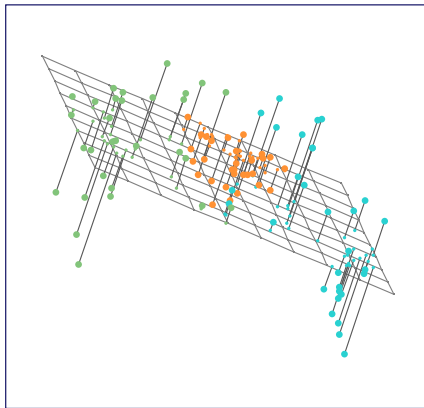
Another 2D example

Take $\mathbf{a} = (1, 0)'$,
 $(0, 1)'$, $(1/\sqrt{2}, 1/\sqrt{2})'$.

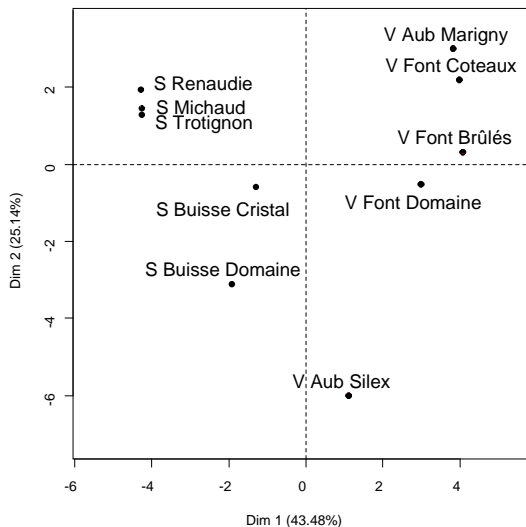
Which
one is better?



3D example

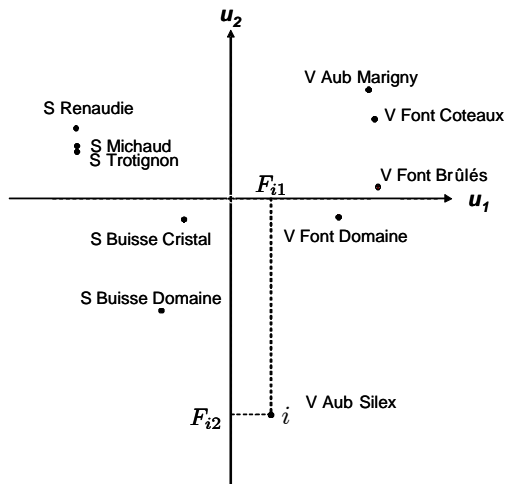


Wine example



- λ_1 and λ_2 account for 43.48% and 25.14% of the total variability => keeping only u_1 and u_2 .
- An approximate 2D depiction of similarity among wines in terms of all the variables.

Coordinates of individuals in the reduced basis



	1	k	K	$F_{\cdot 1}$	$F_{\cdot 2}$
1					
i					
	x_{ik}			F_{i1}	F_{i2}
I					

- $F_{i,1} = \mathbf{u}_1^T \mathbf{x}_{\cdot i}$, $F_{i,2} = \mathbf{u}_2^T \mathbf{x}_{\cdot i}$ are the coordinates of the i -th individual along first two PC directions \mathbf{u}_1 and \mathbf{u}_2 .

Correlation between new and original variables

- Observations of the old variable V_k , $k = 1, \dots, K$ from \mathbf{X} (for I individuals):

$$\mathbf{x}_{\bullet k} = \{x_{1k}, \dots, x_{Ik}\} \quad (6)$$

- Values of the new variable $\tilde{V}_{k'}$, $k' = 1, 2$ (for I individuals):

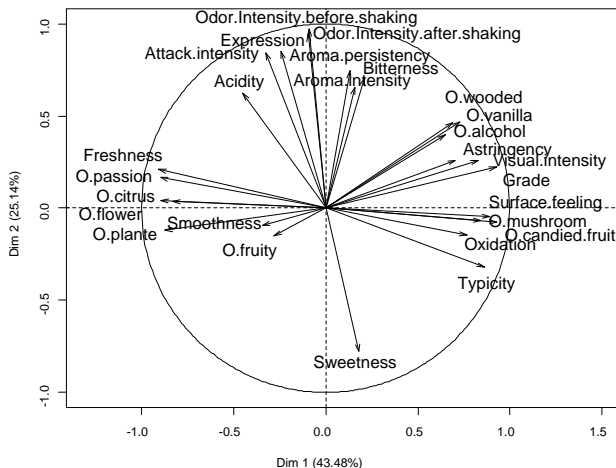
$$\mathbf{F}_{\bullet k'} = \{F_{1k'}, \dots, F_{Ik'}\} \quad (7)$$

- Correlation between V_k and $\tilde{V}_{k'}$ (using (6), (7) data sets):

$$\text{cor}(V_k, \tilde{V}_{k'}) = (\mathbf{u}_{k'})_k \sqrt{\frac{\lambda_{k'}}{s(V_k)}} \quad (8)$$

- **Interpretation:** if (8) 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 26, tends to have a high value of `Visual.Intensity`

Study of variables

- Study of how much variables are alike over individuals.
- Consider the $I \times I$ matrix built from the centered AND standardized data (2)

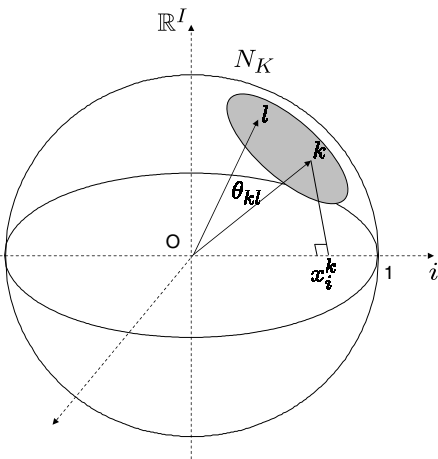
$$\mathbf{XX}^T = \begin{pmatrix} 1 & x_{1\bullet}x_{2\bullet} & \cdots \\ x_{2\bullet}x_{1\bullet} & 1 & \\ \vdots & & \ddots \\ & & & 1 \end{pmatrix}$$

- Apply the PCA to the above matrix, i.e., find the eigenvectors v_1, v_2, \dots, v_I and retain a few that keep enough variability of the data.
- The new variables, $k = 1, \dots, K$:

$$\begin{pmatrix} G_{1k} \\ G_{2k} \\ \vdots \\ G_{Ik} \end{pmatrix} = \begin{pmatrix} v_1 \rightarrow \\ v_2 \rightarrow \\ \vdots \\ v_I \rightarrow \end{pmatrix} \begin{pmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{Ik} \end{pmatrix} \quad (9)$$

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

Study of variables: geometry



- The inner product:

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

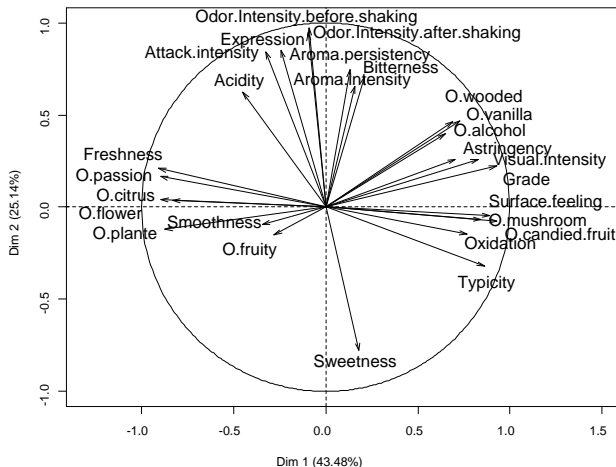
- (Sample) correlation between V_k and V_l :

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

- **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

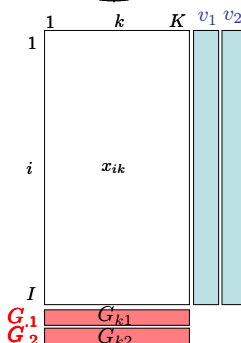
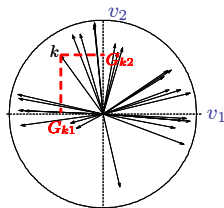
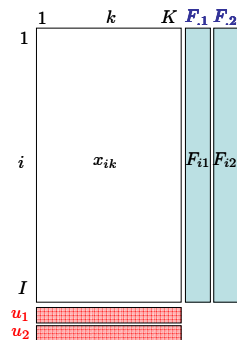
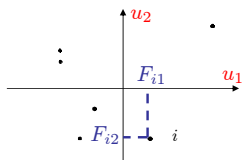


- The same plot as in Slide 29.
- Shows the extent of correlation between the original variables.

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

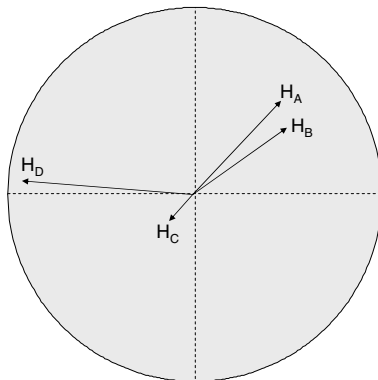
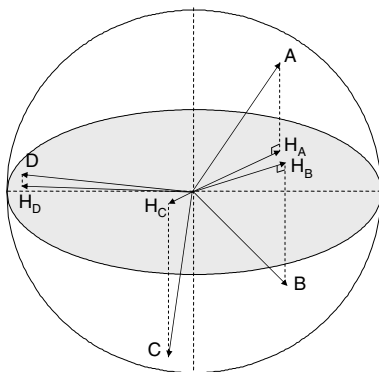
$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k=1}^K x_{ik} G_s(k)$$

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

Quality of representation

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

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



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 s th PC explained by the i th individual:

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

=> 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 s th PC explained by the k th variable:

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

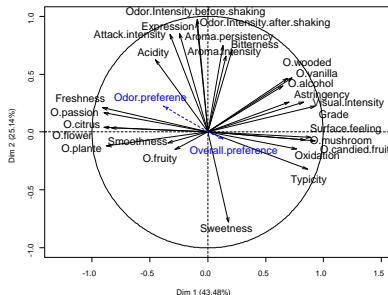
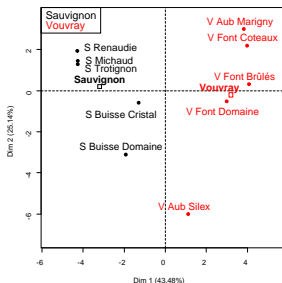
=> variables highly correlated with the s th principal component (i.e., large coordinate along \mathbf{v}_s) contribute the most

- 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

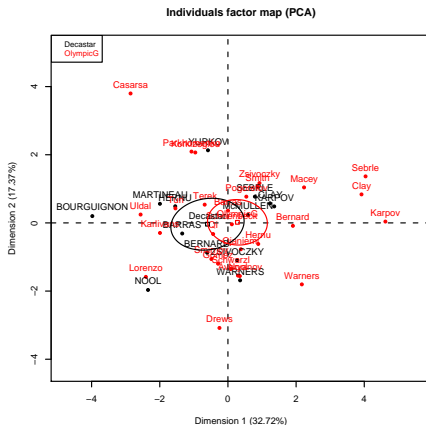
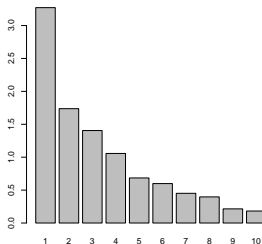


Figure: Confidence ellipses around the barycenter of each category

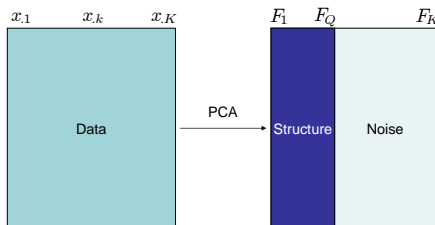
Choosing the number of components

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



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

- ▶ 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
0.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
0.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
0.plante	-0.87	1.0e-03			
0.flower	-0.89	4.9e-04			
0.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
```

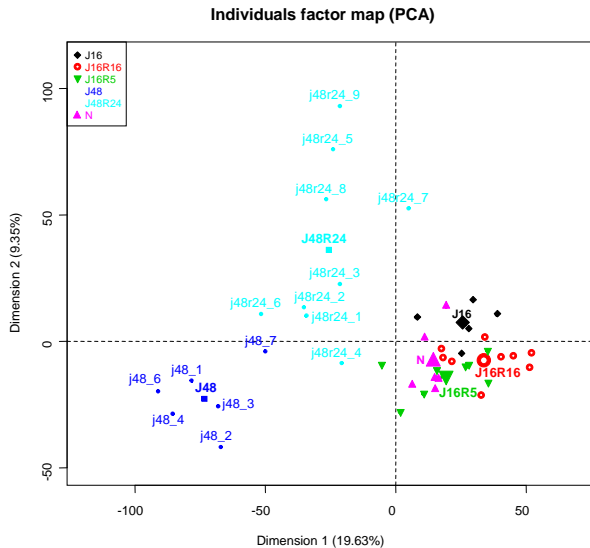
Example

Chicken data:

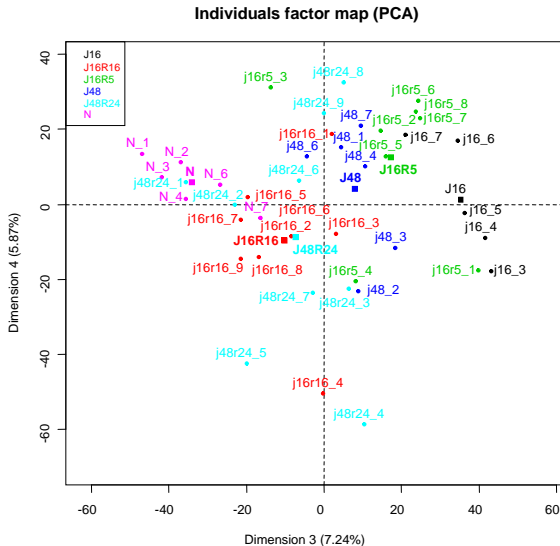
- 43 chickens (individuals)
- 7407 genes (variables)
- One categorical variable: 6 diets corresponding to different stresses
- Do genes differentially expressed from one stress to another?

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

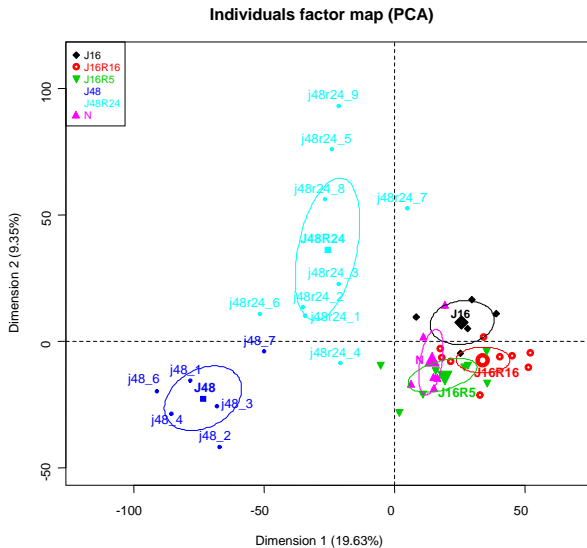
Example (2)



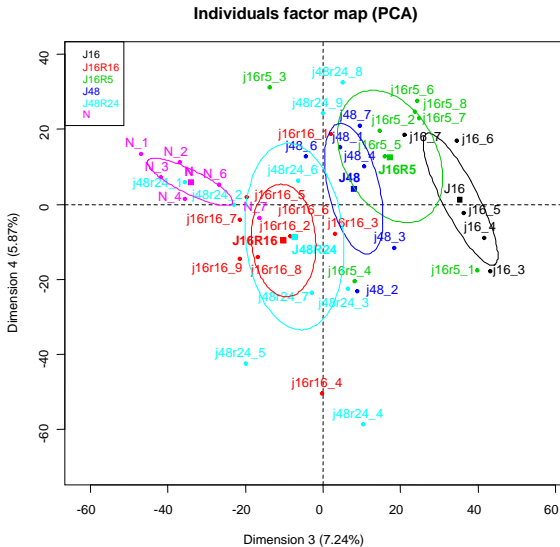
Example (3)



Example (4)



Example (5)



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>