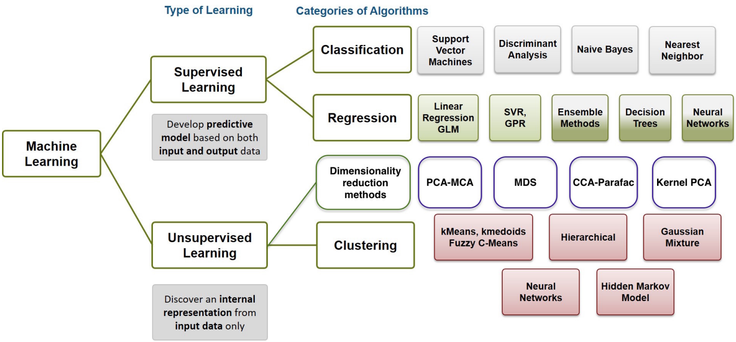
MSc Data Science for Business

Introduction to Machine Learning

Map 534

Julie Josse

Machine Learning



Outline

Supervised learning

- We have training data $D_n = [(x_1, y_1), \dots, (x_n, y_n)]$
- Construct a predictor $\hat{f} : \mathcal{X} \to \mathcal{Y}$ using D_n
- Loss $\ell(y, f(x))$ measures how well $f(x)$ predicts y well
- Aim: minimize the generalization error
  The goal is clear: predict the label y based on features x

Unsupervised learning

- We have training data $D_n = [x_1, \dots, x_n]$
- Loss: ????, Aim: ????
  The goal is less well defined.
  Clustering: construct homogeneous groups
  Dimension reduction: construct a map to visualize the data

Outline

Unsupervised learning

- Dimensionality reduction methods (PCA)
- Clustering (k-means, Hierarchical clustering)
- Handling missing values/ matrix completion
  $\Rightarrow$ Data visualization - exploratory data analysis

Supervised learning

- Theoritical framework - Bayes risk (LDA, SVM)
- Logistic regression
- Optimization

References

G. James, D. Witten, T. Hastie, and R. Tibshirani (2013)
An Introduction to Statistical Learning with Applications in R
Springer Series in Statistics.

MOOC data sciences coursera Johns Hopkins (J. Leek, B. Caiffo, R. D. Peng) - Youtube

Practical information

Time: Wednesday Amphi Bequerel - PC16-17-18

Team: Julie Josse/ Sylvain Lecorff/ Genevieve Robin- Florian Bourgey

Grades:

- 50% Homeworks (2) including case studies.
  Reproducible report (pdf and a .Rmd file) should be submitted. Other homeworks with correction and TA help if needed.
- 50% Final exam (19 december morning)

Teaching Assistant: Genevieve Robin - Florian Bourgey

Office hours Genevieve Robin & Florian Bourgey: Tuesday PC18 - Thursday PC104 from 5 pm. Professors: appointment by email - JJ: office hour Tuesday evening (6pm-).

Outline

Principal Component Analysis

J. Josse

Plan

Principal Component Analysis

J. Josse

Dimensionality reduction methods

⇒ Find a low-dimensional representation that captures the "essence" of high-dimensional data

- compression, denoising, data completion, anomaly detection
- preprocessing before supervised learning (improve performances / regularization to reduce overfitting)

⇒ descriptive methods, data visualization tools to better understand the data (difficult to plot and interpret $> 3d$)

- Principal component Analysis (PCA): continuous data
- Correspondence analysis (CA): contingency table
- Multiple correspondence analysis (MCA): categorical data
- Multiple factor analysis (MFA): multi-table, array data

Exploratory/inferential. "Let the data speak" J.P. Benzecri 1960.

Principal Component Analysis

1. Data - Issues - Preprocessing
2. Observations Study
3. Variables Study
4. Interpretation Tools

Plan

PCA for which data?

PCA deals with continuous variables, but categorical ones can also be included

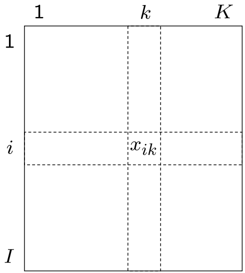


Figure: Data table

Many examples:

- Environmental data: waters - physico-chemical analyses, towns - temperature
- Economy: countries - economic indicators
- Biology: cheeses - microbiological analyses, tumors - genes expression
- etc.

Wine data

- 10 observations (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

Objectives

- Observations study:
  similarity between observations with respect to all the variables
  partition between observations

- Variables study:
  linear relationships between variables
  visualization of the correlation matrix
  find synthetic variables

- Link between the two studies:
  characterization of the groups of observations with variables
  specific observations to understand links between variables

Plan

Preprocessing... (not an appropriate word?)

$\Rightarrow$ Similarity between observations: Euclidean distance

- Choosing active variables

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

- Variables are centered

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

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

- Standardizing variables or not?

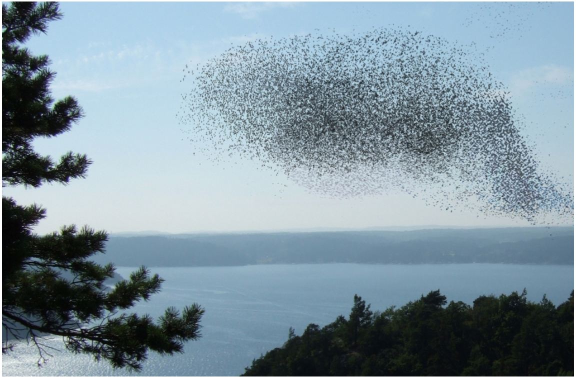
$$d^2(i,i') = \sum_{k=1}^{K}\frac{1}{s_k^2}(x_{ik} - x_{i'k})^2$$

Wine data

- Sensory descriptors as active: only these variables are used to define the dimensions
- 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

Observations cloud



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

Fit the observations cloud

Find the subspace which best sums up the data

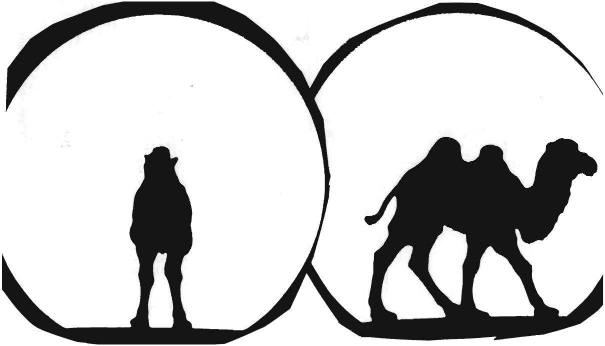


Figure: Camel vs dromedary? (J.P. Fenelon)

⇒ Closest representation by projection

⇒ Best representation of the diversity, variability

Fit the observations cloud

$1d$ subspace identified by a unit vector $\|u_1\|_2^2 = 1$

$$\begin{aligned} P_{u_1}(x_{i.}) &= u_1(u_1'u_1)^{-1}u_1'x_{i.} = u_1 < x_{i.}, u_1 > \\ F_{i1} &= < x_{i.}, u_1 > \end{aligned}$$

Minimize distance between obs and their projections

Maximize the variance (inertia) of the projected data

$$u_1^\star = \underset{u_1 \in \mathbb{R}^K,\ \|u_1\|_2^2=1}{\arg\max} \frac{1}{I}\sum_{i=1}^{I} F_{i1}^2 = \underset{u_1 \in \mathbb{R}^K,\ \|u_1\|_2^2=1}{\arg\max} \frac{1}{I}\sum_{i=1}^{I}(u_1'x_{i.})^2$$

u_1 loadings - $F_{.1}$ principal component (PC), scores

$$\max_{u_1 \in \mathbb{R}^K,\ \|u_1\|_2^2=1} u_1^{'} \left(\sum_{i=1}^{I} \frac{1}{I} x_{i.}x_{i.}^{'}\right) u_1 = \frac{u_1^{'}X^{'}Xu_1}{I}$$

$\Rightarrow u_1^\star$ the first eigenvector of the covariance matrix $S = \frac{X'X}{I}$ associated with the largest eigenvalue λ_1. $\mathrm{var}(F_{.1}) = \lambda_1$

Fit the observations cloud

Additional axes sequentially defined: maximizes the projected inertia among all orthogonal directions. Eigenvectors $u_1,\dots,u_Q$ with $\lambda_1,\dots,\lambda_Q$

$\Rightarrow$ Representation quality (dimensionality reduction loses information):

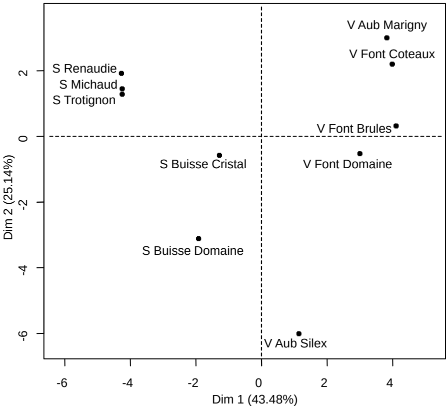
- Total variance of the observations cloud (total inertia):

$$\sum_i p_i d(x_{i.}, G_I)^2 = \frac{1}{I}\sum_i \sum_k (x_{ik})^2 = tr(S) = \sum_{k=1}^{K} \lambda_k \quad (= K)$$

- Variance of the projected observations cloud (Q-dimensional representation): $var(F_{.1}) + var(F_{.2}) + \dots + var(F_{.Q})$

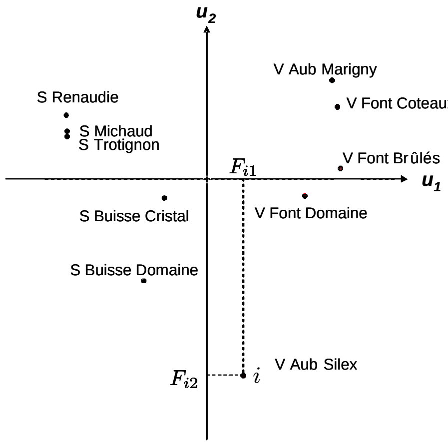
$\Rightarrow$ Percentage of inertia explained: $\frac{\sum_{k=1}^{Q} \lambda_k}{\sum_{k=1}^{K} \lambda_k}$

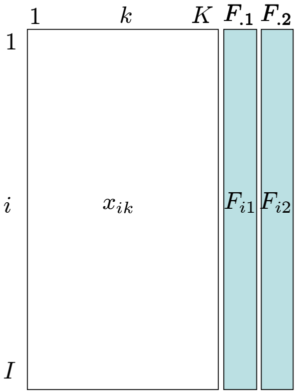
Wine: graph of observations



$\Rightarrow$ Need variables to interpret the dimensions of variability

Observations coordinates considered as variables

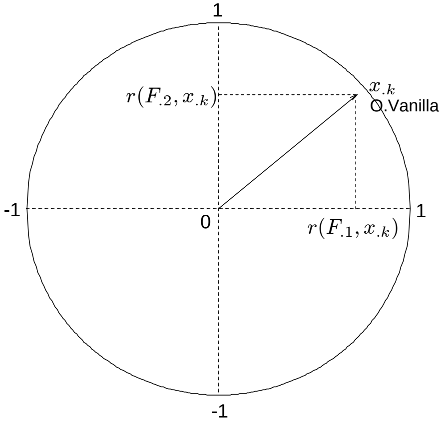




$F = Xu$ (linear combination of variables)

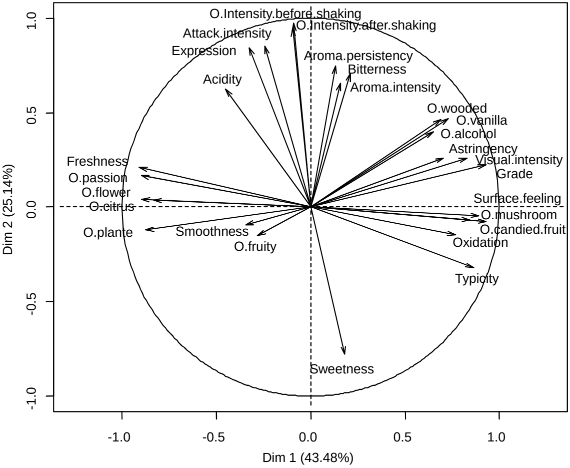
Interpretation of the observations graph with the variables

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



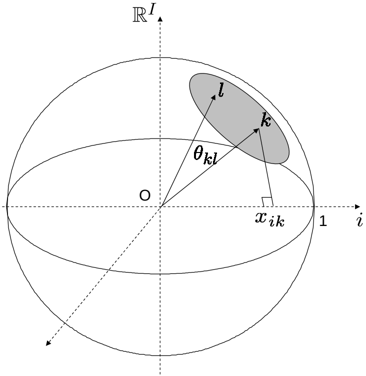
$\Rightarrow$ Correlation circle

Interpretation of the observations graph with the variables



Plan

Cloud of variables

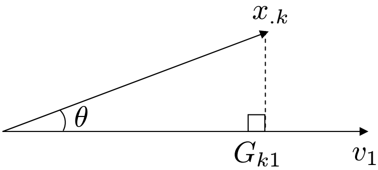


Inner product $< x_{.k}, x_{.l} >_D = \frac{1}{I}\sum_{i=1}^{I} x_{ik}x_{il}$ induces correspondence between geometry and statistics. D diag matrix with $1/I$

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

Fit the variables cloud

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



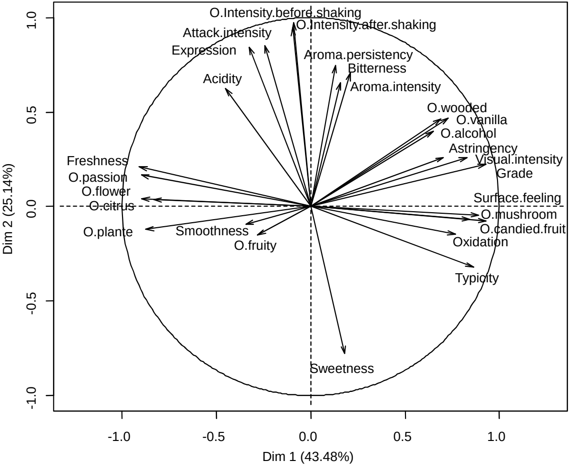
$$
\begin{aligned}
P_{v_1}(x_{.k}) &= v_1 (v_1' D v_1)^{-1} v_1' D x_{.k} \\
G_{k1} &= < x.k, v_1 >_D \\
G_{k1} &= \frac{< v_1, x.k >_D}{\|v_1\|_D \; \|x_{.k}\|_D}
\end{aligned}
$$

$$
\underset{v_1 \in \mathbb{R}^I,\ \|v_1\|_D^2 = 1}{\arg\max} \sum_{i=k}^{K} G_{k1}^2 = \underset{v_1 \in \mathbb{R}^I,\ \|v_1\|_D^2 = 1}{\arg\max} \sum_{i=k}^{K} r(v_1, x_{.k})^2
$$

$\Rightarrow$ v_1 is the best synthetic variable

$\Rightarrow$ $v_1, \ldots, v_Q$ are eigenvectors of $WD = XX'D$ associated with the largest eigenvalues: $WDv_q = \lambda_q v_q$

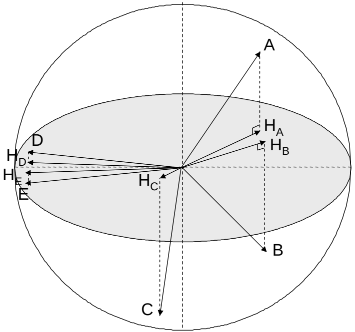
Fit the variables cloud

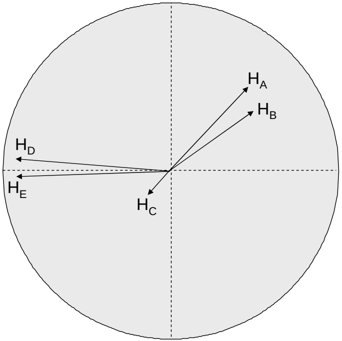


⇒ Same representation

Projections...

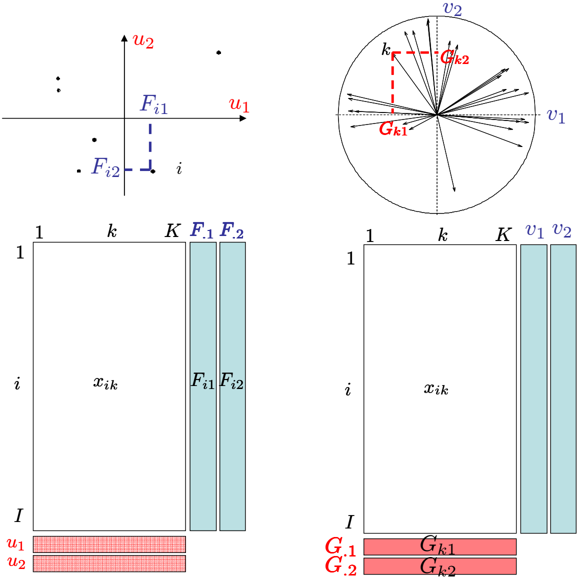
$r(x_{.1}, x_{.2}) = cos(\theta_{x_{.1},x_{.2}})$
$cos(\theta_{x_{.1},x_{.2}}) \approx cos(\theta_{H_1,H_2})$ if variables are well projected





Only well projected variables can be interpreted

Link between the two representations: transition formulae



Link between the two representations: transition formulae

- $Su = (1/I)X'Xu = \lambda u$
- $XX'DXu = X\lambda u \rightarrow WD(Xu) = \lambda(Xu)$
- $WDF = \lambda F$ and $WDv = \lambda v$: F and v are collinear
- $||F||_D^2 = \lambda$ and $||v||_D^2 = 1$:

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

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

$F_{.q}$: principal components (variance=eigenvalues), scores

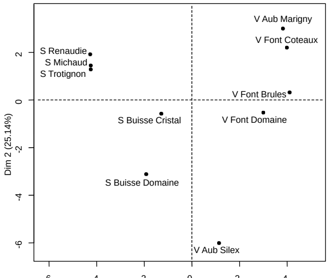
$G_{.q}$: correlations between variables and principal components

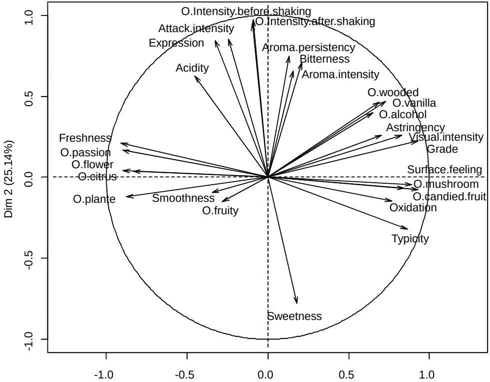
Link between the two representations: transition formulae

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

$$G_{kq} = \frac{1}{\sqrt{\lambda_q}} \sum_{i=1}^{I} (1/I) x_{ik} F_{iq}$$

Observation on the side of the variables where it takes high values

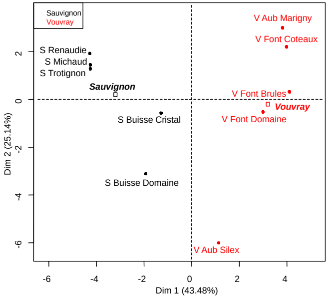


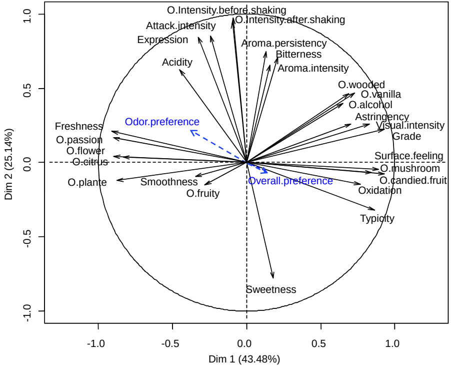


Plan

Supplementary information

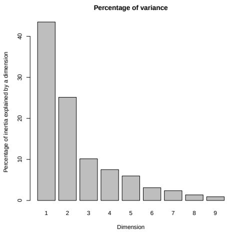
- continuous variables: projection (correlation with dimensions)
- observations: projection
- categorical variables: projection of the categories at the barycentre of the observations which take the categories





Choosing the number of components

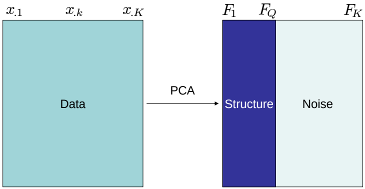
Bar plot, test on eigenvalues, confidence interval, cross-validation, generalized cross-validation, etc.



Objectives:

$\Rightarrow$ Interpretation

$\Rightarrow$ Separate structure and noise



Percentage of inertia obtained under independence

⇒ Is there any structure in the data?

	Number of variables												
Nb obs	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 inertia obtained under independence

	Number of variables												
Nb ind	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: $\cos^2(\theta)$

$\Rightarrow$ Projected inertia of an element / total inertia of the element

- observations: $\frac{F_{iq}^2}{d^2(x_{i.}, G_I)} = \frac{F_{iq}^2}{\sum_{q=1}^{K} F_{iq}^2}$

```
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
```

- variables: squared coordinate

```
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
```

$\Rightarrow$ Only well projected elements can be interpreted

Contribution

$\Rightarrow$ Contribution to the dimension (percentage of variability)

- observation: $\text{Ctr}_q(x_{i.}) = \dfrac{(1/I)F_{iq}^2}{\sum_{i=1}^{I}(1/I)F_{iq}^2} = \dfrac{(1/I)F_{iq}^2}{\lambda_q}$

  $\Rightarrow$ observations with 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
```

- variables: $\text{Ctr}_q(x_{k.}) = \dfrac{G_{kq}^2}{\lambda_q} = \dfrac{r(x_{.k}, v_q)^2}{\lambda_q}$

  $\Rightarrow$ variables with large correlation contribute the most

Description of dimensions

Using continuous variables:

- correlation between variable and the principal components
- sort correlation coefficients and give significant ones (rought tests)

```
> 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 dimensions

Using categorical variables:

- One-way anova with the coordinates of the observations ($F_{.q}$) explained by the categorical variable
  - 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. Select active variables
2. Scale or not the variables
3. Perform PCA
4. Choose the number of dimensions to interpret
5. Simultaneously interpret the observations and variables graphs
6. Use interpretation tools

```
library(FactoMineR)
Expert <- read.table("http://factominer.free.fr/course/doc/data_PCA_ExpertWine.
   header = TRUE, sep = ";", row.names = 1)
res.pca <- PCA(Expert, scale = T, quanti.sup = 29:30, quali.sup = 1)
summary(res.pca)
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)
plotellipses(res.pca, 1)
write.infile(res.pca, file = "my_FactoMineR_results.csv")
```

References PCA

Jolliffe (2002): PCA Springer

Multivariate Analysis - Susan Holmes

A Generalized Least Squares Matrix Decomposition, Allen G. *et al.* JASA (2014)

Multivariate Data Analysis: The French Way (S. Holmes)

References FactoMineR

Exploratory Multivariate Analysis by Example using R, Husson, Le, Pages (2017), Chapman & Hall
Multiple Factor Analysis by Example using R, Pages (2015), CRC Press

Package `FactoMineR`: `http://factominer.free.fr`
Youtube: playlist

Packages - code: ade4, prcomp, pricomp

Degustation

Degustation

Degustation

Plan

To go further

- Low-rank matrix approximation - SVD
- Selecting the number of components
- Inference in PCA

Plan

Minimize the reconstruction error

$\Rightarrow$ Minimize the distance between observations and their projection

A projection of $x_{i.}$ on a Q dimensional subspace: $uu'x_{i.} = uF_{i.}$ for some $u \in \mathbb{R}^{K \times Q}$ with $u'u = I_Q$.

$$u^\star = \underset{u \in \mathbb{R}^{K \times Q},\ u'u = I_Q}{\arg\min} \sum_{i=1}^{I} \|x_{i.} - uu'x_{i.}\|_2^2$$

$\Rightarrow$ Solution given with u the Q leading eigenvectors of S ($K \times K$).

Solution can also be obtained with the diagonalization of the inner-product matrix WD (of size $I \times I$) when K is large

SVD

SVD of $D^{1/2}X$ at order Q

$$U\Lambda^{1/2}V'$$

with $U^tU = V^tV = \mathbb{I}_Q$,

- $U_{I\times Q}$ eigenvectors of $WD = XX^tD$
- $V_{K\times Q}$ eigenvectors of $S = X^tDX$
- Λ diagonal matrix with eigenvalues (of S and WD)

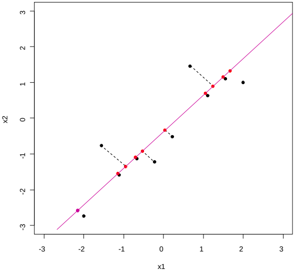
- U standardized principal component (scores) - $F = U\Lambda^{1/2}$ the principal component (scores - variance eigenvalue)
- V the loadings (u)

$\Rightarrow$ Power method to compute the first singular vector.

Fitting a cloud of points

X

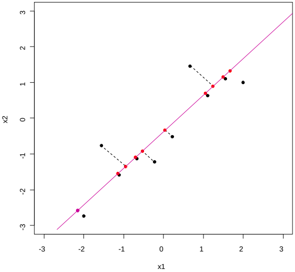
-2.00	-2.74
-1.56	-0.77
-1.11	-1.59
-0.67	-1.13
-0.22	-1.22
0.22	-0.52
0.67	1.46
1.11	0.63
1.56	1.10
2.00	1.00



Fitting a cloud of points

X

-2.00	-2.74
-1.56	-0.77
-1.11	-1.59
-0.67	-1.13
-0.22	-1.22
0.22	-0.52
0.67	1.46
1.11	0.63
1.56	1.10
2.00	1.00



-2.16	-2.58
-0.96	-1.35
-1.15	-1.55
-0.70	-1.09
-0.53	-0.92
0.04	-0.34
1.24	0.89
1.05	0.69
1.50	1.15
1.67	1.33

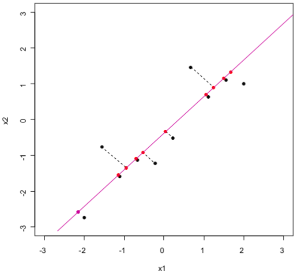
$\widehat{X}$

Fitting a cloud of points

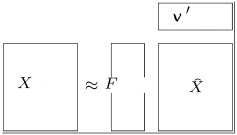
X

-2.00	-2.74
-1.56	-0.77
-1.11	-1.59
-0.67	-1.13
-0.22	-1.22
0.22	-0.52
0.67	1.46
1.11	0.63
1.56	1.10
2.00	1.00

F	u': 0.699	0.715
-3.09	-2.16	-2.58
-1.37	-0.96	-1.35
-1.64	-1.15	-1.55
-1.00	-0.70	-1.09
-0.76	-0.53	-0.92
0.06	0.04	-0.34
1.78	1.24	0.89
1.50	1.05	0.69
2.14	1.50	1.15
2.38	1.67	1.33



$\hat{X} = F u'$



$\hat{X} = Fu^t$

$\Rightarrow$ Approximation of X with a low rank matrix $Q < K$

Low rank matrix approximation - model

$\Rightarrow$ Model: $X \in \mathbb{R}^{I \times K} \sim \mathcal{L}(\mu)$ with $\mathbb{E}\left[X\right] = \mu$ of low-rank Q

$\Rightarrow$ Gaussian noise model: $X_{I \times K} = \mu_{I \times K} + \varepsilon_{I \times K}, \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$

$$\text{argmin}_{\mu} \left\{ \|X - \mu\|_2^2 : \text{rank}\left(\mu\right) \leq Q \right\}$$

$\Rightarrow$ Solution: the truncated SVD of X. Eckart-Young (1936).

$\Rightarrow$ Least squares solution = maximum likelihood.

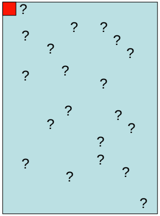
Plan

Many methods

Jolliffe (2002):

- scree-tests
- tests based on distributional assumptions
- computational methods (bootstrap, permutation, cv)

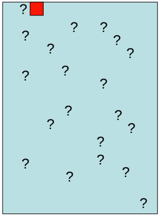
Cross validation



$\Rightarrow$ EM-CV (Bro *et al.* 2008)

$$\text{MSEP}_Q = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} (x_{ik} - \hat{x}_{ik}^{-ik})^2$$

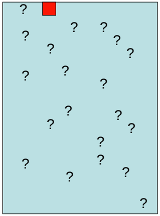
Cross validation



$\Rightarrow$ EM-CV (Bro *et al.* 2008)

$$\mathrm{MSEP}_Q = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} (x_{ik} - \hat{x}_{ik}^{-ik})^2$$

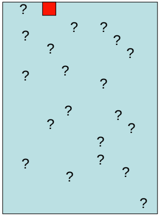
Cross validation



$\Rightarrow$ EM-CV (Bro *et al.* 2008)

$$\text{MSEP}_Q = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}(x_{ik} - \hat{x}_{ik}^{-ik})^2$$

Cross validation



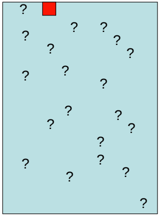
$\Rightarrow$ EM-CV (Bro *et al.* 2008)

$$\text{MSEP}_Q = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} (x_{ik} - \hat{x}_{ik}^{-ik})^2$$

$\Rightarrow$ In regression $\hat{y} = Py$ (Craven & Wahba, 1979):

$$\hat{y}_i^{-i} - y_i = \frac{\hat{y}_i - y_i}{1 - P_{i,i}}$$

Cross validation



$\Rightarrow$ EM-CV (Bro *et al.* 2008)

$$\text{MSEP}_Q = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}(x_{ik} - \hat{x}_{ik}^{-ik})^2$$

$\Rightarrow$ In regression $\hat{y} = Py$ (Craven & Wahba, 1979):

$$\hat{y}_i^{-i} - y_i = \frac{\hat{y}_i - y_i}{1 - P_{i,i}}$$

$\Rightarrow$ Aim: write PCA as $\hat{X} = PX$

$$\hat{x}_{ik}^{-ik} - x_{ik} \simeq \frac{\hat{x}_{ik} - x_{ij}}{1 - P_{ik,ik}}$$

Projection in PCA

$\Rightarrow$ Approximation of X of low rank ($Q < p$):

$\|X_{I\times K} - \hat{X}_{I\times K}\|^2$ SVD: $\hat{X} = U_{I\times Q}\Lambda^{\frac{1}{2}}_{Q\times Q}V'_{K\times Q} = F_{I\times Q}V'_{K\times Q}$

$\Rightarrow$ 2 projection matrices

$$\begin{cases} V' = (F'F)^{-1}F'X & \Rightarrow P_F = F(F'F)^{-1}F' \\ F = XV(V'V)^{-1} & \Rightarrow P_V = V(V'V)^{-1}V' \end{cases}$$

$\hat{X}^{(Q)} = FV' \Rightarrow \hat{X}^{(Q)} = P_F X = XP_V$ Pazman & Denis, 2002; Candes & Tao, 2009

$\hat{\varepsilon} = X - \hat{X}^{(Q)} = (\mathbb{I}_I - P_F)X(\mathbb{I}_K - P_V)$

$\text{vec}(\hat{X}^{(Q)}) = P\text{vec}(X)$ $P_{IK\times IK} = (P'_V \otimes \mathbb{I}_I) + (\mathbb{I}'_K \otimes P_F) - (P'_V \otimes P_F)$

Approximations

$\Rightarrow$ Number of parameters

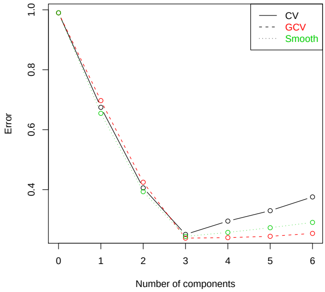
$$\mathrm{tr}\left(P^{(Q)}\right) = IQ - KQ + Q^2$$

$\Rightarrow$ Cross-validation approximation

$$\begin{aligned}
\hat{x}_{ik}^{-ik} - x_{ik} &\simeq \frac{\hat{x}_{ik} - x_{ik}}{1 - P_{ik,ik}} \\
\mathrm{ACV}_Q &= \frac{1}{IK} \sum_{i,j} \left(\frac{\hat{x}_{ik} - x_{ik}}{1 - P_{ik,ik}} \right)^2 \\
\mathrm{GCV}_Q &= \frac{1}{IK} \times \frac{\sum_{i,j} (\hat{x}_{ik} - x_{ik})^2}{(1 - \mathrm{tr}(P^{(Q)})/IK)^2}
\end{aligned}$$

Approximations

```
> nb <- estim_ncp(don)
> nb$ncp
> nb$criterion
        0         1         2         3         4         5
1.2884873 0.8069719 0.6400517 0.7045074 2.2257738 3.0274337
```

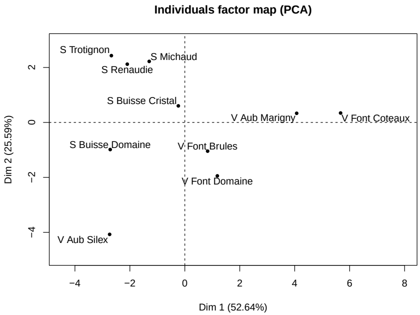


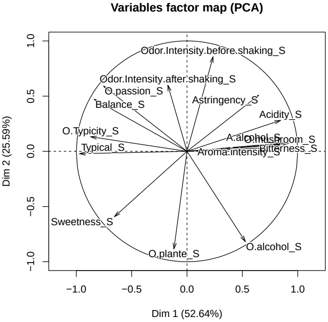
Josse, J. & Husson, F. (2011). Selecting the number of components in PCA using cross-validation approximations. *Computational Statistics and Data Analysis*. 56 (6), pp. 1869-1879.

Plan

Inference in PCA

⇒ PCA on a full population data?

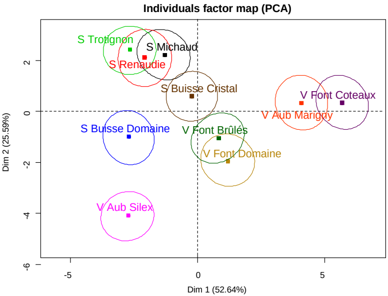


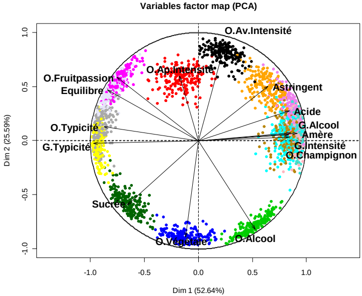


Many examples: plant breeding (genotypes - environments); economy (countries - indicators), climate (cities - temperature) ...

Inference in PCA

⇒ PCA on a full population data?





Many examples: plant breeding (genotypes - environments); economy (countries - indicators), climate (cities - temperature) ...

Inference in PCA

$\Rightarrow$ PCA on a random sample from a population

- Observations bootstrap (Holmes, 1985, Timmerman *et al*, 2007)
- Sampling variability
- Confidence areas around the position of the variables

$\Rightarrow$ PCA on a full population data? $x_{ik} = \mu + \varepsilon_{ik}$

- Residuals bootstrap
- Fluctuations due to the noise
- Confidence areas around the observations and the variables

Same in regression

Bootstrap confidence ellipses

- PCA on $X \Rightarrow F_{I \times Q}$ and $V_{K \times Q}$ (Q dimensions are kept)
- Model matrix $\hat{X} = FV'$ and residuals $\hat{\varepsilon} = X - \hat{X}$
- Bootstrap procedure: repeat B times
  1. residuals are bootstrapped or drawn from $\mathcal{N}(0, \hat{\sigma}^2)$: ε^b
  2. $X^b = \hat{X} + \varepsilon^b$
  3. PCA on X^b to obtain F^b and V^b

$\Rightarrow$ B matrices $\hat{X}^1 = F^1 V^{1'}, ..., \hat{X}^B = F^B V^{B'}$

Bootstrap confidence ellipses

- PCA on $X \Rightarrow F_{I\times Q}$ and $V_{K\times Q}$ (Q dimensions are kept)
- Model matrix $\hat{X} = FV'$ and residuals $\hat{\varepsilon} = X - \hat{X}$
  $\Rightarrow$ Number of dimensions?
- Bootstrap procedure: repeat B times

  1. residuals are bootstrapped or drawn from $\mathcal{N}(0, \hat{\sigma}^2)$: ε^b
     $\Rightarrow$ Under-estimation of the residuals?
  2. $X^b = \hat{X} + \varepsilon^b$
  3. PCA on X^b to obtain F^b and V^b

$\Rightarrow$ B matrices $\hat{X}^1 = F^1 V^{1'}, ..., \hat{X}^B = F^B V^{B'}$
$\Rightarrow$ Visualization?

Eigen values, variance and inertia

- $var(F_{.1}) = \frac{1}{I}\sum_{i=1}^{I} F_{i1}^2 - (\frac{1}{I}\sum_{i=1}^{I} F_{i1})^2$
- From the previous slides $\frac{1}{I}\sum_{i=1}^{I} F_{i1}^2 = \lambda_1$
- $\frac{1}{I}\sum_{i=1}^{I} F_{i1} = \frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K} x_{ik}u_1 = \sum_{k=1}^{K} u_1\bar{x}_k = 0$ (data is centered)
- Why do we have $\sum_{k=1}^{K}\lambda_k = K$?
- $Tr(S) = Tr(\frac{X^TX}{I}) = \sum_{k=1}^{K}\lambda_k$
- And the matrix is standardized : $Tr(S) = \frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K} x_{ik}^2 = \sum_{k=1}^{k} 1 = K$