IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Principal Component Analysis

Lab 1: Illustration of PCA with real datasets

Romain Billot

Please send comment to (version 1.0 – September 30, 2019):
romain.billot@imt-atlantique.fr

1 About the evaluation of labs (travaux pratiques) during the semester

Here are the rules for the semester:

- R and Python are the official programming languages. Some labs (e.g today) are more adapted to one of the two solutions (e.g R for today) but it is always possible to use the other one. Differences lie in the presence of comprehensive libraries for some of the methods used.
- You will work in tandem. Today you choose a lab partner that you will keep during the semester. At the end of each lab, you have to send by email a small report that addresses the questions, show your results and your R or Python code. Notebooks links (e.g Jupyter) are allowed.
- During the semester, your teacher will evaluate two labs. Tomorrow, lab 2 about clustering will be evaluated for all groups. For your second mark, one of the following lab reports will be picked randomly.
- Today is a training. No report has to be sent. Take your time to get familiar with your programming environment and get ready for tomorrow.

2 Tutorial about PCA

To complement the short lecture about PCA, here is some material that could help:

1. Very nice tutorial in French
2. Part I in english
3. Part II in english
4. Part III in english

3 Brest, so hot

You can use your favorite library (FactoMiner in R or Scikitlearn.decomposition in Python) to build a PCA on the dataset villes.txt which contains data about 15 French city. Data represent monthly average temperature data over a period of 30 years (1931 to 1960).

Question 1

Prepare a scaled PCA of the dataset considering the first 12 features as active variables and the last 4 as illustrative (supplementary) variables

Question 2

Analyze the descriptive statistics of each of the variables, the matrix of correlation (which variables are the most correlated?), eigenvalues (what is the percentage of inertia explained by the first axis, the second? how many axes can can we keep?)

Question 3

What is the first principal component? Study the representation of individuals on the first axis. What are the points that make the greatest contribution? Try to get their coordinates.

Question 4

Same job with the second principal component

Question 5

Give a global interpretation of the first correlation circle by taking into account he illustrative variables. Perform a typology of the 15 cities based on this representation

4 Decathlon Data

Question 6

You will carry out a full analysis of the decathlon data set. It is composed of 41 rows and 13 columns that represent the athletes performances at two sports competitions: Decastar (an exhibition meeting) and the 2004 Olympic Games. The last column is a categorical variable corresponding to the name of the competition whereas the columns 1 to 12 are continuous variables: the first ten columns correspond to the performance of the athletes for the 10 events of the decathlon and the columns 11 and 12 correspond respectively to the rank and the points obtained. You can run the analysis at your own pace using the following ressources:

1. A video tutorial in R
2. Static tutorial in R
3. Implementation in Python