Data challenge & SHS: Principal component analysis and clustering in R

Julie Josse, Gaël Varoquaux, and Bénédicte Colnet

February 2022

Abstract

This is the practical class associated with the class 2 on principal component analysis and clustering. In this tutorial, you will learn how to perform a principal component analysis and how to interpret it. You will also learn how to perform a clustering on quantitative data. This notebook makes an intensive use of the package `FactoMineR`. Interpretation of the results remains the most important part of this tutorial.

Contents

Aknowledgments: François Husson class on youtube, the book "R pour la statistique et la science des données".

```r
knitr::opts_chunk$set(echo = TRUE)
# Load all packages needed to execute the job
# If the packages are not installed, write
# install.packages("<name of package>")

# Clear any existing variables
rm(list = ls())

# Set seed for reproducibility
set.seed(123)
```

Principal component analysis

Illustrative example

Before going into details, let us look at a funny example. Imagine that I generate two variables X_1 and X_2 from normal distributions. We want these variables to be linked (correlated) and such that $X_j \sim \mathcal{N}(0, 1)$. The following chunk performs the simulation. You can take the output data frame and explore the data first with univariate analysis. And then with a bivariate plot.

Remark: An outlier is in the dataset. Can you recover it?

```r
library(MASS) # for simulations
Sigma <- matrix(c(1,0.8,1,0.8),2,2)
simulated_data <- mvrnorm(n = 500, mu = c(0,0), Sigma)
output <- data.frame(simulated_data)
names(output) <- c("X1", "X2")
output[501,] <- c("X1" = 2, "X2" = -2) # outlier step
```

General introduction

Context

Principal Component Analysis (usually the shortname is PCA but you can also find ACP in French) focuses on typical data you can find in several domains: *observations* (or individus) in rows, and *variables* in column. Note that the PCA focuses on *quantitative* variables (for example age, or price, but not color or sex). For example we can study the average temperature depending on cities. In that case cities are rows, and in column the average temperature per month.

Typical question an ACP answers

A typical question you may ask on your data is: how much the different observations are close to one another considering the variables? (remember that everything you will conclude depends on these variables that you added in your initial model) You can also see PCA as a way to find a low-dimensional representation that captures the "essence" of high-dimensional data

What can you interpret from data?

The PCA will group similar individuals together. Information are also learned on variables, with the correlated variables (meaning that you have a linear link between two variables), and also which variables synthetize the most the observations, or which variables bring different informations.

Package

In this notebook we propose to use the package `FactoMineR` and the function `PCA`.

An example: the decathlon data set

The data set is based on the decathlon results during the Athene's olympic games and the Décastar (another competition). For each athletes the data set contains the results in the 10 tests, with the total number of points and ranking. The competition in which the athlete participated is also mentioned.

For both competitions, the following information is available for each athlete: performance for each of the 10 events, total number of points (for each event, an athlete earns points based on performance; here the sum of points scored), and final ranking. The events take place in the following order: 100 meters, long jump, shot put, high jump, 400 meters (first day) and 110 meter hurdles, discus, pole vault, javelin, 1500 meters (second day).

The overall objective of this exercice is to characterize the athletes and their differences, and to observe it tests evaluate similar skills or different ones. The aim of conducting PCA on this dataset is to determine profiles for similar performances: are there any athletes who are better at endurance events or those requiring short bursts of energy, etc? And are some of the events similar? If an athlete performs well in one event, will he necessarily perform well in another?

Question 1

First, load the data and inspect the data (for example which variables are quantitative or qualitative?).

Remark: This step is the first step you should do before any data analysis, and not only PCA.

Question 2

Apply a PCA on the data using the function from FactoMineR, and interpret it.

Tips: - First install the package.

- The appropriate function is called PCA.
- You can check if this function does or not the normalization step going in the documentation (`?PCA`).
- Why are normalization and reduction an important step?
- Explain your choices for the active and illustrative variables/individuals? (because you don't have to use all the variables to perform the PCA, you can only run it on a subset of variables that makes more sens.)
- When you interpret the data, you can also do a bar plot of the eigenvectors found by the PCA. For this purpose you can use the result object of the PCA analysis, and look at the `eig` component of this object. You can plot this using the `barplot` function or ggplot2 (which is a little bit more challenging, but a good exercice)

Question 3

What can you say on the variables related to speed (100m and 400m) versus the long jump?

Tips:

- You can first give a general comment looking at the correlation circle
- You can also access to the details of this graph looking at what hides in the variable results `res.PCA`var`else`.

Question 4

What can you say on Carsara athlete, Sebrle and Clay, and also Schoenbeck and Barras?

Question 5

Which variable predict the best the final score?

FactoShiny

`FactoShiny` is a graphical interface to the `FactoMineR` package to plot interactive plots. Therefore the underlying tools are the same as we saw previously. But this graphical interface can help you while working on data, and also to present in a funny way your data to a team. In this part we keep the same decathlon data.

To test it on your own, you can load the `Factoshiny` library and use the command `PCAshiny`.

Tip:

- If you use Mac you don't have a working Tcl/Tk by default, while it is needed for this package. So don't worry if you see an error while installing it! Go in the console and type `brew install tcl-tk` (if you use brew, what we recommend for Mac). The error can be quite complex as explained here[1].

Clustering

Hierarchical Cluster Analysis (HCA)

or Classification Ascendante Hiérarchique (CAH) in French!

Question 1

Explain the general principle. Do you need to scale your data or not? (explain why) Launch a HCA on the decathlon data using the package `cluster` and the function `agnes()`. You can visualize your results using the function `plot()`. Be careful to use the ward distance.

Question 2

As seen in class, a challenge is to cut the tree (corresponds to choosing the number of classes). Using the result you had and the function `as.hclust` you can observe the height where two classes are merged. How can you use this result? Once you know where you want to cut your tree, you can use the function `cutree` with the number of classes you want to keep.

Question 3

Once you decided a certain number of class, how can you describe the class? For example which variables characterize the classes you have?

Tips: you can use the function `catdes()`

You can also use the paragons of the classes to describe the classes.

Hierarchical Clustering on Principal Components (HCPC)

or Classification Hiérarchique sur Composantes Principales in French!

Open question (on your own)

It is also possible to perform a clustering on the variables obtained after a PCA analysis. For this you can use the `HCPC()` function. On your own, try to do this analysis with plot and quantitative analysis.

Try to have the plot(s) you prefer to present your results. You can find a documentation here: http://www.imsbio.co.jp/RGM/R_rdfile?f=FactoMineR/man/plot.HCPC.Rd&d=R_CC

[1] https://swvanderlaan.github.io/post/getting-r-with-tcl-tk-on-my-mac/