Principal Component Analysis

Part 2 - Relevance and intuition beyond PCA

The aim of this course is to give you all the information you need to understand the PCA, as well as give you all tools necessary to put it in place. We will focus on the practical aspects of PCA without ignoring the maths beyond. Each part ends with several exercices to do. Some of them are very easy and other harder. I highly recommend to verify all demonstration presents during the course, for 2 main reasons: verify there is no mistake and help you to understand the objects manipulated. I truly believe that equations manipulation is very helpful in order to have a good intuition of why they are the way they are.

> *"Statistics is about reducing the amount of data."* **R. Fisher**

Why variance matters so much? In statistics, it's very common that we want to explain the more variance possible. It might sound not very clear at first. In order to give the intuition of what it means, we will see concrete examples. The variance is a measure of variability of a variable. The higher (resp. lower) it is, the more (resp. less) *volatile* our variable is.

For example, if we want to make a linear regression between annual income, age, level of education and number of children, we want to explain the more variance possible. Intuitively, if our model (i.e: a simplification of reality) explains most (>90%) of the variance from our initial data, we can be confident with our model predictions. It's because individual are all different that statistics is so relevant.

What is PCA useful for? For the moment, we only have computed correlation between two variables at a time. But in real life data, you will have lot of different variables and calculate correlation for each pair has several problems. First, it's hard to have a global idea of which variable is correlated to which one. In fact, if you have p variables, you have $C(p) = \frac{p!}{2!(p-2)!} = \frac{p(p-1)}{2} = \frac{p^2-p}{2}$ Pearson coefficients correlation to interpret (trust me, not very fun and/or relevant). You can see that C increases quadratically with p.

But, there is a way to make it easy to compute: **matrix correlation**. It's a sheet that contains Pearson coefficients correlation for all numeric variables. It's an easy way to summarize lot of information. In R, it's **pretty easy to compute**:

```
library(corrplot)
data("iris")
data = iris[, 1:4] #remove the qualitative variable
summary(data)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

```
correlation = cor(data)
corrplot(correlation, tl.srt=45, method="number")
```

In this case we have only 4 variables. And, as said before, it's very common to have way more than that. I let you imagine **how hard it is to interpret a correlation matrix for 10, 20 or more variables**.
And that's where PCA appears: it proposes a way to show correlation between multiple variables in a very simple way. It projects our initial variables into a subspace and plots vector of variables. If 2 vectors seems to go in the same direction and our projection had work well (we will see what it means), **that imply that these 2 variables are correlated**.

The main motivation for PCA is that in order to represent/visualize data, **one must restrict oneself to a small number of dimensions**. Indeed, a correlation circle gives, more or less, the same information as a correlation matrix but in a much more intuitive and accessible way.

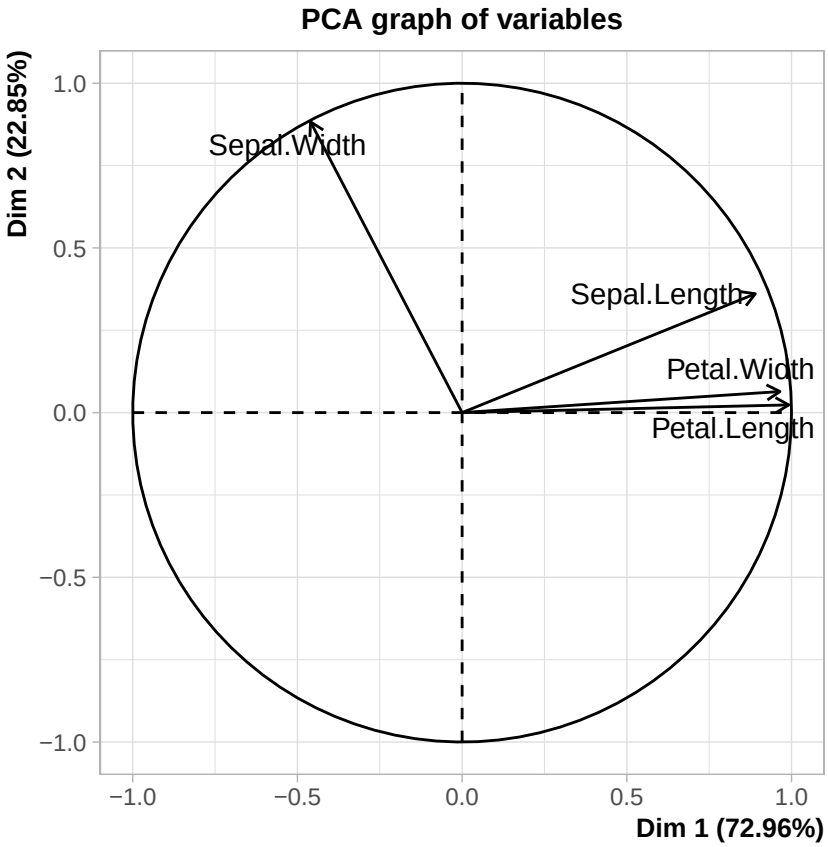
From a data set of more or less correlated quantitative variables, we wish to create new variables, de-correlated, representing **the principal components**. The objective is to keep the maximum amount of information present in our initial data set while reducing the number of dimensions necessary to express it (in our case, generally 2 or 3).

Initially, in a data set with p variables, p dimensions are needed to express all the information contained (one dimension per variable). **PCA seeks to "create" the dimensions with the maximum amount of information**. Expressed more formally, we want to create the principal components that express the most variance (or inertia, a geometric formalization of the variance) possible.
Important: unless our initial variables are perfectly correlated (we can express each variable as a linear combination of one or more others), **it is impossible to keep all the information in our principal**

components. However, we quickly realize that there is no point in keeping co-linear variables from the point of view of data analysis (e.g. having data in two different currencies at a fixed date).

Concretely, what is going on when doing a PCA on a software? Let's do a PCA on the iris data set used before!

```
library(FactoMineR)
pca_results = PCA(data, graph = FALSE)
plot.PCA(pca_results, choix = "var")
```



For the moment, we will only be interested in the **correlation circle plot** (graph of variables). If you remember what's said before, you should have a first idea of which variable is correlated to which one. If you don't, the thing important to know here is the fact that **each variable is associated to a vector**, and if 2 vectors are going in the "same direction", **it means that they are positively correlated**.
Based on this, we can intuitively think of a relationship between correlation and cosinus angle of 2 variables from our correlation circle. In fact, there is, and it's pretty intuitive if you understood the correlation circle:

$$cor(x, y) = cos\theta(x, y)$$

For the moment, it doesn't really matter if you don't understand what the axes are and how to interpret the percentage associated. The only thing that really matter here is the fact we are able to summarize a matrix correlation in a very simple way. If you're not convince, **I highly encourage you** to compare the Pearson coefficients correlation obtained before and the direction of each vector-variable.

PCA is very powerful because it's allow us to **show correlation between a lot of variables** in a 2 dimensional graph although it should require as much dimensions as the numbers of variables to represent

them. You might acquire with the fact that it's **not very intuitive to create a scatter plot with more than 3 dimensions**.

Exercises

1. With R, create a variable which variance is equal to 0 with a sample size of 100.
2. If 2 vector-variables x and y are going in the striclty opposite direction, to what is equal $cos\theta(x, y)$?
3. Using internet (or other), find a statistic that measures the amount of variance explained (there are multiple ones).
4. Why did I remove the qualitative variable of the iris data set before doing the PCA?
5. Choose and explain a concrete case where PCA might be relevant? irelevant?
6. Is it possible that, if x is correlated to y and to z that y and z are not correlated? Why so?
7. Create a data set where: there are 2 variables highly positively correlated, 2 variables highly negatively correlated and 2 variables not correlated at all. Do it creating only 4 different variables.
8. Create the correlation circle of the variables from the question 7. Compare your results to the interpretation with the cosinus angle. Tip: use a high sample size in order to ensure that randomness will not have too much impact.