Practicals 1 : Introduction and PCA

1 Preliminaries

The R software can be used in several ways. You can use the interactive mode (terminal command `R --vanilla`) although this is not particularly adapted to our practical sessions. Indeed, in order to keep a record of your work and to avoid typing the same commands again and again, it is recommended to work with a text file that we can load in R with the command `source`. It also exists numerous softwares (Rcmdr, rgedit, ...) that offer graphical user interface for using R.

To make easier the recovery of the data and to have at our disposal some graphical commands in the sequel, we start by loading the script *tp1.R*,

```
source("http://www.math.univ-toulouse.fr/~xgendre/ens/m2se/tp1.R")
```

2 Toy examples

In this section, we deal with the two data sets introduced during the lectures. The data are briefly explored before proceeding to the principal component analysis (PCA). Finally, we draw attention on some specific tools.

2.1 Grades

The first data set that we consider corresponds to the grades obtained by $n = 9$ students in $p = 4$ subjects. To get the associated matrix data X, execute the command

```
X <- DataNotes()
```

Explore this data set:

- print the matrix X,
- use the command `summary(X)` to view some important informations,
- print the box plots relative to the grades in the 4 subjects (command `boxplot`),
- what do you obtain with `summary(t(X))`?

To do the PCA, we begin by centering the data,

```
Xbar <- scale(X, scale = F)
```

Print the box plots relative to the centered data and remind why we did not normalize these data during the lecures.

Execute the following commands and explain what they do and what you get:

```
Sigma <- t(Xbar) %*% Xbar/nrow(Xbar)
ACP <- eigen(Sigma)
ACP$values
cumsum(ACP$values)/sum(ACP$values)
plot(ACP$values, type = "b")
```

Use the obtained results to answer to the following questions:

- how many axis should you keep for the PCA of this data set?
- what is the part of inertia explained by the principal plane?
- compute the principal components matrix `C`.
- compute the quality of the representation of each student in the principal plane and put the results into a vector `cos2`.
- compute and comment how each student contribute to the principal variables.

With the help of the matrix C, we can represent the data in the principal plane. Execute the following commands and comments the results,

```
plot(C[, 1:2], pch = 2, cex = cos2)
text(C[, 1:2], labels = rownames(C), pos = 3)
```

We can also compute the matrix of correlations between initial variables and principal variables:

```
Rho <- diag(1/sqrt(diag(Sigma))) %*% ACP$vectors %*% diag(sqrt(ACP$values))
rownames(Rho) <- colnames(X)
```

Check that the representations of the variables are in the unit circle. With the help of the command `CercleCor(Rho)`, print and comment the circle of correlations. To conclude, we can simultaneously print the individuals and the variables in the graph called *biplot*. Discuss about the results obtained with the command `Biplot(C,Rho)`.

2.2 Bows and crossbows

The second data set that we have handled in the lectures was the one relative to the bows and the crossbows of the video game "The Elder Scrolls V : Skyrim". It contains the statistics of $n = 14$ weapons related to $p = 4$ variables and can be obtained with the command

```
X <- DataSkyrim()
```

As previously, explore the data set:

- print the matrix X,
- use the command `summary(X)` to visualize some important informations,
- print the box plots associated to the 4 variables.

To proceed the PCA, we start by centering the data,

```
Xbar <- scale(X, scale = F)
```

Print the box plots relative to the centered data. Comment this graph and execute and explain the following lines,

```
S <- t(Xbar) %*% Xbar/nrow(Xbar)
M <- diag(1/diag(S))
Xbar <- Xbar %*% sqrt(M)
Sigma <- t(Xbar) %*% Xbar/nrow(Xbar)
```

In particular, explain the role of the line `Xbar <- Xbar %*% sqrt(M)`.

Take again the questions of the previous example and compute the PCA for these data.

2.3 Other tools

The R software, as many other softwares, offers several tools for directly computing the PCA of a given data set. Among the common commands, we have to mention `prcomp` and `princomp`. These commands only differ on few points and we only focus on `prcomp`.

Let take again our first example and print some graphics,

```
X <- DataNotes()
ACP <- prcomp(X)
plot(ACP)
biplot(ACP)
```

With the help of the documentation about the command `prcomp`, what does contain the variable `ACP`? Compare the matrix of change of basis with the one we obtained. What does it imply for the principal component matrix? Comment the differences with the results of Section 2.1.

Likewise, take the second example and comment the differences with the method seen during the lectures. How can we say to `prcomp` that we want to work with normalized data?

3 Criminality in USA

In this section, we consider statistics about crimes commited in USA in 1977. This data set contains the numbers of crimes for 100 000 inhabitants relative to $p = 7$ kinds of crimes in the $n = 50$ states. To get this data set,

```
X <- DataCrimes()
```

Explore succinctly these data and compute the associated PCA. Print the decreasing curves of the eigenvalues, what number d of variables does this graph suggest to keep? To what part of explained inertia r_d your choice leads to? Compute the contribution of the initial variables to the inertias of the principal variables. Discuss the obtained results with respect to your choice of d.

Display the variables in the circle of correlations with the help of the function `CercleCor`. How do you interpret the axis of the principal plane? Print the individuals in the principal plane with the function `plot`. Is this representation sufficient to properly summarise the information contained in the data set? For $d > 2$, using the command `pairs`, display the data in all the planes generated by any pair of principal variables among the d first.

Compute the contribution of each state to the inertia of the axis. Detect some states that have important contributions. With the help of Figure 1, can you give a geographical interpretation to the axis of the PCA?

Figure 1: Carte des États-Unis

4 Countries of OECD

The last data set of this session is related to statistics from the observatory of the OECD. For each country of OECD and each year 1975, 1977, 1979 and 1981, we have the values taken by the following variables::

- NATA: birth rate,
- CHOM: unemployment rate,
- APRI: percent of jobs in the primary sector of the economy,
- ASEC: percent of jobs in the secondary sector of the economy,
- PIB: gross domestic product (per resident),
- FBCF: gross fixed capital formation (per resident),
- INFL: inflation,
- RECC: cash receipts (per resident),
- MINF: infant mortality,
- PROT: consumption of animal protein (per resident),
- NRJ: consumption of energy (per resident).

To get this data set,

```
Data <- DataOCDE()
```

The first column contains the initials of the country and the next columns contain the $p = 11$ variables.

After looking at this data, discuss about normalizing them or not and compute the PCA. Study the number of axis to keep, the quality of the representation, the contributions to the axis, ...

In the principal plane, print the $n = 68$ points relative to the 17 countries. Highlight the groups of 4 points related to each country and link them in order to visualize the "moving direction" of each country in the principal plane. With the help of your preliminary study, interpret this graph.