Practical work on PCA and unsupervised classification

2023-2024

Data

This project will be based on Breast Cancer Coimbra Data Set http://archive.ics.uci.edu/ml//datasets/Breast+Cancer+Coimbra. The dataset consists of 10 variables:

- Age (years)
- BMI (kg/m2)
- Glucose (mg/dL)
- Insulin (µg/mL)
- HOMA : index informing about insulin resistance and blood sugar regulation
- Leptin (ng/mL) : satiety hormone
- Adiponectin (µg/mL) : hormone produced by adipose tissue whose plasma level is decreased in overweight or obese people and in diabetic patients
- Resistin (ng/mL) : hormone involved in insulin resistance
- MCP-1(pg/dL) : inflammation marker
- Classification (1=Healthy controls, 2=Patients (with cancer))

The objective is to implement exploratory statistical tools to finely characterize the sources of variability present in the data and to describe the individuals by characterizing the main axes of similarity and dissimilarity.

Questions

1. Import the data into **Rstudio** and the **tidyverse** library.

```
library(tidyverse)
cancer <- read_csv(file='dataR2.csv',col_names =T)
cancer <- cancer %>% mutate(Classification = as.factor(Classification))
str(cancer)
```

```
## tibble [116 x 10] (S3: tbl_df/tbl/data.frame)
##  $ Age           : num [1:116] 48 83 82 68 86 49 89 76 73 75 ...
##  $ BMI           : num [1:116] 23.5 20.7 23.1 21.4 21.1 ...
##  $ Glucose       : num [1:116] 70 92 91 77 92 92 77 118 97 83 ...
##  $ Insulin       : num [1:116] 2.71 3.12 4.5 3.23 3.55 ...
##  $ HOMA          : num [1:116] 0.467 0.707 1.01 0.613 0.805 ...
##  $ Leptin        : num [1:116] 8.81 8.84 17.94 9.88 6.7 ...
##  $ Adiponectin   : num [1:116] 9.7 5.43 22.43 7.17 4.82 ...
##  $ Resistin      : num [1:116] 8 4.06 9.28 12.77 10.58 ...
##  $ MCP.1         : num [1:116] 417 469 555 928 774 ...
##  $ Classification: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

2. Provide basic descriptive statistics and briefly comment on the results.

```
cor(cancer[,-10])

table(cancer$Classification)

by(cancer$Age,cancer$Classification,mean)
by(cancer$Age,cancer$Classification,sd)

# Etc...
```

3. Use the following commands to perform a principal component analysis on the cancer data. What do the options **scale.unit = TRUE** and **quali.sup = 10** do? Use the help if necessary.

```
library(FactoMineR)
library(factoextra)

acp.res <- PCA(cancer, scale.unit = TRUE, graph=F, quali.sup = 10)
```

Remark: A supplementary variable is a variable that is not used to build the principal components. It is only used to interpret the results.

4. Based on the following commands, choose the number of components.

```
acp.res$eig
fviz_eig(acp.res)
```

5. Draw the correlation circle on the plan composed of the first two principal components using the following commands:

```
fviz_pca_var(acp.res,axes=c(1,2),repel=T)
```

- Comment on the correlations between variables and interpret the first two principal components.
- Adapt the code in order to provide a meaning to the other selected principal components.

6. Use the following commands to propose a typology of the patients.

```
cos2<-cbind(acp.res$ind$cos2[,1:2],rowSums(acp.res$ind$cos2[,1:2]))
colnames(cos2) <- c("Dim.1","Dim.2","Plan.1.2")
#cos2
fviz_pca_ind(acp.res,axes=c(1,2),repel=F,col.ind = "cos2")
fviz_pca_ind(acp.res,
             col.ind = cancer$Classification,
             addEllipses = TRUE,
             ellipse.type = "confidence",
             legend.title = "Groups",
             repel = FALSE
)
```

7. Propose a classification using the hierarchical ascending classification (HAC) method.

   a. How many classes would you choose?
   b. Represent the classes obtained on the first factorial plane of the PCA and give them an interpretation.

Some useful scripts :

```
# centrer et réduire les variables
tab<-scale(data, center=TRUE,scale=TRUE)
# calculer les distances entre individus
data.dist <- dist(tab,method = "euclidean")
# calculer les distances de Ward
data.ward <- hclust(data.dist,method="ward.D2")
# représentation graphique
plot(data.ward,hang=-1)
# couper le dendrogramme pour obtenir k groupes (valeur de k à préciser)
# et récupérer les classes
cluster <- cutree(data.ward, k = ...)
# visualiser les classes sur le premier plan factoriel de l'ACP
fviz_pca_ind(acp.res,axes=c(1,2),habillage=as.factor(cluster))
```

8. Propose a k-means classification using the same number of classes as with HAC.

   a. Is the new classification identical to the previous one?
   b. Interpret the classes obtained by the k-means method.

Some useful scripts :

```
# centers indique le nombre de classes
res.k.means <- kmeans(data,centers=...)
```