Data Science Method Lab 1: PCA and Clustering

The aim of this tutorial is to practice PCA and clustering technics to analyze the patterns and groups within a dataset.

Data

The following two exercises are based on employment dataset "employment.csv". This dataset is adapted from the LISS panel (Longitudinal Internet Studies for the Social sciences). See https://www.lissdata.nl/about-panel for a detailed description about LISS. Our dataset contains survey information from 1980 employees in year 2015.

The variables used in the analysis are the following:
age: Age
income: Net monthly income in euros
tenure: Years being in the current job
training: Dummy for taking training in the past 12 months
jobsatisfaction: Rating from 0 to 10. The higher the more satisfied.
commuttime: Commuting time in minutes.
female: Dummy for females (0 for men, 1 for women)

Exercise 1: PCA

(Suggested reading before exercises: Page 401 – 404, "10.4 Lab1: PCA")

1. Read the data. Are there missing values in any variable? If yes, drop observations with missing value.
   (hint: **read_csv()**; **na.omit()**)

2. Produce some descriptive statistics (mean, standard deviation, min, max etc.) to get a feeling of the data.
   (hint: **summary()**; **apply()**; **sd**)

3. Perform a principal components analysis using variables "income", "tenure", "training", "jobsatisfaction", "female". What if we do not standardize the variables? Compare the loading vectors in two cases and explain why they differ. Which method do you prefer?
   (hint: **prcomp()**)

4. We focus on the PCA with standardized variables. Use **biplot()** to plot the first two PCs. Note that R will automatically use labels for observations, which is not ideal. Use options to adjust plot properly, e.g.: **xlabs=rep(".", nrow(yourdata))** for dots for each PC score pair, and **col=c("blue","black")** for colors. Give interpretations based on the loading vectors and the plot.

5. Produce a scree plot of PVE and a scree plot of cumulated PVE to determine how many PCs to keep. Motivate your choice.

Exercise 2: Clustering

(Suggested reading before exercises: Page 404 – 407, "10.5 Lab2: Clustering")

Now we try to group employees based on all the 7 features:

6. Because clustering command involves random assignment of the initial clusters, we use **set.seed(5829)** to keep results replicable.

7. We would like work on a random subset of the original dataset. Use **sample()** to draw 60 random individuals to form the new small dataset.

8. Scale the new data with **scale()**. Think it over: should we scale the data? Why and why not?

Questions 9 to 11 are based on the dataset generated in question 8:

9. We want to perform a hierarchical clustering to get a feeling of how the individuals are clustered. Use **hclust()** and "**average**" linkage to do it.

10. Plot the dendrogram. If we cut the dendrogram at the height of 3.6. How many clusters do we get? How many observations are in cluster 3? Which are they? (Hint: **cutree()**)

11. Instead of using the Euclidean distance, we now use the correlation-based distance to redo the hierarchical clustering. Plot the dendrogram. Does the plot change much? If we would like to have 5 clusters, how many observations are in cluster 3? (Hint: First use **cor()** to calculate the 60 by 60 correlation matrix of these individuals. Note that correlation measures "similarity" among individuals. You need to transform it into "dissimilarity". Then use **as.dist()** to transform dissimilarity matrix into distance matrix. To indicate observations in cluster, you may use **which()**)

12. We now come back to the full sample of 1965 employees. We decide to group them into 3 clusters. Scale the variables and perform a K-means clustering with **kmeans()**. Choose **nstart** to be 50. Display the total within-cluster sum of squares. [Note: **nstart**=50 means: run **kmeans()** 50 times with different random initial cluster assignments, and pick the best results (the one with smallest total within-cluster sum of squares) to report. A large **nstart** is recommended.]

13. Plot income against tenure. Use different colors for the three groups. Interpret the three groups.

(Suggested further reading: Page 404 – 413, "10.6 Lab3: NCI60 Data Example")