High-dimensional statistics

Academic Year 2022–2023

Project n°1 : Exploratory data analysis

Two QA sessions are scheduled: on 5/10 and on 19/10

1 Preliminary comment

This project may be done individually or in groups of 2 students (in the latter case, a unique project needs to be handed in, mentioning the two names). Even when working in pairs, it is expected that all parts of the project have been developed in collaboration between the members of the team.

The project, written in English, is due on the 26th of October 2022 (23h59) and needs to be submitted via eCampus. In the main body of the report (8 pages max), only the results, graphics and **interpretations** must be supplied and discussed (additional graphics or tables may be included in an annex). The `R` script used to compute the outputs of the analyses has to be submitted too but as a separated document (the commands should not be included in the report).

2 Data

For this project, a data set needs to be found[1]. The data should contain at least 10 **quantitative** variables and at least one binary indicator ($p \geq 11$). The number of individuals (i.e. the sample size n) should be smaller than 500 (a random selection of the instances or an appropriate and justified choice of a subset of instances needs to be performed if the original data set is bigger), with nevertheless the requirement of $n/p \geq 5$. The missingness rate observed on the data (i.e. the number of missing values divided by the number of cells in the data matrix) should be superior to 1%.

The source (web site, book, scientific paper...) of the data must be provided. Moreover, a text file containing the data must be submitted together with the report and code (there is no need to display the data in the report).

3 Statistical analysis

The following steps are required for this project:

1. Presentation of the data (context, information on the way they were collected, description of the variables) and informal discussion of the potential link between the variables, including with the binary one.

[1]Here are some links that might be of interest: `https://archive.ics.uci.edu/`, `https://web.stanford.edu/~hastie/ElemStatLearn/data.html`, `https://dasl.datadescription.com`, `https://walstat.iweps.be/`, `https://ec.europa.eu/eurostat/data/database`, ...

2. Information on the missing data, together with some discussion on their characteristics and their potential plausible reasons. It is suggested (to simplify the project) to work with the complete-case strategy for the questions 3 and 4 below. However, assuming that one of the "basic" imputation strategies outlined during the lecture could be relevant for the data, explain which one you would choose, apply it on a duplicate of your data and put forward some consequences of the application of the technique.

3. Exploratory analysis of the data in order to derive their main characteristics.

   Among other possible developments, it is compulsory to consider the following items:

   - Statistical and graphical summary of the variables, focusing on the most relevant aspects;
   - Analysis of the correlation structure of the data, with also an emphasis on the potential association between the binary indicator and the quantitative variables.
   - Discussion on potential outlying observations detected using Mahalanobis distances. The discussion must provide the number of detected observations and some insights about their profile (they may be extreme on only one variable or be outlying on all variables, for example).

4. Further exploration via a dimension reduction.

   A 2D plot of the quantitative variables (the binary indicator is left aside) is expected as final topic of this project. Depending on the fact that there is or there isn't some correlation to exploit, a PCA or a t-SNE projection is expected. Priority should be given to a PCA if there are indeed some linear relationships to rely on.

   If using a PCA is relevant for your data, a discussion on the matrix (either covariance or correlation) needs to be provided, together with a scree plot and a discussion on the number of PCs to keep. A discussion on the correlations between the original variables and the selected PCs is also expected. A projection of the data on the first principal plane needs to conclude this topic.

   If there is no correlation to exploit in your data, a t-SNE projection has to be performed. A thorough investigation of the impact of the tuning parameters (perplexity) needs to be done as well as a comparison of the process when using the raw data or the standardized data.

   The binary indicator has been left out of the construction of the projection but the two groups it defines might be distinguished on the 2D plot by using different symbols or colors. Would that be of interest here?