High-dimensional data analysis

Academic Year 2021–2022

Project n°1 : Exploratory data analysis and correlation analysis

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 20th of October 2021 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 as a *complementary information* (no specific mark will be attributed to the code in this project but it will be consulted in case there are some unexpected results in the report).

2 Data

For this project, a data set needs to be found[1]. The data should contain at least 12 **quantitative** variables ($p \geq 12$). 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 data may contain or not missing values.

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, missingness rate and the decision taken to handle these missing values if any...) and informal discussion of the potential link between the variables.

[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. Assess the plausibility of the normality assumption for the different variables of interest by means of graphics and/or some summary statistics (there is no need to provide a formal statistical test). In case a positive variable shows a skewed distribution, it is recommended to apply a log-transform (or $\log(\bullet + 1)$ transform) to the corresponding variable for the project.

3. Perform an outlier detection by comparing robust distances [2] and classic Mahalanobis distances by means of a DD-plot. Compare the different numbers of detected observations and discuss the characteristics of the different types of outliers (those detected by both strategies, or only by the robust approach).

4. Provide an analysis of the correlation structure of the variables, both using the classic correlation matrix but also using a robust estimation of the correlation matrix. Try to put forward any difference that exists between the two correlation matrices and visualize the scatter plots of the most correlated couples. Interpret the most interesting correlations.

5. In order to summarize the main relationships between the variables, a graphical model needs to be represented[3]. Using your knowledge of the data, specify, with arguments, a number of relationships (edges) that should be represented in the graphical model and use this number in order to choose the penalization parameter of the L_1-penalized likelihood function (under the normal model). A plot of the number of edges of the graphical model with respect to the penalization constant needs to be provided, as well as the final graphical model[4]. Interpret the graph in light of the previous correlation analysis.

[2] The use of the robust approach will be repeated in this project: it is required to specify which robust estimator is used and which values were given to the potential tuning parameters. It will be expected that the same robust approach is used throughout the project.

[3] If there is much difference between the classic and robust correlation analysis (exercise 4), it is suggested to work on the subset of observations having a robust distance below the usual cutoff.

[4] An `R` function is available on eCampus in order to represent that graphical model; it can also be represented by other means.