High-dimensional data analysis

Academic Year 2019–2020

Project n°1 : exploratory data analysis

1 Preliminary comment

This project may be done individually or in groups of 2/3 students (in the latter case, a unique project needs to be handed in, mentioning all the names). Even when working in pairs or triples, it is expected that all parts of the project have been developed in collaboration between the members of the team. For each "collaborative project", a specific question on the project will be included in the exam questionnaire of the team members in order to further develop or explain some aspects of it.

The project, written in English, is due on the 16th of October 2019 and a **paper version** must be handed in. 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 sent via email (to G.Haesbroeck@uliege.be) as a complementary information (e.g. valuable details not reported in the text might be obtained by looking at the script).

2 Data

For this project, a data set needs to be found[1]. The number of variables should be greater than 15, with at least 10 **continuous quantitative** variables and at least one binary indicator. 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). 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 2%. The source (web site, book, scientific paper...) of the data must be provided. Moreover, a text file containing the data must be sent by email to G. Haesbroeck on the day of the submission of the project (there is no need to print the data in the report).

If the data are suitable, they could be kept for the other projects based on additional and more specific data analyses.

[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`, ...

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 discussion of at least one scientific question of interest that might be treated thanks to these data.

2. Information on the missing data, together with some discussion on their characteristics and their potential plausible reasons, and presentation of the strategy followed to treat them while considering the exploratory data analysis of part 3.

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 quantitative data;
   - Analysis of the potential impact of the qualitative variable(s) on the quantitative ones;
   - Discussion on potential outlying observations detected using $z-$scores and by means of Mahalanobis distances.