High-dimensional data analysis

Academic Year 2018–2019

Project n°1 : exploratory data analysis and principal component analysis

1 Preliminary comment

This project may be done individually or together with another student of the course (in the latter case, a unique project needs to be handed in, mentioning both names). Even when working in pairs, it is expected that all parts of the project have been developed in collaboration between the two members of the team. A specific question on the project will be included in the exam questionnaire in order to further develop or explain some aspects of it.

The project, written in English, is due on the 5th of December 2018 and a **paper version** must be handed in (max 10 pages). In the main body of the report, only the results, graphics and **interpretations** must be supplied and discussed. It is not compulsory to use the software `R`. However, if available, the `R` script used to compute the outputs of the analyses may 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 10, with at least 6 continuous quantitative variables and at least one binary indicator. If n denotes the sample size and p the number of variables (ie the dimension), the ratio n/p must be greater than 5. 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.

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 a scientific question of interest that might be treated thanks to these data.

2. 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 summary of the variables;

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

- Analysis of the correlation structure of the quantitative data;
- Graphical analysis of the impact of the qualitative variable(s) on the quantitative ones;
- Detection of outlying observations based on a robust multivariate approach (a comparison of the robust distances with the classical Mahalanobis distances is expected via a DDplot).

3. Principal component analysis on the quantitative variables.

   The decision to work on the covariance or the correlation matrix must be justified and the number of principal components that should be kept needs to be discussed. Particular features of the PCA must be highlighted if there are any and some interpretation of the principal components should be provided.

4. Further analysis of the dependance structure of the continuous variables: assuming multivariate normality, construct a graphical model highlighting the main dependencies in the data and interpret the results while taking into account the results of the PCA.