High-dimensional data analysis

Academic year 2019–2020

Project n°2 : Further study of the correlation structure and dimension reduction

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). It is not compulsory to keep the same team as for project 1 and/or to keep working alone if that was the case for project 1. When working in team, it is again expected that all parts of the project have been developed in collaboration between the members of the team. A specific question on the collaborative projects will be included in the exam questionnaire in order to further develop or explain some aspects of these projects.

The project, written in English, is due on the 20th of November 2019 and a **paper version** must be handed in. In the main body of the report (6 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.

2 Data

The same data set as the one used in the first project may be used for this second project. Links with the exploratory analysis performed in the first project are required in this new project.

It is also possible to choose another data set. In such a case, the latter needs to satisfy the same constraints as those specified for Project n°1 and a thorough presentation of the data needs to be done. Moreover, each time some results displayed in Project 1 are quoted in this statement, some information on the behavior of the new data needs to be provided in order to be able to fully answer the questions of this second project. When using a new data set, 3 additional pages are then allowed in order to provide the extra information.

3 Statistical analysis

The project is decomposed into three part.

3.1 Robust outlier detection

An outlier detection procedure has been performed in Project 1. However, the location and dispersion estimators that were used by default were the sample average and covariance matrix, that might have been corrupted by outliers in the data. Perform here a *robust* outlier detection technique[1] and compare the so-called robust distances with the classic Mahalanobis distances by means of a DD-plot. Discuss the characteristics of the different types of outliers (referring, if appropriate, to the discussion already made in Project 1).

3.2 Further investigation of the correlation structure of the *quantitative* variables

1. In project 1, the correlation matrix estimated in the usual way has been discussed and interpreted. Compute a robust estimation of that matrix and compare the results with the classic one.

2. Assuming that the multivariate normality assumption holds for the data base restricted to the quantitative variables, a graphical model[2] allows to visualize the conditional dependence structure of the data. In such a graphical models, the variables are the nodes and an edge is drawn between the two nodes corresponding to X_i and X_j iff the corresponding element (i, j) of the inverse of the covariance matrix is different from 0, meaning that the two variables are dependent, conditionally on all the remaining variables.

   (a) Using the classic covariance matrix estimated on the data, represent the corresponding graphical model and interpret (taking into account the discussion on the correlation structure provided in Project 1).

   (b) In order to facilitate the interpretation of the graph, it would be easier if it was sparse. Use an L_1-regularized estimation of the covariance matrix in order to draw a sparse graphical model. A discussion should be provided concerning the appropriate choice of the regularization parameter.

   (c) The drawings here above are based on the multivariate normality assumption. Discuss the relevance of that assumption, using, as justification, the discussion made on the distributions of the variables in Project 1.

3.3 Visualisation of the quantitative data in 2D

A graphical visualization of the data in 2D needs to be provided using both PCA and tSNE. A discussion on the projection that seems to be the most informative should be included in the report.

Concerning PCA, the following information/justification should be provided:

- the choice of the matrix (correlation or covariance; robust or non robust)
- the variance explained by the 2D projection
- interpretation of the two first PCs (a correlation circle should be represented)

[1] 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.

[2] Among other available packages, the package `qgraph` (and the corresponding function `qgrpah`) offers the possibility to draw such a graphical model.