Multivariate Statistics: Exercise 7

November 21, 2018

Principal component analysis:

We consider the data sets *winequality-red.csv* and *winequality-white.csv*, which are available from the TUWEL course of our exercises. You can also fond a description there (*winequality.txt*). As the names indicate, the data sets describe some characteristica of red and white wines. The data sets originate from:
`https://archive.ics.uci.edu/ml/datasets/Wine+Quality`
Read the data into *R*, and check with `str()` if you have done this correctly. For the following tasks, use the function `PcaHubert()` from the package `rrcov`, which performs a robust PCA - see also latest version of the course notes. Be careful, this function gives an S4 object, and the list entries can be accessed only with the @ symbol.

1. Perform PCA separately on the two data sets and show the biplots. Which variables are associated with "quality"? Are the characteristics similar for red and white wines?

2. Look at diagnostic plots (see course notes – latest version), which show the orthogonal and score distances. Is it possible to find for the most extreme outliers the reason for their outlyingness?

3. Take the PCA result from 1. for the white wine data, and project the red wine data into the plane of the first 2 PCs. This allows for a better comparison of the main differences in the data structure of the two data sets. What are these main differences?

4. Now take the red wine data, but only the observations with high values of "quality" (7 and 8). Perform PCA with these data (without the variable "quality"). Project the data with low "quality" (values 3 and 4) into the PCA plane. How do low and high quality wines differ?

5. Show the diagnostic plot for the high quality data, and "project" the low quality data into this plot. Suppose you take the cutoff values for the distances as a classifier: How many observations from the low quality group would be misclassified? *Hint:* You will have to compute the orthogonal and score distances for the low quality data "by hand".

Save your (successful) R code together with short documentations and interpretations of results in a text file (= R script file), named as *Matrikelnummer_7.R* (no word document, no plots). Submit this file to Exercise 7 of our tuwel course (deadline November 20).