G0O02a: Statistical Data Analysis: Project 1

Prof. Mia Hubert

March 2022

The project consists of analyzing the Melon data set. The data set contains the result of a spectroscopy experiment conducted on $n = 2158$ cantaloupe melons of four different cultivars. Each of the spectra was measured on 256 wavelengths. The last variable `y` indicates the groups number of the cultivar, whereas the variable `cultivar_levels` contains their name.

You first draw an individual data set of random 180 spectra from one random class. The training data set consists of the first 90 observations, the validation set of the remaining 90 ones. You use the following code, where you change 0012345 by your student number.

```
load("Melon.rdata")
set.seed(0012345)
mygroup <- which(rmultinom(1, 1, c(0.25,0.25,0.25,0.25)) == 1)
mysample <- sample(which(y == mygroup), 180)
X_train <- data.frame(X[mysample[1:90], ])
X_valid <- data.frame(X[mysample[91:180], ])
```

You answer the questions by performing an appropriate analysis with R. The discussion of the results and the necessary figures are reported in a written text that consists of a maximum of 5 pages (12pt font size). Only report results and interpretations, do not repeat theory from the course! Include the figures when they are discussed, not at the end of the report. Additionally a separate file with the full R script should be provided.

One single folder containing your report and R script should be uploaded on Toledo before **April 18, 2022, 23h**. This project is graded on 3 points.

Good luck!

1. Plot the spectra. State your main findings.

2. Perform a classical PCA analysis on your **training** data set. Argue why you base the analysis on the correlation or covariance matrix of the data. Explain how you choose the number of components. Plot the loadings, the scores and discuss which spectra (if any) are flagged as outlying.

3. Perform a robust PCA analysis on your **training** data set. Compare the results between the classical and robust results (loadings, scores and outliers).

4. Continue with the PCA analysis you find most appropriate. Consider now the observations from the **validation** set, and compute their scores and predicted values (with respect to the PCA analysis of the training data). Make an outlier map with the observations from both the training and the validation set (use a different symbol or color). Discuss the result.

5. Remove all observations that are flagged as outliers in the training data (exceeding the score and/or the orthogonal distance cutoff). Consider the scores of the regular observations. Investigate whether it can be assumed that they are sampled from a multivariate normal distribution.