DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCE
KU LEUVEN

Report on project about principal component analysis

Statistical Data Analysis

Project 1

Mitja Mandić

April 2022

1 Introduction

For the first project we are analysing the spectroscopy dataset of four different types of cantaloupe melons. Each spectra was measured on 256 wavelengths of 2158 melons. In we plot the spectra of the whole dataset.

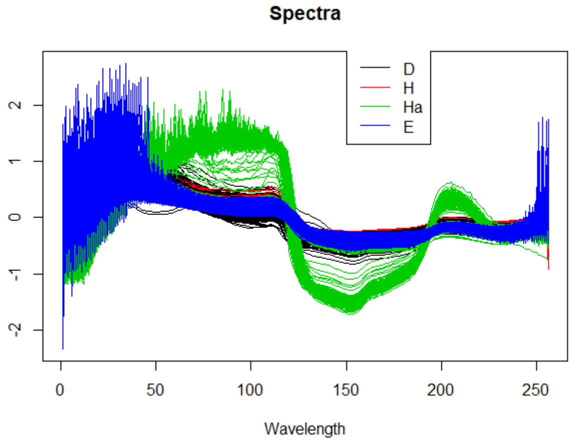


Figure 1: Spectra of the whole dataset

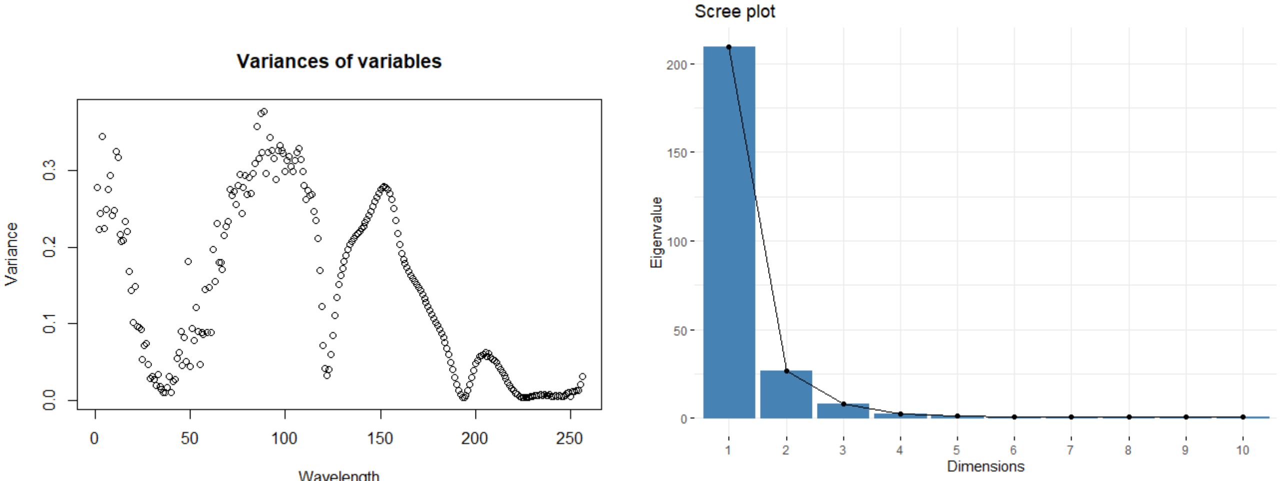
Apart from the cultivar “Ha” we seem to have a rather homogenous dataset, where most of variance seem to be happening in the first 50 variables. In the “Ha” group this extends to around 200 variables.

2 Classical PCA

We perform classical PCA on a part of the original dataset. Out of the main dataset we randomly select 180 (90 for training and 90 for test data) cantaloupes – in our case, rather interestingly, this whole sample consists only of members of group “Ha”. Before we continue with PCA, we check for differences between variances of the variables.

As we can see in figure 2a, there are some variables with much higher variance than others. That is why we will base our further PCA analysis on correlation matrix rather than covariance matrix. Next to it in figure 2b a scree plot is shown - based on that we decide to keep two variables, as they explain over 90% total variance.

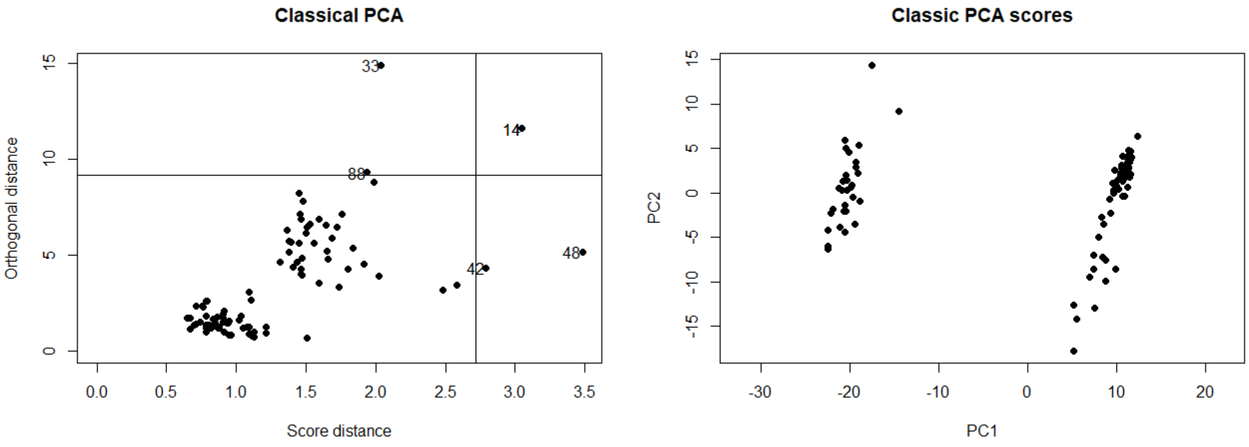
Classic PCA of training data based on two components flags five points as outliers (surpassing the score and/or orthogonal distance cutoff) - these are cantaloupes number 14,33,42,48 and 88. In the outlier plot depicted on 3a we notice formation of two groups, which is confirmed when we plot the scores, which we then see on figure 3b.



(a) Variances of each component vary in size

(b) Scree plot, showing eigengvalues in descending order - first two are much larger than the rest

Figure 2: Figures of variances and scree plot



(a) Outlier plot of PCA based on 2 components

(b) Scores of classic PCA

Figure 3: Outlier plot and scores of classic PCA.

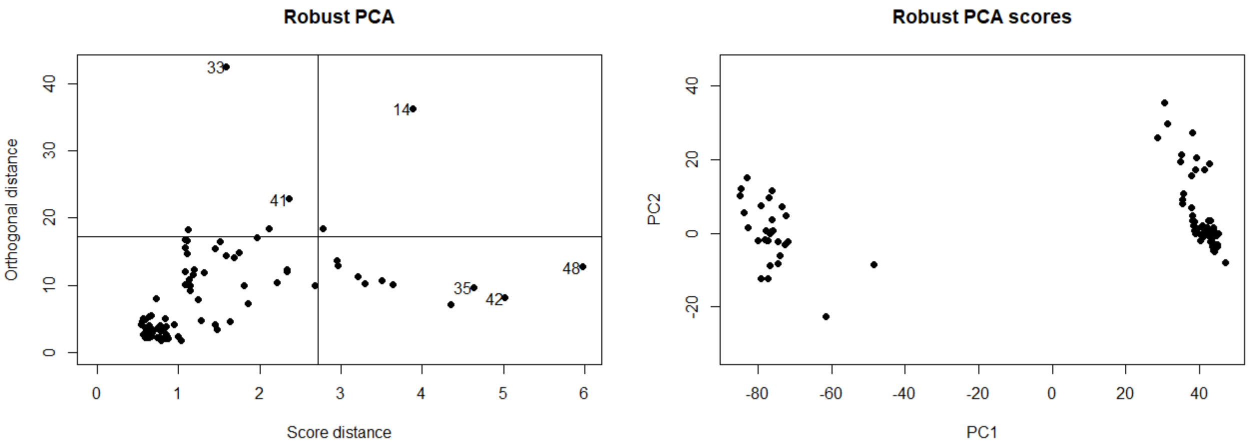
3 Robust PCA

In this section we comment on results obtaining by performing robust principal component analysis of the training data. For this we use function `PcaHubert` from the package `rrcov`, with scale parameter set to `mad`. Interestingly, we obtain quite different results compared to classical PCA analysis. To facilitate the discussion, we recreate the plots seen in section 2.

Firtly we notice that much more points are exceeding the score or orthogonal /distance cutoff. Additionally, one more point has both orthogonal and score distance larger than cutoffs than in the classical case (point 14). Outliers detected by this method are also much further from the main data cloud than those detected in previous section.

Another thing we notice comparing figures 4b and 3b is that in the former case the two groups are much further apart.

Because of such differences between the two methods and rather unusal data, we decide to proceed to further questions with robust PCA.



(a) Outlier plot of PCA based on 2 components

(b) Scores of robust PCA

Figure 4: Outlier plot and scores of robust PCA.

4 Validation set analysis

Now we take a look at the other half of data set we selected in the beginning. We compute scores and predicted values of the validation set with respect to the values obtained by performing robust PCA on training data.