DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCE
KU LEUVEN

KU LEUVEN

Report on project about principal component analysis

Statistical Data Analysis

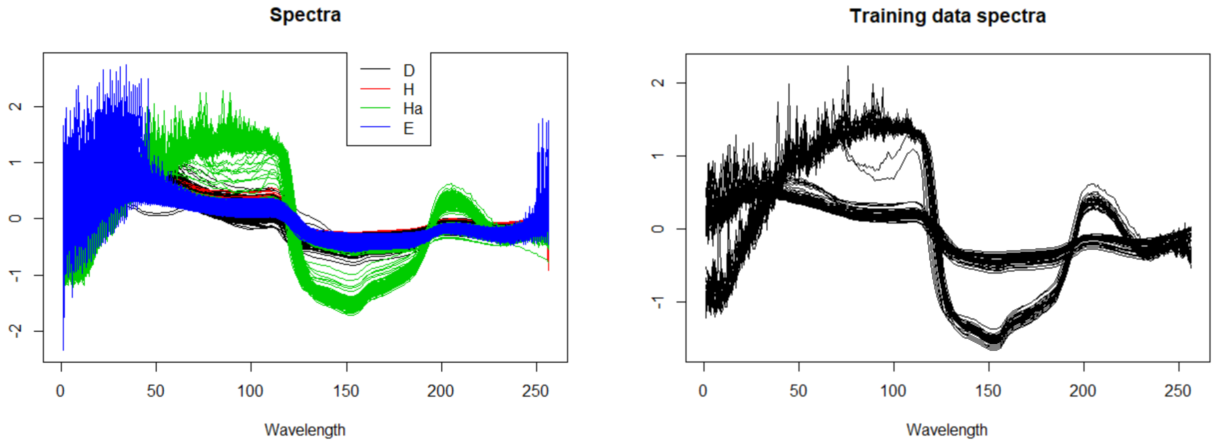
Project 1

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 1a we plot the 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.



(a) Spectral plot of the whole dataset

(b) Spectral plot of the training data

Figure 1: Spectral plots of both the whole and training data datasets

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”.

2 Classical PCA

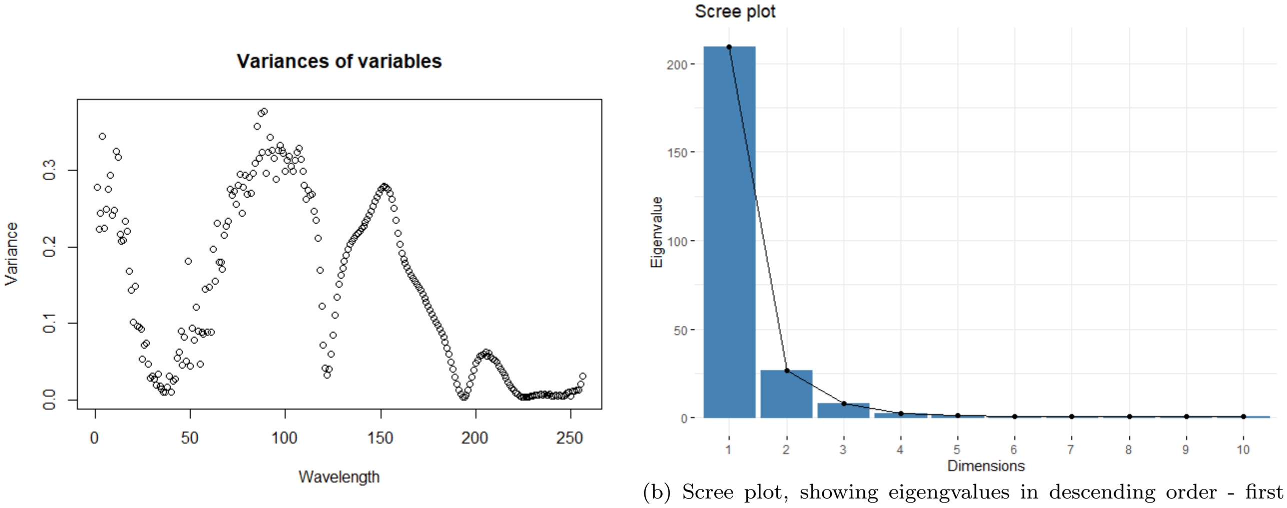
We perform classical PCA on a part of the original dataset. Before we continue, we check for differences between variances of the variables.

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.

3 Robust PCA

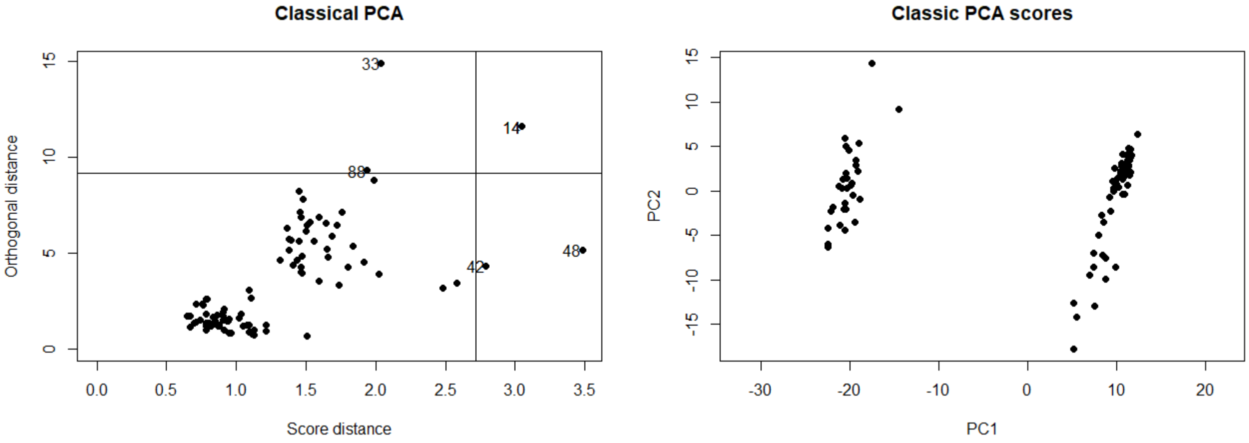
In this section we comment on results obtaining by performing robust principal component analysis of the training data. For this we use function `PcaHubert` from the package `rrcov`,



(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

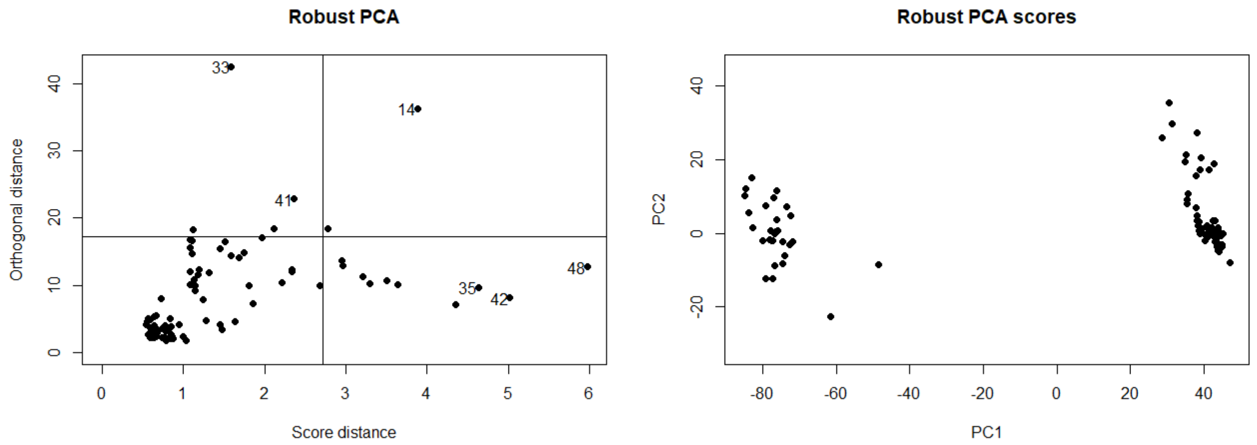
Figure 3: Outlier plot and scores of classic PCA.

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 in figure 4.

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. Let us now try to find the origin of such behaviour. In 1b we see that two groups seem to differ at, for example, 100th wavelength, with one group having values around 0 and the other between 1 and 2. After separating them like this, we see that this are the groups also found by PCA, as seen in 5.

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.

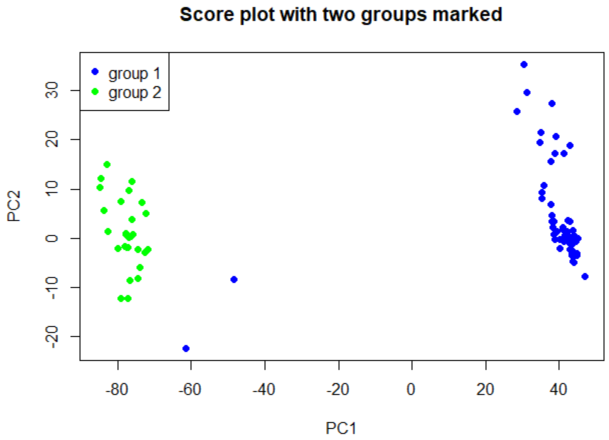


Figure 5: Score plot obtained by robust PCA with two groups marked

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.

Outlier plot, containing values from training data and calculated validation data scores is presented in figure 6.

We see that dataclouds in the bottom left align rather nicely. On the other hands, outliers are more dispersed when considering predicted data, especially in orthogonal distance. We assume that the second data cloud of predicted values might be affected by the two severe orthogonal outliers found in robust PCA (points 14 and 33 seen in figure 4b). Similar amount of outliers are found in score distance.

Plotting the spectra of predicted and actual data gives satisfying results, with plots looking very similar.

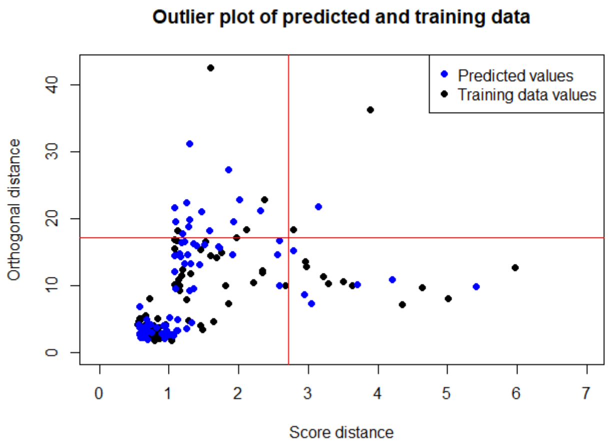
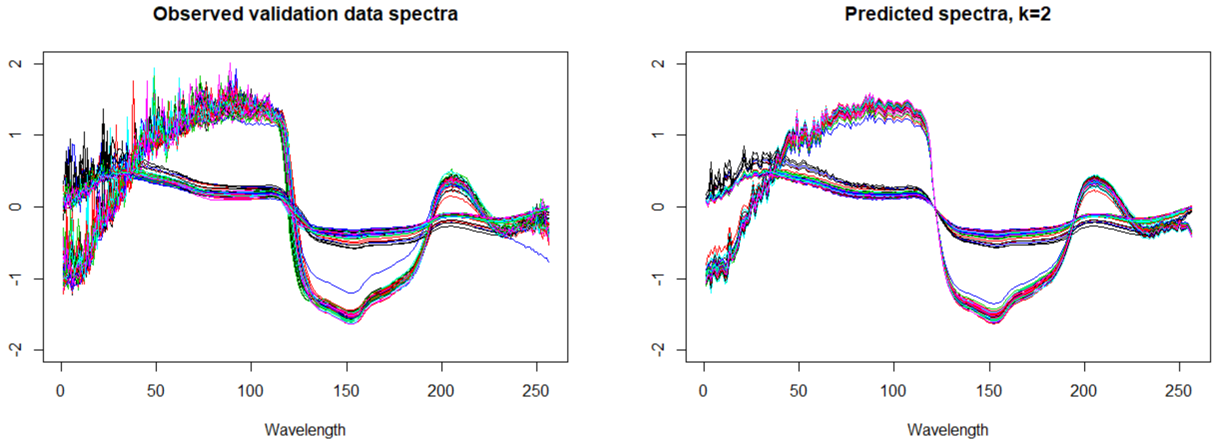


Figure 6: Outlier plot of training and predicted data



(a) Spectral plot of the validation data

(b) Predicted spectral plot

Figure 7: Figures comparing predicted and observed spectra.

5 Normality of training data

In the final section of the report we investigate whether the clean training dataset can be assumed to be sampled from a normal distrbution. From training data we remove outliers found by robust PCA, that is points exceeding orthogonal and/or score distance cutoff.

There are 16 such points found. Shapiro-Wilk test strongly rejects normality with p-value $1.768 \cdot 10^{-12}$. Plotting the regular and removed values we can see that there still are two groups present, which makes it easier to understand such a low p-value. Because of these groups, Mahalanobis distance measures are useless.

Scores in this case certainly are not unimodally normally distributed, as we have found two very clear groups. However, we might consider investigating normality of scores of each group separately. For the group on the left Shapiro-Wilk gives once again a very low p-value, so normality is rejected in this case as well. Similar results are obtained for the other part of the data.

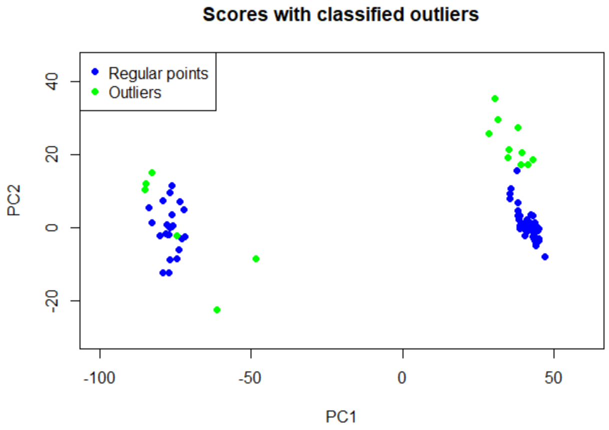


Figure 8: Plot of scores of regular and outlying points.

We conclude that scores of cleaned data are not normally distributed and neither are scores in each of the existing groups.