DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCE
KU LEUVEN

KU LEUVEN

Report on project about clustering

Statistical Data Analysis

Project 2

Mitja Mandić

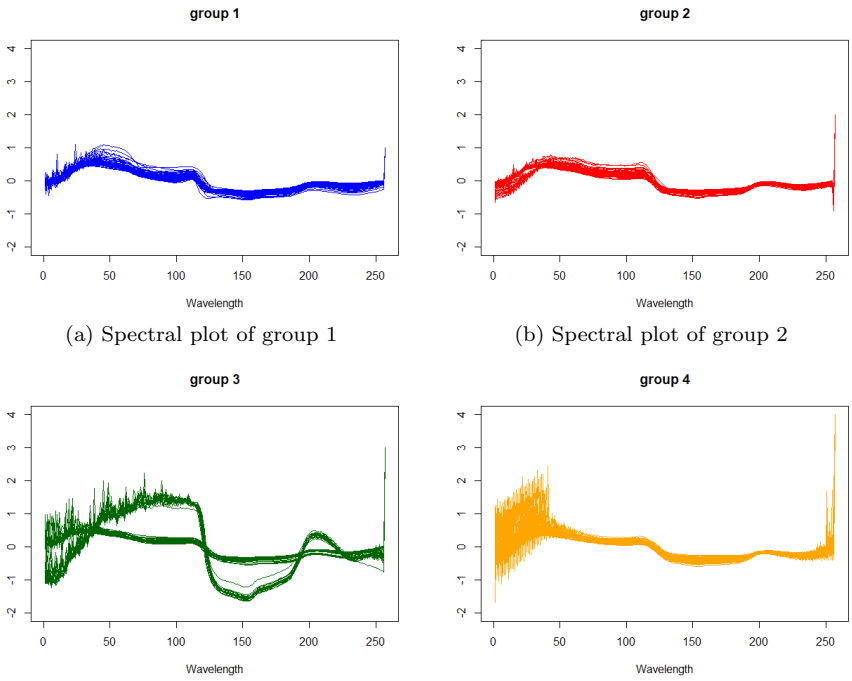
May 2022

1 Introduction

For the second project for the course Statistical Data Analysis we are once again worikng with the dataset of spectral data of four cultivars of canteloupe melons.

Groups we work in this report with were also again chosen randomly, with 50 observations drawn from each of the groups. In figure 1 we see a figure of spectral plots of each of the groups. We see in figures 1a, 1b and 1d that groups 1, 2 and 4 differ only slightly in lower wavelength numbers with group 4 having more variability there, but are quite similar in their behaviour in higher frequencies. Group 3 is the one that clearly stands out; a part of it resembles the behaviour of other three groups, while some observations form a different pattern, visible in 1c.

We predict that the differently behaving part of group three will form a separate cluster. Another will possibly be formed by lower wavelengths of group 4 – other observations seem to be too similar to form different clusters. Similar conclusions can be drawn from the heatmap of the spectra and we therefore not include it here in the name of brevity.



(a) Spectral plot of group 1

(b) Spectral plot of group 2

(c) Spectral plot of group 3

(d) Spectral plot of group 4

Figure 1: Spectral plots by group

2 K-medoids clustering

In the following sections we discuss results obtained by *k-medoids* clustering method. We impose 2 up to 6 clusters to randomly drawn data.

2.1 Evaluating cluster quality

Without using the information we have on which group do individual observations belong to, we look into the quality of clusters obdatined. For this purpose we use *silhouette values and plots*, that help us determine how similar are objects in a given cluster.

Starting off with two clusters already gives quite satisfying results. As seen in image 2, average silhouette width for the whole dataset is 0.76, which indicates we have found strong structure in the data. Some further analysis reveals that silhouette widths for each of the clusters are 0.76 and 0.79, respectively. We also see that the first cluster contains much more elements than the second – 182 compared to 18.

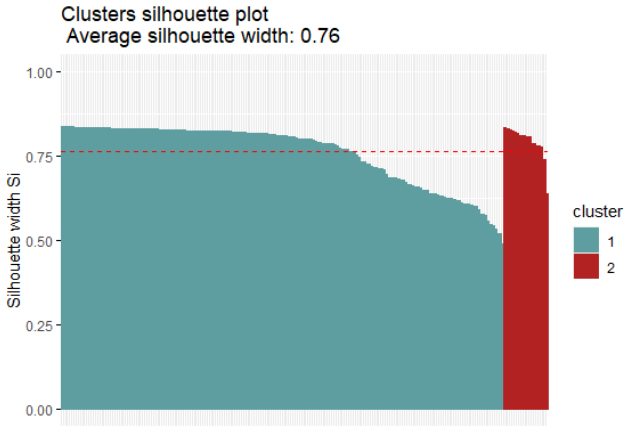


Figure 2: Silhouette plot for two clusters

We continue with the clustering and quickly notice a negative trend, with silhouette values never reaching the heights from those obtained with two clusters. Minimum is reached with four clusters, with the average silhouette width 0.29 and values for some individual observations falling below zero. More clusters also result in one large main cluster and a few smaller ones.

These results can be summarised in figure 3, where we see a clear peak at $k = 2$.

2.2 Comparison of clustering with known groups

Now we will also take into account the information we have on groups and investigate how they relate to our clustering methods.

Confusion matrix for two clusters reveals, that the small second cluster completely consists of observations from group 3 (figure 4a) – as predicted in the introduction. Additionally, making more than 4 clusters results in observations from group 4 being spread out across more clusters and a part of groups 2 and 3 fall in clusters 2 and 3 respectively, as depicted in figures 4b and 4c. We can already see the trend where clustering does not follow the underlying grouping.

Drawing cluster plots of PCA scores gives some additional insight into cluster structures. Plotting the results from two clusters in figure 5a we see one larger group and one smaller, far away from the other. With more clusters generated, the large datacloud on the left gets divided further into smaller clusters, most of which are overlapping significantly, as seen in 5b.

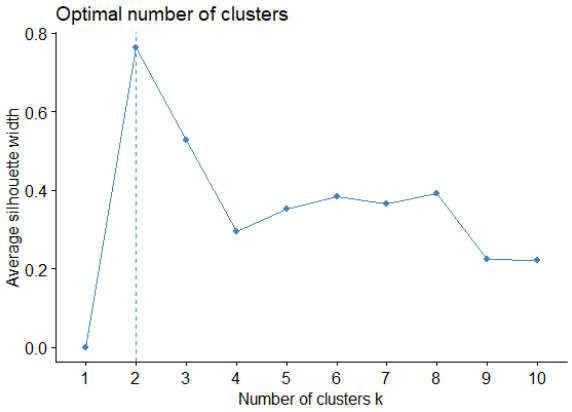
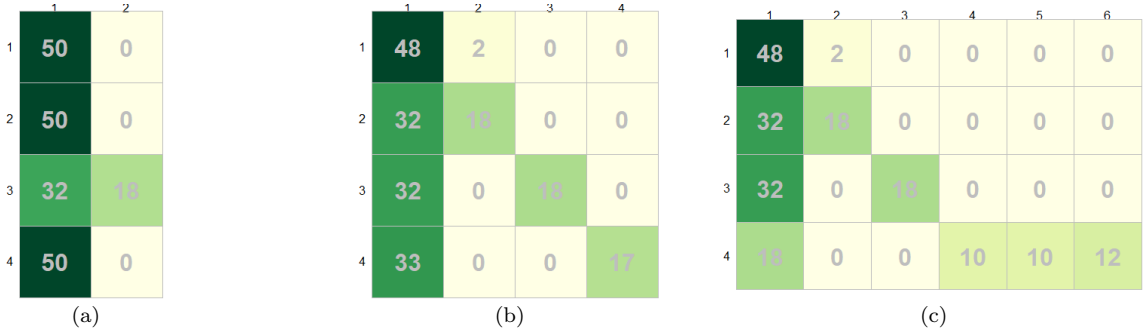


Figure 3: Average silhouette widths with different numbers of clusters



(a) (b) (c)

Figure 4: Confusion matrices for $k = 2$, $k = 4$ and $k = 6$.

Similar conclusion can be drawn by looking at spectral plots, coloured according to the clustering as seen in figure 6. In figure 6a clearly see that with $k = 2$ we get separation between main group and the differently behaving one. Adding more groups divides the main group further, but there is no clear structure.

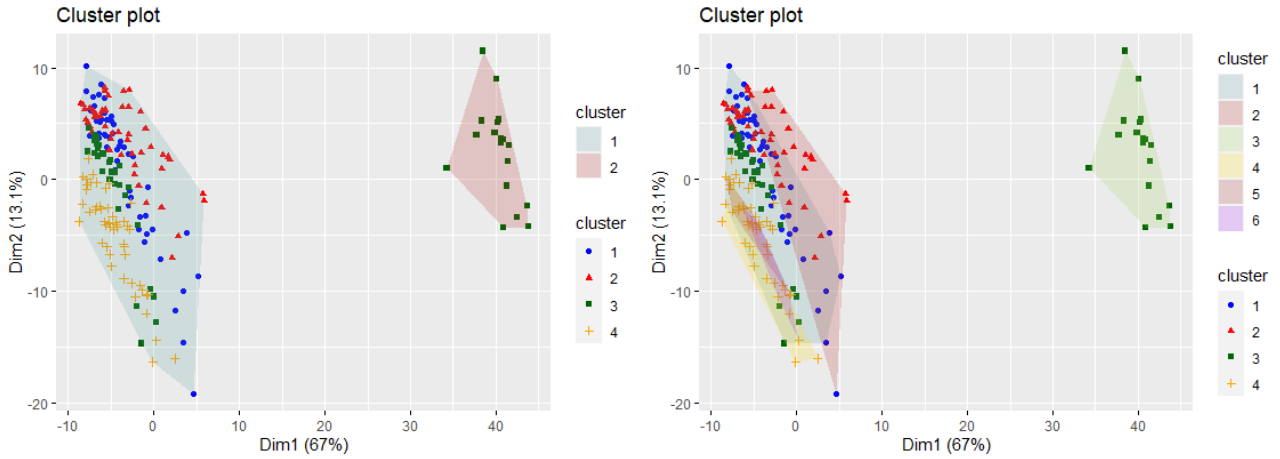
Other ways of analysing, such as heatmaps and mosaic plots do not reveal any new information on the clustering, so we do not include them in the report. Based on all results above we conclude that the optimal number of clusters is $k_{opt} = 2$.

3 Hierarchical clustering

We continue our research by performing agglomerative hierarchical clustering with average and complete linkage. When considering hard clustering with k_{opt} results are exactly equal to those obtained by k-medoids clustering – Rand index of the clusterings is 1. We can attribute this to the very clear structure in the data, with one group being rather far from another.

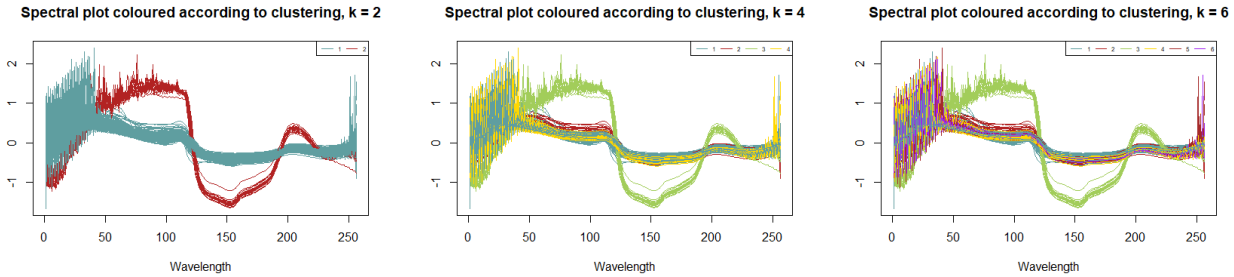
Looking at dendrograms of both clustering methods, we notice a very large jump in height when going from 3 to 2 clusters, and much smaller ones when the number of clusters is larger.

Further, taking complete linkage into account, we investigate a third group that seems to be appearing (visible on the right of dendrogram 7b). Hard clustering with $k = 3$ results in a large drop of average silhouette value, which in this case equals 0.43. Plotting the cluster plot reveals



(a) Cluster plot with $k = 2$ (b) Cluster plot with $k = 6$

Figure 5: Cluster plots



(a) (b) (c)

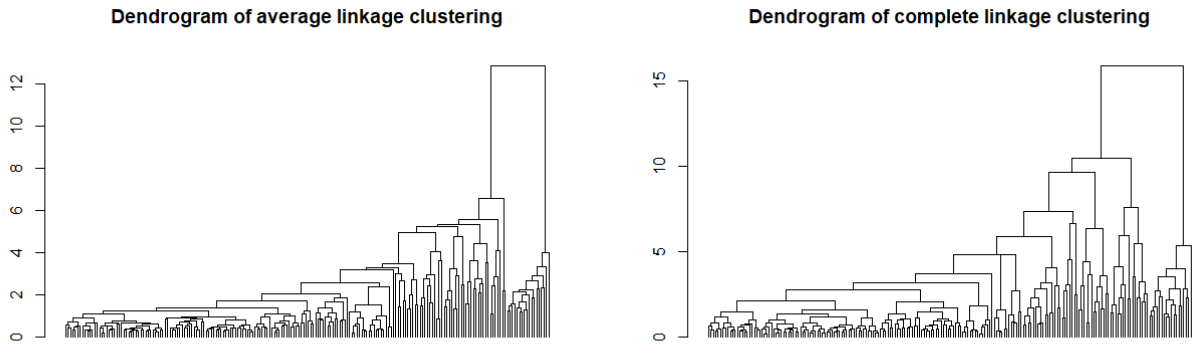
Figure 6: Spectral plots coloured according to clustering

that the new cluster contains mostly data from the 4th group, however it almost completely overlaps with the second cluster. Calculating the contingency table for this clustering we find out that the cluster plot decieves us a bit – only 14 members of the 4th group are in a separate clusters while the others still remain in the large first cluster.

Finally, we compare k-medoids and hierarchical clusterings. In table 1 we see that in most cases clusterings are quite similar, differing the most when it comes to splitting the data into four clusters. In all cases except for 6 clusters, k-medoids is more similar to complete linkage clustering than the two hierarchical methods.

k	complete, average	k-med, average	k-med, complete
2	1	1	1
3	0.881809	0.8741206	0.9555276
4	0.7514573	0.7020101	0.7801005
5	0.829598	0.7288442	0.8322111
6	0.9221608	0.7720101	0.8433166

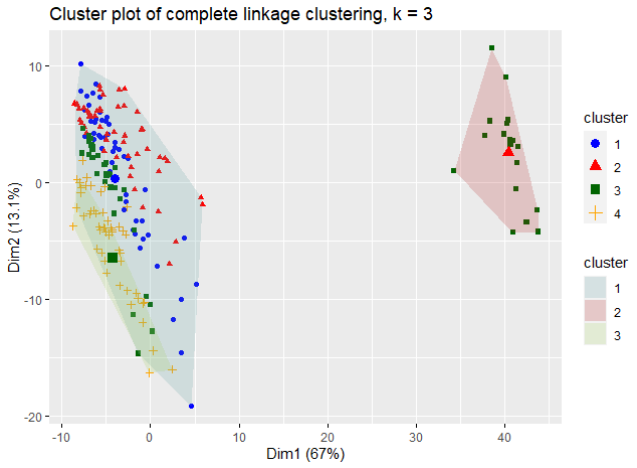
Table 1: Table of Rand indices



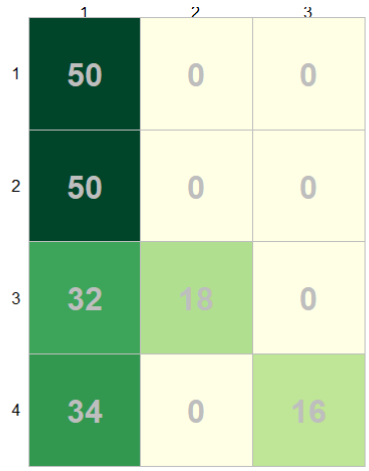
(a) Dendrogram of average linkage clustering

(b) Dendrogram of complete linkage clustering

Figure 7: Dendrograms of hierarchical clusterings



(a) Cluster plot of hierarchical complete linkage clustering with 3 clusters



(b) Confusion matrix of complete linkage clustering with $k = 3$

Figure 8: Analysis of complete linkage clustering with $k = 3$

4 Conclusion

We conclude that setting the number of clusters to two, both hierarchical and k-medoids clustering return same results, so either method is appropriate in this case. This is true because of the data structure, which is clearly divisible in two groups. Splitting the clusters lowers the silhouette values in both cases, which are somewhat higher in hierarchical case as compared to k-medoids (they still fall well below 0.5 mark). Comparing all clustering methods also reveals that they do not differ a lot, with pairwise Rand indices always staying above 0.7.

While clusters do not align perfectly with the groups from data, new clusters (when k is less or equal than 4) consist of observations of separate groups – but the majority of observations still remain in the first, largest cluster. Increasing the number of clusters to five or six results in splitting group 4 into smaller clusters, with other clusters remaining intact.