Multivariate analysis - hw10

Matěj Pešek

May 2, 2022

1 K-means grouping analysis

Our task in this assignment is to use some cluster analysis methods for the data we used in the previous assignments, namely the data containing the consumption of 7 food groups by 12 different french families. We already used the PC and factor analysis, now we will look at the data using the k-means algorithm and some hierarchical aglomerative clustering algorithm, dividing the data into 3 groups in both approaches. Let us start with the K-means algorithm.

From figure 1, we see the found clusters, together with the group means for the groups. From the plot, it seems that the groups can be well separated by the mean consumption, the bread consumption doesn't look like it affects the group belongings well. We shall compare the groups obtained by this process with some other clustering algorithm.

2 Using aglomerative clustering algorithm

Second way we chose to use on the food data was the aglomerative clustering algorithm, using the euclidian distances and the furthest neighbor approach. We can see the dendogram dividing the data into three groups in the figure 2.

The first observation should be that while the K-means algorithm divided the data into groups with almost the same number of observations (3, 4, 5), now we have a group with one observation, and a group with seven observations. We can compare the groups obtained before with the new groups to see, if there are any similarities at all. Before, the first group consisted of four families, coded by names "CA2", "EM5", "MA5" and "EM4". We see that both the first and second pair of families were grouped first in the dendogram, but the pairs ended in different final groups. Second group of families derived by the K-means approach contained the families coded by "CA3", "CA4" and "CA5". We see that the last member ended up being in its own group in the second approach, while the other two were again paired first. The last former group, consisting of five families, was a subset of the first group in the second approach, but two additional families were added.

The groups derived by the K-means algorithm were nicer, both in interpretation and in size of the groups. We could maybe obtain groups closer to them in the second approach, if we used different distances and methods.

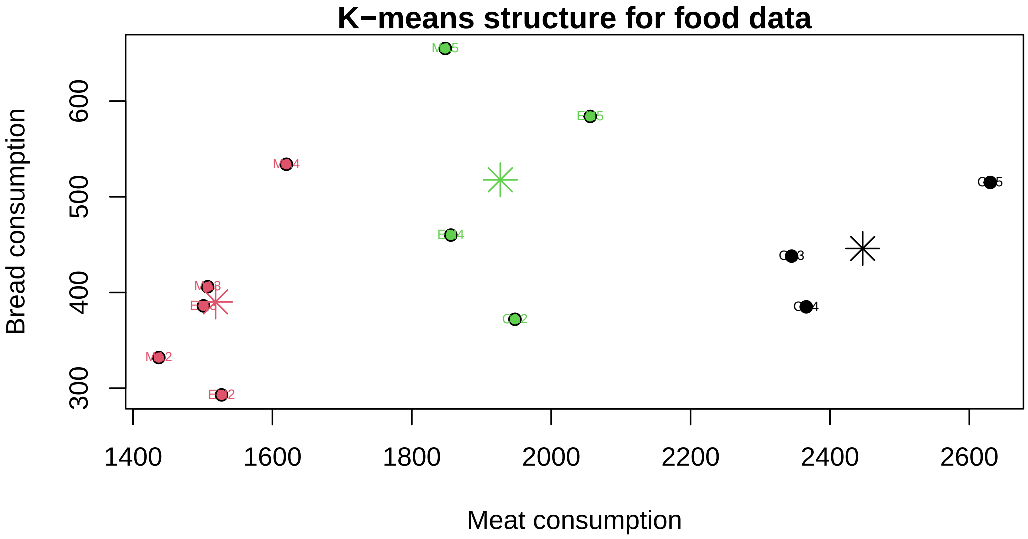

Figure 1: K-means structure found in the food data. The two consumptions which were plotted were found to be close to the first two principal components, which were found to be enough to describe the data well.

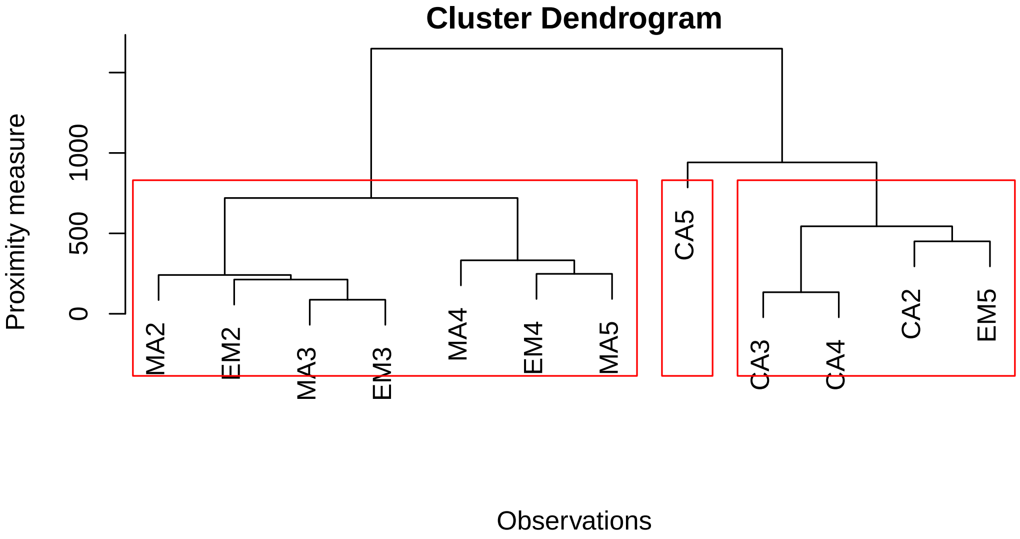

Figure 2: Projection of data on the first 2 principal components, which seem to describe 96 % of the total data variance.