Multivariate analysis - hw11

Matěj Pešek

May 8, 2022

1 Linear discriminant analysis

Our task in this assignment is to use the linear discriminant analysis on the data we used in the previous assignments, namely the data containing the average consumption of 7 food groups by 12 different french families, which differ in the number of kids (2,3,4,5), and in their jobs (manual workers, employees and managers). To make a use of the lda, we will include a new factor variable, describing if the families have 3 and less kids, or 4 and more kids. Note that both options happen exactly 6 times in the dataset.

Using lda, we will study how well we can estimate the number of kids in the family by their consumption. First we can take a look at the average consumption in the two groups divided by the number of children.

Children	bread	vegetables	fruits	meat	poultry	milk	wine
4 and more	522.1667	836	555	2062.667	894.6667	448.6667	365.6667
Less than 4	371.1667	628	455	1710.833	711.6667	267.8333	371.5000

Table 1: Table describing the average consuption of 7 food and drink groups depending on the number of kids in the families.

We see that apart from wine, families with more children had a higher consumption of all food groups. We can use the linear discriminant analysis, to find a one-dimensional line, onto which we will project the consumptions. Then we should see a clear difference between the families with 2 and 3 children, an the families with more children. We can find the projections in figure 1.

We see a clear difference between the two groups, with the families with 5 and 2 children being more distant from the border than the families with 3 and 4 children.

2 Using GLM on the same problem

Second way we can analyze the data is using the logistic regression. We omit the interactions to make the methods comparable. While we cannot compare the fitted values to the LDA projections, as the fitted values are either 0 or 1 (we have a small number of observations), we can compare the coefficients from the model with the scaling from the LDA. Both sets of numbers have a very different meaning, but we can at least compare the signs and the ratios of the numbers to see, which food groups are more important for which method.

Method	bread	vegetables	fruits	meat	poultry	milk	wine
GLM	0.312	0.305	0.157	0.380	-0.882	-0.919	0.106
LDA	0.014	0.03	0.008	0.035	-0.076	-0.081	0.014

Table 2: Table of coefficients of food groups for the two methods used.

By looking at table 2, we see that the coefficients from the GLM are fairly close to 10 times the scaling of the LDA. This might come as a surprise, as both methods use a very different way of dividing the data into two groups. Rather than both methods finding the same underlying structure, it might be a case of a small number of observations, which made both methods find similar and simple results. More observations would be needed to compare the methods better.

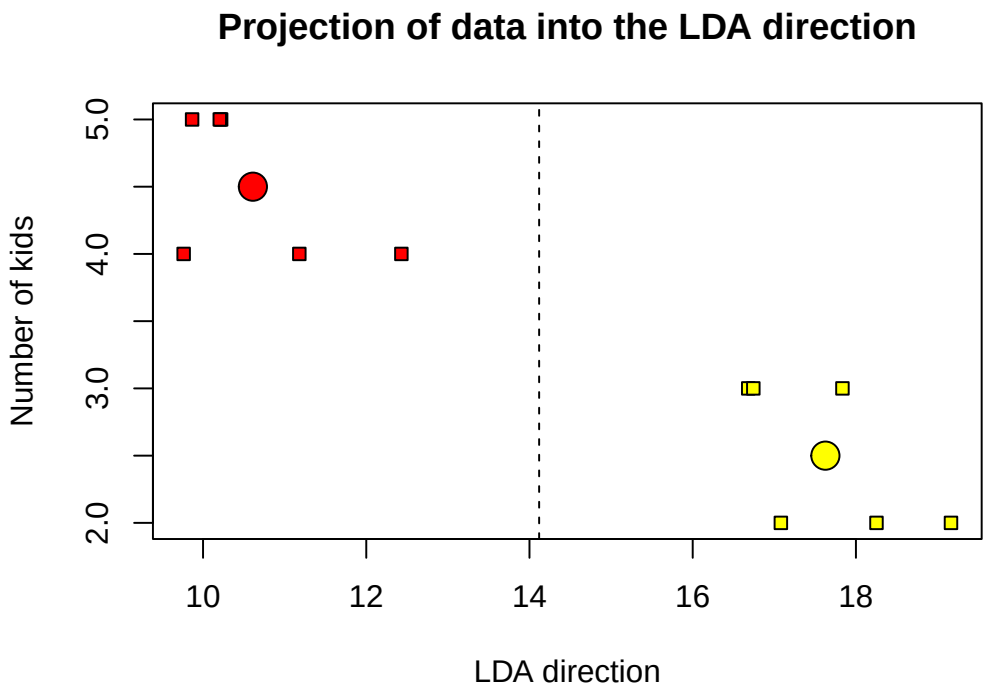

Figure 1: Projections of the different group consumption into the LDA direction. The different families are pictured as squares, circles represent the mean value in the two groups. The dashed line represents the border between the groups, which would be used to predict the number of kids for a new family.