Multivariate analysis - hw6

Matěj Pešek

April 11, 2022

1 Principal component analysis

Our task is to use principal component analysis on food data from r package SMSdata, consisting of consumption by 12 french families of 7 different food and drink types, including bread, milk, poultry and wine. Principal component analysis should allow us to find correlations between the food groups which would allow us to describe the consumption trends by less then seven numbers for each family. At the first glance, the values of consumption for different food types seem to be close to each other, with only meat being the exception, having much higher values than other food groups. But higher values doesn't seem to imply higher variance, and for that reason, we might consider not scaling the data before PCA. We can take a look at correlations between the food groups at figure 1.

We can see that apart from wine, higher consumption of one food group should imply also higher consumption of other food groups.

2 Plotting the PCA

We start by computing the components with highest variance by singular value decomposition of the data matrix (unscaled). We can take a look at the projection of the data on the first 2 principal components on figure 3. By using only 2 PC's, we could describe the consumption of the families by only two numbers, in which consumption of poultry, meat and fruits would go together on one axis, and consuption of bread and milk would go on second axis, with vegetable and wine not being strongly related with either of the groups. Valid question also should be, whether using only the first 2 principal components is enough, or more PC's should be used. We can plot what proportion of variance is described by the PC's, as seen on figure **??**. We see that the first 2 PC's are the only ones with higher than average variance. They also together describe 96 % of the total variance, which should be enough to only use the first 2 principal components.

3 Interpretation

By PCA, we see that values of consumption of poultry, meat and fruits are closely tied together, and can be described with 1 number, without losing a lot of precision. It seems to be the case also with the group of milk and bread consumption. We are not able to obtain the original values of consumption from the first 2 principal components, but we have still have a good idea about it, while at the same time we decreased the number of used variables for each family from 7 to 2.

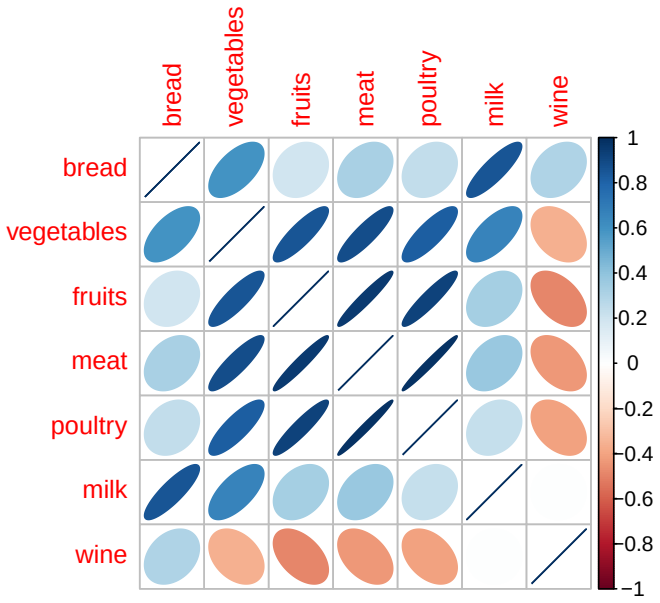

Figure 1: Visualization of correlations between food groups by library corrplot.

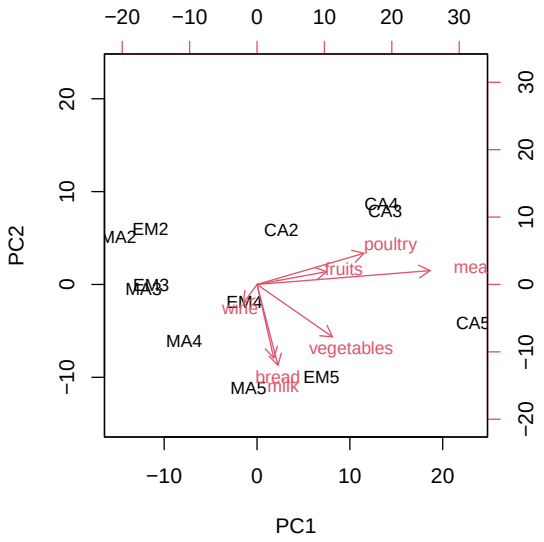

Figure 2: Projection of data on the first 2 principal components, which seem to describe 96 % of the total data variance.

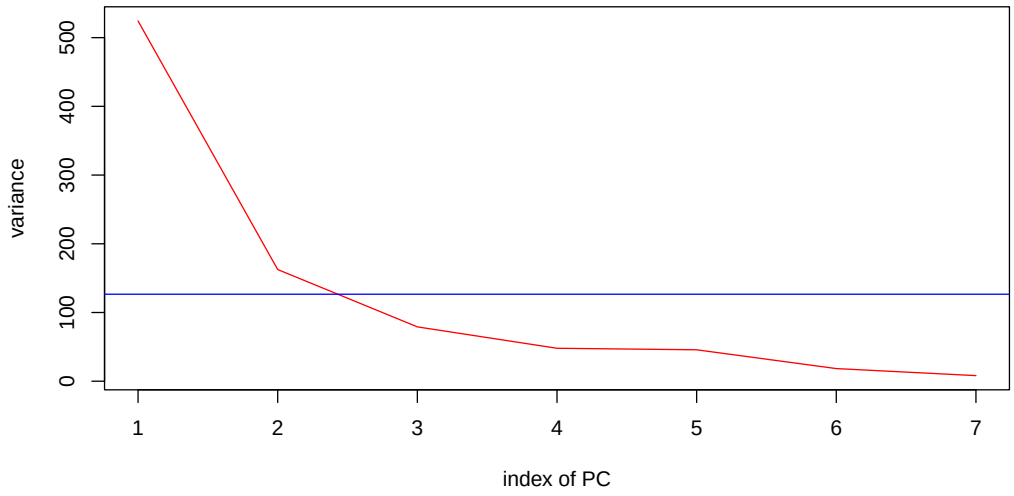

Figure 3: Variance of ordered principal components with average variance plotted by blue line.