MTH6101 Introduction to Machine Learning

Laboratory week four

The intention of this laboratory is to do Principal Component Analysis (PCA) of the olive oil data set. We mostly use the recommended function for PCA which is `prcomp`.

1. Open `RStudio` and create and save a new `R` script file for your commands. If you have `markdown` on your own computer, you are welcomed to use it.

2. In order to load the data, install the library `pdfCluster`. Once installed, load the library and load the data using the `R` command `data(oliveoil)`. Examine the data using commands such as `pairs` and `summary`. Also read the help of the library to find out what the data is about.

3. This dataset is complete in the sense that it has **no** missing values and can be used straightaway. Your analysis will **not** use the first two categorical variables `macro.area` and `region` and thus you will remove them by `X<-oliveoil[,-c(1:2)]`. **The data is now ready for PCA**.

4. Center the data but do not scale it, and compute and examine the variance-covariance matrix, with special emphasis in the diagonal so that you can try and predict what will happen with PCA.

5. Do PCA using the function `prcomp` with these data (use parameters in this function to center but not scale the data). Print a summary of the analysis, select a number of components and interpret them. Do a biplot and examine what is being plotted.

6. Repeat all that you did from 4 but now with the data centered and scaled.

7. Recall the relation seen in lectures $\mathbf{\Lambda} = \frac{1}{n-1}\mathbf{D}^2$ between eigenvalues of the K-L expansion of $\mathbf{\Sigma}$ and eigenvalues of svd of the data $\mathbf{X}$. Numerically check that the relation holds for both analyses you did in 4 and 6.