MTH6101 Introduction to Machine Learning

Laboratory week three

The intention of this laboratory is to do Principal Component Analysis (PCA) of the air quality data of Exercise 9. The tasks involved are loading, preparing the data and analysing it with the Karhunen-Loeve decomposition.

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. Load the data using the `R` command `data(airquality)` and examine the data. To this end, see the data using `airquality`; see what the entries of the variables are and make notes. Also describe the data using commands such as `pairs`, `summary` and `var`.

3. As part of this initial analysis, you will have noted that the first two variables of the data set have **missing values**. For the rest of the analysis, these need to be removed. Use the command `complete.cases(airquality)` to determine which are the values to be removed and **create** a new variable by allocating `airquality[complete.cases(airquality),]` to this new variable, say `X`. Examine `X` and make sure that you have removed missing values.

4. Your analysis will not use the variables `day` nor `month` which you will remove by `X<-X[,-c(5:6)]`. Then center by `X<-scale(x=X,center=TRUE,scale=FALSE)`.

5. **The data is now ready for PCA**. Compute the variance-covariance matrix of `X` and then do the Karhunen-Loeve decomposition of it with by the now well known commands `var` and `eigen`. Store each result in a new variable.

6. Write the eigenvalues you just obtained and compute and interpret the proportional contribution of each eigenvalue to the total variability (sum of eigenvalues which of course equals the trace of the variance covariance matrix).

7. Do a pairs plot of the Principal Components and compare this with your previous use of the command `pairs` for the data `X`. Recall that the Principal Components are the rotated data, computed as `X%*%E$vectors`, assuming that the variable `E` has the results of the K-L decomposition.

8. Redo all from step 4 by scaling the data as well with `scale=TRUE)`. Compare with the previous analysis.