Exercise 2

Problem 1: Principal Component Analysis

Upload the file DECATHLON.txt into your R-workspace. The file contains the results of 48 decathletes from 1973. Familiarize yourself with the data and perform the covariance matrix based PCA transformation. Conduct the analysis without the variables: points, height and weight.

a) Familiarize yourself with the function **princomp**. Visualize the original data.

b) How much of the variation of the original data is explained by k principal components, where $k = 1, 2, ..., 10$.

c) Choose a sufficient amount of principal components and try to interpret them. Visualize the scores of the observations with respect to the first two principal components.

d) Calculate the sample mean and covariance matrix from the score matrix.

Problem 2: Eigendecomposition of a symmetric matrix

Let A be a symmetric matrix with distinct eigenvalues. Show that the eigenvector matrix of A is orthogonal.

Homework Assignment 2: PCA for simulated data

Simulate 100 observations from bivariate normal distribution with parameters:

$$\mu = \begin{pmatrix} 4 \\ 7 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 10 & 6 \\ 6 & 8 \end{pmatrix}.$$

Your report should contain the requested figures for (a), (c) and (g) (note that providing only the R-code is not enough for these parts). Additionally, your report should contain the requested R-code for (b), (e) and an answer to (d), (f).

(a) Plot the data. Label the data points with the corresponding observation number.

(b) Perform the covariance based PCA transformation to the data set.

(c) Plot the score matrix. Use the same scale as in (a) and label the data points with the corresponding observation number. Choose your scale (limits for the x- and y-axis) in a way that all the observations are visible in the figure.

(d) Compare the plots of (a) and (c) and describe the differences.

(e) Calculate the G and Y matrices without using any existing PCA functions. Note that the function **princomp** scales the covariance matrix with $1/n$ (instead of the usual $1/(n-1)$). Attach the R-code to your solution.

(f) Verify that the estimated scores and the loadings are equal (up to signs) in parts (b) and (e). Hint: If parts (b) and (e) are done correctly, the scores and loadings should be the same up to heterogeneous sign changes.

(g) Plot the directions of the first and second principal component to the original data. The function **arrows** might be useful.