G0O01: Statistische modellen en data-analyse
Oefenzitting 4

Exercise 1:

1. Read in `R` the file Dblood.dat. This data (from Jolliffe 2002) consist of 8 blood chemistry variables measured on 72 patients.
2. Compute (and store) the corresponding covariance and correlation matrices of this dataset.
3. Compute and compare the loading matrices obtained by applying PCA on (a) the covariance matrix and (b) the correlation matrix. When looking at the first loading vector, what is the main difference? What do you think causes this?
4. Now consider the PCA analysis that you think is the most appropriate for this dataset. How many components would you choose to retain? Explain.

Exercise 2: bivariate example

1. Load the Headsize data set (headsize.dat). The variables `head1` and `head2` are, respectively, the headsizes of 25 first sons-second sons pairs.
2. Create a matrix composed of the variables `head1` and `head2`.
3. Would you rather perform a PCA analysis on the correlation or the covariance matrix here? Explain.
4. From now on, consider the PCA analysis that you think is the most appropriate for this dataset. Perform a PCA decomposition.
5. Compute the equation (intercept,slope) of the two principal components.
6. Plot the data, using the `asp=1` option in the `plot()` command (this ensures that the axes both have the same scales). Draw the first two principal components.

7. On this plot, add the 97.5% tolerance ellipse and the least squares regression line through the data.

8. Compute the correlation between each of the principal components and the original variables.

Exercise 3: multivariate example

1. Load the heptathlon dataset. This is a dataset of measured performance on the 7 components of olympic heptathlon for the 24 participants of the 1988 Seoul games (the last column is the finale score of the participants). Create a data matrix X composed of the first seven columns of the heptathlon dataset. We will be analysing X.

2. Would you rather perform a PCA analysis on the correlation or the covariance matrix here? Explain.

3. How many principal components k would you recover? Explain.

4. Interpret the first principal component.

5. Make scatter plots of the k scores. Interpret.

6. Make a plot of the `score` variable from the data set versus the first principal score. Interpret.

7. Make a diagnostic plot with the Mahalanobis distances of the k-dimensional scores on the horizontal axis and the orthogonal distances to the PCA subspace on the vertical axis. Discuss the results.

Useful function: the `PcaClassic` function from the `rrcov` library. Make sure you read the associated documentation (and also that of `rrcov::Pca-class`).