Series 3

1. The following table shows the result of the decathlon at the olympic games in Atlanta at 1996. The data is in the dataframe `zehnkampf.dat` and the variables are:

`m100`	100m-sprint	`m400`	400m-sprint	`speer`	javelin
`weit`	long jump	`hurd`	110m-hurdles	`m1500`	1500m-sprint
`kugel`	shotput	`disc`	discus	`punkte`	total number of points
`hoch`	high jump (cm)	`stab`	pole vault (cm)		

Athlet	`m100`	`weit`	`kugel`	`hoch`	`m400`	`hurd`	`disc`	`stab`	`speer`	`m1500`	`punkte`
OBRIEN	10.5	7.57	15.7	207	46.8	13.9	48.8	500	66.9	286	8824
BUSEMANN	10.6	8.07	13.6	204	48.3	13.5	45	480	66.9	271	8706
DVORAK	10.6	7.6	15.8	198	48.3	13.8	46.3	470	70.2	271	8664
...						...					...
AFANASYEV	11.4	6.74	13.4	198	50.8	14.8	43.1	0	55.1	281	6711

a) Make a biplot of the `zehnkampf`-data using the covariance matrix for determing the principal components. Compare it to the biplot using the correlation matrix instead. Which of the two plots seems more advisable? And why?
R-hint: Use the function `biplot` and for the PCA the function `princomp`. Check the help of `princomp` on how to use the correlation instead of the covariance matrix.

b) Take a closer look at the biplots. Answer the following questions and make a short comment about your decision.

- Which discipline has high correlation with the total number of points (i.e. `punkte`)?
- Which variable is displayed badly by the projection?
- State two disciplines with high positiv correlation.
- State two disciplines with high negativ correlation.
- State two disciplines which are uncorrelated.

c) Who is an average athlete? There several answers possible.
R-Hint: `identify()` - after clicking in the plot, right-click to stop `identify` and get the desired output.

Source: The data is from the web-site `http://www.atlanta.olympic.org/` ("Official 1996 Olympic Web Site").

2. In this exercise we will look at eigenvalues and eigenvectors. We consider the data `iris2.dat`. Let $\underline{X}$ be the $p \times n$-matrix with the samples in its columns.

a) Restrict `iris2.dat` to the species *Iris setosa* (`SPECIES=1`) only. Furthermore, we only need length and width of the sepal leaves (`SEP.L` and `SEP.W`).

b) Make a scattor-plot. First center the data `iris.dat`, such that the orgin is at the middle.
R-Hint: `scale()`

c) Determine the covariance-matrix $\mathbf{S}$.
(**R-Hint:** `cov()`)

d) Since $\mathbf{S}$ is symmetric and positive semi-definite, we can decompose the matrix $\mathbf{S}$ according to the eigenvalue problem

$$S = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^\top,$$

where $\mathbf{\Lambda} = \mathrm{diag}[\lambda_1, \ldots, \lambda_n]$ is a diagonal matrix, with $\lambda_i \geq 0$, and $\mathbf{A}$ is an orthogonal matrix, that means $\mathbf{A}\mathbf{A}^\top = 1$ and hence $\mathbf{A}^{-1} = \mathbf{A}^\top$.

The values λ_i are called eigenvalues and the vectors in the columns of $\mathbf{A}$ are called eigenvectors. Find the eigenvalues and eigenvectors of the matrix $\mathbf{S}$. Verify $\mathbf{A}\mathbf{A}^\top = 1$.
(**R-Hint**: `eigen()`):

e) A transformation of the data with $\mathbf{A}$, that means

$$\underline{Z} = \mathbf{A}^\top \underline{X},$$

yields transformed data having a covariance-matrix equal to the diagonal matrix $\mathbf{\Lambda}$. The reason is

$$\widehat{\mathrm{var}}(\underline{Z}) = \widehat{\mathrm{var}}(\mathbf{A}^\top \underline{X}) = \mathbf{A}^\top \widehat{\mathrm{var}}(\underline{X})\mathbf{A} = \mathbf{A}^\top \mathbf{S}\mathbf{A} = \mathbf{\Lambda}.$$

Transform the data according to the mapping above. Make a scatter-plot for $\underline{Z}$. What do you get? What kind of mapping is $\mathbf{A}$ and $\mathbf{A}^{-1} = \mathbf{A}^\top$ respectively?

f) Plot the non-transformed data together with the following lines

$$g_1: \quad \underline{z}_1 = \mathbf{A}^\top \begin{bmatrix} t \\ 0 \end{bmatrix} \quad \text{und} \quad g_2: \quad \underline{z}_2 = \mathbf{A}^\top \begin{bmatrix} 0 \\ t \end{bmatrix},$$

with $t \in [-2, 2]$.

g) Compare the previous results with the output of `princomp` (eigenvalues, eigenvectors, scores). Do they agree?

Preliminary discussion: 26.03.12.

Deadline: No hand-in.