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:

<table>
<thead>
<tr>
<th>Athlet</th>
<th>m100</th>
<th>weit</th>
<th>kugel</th>
<th>hoch</th>
<th>m400</th>
<th>hurd</th>
<th>disc</th>
<th>stab</th>
<th>speer</th>
<th>hoch</th>
<th>m1500</th>
<th>punkte</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBRIEN</td>
<td>10.5</td>
<td>7.37</td>
<td>15.7</td>
<td>207</td>
<td>46.8</td>
<td>13.9</td>
<td>48.8</td>
<td>500</td>
<td>66.9</td>
<td>286</td>
<td>8824</td>
<td></td>
</tr>
<tr>
<td>BUSEMANN</td>
<td>10.6</td>
<td>8.07</td>
<td>13.6</td>
<td>204</td>
<td>48.3</td>
<td>13.5</td>
<td>45</td>
<td>480</td>
<td>66.9</td>
<td>271</td>
<td>8706</td>
<td></td>
</tr>
<tr>
<td>DVORAK</td>
<td>10.6</td>
<td>7.6</td>
<td>15.8</td>
<td>198</td>
<td>48.3</td>
<td>13.8</td>
<td>46.3</td>
<td>470</td>
<td>70.2</td>
<td>271</td>
<td>8664</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFANASYEV</td>
<td>11.4</td>
<td>6.74</td>
<td>13.4</td>
<td>198</td>
<td>50.8</td>
<td>14.8</td>
<td>43.1</td>
<td>0</td>
<td>55.1</td>
<td>281</td>
<td>8711</td>
<td></td>
</tr>
</tbody>
</table>

**a)** Make a biplot of the `zehnkampf`-data using the covariance matrix for determining 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/](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 $X$ 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 scatter-plot. First center the data `iris.dat`, such that the origin is at the middle.  
**R-Hint:** `scale()`  
**c)** Determine the covariance-matrix $S$.  
**R-Hint:** `cov()`  
**d)** Since $S$ is symmetric and positive semi-definite, we can decompose the matrix $S$ according to the eigenvalue problem  

$$S = A \Lambda A^\top,$$

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

The values $\lambda_i$ are called eigenvalues and the vectors in the columns of $A$ are called eigenvectors. Find the eigenvalues and eigenvectors of the matrix $S$. Verify $AA^\top = 1$.  
**R-Hint:** `eigen()`;
e) A transformation of the data with $A$, that means

$$Z = A^\top X,$$

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

$$\text{var}(Z) = \text{var}(A^\top X) = A^\top \text{var}(X) A = A^\top S A = \Lambda.$$  

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

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

$$g_1 : \quad z_1 = A^\top \begin{bmatrix} t \\ 0 \end{bmatrix} \quad \text{und} \quad g_2 : \quad z_2 = A^\top \begin{bmatrix} 0 \\ t \end{bmatrix},$$

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.