Sheet 2

Hand in solution by **March 7, in the lecture room**

1. In this exercise we study the data `banknoten.dat`, which contains measurements of forged and unforged banknotes. We practice how to standardise using the Cholesky decomposition. For the purpose of this exercise we only consider the unforged (`CODE == 0`) banknotes.

Read in the data using the R command:

```
d.banknoten <- read.table("http://www.stat.math.ethz.ch/~stahel/
courses/multivariate/datasets/banknoten.dat", header=TRUE)
```

Extract the unforged notes and get rid of the `CODE` variable:

```
d.banknoten <- d.banknoten[d.banknoten[,"CODE"]==0,-1]
```

a) Determine the mean and the standard deviation of the unforged banknotes for the 6 variables.

b) Determine the covariance and the correlation matrix for the 6 variables. Which variables show high correlation? (**R-Hint:** `var()` and `cor()`)

c) We standardise the data using the Cholesky decomposition:

- Make a Cholesky decomposition of the covariance matrix using `chol()`. To get a lower triangular matrix, as in the lecture, use the transpose function `t()` in R.
- Invert the matrix with `solve()` (`C <- solve(...)`).
- As in the lecture, transform the data $\underline{z}_i = \mathbf{C}(\underline{x}_i - \overline{\underline{x}})$.

  R-Hints: Center the data with `scale(d.banknoten, scale=FALSE)`. Multiply the centered data matrix with the matrix `t(C)`.
- Make a scatter plot of the transformed and untransformed data of `RIGHT` and `LEFT`.

2. For this item, we work with the data set `iris`, which is available in R directly.

a) For the setosa plants, obtain the characteristic measures mean and standard deviation for the 4 variables.

b) For the setosa plants, calculate the covariance and the correlation matrix of the 4 variables. Which variables correlate strongly?

c) (*) Do the calculations for the covariance matrix by following the formula

$$\widehat{\text{var}}(\underline{X}) = \frac{1}{n-1} x_c^T x_c \,, \text{ where } x_c = x - \underline{1}\,\overline{\underline{x}}^T$$

rather than just calling the R function `var`.

3. a) 100 realizations $\underline{x}_1, \ldots, \underline{x}_{100}$ of a 2-dimensional random vector are given. The sample mean is $\overline{\underline{x}} = [-1.5, 0.3]^T$ and the sample covariance matrix is

$$\widehat{\mathrm{var}}(\underline{X}) = \begin{bmatrix} 1.1 & -0.3 \\ -0.3 & 1.3 \end{bmatrix}.$$

Define the linearly transformed observations $\underline{y}_i = \underline{a} + B\underline{x}_i$ where

$$\underline{a} = [-1, 2]^T \text{ and } B = \begin{bmatrix} 2 & 0 \\ 1 & -4 \end{bmatrix}.$$

Compute the sample mean and sample covariance matrix of the transformed data by hand and verify your results in R.

b) Which of the matrices below are covariance matrices. Explain!

$$A = \begin{bmatrix} 5 & -1 \\ -1 & -2 \end{bmatrix} \quad B = \begin{bmatrix} 3 & 2 \\ 1 & 3 \end{bmatrix} \quad C = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \quad D = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad E = \begin{bmatrix} 1 & 4 \\ 4 & 1 \end{bmatrix}$$

Hint: There is no need to compute eigenvalues. You can argue statistically.