Sheet 3

Hand in solution by **March 21, in the lecture room**

1. Figure 1 displays a scatter plot of two variables $X^{(1)}$ and $X^{(2)}$.

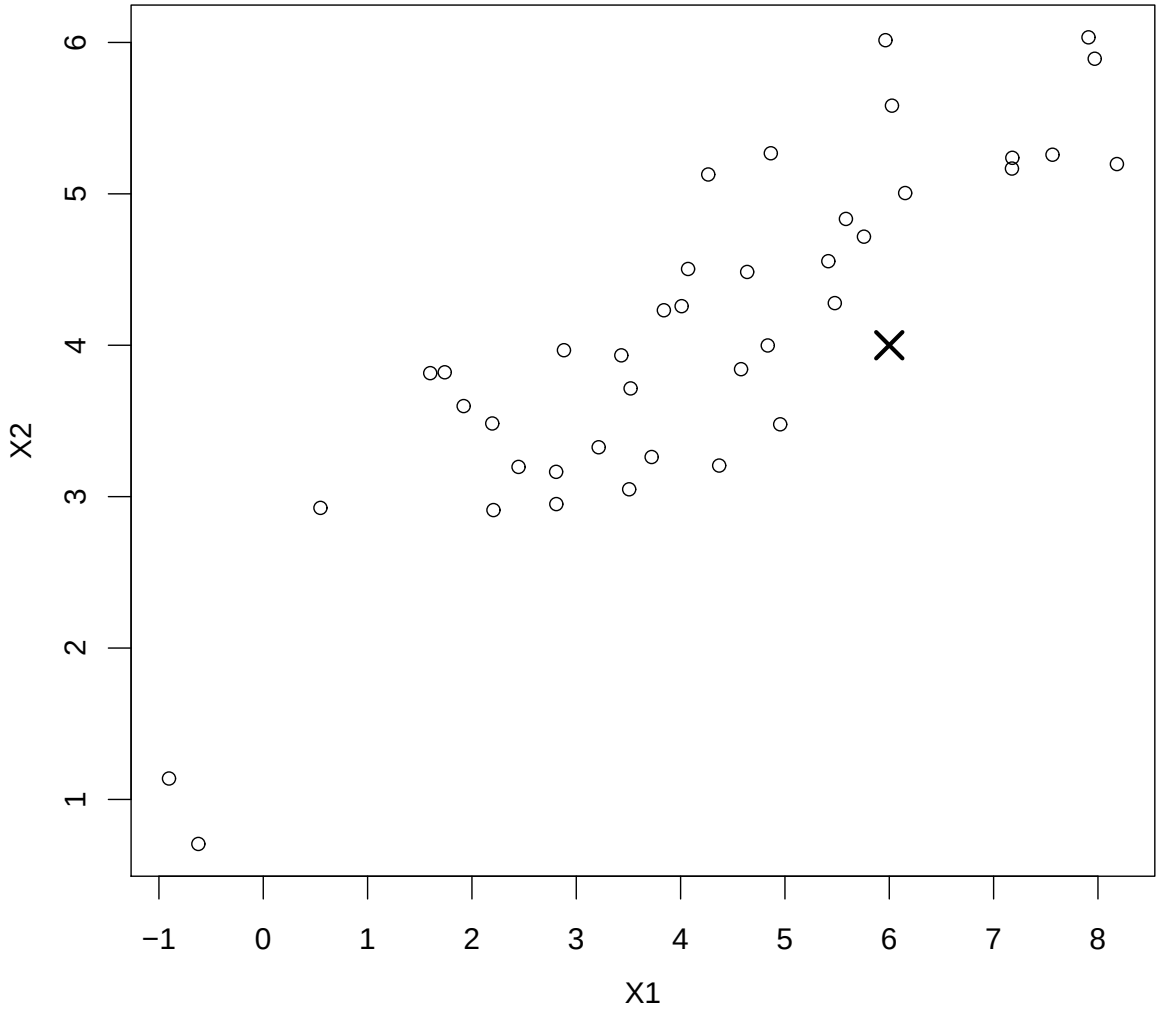

Figure 1: Scatter plot for Exercise 1.

a) Draw by eye the principal component axes into the figure.

b) Estimate by eye the standard deviations for the two principal components, and the two eigenvalues of the covariance matrix.
Hint: For normally distributed univariate data, roughly 2/3 of the data lie within ± 1 standard deviation of the mean.

c) Estimate by eye the values of the principal components for the point marked by $\times$.

2. We consider again the Alp Flix data which comes from a study of an ecosystem of an alp in Grisons. Each observation describes a parcel on this alp. Some variables characterize the soil and land use, and others count the number of individuals of 64 species. The variable named `VegetationGroup` contains the type of vegetation present.

Load the data:

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

Select the species variables (columns 19–82). Two of the species appear nowhere. First find them and exclude the respective variables. This can be done with the following code:

```
t.d <- d.vegenv[,19:82]
t.mn <- mean(t.d)
which(t.mn==0)
t.d <- t.d[,t.mn>0]
```

a) Perform principal component analysis of the species variables and give pairwise scatter plots of the first 3 principal components. Force standardization of variables by the argument `scale = TRUE` in `prcomp`.
Hint: To better judge the utility of this plot, you can set the plotting character `pch` in the function `pairs` to be the variable `d.vegenv$VegetationGroup`. The object `r.pca <- prcomp(...)` is a list, you can extract the rotated data via `r.pca$x`.

b) Do the same for the square root-transformed data, with `scale = TRUE` and with `scale = FALSE`. Which of these plot separates the data best?

c) What proportion of total variance is explained by the first 3 principal components? How many principal components would be needed to explain at least 90% of the variance? Compare the different versions.
Hint: Use `summary(r.pca)`.

3. We consider biplots for the count data from the previous question using the function `g.biplot`. The function `g.biplot` is not built-in in R but can be made available with `source("http://stat.ethz.ch/~stahel/courses/multivariate/g.biplot.R")`.

a) Apply the function `g.biplot` to the result of the principal component analysis of the *square root-transformed* data. The result is a bit of a mess.

b) From the non-transformed data `t.d`, select only those variables whose mean is at least 1, and *square root-transform* these variables. Repeat the PCA (forcing standardization!) on the reduced data and give the biplot.
Hint: The appropriate variables can be selected via `t.d[,mean(t.d)>1]`.

In the following tasks, we look at the biplot from **b)**.

c) Which species is best represented in the biplot?

d) Draw by eye the projection of the observations 30, 34 and 46 onto the arrow of the species `Nardstri`. According to these projections, which observation is expected to have the highest count of `Nardstri`, which one the lowest? Compare with the true counts of `Nardstri`.