Sheet 7

Hand in solution by May 30 in the lecture room

1. **Multidimensional scaling**. Select the species variables (columns 19–82) in the dataset `vegenv`, exclude the two species that appear nowhere (as in Exercise 2 on Sheet 3), and *square root-transform* the counts:

```
t.url <-
  "http://stat.ethz.ch/~stahel/courses/multivariate/datasets/vegenv.dat"
d.vegenv <- read.table(t.url, header=TRUE)
t.d <- d.vegenv[,19:82]
t.mn <- mean(t.d)
t.d <- sqrt(t.d[,t.mn>0])
```

To perform multidimensional scaling, you need the function `isoMDS` from `library(MASS)`.

a) Perform multidimensional scaling with the Manhatten distance as measure of dissimilarity. Plot the result using different symbols for the different vegetation types. Are the vegetation types well seperated?
Hints:

```
t.dist <- dist(?, method = "manhattan")
t.r <- isoMDS(t.dist)
plot(t.r$points, pch = d.vegenv$VegetationGroup)
```

b) Repeat **a)**, but this time with standardized (square root-transformed) counts. Comments?

c) Compare the result from **b)** with PCA.

2. **Hierarchical clustering.** Use the log-transformed petal and sepal measurements from the `iris` dataset as raw data for clustering, and check your results by comparing with `iris$Species`.

a) Carry out hierarchical clustering using the function `hclust`. Use “average linkage”, which is expected to produce something between round and elongated clusters. From the tree you obtain, extract the subdivision of the data into 2 clusters. Provide a `pairs` plot to check whether one of the iris species has been separated from the others (use different colors for the different clusters, and different point characters for the different species).
Hints: Construct the matrix of dissimilarities using
`t.dist <- dist(scale(log(iris[,1:4])),method="euclidian");`
Cluster the data with
`t.hcl1 <- hclust(t.dist,method="average");`
Extract the memberships of the observations for a subdivision into 2 clusters via
`t.gp <- cutree(hcl1,k=2).`

b) Now use the same clustering tree to extract the subdivision into 3 clusters, and again check whether the species are correctly distinguished in this way. How to explain the result you see?

c) (*) Try other linkages (e.g. `single`, `complete`, `ward` or `centroid`). Which methods produce acceptable divisions into 3 clusters?

3. K-means clustering. We use the same raw data for clustering as in the previous exercise.

a) Perform K-means clustering with 3 groups without giving initial cluster centers (in this case, 3 distinct observations are randomly chosen as initial centers). Use a `pairs` plot to check the quality of the obtained clusters. Repeat the same commands several times and observe the changes in cluster assignment.
Hint: `t.km1 <- kmeans(scale(log(iris[,1:4])),3)`.

b) Now use the logarithms of the means of the 3 species as initial centers. Are the clusters defined by the 3 iris species correctly identified?
Hint: `t.km2 <- kmeans(scale(log(iris[,1:4])),centers=log(cent))`, where `cent` is a matrix (with 3 rows) that stores the 3 initial centers.

Information about time and place for the **Ferienpräsenz** can be found on the lecture homepage `http://stat.ethz.ch/education/semesters/ss2011/ams`.