7 Principal component analysis

There are several functions in R that can be used for principal component analysis. Below we will use the `princomp` function as working horse. Other possibilities include the `rda` function from the `vegan` package, which has advanced options for scaling, among others.

7.1 PCA: Physical measurements of crabs

Prerequisites: Exercise 3.3 (working with datasets)

The data for this exercise come from 200 specimens of a certain type of crabs. The crabs come in two colours (blue and orange). In the experiment 100 of each type were collected, 50 males and 50 females, and for each of the 200 crabs, five quantities were measured: The carapace/shell length (`CL`), carapace/shell width (`CW`), size of frontal lobe (`FL`), rear width (`RW`), and body depth (`BD`). The experimenters are interested in characterization of the colour types (and sexes) in terms of the variables.

The data are available as the dataset `crab` in the `MASS` package.

1. We are going to work on the log-measurements. Try the following commands and explain what happens:

   ```
   library(MASS)                  # Load package
   head(crabs)                    # Just looking at the data
   logcrabs <- log(crabs[,4:8])   # Dataset with log-values
   head(logcrabs)                 # The log-dataset

   group <- crabs$sex : crabs$sp  # Group variable
   group
   plot(logcrabs, col=group)
   ```

   In particular, does the raw data make it possible to easily distinguish between the four groups? Notice that black/red/green/blue corresponds to Female-Blue/Female-Orange/Male-Blue/Male-Orange.

2. A PCA can be carried out with the `princomp` function. If we use the option `cor=T`, then the correlation matrix (rather than the covariance function) is used. This correspods to a scaling of the variables. Try the following commands and discuss the output:

   ```
   pca <- princomp(logcrabs, cor=T)
   pca
   summary(pca)
   plot(pca)
   loadings(pca)
   pca$scores
   ```

3. Try the following commands and discuss the graphs. I particular, is it possible to use the principal components (the scores) to distinguish between the four groups? Which

aspects of crab characteristics relate to the three first components (recall the association between colours in the graphs, and colour/sex of the crabs from question 1)?

```
scorData <- data.frame(pca$scores)
plot(scorData)
plot(scorData, col=group)
plot(scorData[,1:3], col=group)
plot(logcrabs[,1], scorData[,1], col=group)
```

7.2 PCA: Ecological zones along the Doubs River

Prerequisites: Exercise 7.1 (PCA)

As part of a large project on characterization of ecological zones, 11 environmental variables were measures at 30 sites along the Doubs River. The variables were distance from the source, *i.e.* from the start location (`das`), altitude (`alt`), slope (`pen`), mean minimum discharge (`deb`), pH of water (`pH`), concentration of calcium, phosphate, nitrate, ammonium, respectively (`dur`, `pho`, `nit`, `amm`), dissolved oxygene (`oxy`), biological oxygen demand (`dbo`). See Borcard *et al.* (2011) for more details.

The data is saved in the files `DoubsEnv.xlsx` and `DoubsEnv.csv`.

1. Read the data into an R dataset called `doubs`.

2. The first variable in the dataset, `das`, is not an environmental variable, so we will only use the remaining variables. Run a PCA on the those data, *i.e.* on `doubs[,-1]`.

3. How many components are needed to explain 75% and 90%, respectively, of the total variation? Make a plot of the first two principal components.

4. One of the aims of the study was to identify ecological zones, *i.e.* groups of sites that are similar in certain senses. Based on the abundance of different fish species, the 30 sites were allocated to four clusters. This allocation is given by the variable `clus4` below.

   ```
   clus4 <- c(rep(1,10), rep(2,9), rep(3,3), rep(4,3), rep(3,5))
   clus4
   ```

   For example,the first site in included in cluster 1, whereas the last site is included in cluster 3.

   Make the plot of the two first principal components again, this time with the points coloured according to the clusters. Do the first two principal components contain information about the clusters?