

MULTIVAR STATS IN ECOL AND GENETICS — EXERCISES, SHEET 5

Exercise 1 This exercise trains you in reading biplots. Consider the correlation biplot and the distance biplot of the height and weight data of the lecture.

- First execute the commands from the script to produce the biplots yourself. You find the file `HeightShoeWeight.txt` on the course homepage.
- In the distance biplot 'weight' and 'height' are orthogonal to each other. Does that mean that they are uncorrelated? You might want to confirm your answer with the `cor()` command.
- From the two biplots read off the approximate (Euclidian) distance between data point '211' and data point '103'.
- Is the data point '211' a student with normal weight, overweight or underweight?
- Guess the sex of student '121'.

Exercise 2 Let us continue as in Exercise 1. Consider the two biplots of the height and weight data of the lecture.

- By looking at the biplots, order the students '104', '106' and '226' by increasing weight. Hint: You need to use projections.
- Due the projection in the plane of the first two principal components, the angle between lines in the correlation biplot might not well represent the correlation. Use the command `cor()` to find out which of the correlations between `weight`, `shoe` and `height` is represented poorly.
- By looking at the biplots, order the students '98', '103' and '106' by increasing shoe sizes.

Exercise 3 Let's see how PCA rotates 2-dimensional data.

- Sample 500 values from a multivariate normal distribution mean vector $(0, 0)^T$ and with covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix} \quad (1)$$

For this load the library `mvtnorm` and sample with the R command `rmvnorm()` as in the lecture.

- Look at the data cloud (set the x-range and the y-range suitably: the cloud looks roundish if the ranges are not fixed). Guess how PCA is going to rotate the cloud.
- Apply `prcomp(, scale=FALSE)` to your sampled data and store the returned object as `mypca`. Have a look on the object `mypca` with `unclass(mypca)`. Add the transformed data cloud to the plot of the original cloud with `points(mypca$x, col="red")`. Consider the rotation matrix. Rounding its entries, which matrix does the rotation matrix resemble?
- Have a look on the distance biplot and on the correlation biplot.

Exercise 4 Let us see how 5-dimensional data is visualized. In order to know the result in advance, we start with 2-dimensional data and transform it into \mathbb{R}^5 .

- Sample 500 values from a multivariate normal distribution mean vector $(0, 0)^T$ and with covariance matrix

$$\Sigma = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix} \quad (2)$$

Denote the data matrix as \times . Look at the data cloud (set the x-range and the y-range suitably: the cloud looks roundish if the ranges are not fixed).

- Multiply (matrix multiplication `%*%`) \times with the matrix

$$\begin{pmatrix} 0.5 & 0 & 0.5 & 0.5 & 0.5 \\ 0 & -1/\sqrt{2} & 0.5 & -0.5 & 0 \end{pmatrix} \quad (3)$$

- Now apply PCA. Try to plot the transformed data into the same plot as \times such that the points match. You might have to rotate or flip the transformed data (e.g. multiply the x-coordinate with -1). For this note that the transformed data is 5-dimensional and that you only need the first two columns.
- Have a look on the distance biplot and on the correlation biplot.

Exercise 5 Download the data set 'EWU.txt' from the web page. The variables X1, X2, X3 and X4 have been explained in the lecture.

- Apply PCA to the data set and produce the distance biplot and the correlation biplot.
- Sort the countries *GB*, *D*, *I* and *IRL* according to their public dept level in 1997. Which of these countries have a debt level below the average?
- Apply all three rules from the lecture in order to decide which of the principal components are important.