Multivariate Data Analysis
Prof. Dr. Christina Andersson

Exercise Sheet 1: Principal Component Analysis

Computer Problems:

1. The data set *USArrests* contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas. USArrests data sets comes with basic dataset with 4 variables. Using PCA, we are going to find linear combinations of the variables that maxmimal variance and mutually uncorrelated.

   (a) Which variables are contained in the data set?

   (b) Create scatterplots of the different variable combinations.

   (c) Perform a standard PCA!

   (d) How many principal components are returned by the standard PCA?

   (e) How much of the total variance is explaned by the different principal components?

   (f) Take a look at the scree plot and interpret the results.

   (g) Produce a plot of the first two principal components

   (h) Interpret the loadings for the first two principal components.

2. The famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

   (a) Which variables are contained in the data set?

   (b) How many observations are contained in the data set?

   (c) Create scatterplots of the different variable combinations.

   (d) Perform a standard PCA for the quantitative variables in the data set!

   (e) How many principal components are returned by the standard PCA?

   (f) How much of the total variance is explaned by the different principal components?

(g) Take a look at the scree plot and interpret the results.

(h) Produce a plot of the first two principal components

(i) Interpret the loadings for the first two principal components.