Exercise 4

Problem 1: Robust PCA

The data set in this exercise is a modified version from Draper and Smith (1966) and it was used to determine the influence of anatomical factors on wood specific gravity, with five explanatory variables. The data is contaminated by replacing a few observations with outliers.

a) Install the package *rrcov*. Download the data "wood.txt" into your R-workspace.

b) Plot the variables pairwise. Do you see any outliers?

c) Estimate the covariance matrix using the classical sample covariance matrix and using the Minimum Covariance Determinant (MCD) method. What differences do you see between the two covariance matrices?

d) Calculate the robust and classical Mahalanobis distances for the data set and identify the outliers. Mahalanobis distance is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)},$$

where $x = (x_1, x_2, ..., x_N)$ contains all the observations, μ is the mean vector and S is the chosen covariance matrix estimate.

e) Assume that the original data is normally distributed. Perform the PCA transformation using both of the covariance matrix estimates, that were calculated in (c). Should we use covariance or correlation based PCA? Compare the loadings of the different approaches.

f) Plot the principal components pairwise.

Problem 2: Covariance Estimators

Simulate 200 observations from the following bivariate distributions:

a) Bivariate standard normal distribution.

b) Bivariate t-distribution where the degree of freedom is 5 and the scale matrix is an identity matrix.

c) Bivariate distribution where the first component follows the Weibull distribution with parameters $a = 1$ and $b = 2$ and the second component follows the Gamma distribution with the parameters $\alpha = 2$ and $\beta = 1$.

Visualize the simulated variables. Calculate the MCD and the regular sample covariance from the distributions. Compare the estimates.

Home Exercise 4: Influence functions and breakdown points

a) Derive the asymptotic and sample breakdown point of sample median.

b) Plot the empirical influence function of the median. You can use the code under lecture 4 as a starting point.

c) According to (a) and (b), is sample median a robust estimate?