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**. Upload the data "wood.txt" into your R-workspace.

b) Plot the variables pairwise. Can you detect any outliers?

c) Estimate the covariance matrix using the regular sample covariance matrix and using the Minimum Covariance Determinant (MCD) method. Are there differences between the estimates?

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

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

where $T(\cdot)$ is an estimator of location and $S(\cdot)$ is an estimator of scatter. By choosing T as the sample mean and S as the regular covariance matrix, we get the regular Mahalanobis distances. Likewise, by choosing the MCD as S and the corresponding estimator of location as T, we get the robust version.

e) Assume that the original data is normally distributed. Perform the PCA transformation using the regular covariance matrix and MCD. Should we use covariance or correlation based PCA? Compare the loadings of the different approaches. Plot the components of the score matrix pairwise.

Problem 2: Estimators of scatter

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 data. Estimate the MCD and the regular sample covariance for the simulated data in (a) - (c). Compare the estimates.

Homework Assignment 4: Influence functions and breakdown points

Provide the requested proofs and figures in your report.

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

b) Simulate a sample of your choice. Use the sample to plot the empirical influence function of the median. When plotting, make sure that the limits of the x-axis are chosen in such a way that it covers the values in your original sample.

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