Statistics and Numerics

Lecture Prof. Dr. Jens Timmer
Exercises Helge Hass, Mirjam Fehling-Kaschek

Exercise Sheet Nr. 8

Exercise 1: Robust linear regression

Consider the model

$$y_i = a + bx_i + \varepsilon_i, \quad i \in \{1, \ldots, N\} \tag{1}$$

for $x_i \sim N(0,1)$, $a = 0$, $b = 1$, $N = 1000$ and Laplace-distributed random noise ε_i, $p(\varepsilon_i) = \frac{1}{2}e^{-|\varepsilon_i|}$.

a) Generate $M = 200$ realizations of simulated data with $N = 1000$ data points each. *Hint:* You can generate the Laplace distributed noise via $\varepsilon_i \sim -\log(U(0,1)/U(0,1))$.

b) Estimate the parameters a and b using the least squares estimator (LSE): write a function `LSE_ab(x,y)` that returns a_{est} and b_{est} for a given realization $\{x_i, y_i\}_{i=1,\ldots,N}$:

$$b_{est} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$a_{est} = \bar{y} - \hat{\beta}\,\bar{x},$$

with $\bar{x}$ and $\bar{y}$ as averages of x_i and y_i, respectively.

c) Estimate the parameters a and b again, this time using the maximum likelihood estimator (MLE) for double-exponential distributed noise: implement the MLE function `MLE(x,y)`. See end of chapter 10.1 of the script for an iterative procedure via bisection root search.

d) Compare the results of both estimators: plot histograms of the estimated parameters for each of the M realisations and compute their mean values and variances. Which estimator is performing better?

e) Compute the efficiency of the least squares estimator, $eff = \frac{Var(\Theta_{MLE})}{Var(\Theta_{LSE})}$. Analyze the dependency of the variances and efficiency on N.

f) Repeat the exercise by simulating data with normally distributed noise ($\varepsilon_i \sim N(0, \sigma^2)$). Do not change the implementation of the estimators. Explain the different behaviour of the two estimators for normally distributed noise.

g) Repeat the exercise by simulating data with Cauchy distributed noise. Can you explain why the MLE estimator of (c) still yields good results even though it is based on exponential noise? And how reliable is the LSE estimator now?