Computational Statistics 732A90 – Fall 2023
Computer Lab 1

Krzysztof Bartoszek, Bayu Brahmantio, Frank Miller, Héctor Rodriguez Déniz
Department of Computer and Information Science (IDA), Linköpings University

October 31, 2023

This computer laboratory is part of the examination for the Computational Statistics course. Create a group report, (that is directly presentable, if you are a presenting group), on the solutions to the lab as a .PDF file. Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments.

All R code should be included as an appendix into your report.

A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes. In the report reference all consulted sources and disclose all collaborations.

The report should be handed in via LISAM (or alternatively in case of problems by email), by **23:59 7 November 2023** at latest. Notice there is a deadline for corrections **23:59 21 January 2024** and a final deadline of **23:59 11 February 2024** after which no submissions nor corrections will be considered and you will have to redo the missing labs next year. The seminar for this lab will take place **22 November 2023**.

The report has to be written in English.

Question 1: Maximization of a likelihood function in one variable

We have independent data $x_1, \ldots, x_n$ from a Cauchy-distribution with unknown location parameter θ and known scale parameter 1. The log likelihood function is

$$-n\log(\pi) - \sum_{i=1}^{n} \log(1 + (x_i - \theta)^2),$$

and it's derivative with respect to θ is

$$\sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2}.$$

Data of size $n = 5$ is given: $x = (-2.8, 3.4, 1.2, -0.3, -2.6)$.

a. Plot the log likelihood function for the given data in the range from -4 to 4. Plot the derivative in the same range and check visually how often the derivative is equal to 0.

b. Choose **one** of the following methods: bisection, secant, or Newton-Raphson. Write your own code to optimize with the chosen method. If you have chosen Newton-Raphson, describe how you derived the second derivative.

c. Choose suitable starting values based on your plots to identify all local maxima of the likelihood function. Note that you need pairs of interval boundaries for the bisection or pairs of starting values for secant. Are there any (pairs of) starting values which do not lead to a local maximum? Decide which is the global maximum based on programming results.

d. Assume now that you are in a situation where you cannot choose starting values based on a plot since the program should run automatised. How could you modify your program to identify the global maximum even in the presence of other local optima?

Question 2: Computer arithmetics (variance)

A known formula for estimating the variance based on a vector of n observations is

$$\mathrm{Var}(\vec{x}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

a. Write your own R function, `myvar`, to estimate the variance in this way.

b. Generate a vector $x = (x_1, \ldots, x_{10000})$ with 10000 random numbers with mean 10^8 and variance 1.

c. For each subset $X_i = \{x_1, \ldots, x_i\}$, $i = 1, \ldots, 10000$ compute the difference $Y_i =$ `myvar`$(X_i) -$ `var`(X_i), where `var`(X_i) is the standard variance estimation function in R. Plot the dependence Y_i on i. Draw conclusions from this plot. How well does your function work? Can you explain the behaviour?

d. How can you better implement a variance estimator? Find and implement a formula that will give the same results as `var()`?