LINKÖPING UNIVERSITY

Computational Statistics
Exam, Spring 2025

Department of Computer and Information Science (IDA), Linköping University

March 24, 2025, 8:00-13:00

Course:	732A89, 732A90, 732A72 Computational Statistics
Teacher and examiner:	Frank Miller
Allowed aids:	Printed course books, 1 handwritten page (A4, front page, only) with notes
Provided aids:	Helpfiles (lecture slides and some chapters from Givens and Hoeting)
Grades:	A = [36, 40] points, B = [32, 36) points, C = [28, 32) points, D = [24, 28) points, E = [20, 24) points, F = [0, 20) points.
Instructions:	Provide a detailed report that includes plots, conclusions and interpretations. If you are unable to include a plot in your solution file clearly indicate the section of R code that generates it. Give motivated answers to the questions. If an answer is not motivated, the points are reduced. Provide all necessary codes in an appendix. In a number of questions, you are asked to do plots. Make sure that they are informative, have correctly labelled axes, informative axes limits and are correctly described. Points may be deducted for poorly done graphs. If you have problems with creating a pdf, you may submit your solutions in text files with unambiguous references to graphics and code that are saved in separate files. There are THREE assignments (with sub–questions) to solve. Provide a separate solution file for each assignment. Include all R code that was used to obtain your answers in your solution files. Make sure it is clear which code section corresponds to which question. If you also need to provide some hand–written derivations, please note the number of the question on each page. Name your solution files as: [your exam account id] [own file description].[format]

Note: If you are not able to solve a part of a question, you can anyway try to show how you would solve the subsequent parts with explaining and providing code examples and you might receive partial points.

1 Optimization (12 points)

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. Program **one** of the following methods: bisection, secant, or Newton-Raphson. Choose suitable starting values (based on your plots) to identify all local maxima of the likelihood function. Decide which is the global maximum.

c. Mention differences (in general) between bisection, secant, and Newton-Raphson methods with regard to convergence speed, sensitivity to starting values, requirements of available derivatives, and necessary programming effort.

2 Rejection sampling (14 points)

Consider the following density:

$$f(x) = \begin{cases} 0, & \text{if } x < -3 \text{ or } x > 3, \\ (3 + x)/8, & \text{if } -3 \leq x \leq -1, \\ 1/4, & \text{if } -1 < x \leq 1, \\ (3 - x)/8, & \text{if } 1 < x \leq 3. \end{cases}$$

We are interested to generate draws of a random variable X with this density using rejection sampling.

a. Choose an appropriate envelope $e_1(x)$ based on a uniform distribution and an appropriate envelope $e_2(x)$ based on a normal distribution for the density.

b. Write a function in `R` for the density $f(x)$ and for the two envelopes $e_1(x)$ and $e_2(x)$ and plot the three functions over the range $-4 \leq x \leq 4$. Which of the two envelopes is better for rejection sampling and why?

c. Program a random generator for X using rejection sampling for the envelope e_i chosen in b. Generate 10000 random variables and plot a histogram. Provide an estimate for the standard deviation of X.

3 Bootstrap and permutation test for regression (14 points)

The dataset `tempLink.csv` contains the yearly average temperatures measured at the station Linköping-Malmslätt from 1961 to 2024 in degrees Celcius (64-data points; data source from SMHI, yearly averages calculated as weighted means from monthly averages). The first value in each row is the year and the second value after a comma is the average temperature. Alternatively, the object `data` in the file `tempLink.Rdata` contains the same data.

You are supposed to fit a linear regression for temperature y_i as dependent variable and year x_i as independent variable, $y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$, assuming independent errors $\epsilon_i, i = 1, \ldots, 64$.

a. Read in the dataset, fit a linear regression and estimate the slope β_1 of the regression line together with a 95%-confidence interval using the `R`-function `lm`. Create a plot for average temperature vs. year and add the estimated regression line to the plot.

b. Derive a 95%-bootstrap confidence interval for the slope based on the percentile method. Do not use a bootstrap package for this calculation; program the bootstrap on your own. Use 5000 bootstrap replicates. Plot a histogram with the bootstrap distribution. Check if the confidence interval here and in part a. are similar or not and comment on it. What is the interpretation of these results?

c. Suppose that you should test the null hypothesis H_0: "the slope $\beta_1 = 0$" versus the alternative H_1: "the slope $\beta_1 > 0$" with a permutation test. Generate a null distribution and argue whether the p-value of the permutation test is < 0.001 or not.