LINKÖPING UNIVERSITY

Computational Statistics 732A89 – Spring 2025
Computer Lab 6

Martin Andrae, Bayu Brahmantio, Frank Miller
Department of Computer and Information Science (IDA), Linköpings University

February 25, 2025

This computer laboratory is part of the examination for the Computational Statistics course. Create a group report (which 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 to your report.
A typical lab report should contain 2-4 pages of text plus some figures plus an appendix with codes. In the report, refer to 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 March 5, 2025** at the latest. Notice that there is a deadline for corrections 23:59 08 April 2025 and a final deadline of 23:59 29 April 2025 after which no submissions or corrections will be considered, and you will have to redo the missing labs next year. The seminar for this lab will take place **March 11, 2025**.
The report has to be written in English.

Question 1: EM algorithm

In the lecture, an EM algorithm was presented for the case of a univariate normal mixture model with two components; you can find it also in the file `emalg.r` on the course homepage.

a. Use the algorithm from the lecture as start and modify it for the case of three components, i.e. the mixture of three normal distributions. Important: Use this provided algorithm to start with and generalize it; do not write completely new code.

b. Modify the stopping criterion such that the stopping does not depend on a scaling of data. This means that if data in another unit is analyzed, for example if numbers are multiplied or divided by 1000, the results after stopping should be the same.

c. Use the data `dat3p` in `threepops.Rdata` on the course homepage which contains $n = 230$ observations. Create first a histogram of the data. Fit then a normal mixture model with three components using your program from the a. and b. parts. Which estimates do you get for the model parameters?

d. Provide plots of current estimates for each model parameter versus the iteration-number. Do they support that convergence has been achieved for each parameter?

Question 2: Simulated annealing

The dataset `bankdata` in `bankdata.Rdata` contains a part of a register of clients who should be contacted to offer a new term deposit. We have here 4364 clients and focus on only two features, the clients' age in years and the logarithm of their account balance (and then multiplied by 6 such that both variables have a similar range, i.e., you would get the original account balance by exp(`balance`/6), but you do not need it for this question, use the two features as provided).

The bank wants to investigate how the willingness to agree to the bank's offer depends on the client's age and their account balance. To obtain a first idea about the dependence, 22 clients of these 4364 should be contacted. The bank wants to have a good spread in age and balance in the subsample. Therefore, the bank wants to select the 22 clients in the following way:

For each client, we compute the minimal distance to one of the 22 selected clients, where we use the Euclidean distance in the two-dimensional feature space. The sum of these minimal distances for all clients should be minimized. A function `crit` is available in the file `bankcrit.r` on the course homepage which calculates this sum of minimal distances and which you can use.

a. Plot the data, select randomly 22 clients and mark them in the plot (this should be your starting subsample for the algorithm in b.).

b. Program your own simulated annealing algorithm to minimize the criterion function. Think about how to select candidates based on the current subset: To ensure that you always have subsets with 22 clients, you might remove and add equally many clients to create the next candidate subset (this means, think about exchanging individuals).

c. Investigate several cooling schedules, starting temperatures, and number of iterations for the simulated annealing algorithm. Discuss at least two combinations of schedules+temperatures+iterations in your report (for example, a good and a bad choice). Compare plots of the criterion-value versus the iteration number. Show also plots of the data with the clients in the final subset marked.