Computer Lab 6
Computational Statistics

Linköpings Universitet, IDA, Statistik

2020/02/13

Kurskod och namn:	732A90 Computational Statistics
Datum:	2020/02/10–2020/03/06 (lab session 28 February 2020)
Delmomentsansvarig:	Krzysztof Bartoszek, Hao Chi Kiang
Instruktioner:	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 e–mailed to krzysztof.bartoszek@liu.se or hao.chi.kiang@liu.se), by **23:59 6 March 2020** at latest. Notice there is a final deadline of **23:59 5 April 2020** 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 **11 March 2020**. The report has to be written in English.

Question 1: Genetic algorithm

In this assignment, you will try to perform one-dimensional maximization with the help of a genetic algorithm.

1. Define the function

$$f(x) := \frac{x^2}{e^x} - 2\exp(-(9\sin x)/(x^2 + x + 1))$$

2. Define the function `crossover()`: for two scalars x and y it returns their "kid" as $(x+y)/2$.

3. Define the function `mutate()` that for a scalar x returns the result of the integer division $x^2 \mod 30$. (Operation $\mod$ is denoted in `R` as `%%`).

4. Write a function that depends on the parameters `maxiter` and `mutprob` and:

   (a) Plots function f in the range from 0 to 30. Do you see any maximum value?

   (b) Defines an initial population for the genetic algorithm as $X = (0, 5, 10, 15, \ldots, 30)$.

   (c) Computes vector `Values` that contains the function values for each population point.

   (d) Performs `maxiter` iterations where at each iteration

      i. Two indexes are randomly sampled from the current population, they are further used as parents (use `sample()`).

      ii. One index with the smallest objective function is selected from the current population, the point is referred to as victim (use `order()`).

      iii. Parents are used to produce a new kid by crossover. Mutate this kid with probability `mutprob` (use `crossover()`, `mutate()`).

      iv. The victim is replaced by the kid in the population and the vector `Values` is updated.

      v. The current maximal value of the objective function is saved.

   (e) Add the final observations to the current plot in another colour.

5. Run your code with different combinations of `maxiter`$= 10, 100$ and `mutprob`$= 0.1, 0.5, 0.9$. Observe the initial population and final population. Conclusions?

Question 2: EM algorithm

The data file `physical.csv` describes a behavior of two related physical processes $Y = Y(X)$ and $Z = Z(X)$.

1. Make a time series plot describing dependence of Z and Y versus X. Does it seem that two processes are related to each other? What can you say about the variation of the response values with respect to X?

2. Note that there are some missing values of Z in the data which implies problems in estimating models by maximum likelihood. Use the following model

$$Y_i \sim \exp\left(X_i/\lambda\right), \qquad Z_i \sim \exp\left(X_i/(2\lambda)\right)$$

   where λ is some unknown parameter.
   The goal is to derive an EM algorithm that estimates λ.

3. Implement this algorithm in `R`, use $\lambda_0 = 100$ and convergence criterion "stop if the change in λ is less than 0.001". What is the optimal λ and how many iterations were required to compute it?

4. Plot $\mathrm{E}\left[Y\right]$ and $\mathrm{E}\left[Z\right]$ versus X in the same plot as Y and Z versus X. Comment whether the computed λ seems to be reasonable.