Computer Lab 4
Computational Statistics

Linköpings Universitet, IDA, Statistik

2022 XI 30

Kurskod och namn:	732A90 Computational Statistics
Datum:	2022 XI 28—2022 XII 06 (lab session 30 XI 2022)
Delmomentsansvarig:	Krzysztof Bartoszek, Bayu Brahmantio, Jaskirat Marar, Shashi Nagarajan
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 bayu_dot_brahmantio_at sign_liu_dot_se, jasma356_at sign_student_dot_liu_dot_se, shashi_dot_nagarajan_at sign_@liu_dot_se, or krzysztof_dot_bartoszek_at sign_liu_dot_se), by **23:59 6 December 2022** at latest. Notice there is a deadline for corrections **23:59 22 January 2023** and a final deadline of **23:59 12 February 2023** 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 **6 December 2022**. The report has to be written in English.

Question 1: Computations with Metropolis–Hastings

Consider the following probability density function:

$$f(x) \propto x^5 e^{-x}, \quad x > 0.$$

You can see that the distribution is known up to some constant of proportionality. If you are interested (**NOT** part of the Lab) this constant can be found by applying integration by parts multiple times and equals 120.

1. Use Metropolis–Hastings algorithm to generate samples from this distribution by using proposal distribution as log–normal $LN(X_t, 1)$, take some starting point. Plot the chain you obtained as a time series plot. What can you guess about the convergence of the chain? If there is a burn–in period, what can be the size of this period?

2. . Perform Step 1 by using the chi–square distribution $\chi^2(\lfloor X_t + 1 \rfloor)$ as a proposal distribution, where $\lfloor x \rfloor$ is the floor function, meaning the integer part of x for positive x, i.e. $\lfloor 2.95 \rfloor = 2$

3. Compare the results of Steps 1 and 2 and make conclusions.

4. Generate 10 MCMC sequences using the generator from Step 2 and starting points $1, 2, \ldots,$ or 10. Use the Gelman–Rubin method to analyze convergence of these sequences.

5. Estimate
$$\int_0^\infty x f(x) \mathrm{d}x$$
using the samples from Steps 1 and 2.

6. The distribution generated is in fact a gamma distribution. Look in the literature and define the actual value of the integral. Compare it with the one you obtained.

Question 2: Gibbs sampling

A concentration of a certain chemical was measured in a water sample, and the result was stored in the data `chemical.RData` having the following variables:

- `X`: day of the measurement
- `Y`: measured concentration of the chemical.

The instrument used to measure the concentration had certain accuracy; this is why the measurements can be treated as noisy. Your purpose is to restore the expected concentration values.

1. Import the data to `R` and plot the dependence of `Y` on `X`. What kind of model is reasonable to use here?

2. A researcher has decided to use the following (random–walk) Bayesian model (n=number of observations, $\vec{\mu} = (\mu_1, \dots, \mu_n)$ are unknown parameters):

$$Y_i \sim \mathcal{N}(\mu_i, \text{ variance} = 0.2), \quad i = 1, \dots, n$$

where the prior is

$$p(\mu_1) = 1$$
$$p(\mu_{i+1}|\mu_i) = \mathcal{N}(\mu_i, 0.2), i = 1, \dots, n1 \quad .$$

Present the formulae showing the likelihood $p(\vec{Y}|\vec{\mu})$ and the prior $p(\vec{\mu})$. **Hint**: a chain rule can be used here $p(\vec{\mu}) = p(\mu_1)p(\mu_2|\mu_1)p(\mu_3|\mu_2)\dots p(\mu_n|\mu_{n1})$.

3. Use Bayes' Theorem to get the posterior up to a constant proportionality, and then find out the distributions of $(\mu_i|\vec{\mu}_{-i}, \vec{Y})$, where $\vec{\mu}_{-i}$ is a vector containing all μ values except of μ_i.

Hint A: consider for separate formulae for $(\mu_1|\vec{\mu}_{-1}, \vec{Y})$, $(\mu_n|\vec{\mu}_{-n}, \vec{Y})$ and then a formula for all remaining $(\mu_i|\vec{\mu}_{-i}, \vec{Y})$.

Hint B:

$$\exp\left(-\frac{1}{d}\left((x-a)^2 + (x-b)^2\right)\right) \propto \exp\left(-\frac{(x-(a+b)/2)^2}{d/2}\right)$$

Hint C:

$$\exp\left(-\frac{1}{d}\left((x-a)^2 + (x-b)^2 + (x-c)^2\right)\right) \propto \exp\left(-\frac{(x-(a+b+c)/3)^2}{d/3}\right)$$

4. Use the distributions derived in Step 3 to implement a Gibbs sampler that uses $\vec{\mu}^0 = (0, \dots, 0)$ as a starting point. Run the Gibbs sampler to obtain 1000 values of $\vec{\mu}$ and then compute the expected value of $\vec{\mu}$ by using a Monte Carlo approach. Plot the expected value of $\vec{\mu}$ versus X and Y versus X in the same graph. Does it seem that you have managed to remove the noise? Does it seem that the expected value of $\vec{\mu}$ can catch the true underlying dependence between Y and X?

5. Make a trace plot for μ_n and comment on the burn–in period and convergence.