Computer Lab 4
Computational Statistics

Linköpings Universitet, IDA, Statistik

November 19, 2021

Course code and name:	732A90 Computational Statistics
Lab session:	22.11, 8-10
Submission deadline:	26.11, 23:59
Resubmission deadlines:	resubmission 1: 10.12, 23:59; resubmission 2 for labs 1-4: 14.1
Seminar:	Seminar 2 (second part) on 14.12
Teachers:	Maryna Prus, Filip Ekström, Joel Oskarsson, Martynas Lukosevicius, Shashi Nagarajan, Yifan Ding
Instructions:	This computer laboratory is a part of the examination Create a group report (in English) on the solutions to the lab as a .PDF file. All R codes should be included as an appendix into your report. 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 your teacher - see file ”lab groups” on lisam).

Exercises originally developed by Krzysztof Bartoszek

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?

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. $\lfloor 2.95 \rfloor = 2$. What can you guess about the convergence of the chain?

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

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

5. The distribution generated is in fact a gamma distribution ($\Gamma(6, 1)$). Determine the actual value of the integral. Compare it with the one you obtained in the previous step.

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

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

   where the prior is

   $$p(\mu_1) = 1$$
   $$p(\mu_{i+1}|\mu_i) = \mathrm{N}(\mu_i, 0.2),\ i = 1, \ldots, n-1.$$

   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 $p(\vec{\mu}) = p(\mu_1)p(\mu_2|\mu_1)p(\mu_3|\mu_2)\ldots p(\mu_n|\mu_{n-1})$.

2. Use Bayes' Theorem to get the posterior up to a constant proportionality (since the integral is constant you don't need to compute it), 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:

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

In Hint B and Hint C "$\propto$" denotes proportionality.

3. Use the distributions derived in Step 2 to implement a Gibbs sampler that uses $\vec{\mu}^0 = (0, \ldots, 0)$ 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?

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