Compulsory exercise for STK4051/9051 - Computational statistics

Spring 2024

Part 2 (of 2)

This is the second part of the compulsory exercise for STK4051/9051, spring semester 2024. The deadline for the complete compulsory exercise (including part 1 and 2) is May 2nd . Note that even if you have delivered parts of the exercise for feedback earlier, you need to include both parts in the May 2nd delivery. The exercise has to be delivered within the Canvas system. Reports may be written in Norwegian or English, and should be text processed (LaTeX, Word). Write concisely. Relevant figures need to be included in the report. Copies of relevant parts of machine programs used (in R, python, or similar) are also to be included, perhaps as an appendix to the report. Within these exercises there are some choices you need to make when designing algorithms. Part of the evaluation will be on your choices, but more importantly are your arguments on why you have made the specific choices!

This second part contains three exercises and comprises five pages (including this front page). Some R-code is available from the course web-page. You are free to use other software but would then need to translate or write your own code for that part included in the R-script. Data sets to be used are available on the course web-page, in a standard R save file. Read the corresponding .txt file to understand the structure of data.

TGsim.dat	Exercise 2
gambia.dat	Exercise 3

There will be a Q and A, with respect to the compulsory exercise on the course webpage. The page is updated when questions arise.

In addition, write a short note about how much time you have spent on the exercise and how hard/difficult you consider the exercise to be. Comment as detailed as you like. Are there parts that you do not think has been covered sufficiently well in class? Include both part 1 and 2 in your evaluation.

Exercise 1 (Simulation from 1D distribution) We will start this exercise by investigating some methods for sampling a univariate distribution. We will assume that we easily obtain samples from a uniform distribution, i.e. $U \sim \text{Unif}[0, 1]$.

a) Describe the standard transformation rule to sample from a univariate distribution, and derive the expression for sampling from an exponential distribution.

When the negative log density is convex we can use adaptive rejection sampling to build an approximation to the density. We will now use this to sample from a standard normal distribution. We shall use the sub-gradient of the convex function, i.e. for a continuously differentiable convex function $k(x)$ we have:

$$k(x) \geq k(x_0) + k'(x_0)(x - x_0) \tag{1}$$

b) Describe how we can use the convex property of the negative log density to find an upper bound on the distribution. Use the linearization around the points -1, 0, and 1, to derive a bounding function for the standard normal. Illustrate the bounding function. Does the bounding function integrate to 1? Show that the probability distribution corresponding to the bounding function is:

$$g(x) = \begin{cases} \frac{1}{3} & \text{if} \quad -0.5 < x \leq 0.5 \\ \frac{\exp(-|x|+0.5)}{3} & \text{else} \end{cases} \tag{2}$$

c) Algorithm 1 below uses rejection sampling to sample the standard normal distribution. Show the relation to the distribution in (2), compute the average acceptance rate, and implement the algorithm.

Algorithm 1:

```
While not accepted
    i.   Sample $U_1, U_2$ iid from Unif[0,1]
    ii.  Sample $i \in \{1,2,3\}$, with probability $p_0 = \left\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right\}$
    iii. $x_p = \begin{cases} U_1 - 0.5, & \text{if } i = 1 \\ \ln(1 - U_1) - 0.5, & \text{if } i = 2 \\ -\ln(1 - U_1) + 0.5, & \text{if } i = 3 \end{cases}$
    iv.  If $U_2 < \frac{\sqrt{2\pi}\cdot\phi(x)}{3\cdot g(x)}$, accept proposal and return $x_p$.
```

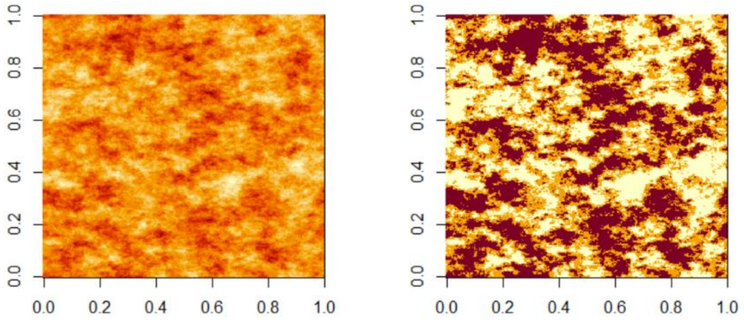
where $\phi(\cdot)$ is the density for a standard normal variable.

d) An alternative to rejection sampling is importance sampling. Implement an importance sampler based on $g(x)$. Give arguments for choices you make when setting up the importance sampler.

e) Discuss the differences between rejection sampling and importance sampling. Compare the two approaches in a numerical study where you quantify the Monte Carlo variance in the results. For each rerun of the estimation should be based on 1000 samples from $g(x)$. In the numerical study you should estimate $E(h(x))$, where:

$$h(x) = x^2 I(x > 0) \tag{4}$$

Exercise 2 (Sequential Monte Carlo) In spatial statistics there are many different approaches for modelling continuity of discrete properties in a region. One method which is popular due to the intuitive interpretation is the truncated Gaussian model. In this model one generates a Gaussian random field in $\mathbb{R}^d$ and divide the value set into intervals to define the classes. When applying the model, it is challenging to get the parameters of the model right. One way to get the parameters correct is to estimate them from training data. In this case we have a map of discrete values and want to derive the properties of the underlying Gaussian random field. This is the situation we will investigate below. We will however restrict focus to 1D in this exercise.



The model is given as:

$$\begin{aligned} &x_1 \sim N(0,1) \\ &x_t = a x_{t-1} + \lambda + \varepsilon_t, t = 2,3,\dots \\ &\varepsilon_t \sim N(0, \sigma^2), \qquad t = 2,3,\dots \\ &y_t = \begin{cases} 1 & \text{if} \quad x_t < -0.5 \\ 2 & \text{if} \; -0.5 \leq x_t < 0.5 \\ 3 & \text{if} \quad x_t \geq 0.5 \end{cases} \quad t = 1,2,3,\dots \end{aligned} \tag{5}$$

The data $\boldsymbol{y}_{1:251}$, is given in the file TGsim.dat on the course webpage.

a) Design a sequential Monte Carlo algorithm for inference about $p(x_t|\boldsymbol{y}_{1:t}), t = 1,\dots,n$, assuming that $(a, \lambda, \sigma^2) = (0.85, 0, 0.5^2)$. Display plot of $\mathrm{E}(x_t|\boldsymbol{y}_{1:t})$, and $\mathrm{Var}(x_t|\boldsymbol{y}_{1:t})$. Give arguments for the choices you make in the construction.

We will use sequential Monte Carlo to perform online learning algorithm for the parameter a. For simplicity we will assume that $(\lambda, \sigma^2) = (0, 0.5^2)$, is fixed and known.

b) Consider a prior model for a which is uniform on the interval $[0, 1]$. Design and implement a sequential Monte Carlo algorithm for inference about a, i.e. estimate $p(a|\boldsymbol{y}_{1:251})$. Use a static approach, i.e. sample a initially from the prior distribution, and update it by weighting/resampling throughout the simulation. Why is this approach in general not recommended? Comment on your results, are these acceptable?

c) STK 9051 only. To overcome the weaknesses with the estimation procedure in b, we can utilize that there are sufficient statistics for the parameter a. Describe how this can be utilized and how it mitigates some of the problems with the method used in b. Implement this method and compare the results to those from b, you can choose your prior distribution for a as it suits you, but give an argument for your choice. (Hint:You can modify the script "smc_lin_bin_parest_suff.r" available on the course webpage)

Exercise 3 (McMC – in Bayesian analysis) We will in this exercise consider a Bayesian generalized linear model for identifying influential factors for presence vs absence of malaria in blood samples taken from children in Gambia. The data set "gambia" from the geoR package is available on the course page. A description of the data is given in an associated file. These data are often used for spatial analysis, but we will not consider that aspect here. We will assume all observations to be independent and investigate a probit-link between the explanatory variables $\boldsymbol{x}$ and a binary response variable y. For person i the probit-link between explanatory variables and the response is defined as:

$$P(y_i = 1) = \Phi(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i), \tag{6}$$

where Φ is the cumulative distribution for a standard normal variable. In the Bayesian analysis below we will use the improper prior $p(\boldsymbol{\beta}) \propto 1$, in which case the posterior $p(\boldsymbol{\beta}|\boldsymbol{y})$ is proportional to the likelihood $L(\boldsymbol{\beta}|\boldsymbol{y})$.

a) Define $p_i = \Phi(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)$, and derive the likelihood function:

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{7}$$

Discuss how you can implement a numerically robust evaluation of this likelihood. How should you handle the situation where $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i$ has a large absolute value? In a Metropolis Hastings algorithm, you are asked to evaluate the ratio of two likelihoods, how is this done in a numerically stable way?

b) Below you will be asked to implement a random walk using a random scan Metropolis Hastings algorithm to investigate the posterior distribution. From the general expression for the Metropolis Hastings (M-H) ratio, deduce the M-H ratio for the random walk. Which criteria need to be met in order for the Markov chain to converge to the stationary distribution? Which of these criteria does the M-H ratio help you to fulfill?

c) For the Gambia data: Implement a random walk algorithm with a random scan to sample from the posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y})$, and apply it. Use a model containing the explanatory variables: age, netuse, treated, green, and phc, in addition to the constant term. (Hint: 1-Remember to standardize the design matrix, 2- put some effort into a robust evaluation of the likelihood ratio, see a)

d) Display plots which illustrate the convergence properties of the method. Comment on the convergence properties of your algorithm. If there are any obvious problems, suggest modifications and revisit c to improve convergence.

A common way to sample the distribution above is to introduce a latent variable, z_i, which is defined such that $\{y_i = 1\} \Leftrightarrow \{z_i > 0\}$, and

$$z_i \sim N(\boldsymbol{\beta}^T \boldsymbol{x}_i, 1^2) \tag{8}$$

e) Argue that the expanded posterior probability distribution:

$$p(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{y}) \propto \prod_{i=1}^{n} [I(z_i > 0) \cdot I(y_i = 1) + I(z_i \le 0) I(y_i = 0)] \phi(z_i - \boldsymbol{\beta}^T \boldsymbol{x}_i) \tag{9}$$

will have the prescribed marginal posterior, $p(\boldsymbol{\beta}|\boldsymbol{y})$ from expression (7). (Hint: consider one data point first and integrate out the latent variable.)

f) You shall now derive the conditional distributions needed for Gibbs sampling. Show that:

$$p(z_i|\boldsymbol{z}_{-i}, \boldsymbol{\beta}, \boldsymbol{y}) \propto \begin{cases} I(z_i \le 0)\phi(z_i - \boldsymbol{\beta}^T \boldsymbol{x}_i), & \text{if } y_i = 0 \\ I(z_i > 0)\phi(z_i - \boldsymbol{\beta}^T \boldsymbol{x}_i), & \text{if } y_i = 1 \end{cases} \tag{10}$$

and

$$p(\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{y}) \propto \exp\left(-0.5 \sum_{i=1}^{n} (z_i - \boldsymbol{\beta}^T \boldsymbol{x}_i)^2\right) \tag{11}$$

Use results from multi linear regression to deduce that the distribution $p(\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{y})$, is multi-normal with parameters:

$$E(\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{y}) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{z} \tag{12}$$

and

$$\mathrm{Cov}(\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{y}) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \tag{13}$$

State the result from multivariate linear regression you are using and show how you use the connection to derive the results.

g) Implement the Gibbs sampler, and use this to solve the inference of Gambia data. Use a block update to sample $p(\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{y})$. Make illustrations of the convergence of this chain as well (as done for the random walk approach in in d). (Hint: to sample from the multi-normal distribution use code from a library, e.g. `rmvnorm` from `mvtnorm`)

h) Compare the two approaches you have implemented for evaluating the posterior. Compare implementation, runtime effects, and results e.g. burn in, convergence, mixing etc.