

MCMC2: Lab Session 1: Simulation

Theo Kypraios and Phil O'Neill

Background

Preliminaries

In the first lecture we introduced the concept of simulation by which we meant the procedure of *producing a realisation of the model*, or in other words a possible outcome. In this lab session we will be modifying existing R functions and by the end of this session, among other things, we will

- draw samples from the distributions of the *final size* and the *duration* of the epidemic (see below for definitions of these quantities);
- be able to simulate from an epidemic model where the infectious period follows a Weibull distribution;
- be able to simulate from a stochastic Susceptible-Exposed-Infective-Removed (SEIR) epidemic model.

Start by downloading the file `simulation.R` from <http://www.maths.nott.ac.uk/personal/tk/files/MCMC2-Seattle/Lab-Sessions/1/> and save it in your workspace. This file `simulation.R` contains four different functions which will enable us to simulate realisations from various stochastic epidemic models:

- `simSIR.Markov`: This function produces realisations from a Markovian SIR model with infection and removal rate, β/N and γ respectively. The procedure (as described in Section 2 in the lecture) is the following: i) first simulate the *time to the next event* and then ii) decide the *type of the event* (infection or removal).
- `simSIR.Markov.alternative`: This function also produces realisations from a Markovian SIR model with infection and removal rate, β/N and γ respectively but with a slightly different algorithm to the one described above. The procedure here is as follows: i) generate a possible time to the next infection and the possible time to the next removal; then ii) the event which happens first determines the type of the next event.
- `simSIR.Non.Markov.constant` and `simSIR.Non.Markov.gamma`: These two functions allows us to simulate realisations from non-Markov SIR models; with constant and Gamma infectious period respectively. The procedure is similar to the one used in `simSIR.Markov.alternative` based on Section 5 in Lecture 1.

Definitions

In this section we give definitions of two quantities for which samples from their distribution will be drawn.

The threshold parameter R_0

An appealing feature of an epidemic model is that it embodies a threshold parameter which can be utilised as a severity measure of the outbreak. Stochastic models such as epidemics and branching processes typically generate bimodal realisations where an epidemic may or may not die out quickly, depending on the value of a threshold parameter R_0 often referred to as the basic reproduction number.

This is the most important parameter in epidemic theory and it is defined as *the expected number of infections generated by a typical infective in an infinite susceptible population*.

In the case of a homogeneously mixing stochastic epidemic, it holds that

$$R_0 = \beta \mathbb{E}[I]$$

where $\mathbb{E}[I]$ denotes the expected infectious period. For instance, in the case of the Markovian SIR model,

$$R_0 = \frac{\beta}{\gamma},$$

since the individuals remain infectious for an average period of length $1/\gamma$.

The final size of an epidemic

The final size of an epidemic is the total number of initial susceptibles who contracted the disease by the end of the outbreak.

The duration of the epidemic

Another quantity of interest is the duration of the epidemic and this is defined as the *time that elapsed from the initial infection until the time of the last removal*.

Exercises

Download the file `simulation.R` from <http://www.maths.nott.ac.uk/personal/tk/files/MCMC2-Seattle/Lab-Sessions/1/> and load all the functions into R by using the command `source`:

```
source("simulation.R")
```

Exercise 1

Go through the lines of each function and make sure that you follow the logic behind.

Exercise 2

Simulate realisations from a Markovian SIR using the function `simSIR.Markov` and make sure that you understand the output. You may assume that size of the population size is $N = 21$ (i.e. 20 susceptibles and 1 initial infective). In addition, you could try different values for (β, γ) , e.g. $(0.9, 1)$, $(2, 1)$ and $(4, 1)$.

Exercise 3

Modify the existing functions in `simulation.R` to record the *final size* and the *duration* of the epidemic as part of the functions' output.

Exercise 4

Derive a simulation-based estimate of the distribution of the *final size* of a Markovian SIR model for different values of R_0 , e.g. $R_0 = 0.9$, $R_0 = 1.5$ and $R_0 = 4$. Furthermore, do the same for the non-Markovian models, e.g. for a constant and a Gamma infectious period. **Hint:** You may find it useful to write a loop which will iterate the following steps for a number of times:

- a. Simulate a realisation from the epidemic model;
- b. Store the final size

At the end you should have a collection of *final sizes* for which then you can plot a histogram as your estimate of the true distribution of the final size.

Exercise 5

Repeat the above exercise but derive, by simulation, the distribution of the *duration* of the epidemic instead of the *final size*.

Exercise 6

Write a function in R to simulate from a non-Markovian stochastic epidemic model where the infectious period is assumed to follow a Weibull distribution. **Hint:** The probability density function (pdf) of the Weibull distribution is as follows:

$$f(x) = (a/b)(x/b)^{(a-1)} \exp(-(x/b)^a), \quad x > 0, \quad a > 0, \quad b > 0$$

Type

```
?Weibull
```

to find out how to simulate from a Weibull distribution.

Exercise 7

Write a function to simulate from an epidemic model which involves a fixed latent period, i.e. write a function to simulate from a stochastic SEIR model.

MCMC2: Lab Session 2: Coding and Output

Theo Kypraios and Phil O'Neill

Background

In Lecture 2 we discussed how Bayesian inference can be drawn for the parameters of a stochastic epidemic model using Markov Chain Monte Carlo algorithms. In this lab session we will look at these algorithms in more detail.

1. Start by downloading the file `coding.R` from <http://www.maths.nott.ac.uk/personal/tk/files/MCMC2-Seattle/Lab-Sessions/2/> and save it in your workspace. This file contains three functions which are used to calculate the likelihood of the augmented data (infection and removal times). These functions are needed to build an MCMC algorithm later on (see file `mcmc-Markov.R`):

`count.no.inf`: This function counts the number of infective individuals just before any (arbitrary) time t . Note that an individual, labeled as i , is infective just before time t if $(I_i < t < R_i)$ where I_i and R_i denote their infection and removal time respectively. Therefore if we go through each (ever infected) individual and count how many of them satisfy this condition then we have the desired number.

`compute.total.pressure`: This function computes the integral $\int S_t I_t dt$ by making use of the fact that it can be re-written as a double sum as described in Lecture 2.

`compute.log.prod`: This function computes the product which is required for the calculation of the likelihood:

$$\log \left\{ \prod_{j \neq a} I_{i_j^-} \right\}$$

Note that $I_{i_j^-}$ denotes the number of infected individuals just before the j_{th} infection time (i_j^-) and a denotes the label of the initial infective. Note that this function uses the function `count.no.inf`.

2. Download the file `mcmc-Markov.R` from <http://www.maths.nott.ac.uk/personal/tk/files/MCMC2-Seattle/Lab-Sessions/2/> and save it in your workspace.

This file contains one function only, `mcmcSIR.Markov`, which is used to draw posterior samples of the parameters a Markovian SIR model by making use of the functions in the file `coding.R`. In brief, a pseudo-code of this MCMC algorithm is given below:

- i. Initialisation;
- ii. Choose one infection and update it using a Metropolis-Hastings step;
- iii. Update β and γ by drawing from their conditional distributions using a Gibbs step;
- iv. Goto to Step (i);

Note that:

- the infection rate (β) and the removal rate (γ) are assumed to be independent and follow *a priori* Gamma distributions with parameters $(\lambda_\beta, \nu_\beta)$ and $(\lambda_\gamma, \nu_\gamma)$, i.e.

$$\pi(\beta) \propto \beta^{\lambda_\beta - 1} \exp \{-\beta \nu_\beta\}$$

- the infection times are updated using a M-H algorithm and in particular, using an independence sampler, i.e. by proposing a candidate value, I_i^{can} ,

$$R_i - I_i^{\text{can}} \sim \text{Exp}(\gamma)$$

Therefore, the q-ratio in the accept/reject probability in the M-H step is given by

$$\frac{(R_i - I_i^{\text{cur}}) \exp\{-\gamma(R_i - I_i^{\text{cur}})\}}{(R_i - I_i^{\text{can}}) \exp\{-\gamma(R_i - I_i^{\text{can}})\}}$$

where I_i^{cur} denote the current value of I_i .

Exercises

Start by loading the functions in the files `coding.R` and `mcmc-Markov.R` by using the command `source`, i.e.

```
source("coding.R")
source("mcmc-Markov.R")
```

The function `mcmcSIR.Markov` requires as input a data frame (or matrix) of size $N \times 2$ where N is the size of the population (including the initial infective!). The first column should contain the labels of the individuals and the second column should contain their corresponding removal time. In other words, each row represents an individual. Note that if an individual is known to be susceptible at the end of the epidemic, then their removal and infection time are assumed to be ∞ .

Exercise 1

Have a look at the function `mcmc-Markov.R` and **make sure you understand** what is going on.

Exercise 2

Download some simulated data from <http://www.maths.nott.ac.uk/personal/tk/files/MCMC2-Seattle/Lab-Sessions/2/> and read them into R by using the command:

```
data <- read.table("data.txt", header=TRUE)
```

Make sure that the data have been read properly and the format is appropriate such that they can be used as an (input) argument to the function `mcmcSIR.Markov`.

Exercise 3

The purpose of this exercise is to fit a non-Markovian model to the observed data where the infectious period is assumed to be Gamma distributed with parameters (α, γ) , i.e.

$$R_i - I_i \sim \text{Gamma}(\alpha, \gamma)$$

with $E[R_i - I_i] = \alpha/\gamma$. The parameter α is treated as fixed and known and assume that *a priori* β and γ follow (independent) Gamma distributions as described in Section 1 above.

Before writing any R code we need to do write down the density of the posterior distribution we want to draw samples from:

- i. Derive the density of the (joint) posterior distribution of the parameters and the infection times given the removal times up to proportionality,

$$\pi(\beta, \gamma, \mathbf{I}|\mathbf{R}) \propto \pi(\mathbf{I}, \mathbf{R}|\beta, \gamma) \times \pi(\beta) \times \pi(\gamma)$$

- ii. Derive the densities of the full conditional distributions for the parameters and the unobserved infection times up to proportionality, i.e. $\pi(\beta|\gamma, \mathbf{I}, \mathbf{R})$, $\pi(\gamma|\beta, \mathbf{I}, \mathbf{R})$ and $\pi(\mathbf{I}|\mathbf{R}, \beta, \gamma)$. To do this, simply look at the density of joint posterior distribution, and for example, to derive the density of the full conditional distribution of β then only put together the terms that involve β . What you end up with is the full conditional density you are after (up to proportionality).

Exercise 4

Write a function in R which will draw samples from the posterior distribution of interest, $\pi(\beta, \gamma, \mathbf{I}|\mathbf{R})$, using MCMC and by making use of the function in the file `coding.R`.

Hint 1: Your code should iterate the following steps

- i. Choose one infection time and update it using a M-H algorithm;
- ii. Update β ;
- iii. Update γ ;
- iv. Goto Step (i);

Hint 2: Note that you do not have to write this function from scratch but you can modify the existing function `mcmcSIR.Markov`. However, it might be better to create a new file called `mcmcSIR.gamma` and copy-paste the parts which you can use straight away from `mcmcSIR.Markov`.

Exercise 5

Use the function that you have written in Exercise 4 and use it to fit the above SIR model with Gamma distributed infectious periods to the observed data and draw 10,000 samples from the posterior distribution of the parameters using the function you wrote in part 2).

Assume that $\alpha = 2$. In addition, we assume that we have weak prior information for the parameters and therefore we choose $\lambda_\beta = \lambda_\gamma = 1$ and $\nu_\beta = \nu_\gamma = 10^{-3}$.

- Look at the output by plotting the trace plots of the parameters β , γ and the sum of the infection times $\sum_i I_i$.
- Look at the posterior correlation between β and γ by drawing a scatter plot of the samples against axes β and γ . Furthermore, look at the correlation between the (sum of the) infection times and γ .
- How does the mixing of β and γ compare with the mixing of $R_0 = \beta\alpha/\gamma$? Why?
- Draw a histogram of the posterior distribution of R_0

MCMC2: Lab Session 3: SIR-Topics

Theo Kypraios and Phil O'Neill

Background

In Lecture 3 we discussed different topics with regards to the SIR models and in particular we focused on aspects such as what can or cannot be estimated from the data as well as different techniques to improve the mixing of standard MCMC algorithms.

In this lab session we will see in more detail how the inference of the model parameters is affected by the number of removal times which are observed. In addition, we will also implement MCMC algorithms in which we have integrated out the model parameters and investigate if such a strategy improves the mixing of this MCMC algorithm.

Exercises

Exercise 1

Start by downloading from <http://www.maths.nott.ac.uk/personal/tk/files/MCMC2-Seattle/Lab-Sessions/3/> two different datasets:

```
dataset_min.txt
dataset_max.txt
```

The first dataset refers to an outbreak where none of the initially susceptible individuals become infected. The second dataset refers to an outbreak where all the initially susceptible individuals became infected some time during the outbreak.

Exercise 2

Fit an SIR model to these datasets assuming that the infectious period follows a $\text{Gamma}(2, \gamma)$ as we did in “Lab Session 3”. In addition, assume that *a priori*:

$$\beta \sim \text{Gamma}(1, 10^{-3})$$

$$\gamma \sim \text{Gamma}(1, 10^{-3})$$

Draw samples from the posterior distribution of the parameters β and γ . What do you observe? How do your posterior inferences differ from your prior knowledge in both cases? Comment on your results.

What happens if we assume that *a priori* $\beta \sim \text{Exp}(1)$?

Exercise 3

We saw in the lecture that by integrating both parameters out and having this distribution as our target (i.e. the distribution we want to draw samples from), this could lead into a more efficient MCMC algorithm.

- i. Derive an expression for $\pi(\mathbf{I}|\mathbf{R})$ where

$$\pi(\mathbf{I}|\mathbf{R}) \propto \int_{\beta} \int_{\gamma} \pi(\beta, \gamma, \mathbf{I}|\mathbf{R}) \, d\gamma \, d\beta$$

- ii. Write a function in **R** which first draws samples from the above target density $\pi(\mathbf{I}|\mathbf{R})$ using MCMC. Then modify this function in order to derive posterior samples from $\pi(\beta|\mathbf{R})$ and $\pi(\gamma|\mathbf{R})$.
- iii. Fit an SIR model with Gamma infectious period to the same dataset that we used in Lab Session 2 (Questions 4 and 5). Compare the posterior samples with the samples obtained in Lab Session 2. The answers (eg. the posterior distributions) should be the same. Why?
- iv. Does such a strategy improve mixing as compared to the algorithm which does not involve integrating the parameters out? **Hint:** One way to compare the efficiency is to compare the ACF plots of the samples obtained by two different algorithms.

MCMC2: Lab Session 4: Households

Phil O'Neill and Theo Kypraios

Background

In Lecture 4 we were concerned with models for disease transmission which incorporate households. Although we looked at two distinct cases with regards to what sort of data are available, in this lab session we will focus on the case where there is only *final outcome* data available. In other words, the available data consist only of the final number of cases in each household.

The main objective of this lab session is to draw Bayesian inference for the model parameters by implementing a Markov Chain Monte Carlo algorithm. Below, there is a brief description of the model (see the lecture notes for more details).

The model, the data and the likelihood

We assume that a population of N individuals is partitioned into households. Each individual in the population has, independently, a constant “risk” per unit time of becoming infected from the community. In addition, within a household, the disease spreads according to the mechanism of an SIR model.

Suppose that the data consist of the set of numbers $\mathbf{n} = \{n(j, k)\}$ where $n(j, k)$ = number of households in which j out of k initial susceptibles become infected. Therefore, the likelihood takes the form

$$\pi(\mathbf{n}|\alpha, p) = \prod_j \prod_k q(j, k)^{n(j, k)}$$

where $q(j, k)$ is the probability that in a household of size k the number of individuals who ever become infected is j .

In the lecture we have derived an expression for the probability $q(j, k)$ and showed that it depends on two parameters:

- p which is the probability that an individual avoids infection from outside the household and
- α which is the person-to-person infection rate between individuals within the same household.

Preliminaries

Download the file `households.R` from <http://www.maths.nott.ac.uk/personal/tk/files/MCMC2-Seattle/Lab-Sessions/4/>

The file contains two functions:

- `compute.conditional.prob` which computes the probability that j susceptibles become infected in an SIR model with constant infectious period of length c , a initial infectives and m initial susceptibles.

```
# this is a function to compute the final size P(T=k|Y=y) = probability
# that k-y susceptibles become infected with $y$ initial infectives
# and n-y susceptibles

# m initial susceptibles
# a initial infectives
# alpha infection rate
```

```

# the length of the fixed infectious period
compute.conditional.prob <- function(m, j, a, alpha, c){
  if (j==0) {
    res <- (exp(-alpha*m*c))^a
  }
  else {

    part.one <- exp(-alpha*(m-j)*c)^(j+a)
    part.two <- choose(m, j)

    sum <- 0;
    for (k in 0:(j-1)) {
      sum <- sum + choose(m-k, j-k)*compute.conditional.prob(m,k,a,alpha,c)/(exp(-alpha*(m-j)*c)^(k+a))
    }
    res <- part.one*(part.two - sum)
  }
  return(res)
}

```

- compute.marginal.probab which computes the probability that in a household of size n the number of individuals who ever become infected is k .

```

# We wish to compute  $P(T = k)$ , for  $k = 0, \dots, n$ 
# "n" is the size of the household
# "alpha" is the person to person infection rate
# "c" is the length of the (fixed) infectious period
compute.marginal.prob <- function(k, n, c, alpha, p) {
  prob <- 0;
  for (y in 0:n) {
    if (k >= y) {
      cond.prob <- compute.conditional.prob(n - y, k - y, y, alpha, c)
    }
    else {
      cond.prob <- 0
    }
    prob <- prob + cond.prob*dbinom(y, n, 1 - p)
  }
  prob
}

```

In addition, Table 1 presents some Asian influenza epidemic household data taken from Longini & Koopman (1982). There is a community which consists of 42 households each of them of size 3.

```

cases <- c(0,1,2,3)
households <- c(29,9,2,2)
mat <- matrix(NA, nrow=4, 2)
mat[,1] <- cases
mat[,2] <- households
mat <- as.data.frame(mat)
colnames(mat) <- c("No.of.Cases", "No.of.Households")
mat

```

```

##   No.of.Cases No.of.Households
## 1           0                29
## 2           1                 9
## 3           2                 2

```

Exercises

Exercise 1

Go through each of the functions in the file `households.R` and make sure that you understand what is going on.

Exercise 2

Write a function in R which computes the log-likelihood of the data given the parameters p and α assuming that the infectious period is constant and of length one unit.

Exercise 3

Suppose that we are interested in drawing Bayesian inference for the model parameters and therefore, we could employ an MCMC algorithm to draw samples from the posterior distribution:

$$\pi(p, \alpha | \mathbf{n}) \propto \pi(\mathbf{n} | \alpha, p) \times \pi(p) \times \pi(\alpha).$$

Assume that *a priori*

$$p \sim \text{Beta}(\mu_p, \nu_p)$$

and

$$\alpha \sim \text{Gamma}(\mu_\alpha, \nu_\alpha).$$

What values should we choose for the hyper-parameters μ_p, ν_p, μ_α and ν_α such that we end up with fairly uninformative priors for the model parameters p and α ?

Exercise 4

Write a function in R which draws samples from $\pi(p, \alpha | \mathbf{n})$ using MCMC and by adopting uninformative priors for the parameters of interest p and α .

Hint: In the lecture we discussed two different ways to update the parameter p . Either using an independence sampler or a random walk Metropolis. For the purposes of this exercise, update p by an independence sampler.

Exercise 5

Investigate how correlated the posterior samples of α and p are.

Exercise 6

Draw samples from the posterior distribution of the probability that a susceptible individual avoids infection from one infected household member.

Exercise 7

In the current version of your MCMC algorithm the parameter p is updated via an independence sampler. Modify your code such that p is now updated using a random walk Metropolis. Does the mixing of your algorithm improve?