Practical:
Data simulation and parameter estimation from complete data for a recurrent infection

Instructors: Kari Auranen, Elizabeth Halloran, Vladimir Minin
July 17– July 19, 2017

Background

In the following exercises we try out Markov chain Monte Carlo methods in the Bayesian data analysis for recurrent infections. The model of infection is taken to be a binary Markov process, where at any given time the epidemiological state for an individual is either 0 (susceptible) or 1 (infected). This is the simplest stochastic "SIS" model (susceptible-infected-susceptible).

To familiarize ourselves with the computational approaches, using the Metropolis-Hastings algorithm with reversible jumps to augment unobserved events, we consider (statistically) independent individuals, omitting thus questions about transmission. This makes the likelihood computations easier and faster.

The binary Markov process is considered from time 0 to time T, at which the process is censored. The model has three parameters: (λ, μ, π), where λ is the per capita rate (force) of infection, μ is the rate of clearing infection and π is the proportion of those that are infected at time 0.

For N independent individuals, the *complete data* comprise the times $(T_{sr}^{(ik)})$ of all transitions between states 0 and 1 that occur between time 0 and the censoring time T (see lectures). In more realistic situations, however, we could not hope to observe complete data. Instead, the process can usually only be observed at some pre-defined times. To apply the complete data likelihood, unobserved event times and states should be augmented. The computations then rely on the reversible jump Markov chain Monte Carlo methodology. However, this problem falls outside the scope of the current exercise.

Exercises

1. **Simulation of complete (event-history) data.** Download the source code of an R function **simulateSIS_N.R**. Then simulate complete data from the binary Markov model ("susceptible-infected-susceptible"):

```
complete_data = simulateSIS_N(N=100,la=0.45,mu=0.67,initprob=0.40,T=12)
```

   The function samples binary processes for N=100 individuals from time 0 to time T=12 (time units). The transition rates are $\lambda = 0.45$ (force of infection, per time unit per capita) and $\mu = 0.67$ (rate of clearing infection, per time unit per capita). The proportion of those that are infected at time 0 is π=0.40 (initprob). The output is a list of N arguments, each containing the event times (times of transition) and the epidemiological states (after each transition) for one individual.

   These data might describe a 12 month follow-up of acquisition and clearance of nasopharyngeal carriage of pneumococci (a recurrent asymptomatic infection), with mean duration of carriage $1/\mu = 1.5$ months and the stationary prevalence of $\lambda/(\lambda + \mu) = 0.40$.

2. **Estimation of model parameters from completely observed data.** You can realize numerical samples from the joint posterior distribution of the three model parameters (λ, μ, π) with the R function **MH_SIS.R**. This function applies a component-wise Metropolis-Hastings algorithm to update each of the parameters in turn. It uses subroutines **likelihoodSIS.R** (to calculate values of the log-likelihood from the observed event histories) and **update_parameters.R** (to perform the actual updating). These routines are in the same source file as the main program.

   To perform M=1500 MCMC iterations, the program is called as follows:

```
par = MH_SIS(complete_data,M=1500)
```

   The output `par` is a list of three parameter vectors, each of length M. These are the MCMC samples from the joint posterior distribution of the model parameters.

(a) Plot the sample paths of each of the parameters. Does it appear that the sampling algorithm has converged? For the rate of acquisition, for example:

```
plot(par[[1]],type="l",xlab="iteration",ylab="rate of acquisition
(per mo)")
```

(b) Calculate the posterior mean and the 90% posterior intervals for the three model parameters. For example:

```
la_samples = par[[1]][501:1500]
la_samples2 = sort(la_samples)
mean(la_samples2)
la_samples2[50] # 5% quantile of the marginal posterior
la_samples2[950]# 95% quantile of the marginal posterior
```

(c) Are there any correlation between rates λ and μ in their joint posterior distribution? For a visual inspection, you can draw the scatter plot of the joint posterior:

```
la_samples = par[[1]][501:1500]
mu_samples = par[[2]][501:1500]
plot(la_samples,mu_samples,type='p')
```

(d) The rate parameters were given (independent) Gamma(ν_1, ν_2) priors with $\nu_1 = \nu_2 = 0.00001$ (see the program code in the subroutine update_parameters.R). With the amount of data, the analysis is quite robust to the choice of prior. However, try how the posterior is affected by a more informative choice of the prior distributions (e.g, by choosing hyperparameters $\nu_1 = 1$ and $\nu_2 = 20$) when $N = 10$.