Practical:
Parameter estimation with data augmentation in the SIR model

Instructors: Kari Auranen, Elizabeth Halloran, Vladimir Minin
July 13 – July 15, 2015

Background

In this exercise we fit the general epidemic model to the Abakaliki smallpox data using Bayesian data augmentation. The data originate from a smallpox outbreak in a community of $M = 120$ initially susceptible individuals. There is one introductory case and 29 subsequent cases so that the total number of cases is $n = 30$. The observed 29 time intervals (Δ) between the n removals, i.e., between the detection of cases are:

$$13, 7, 2, 3, 0, 0, 1, 4, 5, 3, 2, 0, 2, 0, 5, 3, 1, 4, 0, 1, 1, 1, 2, 0, 1, 5, 0, 5, 5 \text{ (days)}.$$

A zero means that symptoms appeared the same day as for the preceding case. After the last removal there were no more cases. To fix the the time origin we assume that the introductory (index) case became infectious at time 0 and was removed at time 14 days (this appears as a long duration of infectiousness but agrees with the interpretation made in [1]). With this assumption, we can calculate the removal times $\boldsymbol{r}$ with respect to the time origin (see exercise 2 below). The total duration of the outbreak is $T = 90$ days ($= 14 + \sum_{i=1}^{29} \Delta_i$).

We explore the joint posterior distribution of the infection rate β and the removal rate γ. The unknown infection times $(i_2, \ldots, i_{30})$ are augmented, i.e., treated as additional model unknowns. All infection times together are denoted by $\boldsymbol{i}$.

The example program is implemented using *individual-based* event histories (see the lectures). The indices thus refer to individuals. In particular, (i_k, r_k) are the infection and removal times for the *same* individual k. This affects the choice of the likelihood function as explained in the lectures. The appropriate expression is:

$$\gamma^n \prod_{k=2}^{n} \{\beta I(i_k)\} \exp\left(- \int_0^T (\gamma I(u) + (\beta/M) I(u) S(u)) du \right).$$

In actual computations, it is more convenient to use the logarithm of the likelihood function:

$$n \log(\gamma) + (n-1) \log(\beta) + \sum_{k=2}^{n} \log I(i_k) - \int_0^T (\gamma I(u) + (\beta/M) I(u) S(u)) du.$$

N.B. The following is not intended to be a comprehensive analysis of the Abakaliki smallpox data. More appropriate analyses are possible. For example, in reference [2], the time of infection of the index case was included in the model unknowns. No adjustments were made to the original data. In [3], heterogeneity across individuals in their susceptibility to infection and a latent period were allowed.

Exercises

1. Download all required source codes by executing **SIRaugmentation_reduced.R**. The complete code will be provided once we have tried to complete the "reduced" version of the sampling routine (see below).

2. **Read the data.** The observed data in the Abakaliki smallpox outbreak include only the time intervals between removal times in the 30 infected individuals (therefore 29 intervals) and the fact that 90 individuals remained uninfected throughout the outbreak. Function **readdata.R** can be used to read in the time intervals of removals:

   ```
   intervals = readdata()
   ```

   The time intervals are in days. Note that the output vector does not include the piece of information that 90 individuals remained uninfected. This has to be input to the estimation routine separately (see below).

3. **Calculate the removal times.** The removal times can be calculated on the basis of the time intervals between them. This requires fixing a time origin. We make the assumption that the index case became infected at time $t = 0$ and was removed at time 14 (see above). These assumptions are "hardwired" in the program **removaltimes.R** (but can be changed easily for other contexts):

```
remtimes = removaltimes(intervals)
```

4. **Implementing the sampling algorithm.** The steps are

   (a) Reserve space for vectors of length K for the two model parameters β and γ (for an MCMC sample of size K; in the actual R code, K = mcmc.size). Samples of the unknown infections times need not be stored but a (vector) variable is needed to store the current iterates.

   (b) Initialise the model unknowns $\beta[1]$ and $\gamma[1]$. The unknown infection times need to be initialized as well. To do this, you can use routine **initializedata.R** which creates a complete data matrix with two columns (infection times and removal times). Each row corresponds to an infected individuals in the data; the index case is on the first row.

```
completedata = initializedata(remtimes)
```

   (c) Update β from its full conditional distribution in a Gibbs step:

   $$\beta[k+1] \mid \boldsymbol{i}[k-1], \boldsymbol{r} \ \sim \ \Gamma(n-1+\nu_\beta, (1/M)\int_0^T I(u)S(u)du + \lambda_\beta)$$

   (d) Update γ from its full conditional distribution in a Gibbs step:

   $$\gamma[k] \mid \boldsymbol{i}[k-1], \boldsymbol{r} \ \sim \ \Gamma(n+\nu_\gamma, \int_0^T I(u)du + \lambda_\gamma)$$

   (e) Update infection times $(i_2, \ldots, i_n)$ using Metropolis-Hastings steps (cf. the lecture). This creates a new vector of infection times $\boldsymbol{i}$[k] (the first element is always fixed by our assumption).

   (f) Repeat steps (c)–(e) K times, storing the samples $(\beta[k], \gamma[k])$, $k = 1, \ldots K$.

The sampling routine is implemented in **sampleSIR_reduced.R**. It requires as input the removal times ($\boldsymbol{r}$), the total number of individuals (M) and the number of iterations (K). The program uses a number of subroutines (with obvious tasks to perform): **initializedata.R**, **update_beta.R**, **update_gamma.R**, **update_inftimes.R**, **loglikelihood.R**, **totaltime_infpressure.R**, and **totaltime_infected.R**.

The subroutines **update_beta.R** and **update_gamma.R** are reduced, so your task is to complete those. These corresponds to steps (c) and (d) above.

5. **Sampling the posterior distribution.** Use the compeleted sampling routine (or **sampleSIR.R**) to realize an MCMC sample from the joint distribution of the model two parameters:

   ```
   mcmc.sample = sampleSIR(remtimes,M=120,mcmc.size=600)
   ```

   Plot the sample paths of the two model parameters (β and γ). For example, for parameter β:

   ```
   plot(mcmc.sample$beta,type="l",xlab="iteration",ylab="beta")
   ```

   Then explore the marginal and joint distributions of the model parameters.

6. **The effect of priors.** The program applied uninformative priors with $(\nu_\beta, \lambda_\beta) = (0.0001, 0.0001)$ and $(\nu_\gamma, \lambda_\gamma) = (0.0001, 0.0001)$ (see functions update_beta.R and update_gamma.R. Try how sensitive the posterior estimates are to a more informative choice of the prior, e.g. $(\nu_\beta, \lambda_\beta) = (10, 100)$ and $(\nu_\gamma, \lambda_\gamma) = (10, 100)$.

7. **The number of secondary cases.** What is the expected number of secondary cases for the index case, that is, calculate the posterior expectation of β/γ.

References:

[1] Becker N. Analysis of infectious diseases data. Chapman and Hall, 1989.

[2] O'Neill Ph. and Roberts G. Bayesian inference for partially observed stochastic processes. Journal of the Royal Statistical Society, Series A, **162**, 121–129 (1999).

[3] O'Neill Ph. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. Mathematical Biosciences **180**, 103-114 (2002).