

## MCMC I Practical session: Chain binomial model I: Gibbs sampler

### Background

In this computer lab, we apply Gibbs sampling to incompletely observed data in a chain binomial model. The observations are based on outbreaks of measles in Rhode Island during the years 1929–1934 [1]. We restrict the analysis to families with 3 susceptible individuals at the onset of the outbreak. This example is based on references [1]–[4].

We assume that there is a single index case that introduces infection to the family. Thus, possible epidemic chains are 1,  $1 \rightarrow 1$ ,  $1 \rightarrow 1 \rightarrow 1$  and  $1 \rightarrow 2$ . Denote the probability for a susceptible to escape infection when exposed to one infective in the family by  $q$  (and  $p = 1 - q$ ). The following table lists chain probabilities, with the actually observed frequencies of the size of epidemic:

chain	prob.	frequency	observed frequency
1	$q^2$	$n_1$	34
$1 \rightarrow 1$	$2q^2p$	$n_{11}$	25
$1 \rightarrow 1 \rightarrow 1$	$2qp^2$	$n_{111}$	not observed
$1 \rightarrow 2$	$p^2$	$n_{12}$	not observed

In this exercise, we assume that frequencies  $n_{111}$  and  $n_{12}$  have not been observed. Only their sum  $N_3 = n_{111} + n_{12} = 275$  is known.

The estimation problem concerns the escape probability  $q$ , so that there is basically only one unknown parameter in the model. However, the fact that not all frequencies have been observed creates a computational problem that can be solved by Bayesian data augmentation and Gibbs sampling [2].

**Marginal likelihood.** The joint probability of the *complete data*  $(n_1, n_{11}, N_3, n_{111})$  is proportional to a multinomial probability:

$$\begin{aligned}
 f(n_1, n_{11}, N_3, n_{111} | q) &= (q^2)^{n_1} (2q^2p)^{n_{11}} (2qp^2)^{n_{111}} (p^2)^{N_3 - n_{111}} \\
 &= \text{constant} \times q^{2n_1 + 2n_{11} + n_{111}} p^{n_{11} + 2N_3}.
 \end{aligned} \tag{1}$$

The marginal likelihood  $f(n_1, n_{11}, N_3 | q)$  would be obtained by summing up expressions (1) with  $n_{111}$  running from 0 to  $N_3$ .

**The Bayesian approach.** Instead of using the marginal likelihood, we will treat frequency  $n_{111}$  as a model unknown in addition to parameter  $q$ . The joint distribution of the observations

$(n_1, n_{11}, N_3)$  and the model unknowns  $(n_{111}, q)$  is

$$f(n_1, n_{11}, N_3, n_{111}, q) = f(n_1, n_{11}, N_3, n_{111} | q) f(q). \quad (2)$$

The first term in is the complete data likelihood (see (1)), based on the augmented data (i.e. the data are augmented with the unknown frequency  $n_{111}$ ).

The second term is the prior density of probability  $q$ . We choose a Beta prior for parameter  $q$ :  $q \sim \text{Beta}(\alpha, \beta)$  so that  $f(q) \propto q^{\alpha-1}(1-q)^{\beta-1}$ . With the choice  $\alpha = \beta = 1$ , this is uniform prior on  $[0,1]$ .

The joint posterior distribution of the model unknowns is  $f(q, n_{111} | n_1, n_{11}, N_3)$ .

**Gibbs sampling.** In the lecture we demonstrated that the joint posterior distribution of the model unknowns  $n_{111}$  and  $q$  can be investigated by Gibbs sampling. This means making a numerical sample from the posterior distribution by drawing samples of  $n_{111}$  and  $q$  in turn from their full conditional posterior distributions:

$$f(q | n_1, n_{11}, N_3, n_{111}) \quad \text{and} \quad f(n_{111} | n_1, n_{11}, N_3, q).$$

These were found to be

$$q | n_1, n_{11}, N_3, n_{111} \sim \text{Beta}(2n_1 + 2n_{11} + n_{111} + \alpha, n_{11} + 2N_3 + \beta) \quad (3)$$

and

$$n_{111} | n_1, n_{11}, N_3, q \sim \text{Binomial}(N_3, 2q/(2q + 1)). \quad (4)$$

## Exercises

1. **Gibbs sampling.** The R program (**chainGibbs\_reduced.R**) contains a function `chainGibbs(mcmc.size,  $\alpha, \beta$ )` that draws samples from the joint posterior distribution of  $q$  and  $n_{111}$ . The function has this particular data set “hardwired” within the program. Starting with the initial values  $(q^{(1)}, n_{111}^{(1)}) = (0.5, 275 * (2 * 0.5) / (2 * 0.5 + 1))$ , it iterates between sampling

$$q^{(i)} | n_1, n_{11}, N_3, n_{111}^{(i-1)} \quad \text{and} \quad n_{111}^{(i)} | n_1, n_{11}, N_3, q^{(i)}, \quad i = 2, \dots, \text{mcmc.size}.$$

This creates a sample  $(q^{(i)}, n_{111}^{(i)})$ ,  $i = 1, \dots, \text{mcmc.size}$ .

**Your task** is to complete the Gibbs sampler by programming the two lines of code that draw samples in turn from the full conditional distributions (3) and (4).

2. **Writing a more general Gibbs sampler function**

- (a) You might like to write a function `mychainGibbs(n1,n11,N3,mcmc.size, $\alpha$ , $\beta$ )` that allows you to do inference on other data sets with observed  $(n_1, n_{11}, N_3)$ .
  - (b) If you write such a function, try altering the value of  $N_3$ . How do larger and smaller values of  $N_3$  alter the posterior distribution of  $q$ ?
3. **Sensitivity to the choice of prior.** Assess how the choice of the prior distribution affects estimation of the escape probability. Use the  $\text{Beta}(\alpha, \beta)$  prior with different values of  $\alpha$  and  $\beta$ . Note that both parameters can be given as input to the function `chainGibbs(mcmc.size, $\alpha$ , $\beta$ )` in **chainGibbs.R**.
4. **Write your own Gibbs sampler:** Alternatively to `chainGibbs.R`, you might like to try writing your own Gibbs sampler for the chain binomial problem. Assume you will run `mcmc.size` iterations.
- (a) Reserve space for the `mcmc.size`-vector of  $q$  and  $n_{111}$  values.
  - (b) Initialize the model unknowns `q[1]` and `n111[1]` (round the `n111[1]`)
  - (c) Enter the data `n1`, `n11`, `N3`
  - (d) Draw the MCMC samples `2:mcmc.size` using the `rbeta()` and `rbinom()` functions

**Posterior inferences.** By discarding a number of "burn-in" samples, you can use the rest of the numerical sample to explore the posterior of escape probability  $q$ . It is enough to discard a few hundred first samples, say 500, in this simple model.

- (a) Make a histogram of the samples `501:mcmc.size` of `q` and `n111`.
- (b) Use the `summary()` function to get summaries the samples `501:mcmc.size` of `q` and `n111`.

Now you can convert the program to a function similar to the function in the file **chain-Gibbs.R** `chainGibbs(mcmc.size, $\alpha$ , $\beta$ )`.

## References:

- [1] Bailey T.J.N. "The Mathematical Theory of Infectious Diseases", Charles Griffiths and Company, London 1975.
- [2] O'Neill P. and Roberts G. "Bayesian inference for partially observed stochastic processes", *Journal of the Royal Statistical Society, Series A*, **162**, 121–129 (1999).
- [3] Becker N. Analysis of infectious disease data. Chapman and Hall, New York 1989.
- [4] O'Neill P. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathematical Biosciences* 2002; 180:103-114.