Practical 6: Hierarchical chain binomial model

Instructors: Vladimir Minin, Kari Auranen, Elizabeth Halloran
July 15–17, 2019

Background

In this computer class, we re-analyze the data about outbreaks of measles in households. The analysis is restricted to households with 3 susceptible individuals at the onset of the outbreak. We assume that there is a single index case that introduces infection to the household. The possible chains of infection then are 1, $1 \to 1$, $1 \to 1 \to 1$, and $1 \to 2$.

In this example, the probabilities for a susceptible to escape infection when exposed to one infective in the household are allowed to be different in different households. These probabilities are denoted by q_j (and $p_j = 1 - q_j$), $j = 1, \ldots, 334$. The following table expresses the chain probabilities in terms of the escape probability q_j. The observed frequency is the number of households with the respective chain.

chain	prob.	frequency	observed frequency
1	q_j^2	n_1	34
$1\to1$	$2q_j^2 p_j$	n_{11}	25
$1\to1\to1$	$2q_j p_j^2$	n_{111}	not observed
$1\to 2$	p_j^2	n_{12}	not observed

The frequencies n_{111} and n_{12} have not been observed. Only their sum $N_3 = n_{111} + n_{12} = 275$ is known.

The hierarchical model was defined in the lecture notes. The joint distribution of parameters α and β, the household-specific escape probabilities and the chain frequencies is

$$\prod_{j=1}^{334}\left(f(n_1^{(j)}, n_{11}^{(j)}, n_{111}^{(j)}, n_{12}^{(j)}|q_j)f(q_j|\alpha,\beta)\right)f(\alpha,\beta),$$

where

$$(n_1^{(j)}, n_{11}^{(j)}, n_{111}^{(j)}, n_{12}^{(j)})|q_j \sim \text{Multinomial}(1,(q_j^2, 2q_j^2p_j, 2q_jp_j^2, p_j^2)),$$
$$q_j|\alpha,\beta \sim \text{Beta}(\alpha,\beta),$$
$$(\alpha,\beta) \propto (\alpha+\beta)^{-5/2}.$$

N.B. The household-specific chain frequencies are vectors in which only one of the elements is 1, all other elements being 0.

N.B. The joint prior distribution of the parameters of the Beta distribution, α and β, is proportional to $(\alpha+\beta)^{-5/2}$. This is derived on the basis of assuming independent uniform priors for $\alpha/(\alpha+\beta)$ (the expectation of the Beta distribution) and $1/(\alpha+\beta)$ (an approximation to the the standard deviation of the Beta distribution). See Chapter 5.3 in Gelman et al.

We index the households with chain 1 as 1,...,34, and households with chain $1 \to 1$ as 35,...,59, and households with chain $1 \to 1 \to 1$ or $1 \to 2$ as 60,...,334. The model unknowns are α, β, frequencies $n_{111}^{(j)}$ for $j = 60, \ldots, 334$ (i.e., for all 275 households with the final number of infected 3) and q_j for $j = 1, \ldots, 334$ (all households).

In this exercise we apply a combined Gibbs and Metropolis algorithm to draw samples from the posterior distribution of the model unknowns. Before that, we explore the fit of the simple model with $q_j = q$ for all j.

Exercises

1. The simple chain binomial model. Using R routine **chainGibbs.R** (or **mychainGibbs**), i.e., repeating the earlier exercise, realize an MCMC sample from the posterior distribution of the escape probability q in the simple model in which this probability is the same across all households.

2. Model checking (simple model). Based on the posterior sample of parameter q, draw samples from the posterior predictive distribution of frequencies (n_1, n_{11}). Compare the sample to the actually observed value (34,25). The algorithm to do this is as follows:

(a) Discard a number of "burn-in" samples in the posterior sample of parameter q, as realised in exercise (1) above.

(b) When the size of the retained sample is K, reserve space for the Kx4 matrix of predicted frequencies for n_1, n_{11}, n_{111} and n_{12}.

(c) Based on the retained part of the posterior sample, take the kth sample $q^{(k)}$.

(d) Draw a sample of frequencies $(n_1^{(k)}, n_{11}^{(k)}, n_{111}^{(k)}, n_{12}^{(k)})$
from Multinomial(334,$((q^{(k)})^2, 2(q^{(k)})^2 p^{(k)}, 2q^{(k)}(p^{(k)})^2, (p^{(k)})^2)$) using the rmultinom() function in R.

(e) Repeat steps (c) and (d) K times, storing the sample of frequencies after each step (d).

(f) Plot the samples of pairs $(n_1^{(k)}, n_{11}^{(k)})$, $k = 1, ..., K$, and compare to the observed point (34,25).

The R routine covering steps (a)-(f) is provided in the script **checkmodel_reduced.R**, except for step (d). Complete step (d) and check the model fit:

```
mcmc.sample = chainGibbs(5000,1,1)
checkmodel_reduced(mcmc.sample,1000)
```

The complete R routine (**checkmodel.R**) will be provided once you have tried writing your own code.

3. A hierarchical chain binomial model. Samples from the joint posterior distribution of the unknowns in the hierarchical (beta-binomial) chain model can be sampled using the following algorithm, applying both Gibbs and Metropolis-Hastings updatings steps (superscript k refers to the kth

MCMC step):

(a) Reserve space for all model unknowns (parameters α and β as well as the 275 unknown frequencies $n_{111}^{(j)}$).

(b) Initialize the model unknowns.

(c) Update all household-specific escape probabilities from their full conditionals:

$$q_j^{(k)}|\alpha^{(k-1)}, \beta^{(k-1)} \sim \text{Beta}(2 + \alpha^{(k-1)}, \beta^{(k-1)}),\ j = 1, \dots, 34$$

$$q_j^{(k)}|\alpha^{(k-1)}, \beta^{(k-1)} \sim \text{Beta}(2 + \alpha^{(k-1)}, 1 + \beta^{(k-1)}),\ j = 35, \dots, 59$$

$$q_j^{(k)}|\alpha^{(k-1)}, \beta^{(k-1)}, n_{111}^{(j,k-1)} \sim \text{Beta}(n_{111}^{(j,k-1)} + \alpha^{(k-1)}, 2 + \beta^{(k-1)}),\ j = 60, \dots, 334$$

(d) Update the unknown binary variables $n_{111}^{(j)}$ ($j = 60, \dots, 334$) from their full conditionals:

$$n_{111}^{(j,k)}|q_j^{(k)} \sim \text{Binomial}(1, 2q_j^{(k)}/(2q_j^{(k)} + 1))$$

(e) Sample $\alpha^{(k)}$ using a Metropolis-Hastings step (see the program code)

(f) Sample $\beta^{(k)}$ using a Metropolis-Hastings step (see the program code)

(g) Repeat steps (b)–(f) K times (in the R code, K=mcmc.size).

The above algorithm is written in the R script **chain_hierarchical_reduced.R**, except for parts of step (c). Complete the code and draw a posterior sample of all model unknowns. Note that the data set and the prior distributions are hardwired within the given program code.

The complete routine (**chain_hierarchical.R**) will be provided once you have tried your own solution.

4. Posterior inferences. Plot the marginal posterior distributions of the parameters α and β. You can also check how their joint posterior distribution looks like. Draw a histogram of the posterior distribution of $\alpha/(\alpha+\beta)$, the expected escape probability (= the expectation of the Beta distribution).

Using output from program **chain_hierarchical.R**, the above plots can be done as follows (based on 2000 samples with the first 500 as burn-in samples):

```
mcmc.size = 10000
mcmc.sample = chain_hierarchical(mcmc.size)
mcmc.al = mcmc.sample$al
mcmc.be = mcmc.sample$be

burn.in = 2000
mcmc.al = mcmc.al[(burn.in+1):mcmc.size]
mcmc.be = mcmc.be[(burn.in+1):mcmc.size]

# The marginal posterior distributions of parameters alpha and beta
par(mfrow=c(1,2))
hist(mcmc.al,xlab='alpha',main='')
hist(mcmc.be,xlab='beta',main='')

# The joint posterior distribution of alpha and beta
par(mfrow=c(1,1))
plot(mcmc.al,mcmc.be,xlab='alpha',ylab='beta')

#  The posterior distribution  of the expected escape probability
hist(mcmc.al/(mcmc.al+ mcmc.be),breaks=20,
  xlab='expected escape probability',main='',xlim=c(0.1,0.35))
```

You can still plot the posteriori predictive distribution of the escape probability: see the programme code.

```
qpost = rbeta((mcmc.size-burn.in),mcmc.al,mcmc.be)
hist(qpost,main="posterior predictive distribution of the escape probability",
   cex.main=1,xlab="predictive q",breaks=20)
```

5. Model checking (hierarchical model). Check the fit of the hierarchical model with the R program **check_hierarchical.R**. The program draws samples from the posterior predictive distribution of the chain frequencies and plots the these samples for frequencies n_1 and n_{11} with the actually observed point (34,25).

```
check_hierarchical(mcmc.sample,mcmc.burnin=500)
```

N.B. Unlike we pretended in the preceding exercises, the original data actually record the frequencies $n_{12} = 239$ and $n_{111} = 36$. You can now check the model fit with respect to these frequencies.

References:

[1] Bailey T.J.N. "The Mathematical Theory of Infectious Diseases", Charles Griffiths and Company, London 1975.

[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] Becker N. Analysis of infectious disease data. Chapman and Hall, New York 1989.

[4] O'Neil Ph. A tutorial introduction to Bayesian inference for stchastic epidemic models using Markov chain Monte Carlo methods. Mathematical Biosciences 2002; 180:103-114.

[5] Gelman, Carlin, Stern, Rubin. Bayesian Data Analysis, Chapman and Hall, London 2004.