数理解析研究所講究録 1252 巻 2002 年 103-109

Bayesian inference for an epidemic model with several kinds of susceptibles

Yu Hayakawa*, Philip D. O'Neill†, Darren Upton‡and Paul S.F. Yip§

Summary. An epidemic model with several kinds of susceptible is analysed from a Bayesian perspective. The posterior distribution of the parameters of the model is explored via Markov chain Monte Carlo methods. The methods are illustrated using data from a real life respiratory disease epidemic, and the results compared with those obtained using martingale estimating equations.

1 Introduction

We shall consider a simple Markov model in which susceptibles are one of k different types, corresponding to different levels of susceptibility to the disease in question. Such differences in susceptibility typically arise due to age, although other factors can be important in specific scenarios (e.g. the effects of vaccination). Epidemic models of this kind have been considered by a number of authors (e.g. Ball [1], Becker [2], Britton [4], and Yip and Chen [7]), although it should be noted that the context of Bayesian inference has not previously been explored.

In general, epidemic outbreak data are incomplete in that the times of infections are unknown. As a consequence, the analysis of such data generally requires the imputation of missing information. O'Neill and Roberts [6] describe a Markov chain Monte Carlo (MCMC) method approach for the Bayesian analysis of a homogeneous epidemic model, and here we extend their basic method to cater for a non-homogeneous setting. For another type of extension, see O'Neill and Becker [5].

The outline of the paper is as follows. In Section 2, the model and data are described, along with the Bayesian framework and inferential objectives. Section 3 presents the procedure used to perform inference for the parameters, in which a Gibbs sampler incorporating a Metropolis-Hastings step is defined. Our method is used to analyse data from a real epidemic in Section 4, and some of the results are compared with classical estimates obtained via martingale equations in Yip and Chen [7]. Further studies are mentioned in Section 5.

*School of Mathematical and Computing Sciences, Victoria University of Wellington, PO Box 600, Wellington, New Zealand, *e-mail*: Yu.Hayakawa@vuw.ac.nz

†School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, England NG7 2RD, *e-mail*: Philip.O'Neill@nottingham.ac.uk

‡Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge CB3 0WB, UK, *e-mail*: D.S.Upton@statslab.cam.ac.uk

§Department of Statistics & Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China, *e-mail*: sfpyip@hku.hk

2 Model, data, notation and Bayesian framework

2.1 Model

Consider a population consisting initially of k groups of susceptible individuals, where the groups are labelled $1, \ldots, k$ and group i contains N_i susceptibles, $i = 1, \ldots, k$. An epidemic is initiated in the population by one of the susceptibles becoming infected, this infection being assumed to occur via some process external to the population. As we shall see later, the identity of the initial infective will not be included as part of the available data. For $t \geq 0$ and $i = 1, \ldots, k$, define $X_i(t)$ and $Y_i(t)$ as the numbers of susceptible and infective individuals, respectively, in group i at time t. Furthermore, let $Y(t) = \sum_{i=1}^{k} Y_i(t)$ denote the total number of infectives in the population at time $t \geq 0$. The epidemic is then defined according to the following infinitesimal transition probabilities, the transitions themselves corresponding respectively to an infection and a removal:

$$\begin{aligned} \Pr\{(X_i(t+\delta t), Y(t+\delta t)) = (x-1, y+1) | (X_i(t), Y(t)) = (x,y)\} &= \beta_i xy + o(\delta t), \\ \Pr\{(X_i(t+\delta t), Y(t+\delta t)) = (x, y-1) | (X_i(t), Y(t)) = (x,y)\} &= \gamma y + o(\delta t), \end{aligned}$$

all other transitions having probability $o(\delta t)$.

2.2 Data and notation

It is assumed that the type and removal time of each individual is observable, but that the infection times are not. These assumptions are motivated by the fact that real-life disease outbreak data almost never includes infection times, but may include the times at which individuals are detected. The infection times are treated as unknown parameters as in O'Neill and Roberts [6]. For convenience, we define a time origin by setting the first observed removal at time zero; the data then consist of the times τ_{ij}, where τ_{ij} is the time of the jth removal in the ith group. The removal times are observed up to a time $T > 0$ and thus $0 \leq \tau_{ij} \leq T$ for all τ_{ij}. Note that if the epidemic has not been completed by time T, then the number of infections occurring up until time T is unknown and $n_i \leq m_i \leq N_i$, $i = 1, \ldots, k$.

Let $\tau_{i\cdot} = (\tau_{i1}, \tau_{i2}, \cdots, \tau_{in_i})$ and $I_{i\cdot} = (I_{i1}, I_{i2}, \cdots, I_{im_i})$ represent the vector of ordered removal times of type i and of ordered infection times of type i, respectively. Furthermore, define $I_{\cdot\cdot} = (I_{1\cdot}, \cdots, I_{k\cdot})$, $\tau_{\cdot\cdot} = (\tau_{1\cdot}, \cdots, \tau_{k\cdot})$, and $\beta_{\cdot} = (\beta_1, \cdots, \beta_k)$. The following is a list of additional notation that we shall use.

I_{i1} :	the time of the first infection in the ith group
$I_{\min}$:	$\min\{I_{11}, I_{21}, \cdots, I_{k1}\} \equiv$ the time of the first infection in the total population
$i_{\min}$:	type i for which $I_{i1} = I_{\min}$
n_i :	the observed total number of removals of type i
m_i :	the unobserved total number of infections of type i
N :	$\sum_{i=1}^{k} N_i \equiv$ the initial number of susceptibles in the total population
$\tilde{I}_{\cdot\cdot}$:	$I_{\cdot\cdot} \setminus I_{min}$

2.3 Bayesian inference

Our objective is to explore the posterior density of the parameters given the data, i.e. $\pi(\beta., \gamma|\tau..)$. However, the likelihood $\pi_L(\tau..|\beta., \gamma)$ is essentially intractable, since it involves integrating over all possible values for the unknown infection times. Consequently we instead work with the augmented likelihood $\pi_L(\tilde{I}.., \tau..|\beta., \gamma, I_{min}, i_{min})$, which is given below (equation (2)). Then, by Bayes' Theorem,

$$\pi(\beta., \gamma, I_{min}, i_{min}|\tilde{I}.., \tau..) \propto \pi_L(\tilde{I}.., \tau..|\beta., \gamma, I_{min}, i_{min})\pi(\beta., \gamma, I_{min}, i_{min}), \tag{1}$$

where $\pi(\beta., \gamma, I_{min}, i_{min})$ denotes the prior density of $\beta., \gamma, I_{min}$ and i_{min}. The MCMC algorithm described below will enable us to obtain samples from $\pi(\beta., \gamma, I_{min}, i_{min}|\tilde{I}.., \tau..)$, and so, by ignoring the values of I_{min}, i_{min} and $\tilde{I}..$ we thus obtain samples from the required posterior density $\pi(\beta., \gamma|\tau..)$.

The augmented likelihood that we require is given by

$$\begin{aligned}\pi_L(\tilde{I}.., \tau..|\beta., \gamma, I_{min}, i_{min}) = \\ \left\{\prod_{i=1}^{k}\left[\prod_{j=1}^{n_i}\gamma Y_i(\tau_{ij}^-)\right]\left[\prod_{l=2}^{m_i}\beta_i X_i(I_{il}^-)Y(I_{il}^-)\right]\right\}\cdot\left\{\prod_{i=1, i\neq i_{\min}}^{k}\beta_i X_i(I_{i1}^-)Y(I_{i1}^-)\right\}\cdot \\ \exp\left\{-\sum_{i=1}^{k}\left[\int_{I_{\min}}^{T}\beta_i X_i(t)Y(t)dt + \int_{I_{i1}}^{T}\gamma Y_i(t)dt\right]\right\},\end{aligned} \tag{2}$$

where τ_{ij}^- denotes the time just prior to the jth removal time of type i.

We assume a priori that $\beta_i \sim \Gamma(\nu_{\beta_i}, \lambda_{\beta_i}), i = 1, \cdots, k$, i.e., β_i has a gamma prior with shape parameter ν_{β_i} and scale parameter λ_{β_i}. Likewise, we set a prior for γ as $\gamma \sim \Gamma(\nu_\gamma, \lambda_\gamma)$. The prior for I_{min} is assumed to be an improper uniform on $(-\infty, 0)$, and the prior for i_{min} uniform on the set $\{1, \ldots, k\}$. These priors are assumed to be mutually independent.

As well as the basic model parameters $\beta.$ and γ, we shall also consider the following quantities of interest.

1. $\Delta_{ij} = \beta_i - \beta_j$, $i, j = 1, \ldots, k$ and $i \neq j$;
2. $\phi_{ij} = \beta_i/\beta_j$, $i, j = 1, \ldots, k$;
3. $\theta_i = N\beta_i/\gamma$, $i = 1, \ldots, k$;
4. $R_0 = \sum_{i=1}^{k}\theta_i N_i$.

The parameter R_0, as described in Yip and Chen [7], is a threshold parameter for the epidemic. Essentially, in a large population then epidemics die out quickly with probability 1 if $R_0 \leq 1$, while if $R_0 > 1$ then there is some probability of a major epidemic (see Ball [1]). Finally, we note that θ_i, and thus ϕ_{ij} and R_0, are estimable from final size data alone (i.e. the final numbers infected), see for example Britton [4].

3 Inference procedures

The MCMC algorithm described below is an extension of one suggested by O'Neill and Roberts [6] for a homogeneous population epidemic model.

First note that, from (1) and (2),

$$\pi(\beta_i|\beta_{-i},\gamma,\tau..,I..) \sim \begin{cases} \Gamma\left(\nu_{\beta_i}+m_i-1,\lambda_{\beta_i}+\int_{I_{\min}}^{T}X_i(t)Y(t)dt\right), & \text{if } i=i_{\min}, \\ \Gamma\left(\nu_{\beta_i}+m_i,\lambda_{\beta_i}+\int_{I_{\min}}^{T}X_i(t)Y(t)dt\right), & \text{otherwise,} \end{cases}$$

$$\pi(\gamma|\beta.,\tau..,I..) \sim \Gamma\left(\nu_\gamma+\sum_{i=1}^{k}n_i,\lambda_\gamma+\sum_{i=1}^{k}\int_{I_{i1}}^{T}Y_i(t)dt\right),$$

where $\beta_{-i}=(\beta_1,\cdots,\beta_{i-1},\beta_{i+1},\cdots,\beta_k)$.

Generating samples from the full conditional distribution of $I..$ is achieved using a Metropolis-Hastings algorithm, along the lines suggested in O'Neill and Roberts [6]. Note that this procedure also updates I_{min} and i_{min}. Specifically, there are three possible moves for $I..$: 1) moving an infection time of type i; 2) removing an infection time of type i; 3) adding a new infection time of type i, and one of these is chosen uniformly at each iteration. Then a type i is selected according to some probability mass function. Where necessary, a candidate infection time, y, is generated according to the density function

$$g(y)=\begin{cases} c\theta, & y\in(l,T], \\ c\theta e^{\theta(y-l)}, & y\in(-\infty,l], \end{cases}$$

where $\theta>0$ and $l<0$ have prespecified values, and where $c=[\theta(T-l)+1]^{-1}$.

The density g is such that new infection times are proposed uniformly on $(l,T]$ and according to an exponential distribution on $(-\infty,l]$. In particular, if $|l|$ is chosen to be something like a typical infectious period length, then new infection times will be approximately uniformly proposed on an interval where they might plausibly be found.

4 An epidemic of respiratory disease

In this section the methods described above are applied to a particular dataset.

4.1 Data

The dataset given in Table 1 corresponds to diagnosis times of individuals with a respiratory disease which occurred between October and November of 1967 on the island of Tristan da Cunha in the South Atlantic (Becker and Hopper [3]). We assume that the time of diagnosis is the same as that of removal. The total population of the island of 255 was partitioned into 3 groups: (1) Infants [age 0-4], (2) Children [age 5-14], (3) Adults [age 15 or above]. There was one unidentified case, hence $N=254$. The groups' initial population sizes were 25, 36 and 193, respectively. The time is discretised and the number in each column of the table is the number of infective individuals of each age group removed at the beginning of the day.

Table 1: 1967 epidemic of respiratory disease on Tristan da Cunha
(taken from Becker and Hopper [3])

Day	1	8	10	11	12	13	15	16	17	18	19	20	21	22	29	30	Total
Infants	0	0	0	3	1	3	1	0	0	1	0	0	0	0	0	0	9
Children	0	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	6
Adults	1	1	1	0	2	3	1	4	1	1	3	2	1	2	1	1	25

4.2 MCMC implementation and priors

We assume that the data describe the entire outbreak, so that the epidemic was contained after the infective adult on Day 30 was removed. Thus the Metropolis-Hastings step used to update infection times in the MCMC algorithm will only have one possible move, namely that of moving infection times. The probability mass function for selecting a group for this move was set by equating p_i with the ratio of the number of removals of type i to the total number of removals of all the groups, e.g., $p_1 = 9/40$. Hence the infection times, regardless of the type of group, are equally likely to be selected. For the candidate generating density function $g(y)$ for the Metropolis-Hastings step we set $\theta = 1, l = -10$.

Two sets of priors are chosen for the β_i's and γ, namely non-informative and informative. Note that we consider informative priors mainly for the purpose of illustrating the MCMC algorithm, although it is also a means of investigating the robustness of the posterior estimates. The hyperparameter values used for the informative case are listed in Table 2. In the non-informative case, we assume that the β_i's and γ are distributed uniformly over the positive real line (improper priors).

Table 2: Informative priors: hyperparameters

ν_{β_1}	λ_{β_1}	ν_{β_2}	λ_{β_2}	ν_{β_3}	λ_{β_3}	ν_γ	λ_γ
1	1000	1	1000	1	1000	1	10

4.3 Results

Table 3 contains the posterior mean and standard deviation of each parameter of interest under two prior selections of hyperparameter values. Figures 1-2 contain scatterplots of β_1 against γ for both non-informative and informative priors. Pairs of those of β_i against γ, $i = 2, 3$, are similar to Figures 1-2, hence are omitted.

Regarding the marginal posterior density estimates of $\beta_1, \beta_2, \beta_3, \gamma$ and R_0, they are more sharply peaked for the informative than the non-informative cases. For both non-informative and informative prior cases, the posterior mean of β_1 is the largest and that of β_3 is the smallest. The posterior estimate of the excess susceptibility rates between the infants and the adults is 0.003206 for the non-informative prior case and 0.002093 for the informative. Hence, the infants are most prone to the disease but the adults are least likely to contract the disease. This implies that the estimate of the susceptibility rate of the infants relative to that of the adults is the largest. This relative rate has mean 3.637 with standard deviation of 1.487 in the non-informative case and mean 2.553 with standard deviation of 0.8689 in the informative prior case. Both cases of prior parameters produced similar posterior mean values of the removal rate.

Table 3: Output from MCMC algorithm

	non-informative prior		informative prior	
	mean	std	mean	std
β_1	0.004499	0.001754	0.003542	0.0009623
β_2	0.00179	0.0008092	0.002077	0.0006461
β_3	0.001294	0.000371	0.001449	0.0003217
γ	0.3638	0.09304	0.376	0.06632
Δ_{12}	0.002709	0.001775	0.001465	0.00112
Δ_{13}	0.003206	0.001674	0.002093	0.0009788
Δ_{23}	0.0004966	0.00079	0.0006286	0.0006755
ϕ_{12}	3.017	1.902	1.879	0.8368
ϕ_{13}	3.637	1.487	2.553	0.8689
ϕ_{23}	1.447	0.6787	1.49	0.5374
θ_1	3.228	1.221	2.449	0.7387
θ_2	1.283	0.5617	1.434	0.4776
θ_3	0.9271	0.2464	0.9953	0.2297
R_0	1.204	0.2773	1.201	0.2326

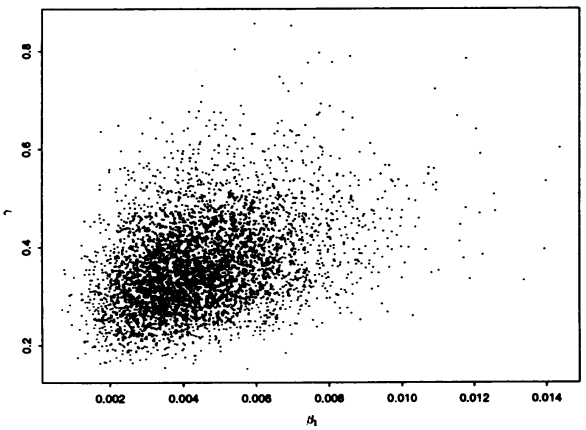

Figure 1: β_1 vs γ (non-informative)

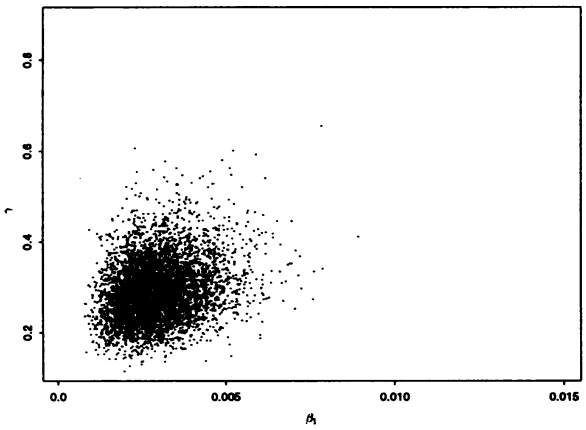

Figure 2: β_1 vs γ (informative)

Yip and Chen [7] used a martingale approach to derive maximum likelihood estimates for the ϕ_{ij}'s, θ_i's and R_0. We now briefly compare their results to ours, although a direct comparison requires caution due to the effects of priors, particularly in the informative case. First, the ranking of the posterior means of the ϕ_{ij}'s agrees with Yip and Chen's corresponding ranking. Second, the posterior means of the θ_i's are larger than the corresponding maximum likelihood estimates obtained by Yip and Chen for both non-informative and informative prior cases, apart from the estimate of θ_1 for the informative prior case. However, since the marginal posterior density for each θ_i is right-skewed, the posterior mean is likely to be greater than the modal value. Thirdly, both posterior means for R_0 are about 1.2, which is comparable to the maximum likelihood estimate of 1.1 obtained by Yip and Chen. Regarding the scatterplots, those for informative priors are, as would be expected, more tightly clustered than those for uninformative priors. Finally, the plots reveal slight positive correlation between each β_i and γ. This is to be expected, essentially because data consisting only of removals make it hard to distinguish between short infectious periods with high susceptibility (*i.e.*, large γ, large β_i) and the converse situation (small γ, small β_i). Thus estimation of $\theta_i = \beta_i/\gamma$ is likely to be more

precise than individual estimation of β_i and γ, as is reflected in Table 3, and β_i and γ are likely to exhibit positive correlation.

5 Further studies

The methodology that we have considered here can, in principle, be adapted for variants of the basic model. These include models with non-exponential infectious periods, or with latent periods. Another possibility is the situation where the actual numbers of initially susceptible individuals are unknown; in our notation, N_i would then become another model parameter. These and other topics are the subject of current research.

6 Acknowledgements

This work was supported by an SFLGC grant of Victoria University of Wellington (Hayakawa and Upton) and a RGC grant of the University of Hong Kong (Yip).

References

[1] Ball, F.G. Deterministic and stochastic epidemics with several kinds of susceptibles. *Advances in Applied Probability*, 17:1–22, 1985.

[2] Becker, N.G. *Analysis of Infectious Disease Data.* Chapman and Hall, London, 1989.

[3] Becker, N.G. and Hopper, J.L. Assessing the heterogeneity of disease spread through a community. *American Journal of Epidemiology*, 117:362–374, 1983.

[4] Britton, T. Estimation in multitype epidemics. *Journal of the Royal Statistical Society, Series B*, 60:663–679, 1998.

[5] O'Neill, P.D. and Becker, N.G. Inference for an epidemic when susceptibility varies. *Biostatistics*, 2:99–108, 2001.

[6] O'Neill, P.D. and Roberts, G.O. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society, Series A*, 162, Part 1:121–129, 1999.

[7] Yip, P.S.F. and Chen, Q. Statistical inference for a multitype epidemic model. *Journal of Statistical Planning and Inference*, 71:229–244, 1998.