J. R. Statist. Soc. A (1999)
162, *Part* 1, *pp.*121–129

Bayesian inference for partially observed stochastic epidemics

Philip D. O'Neill

University of Bradford, UK

and Gareth O. Roberts

University of Cambridge, UK

[Received July 1997. Revised April 1998]

Summary. The analysis of infectious disease data is usually complicated by the fact that real life epidemics are only partially observed. In particular, data concerning the process of infection are seldom available. Consequently, standard statistical techniques can become too complicated to implement effectively. In this paper Markov chain Monte Carlo methods are used to make inferences about the missing data as well as the unknown parameters of interest in a Bayesian framework. The methods are applied to real life data from disease outbreaks.

Keywords: Bayesian inference; Epidemic; General stochastic epidemic; Gibbs sampler; Hastings algorithm; Markov chain Monte Carlo methods; Reed–Frost epidemic

1. Introduction

The kinds of data that are available from real life epidemics are typically rather less complete than would be desirable for modelling and inference. For example, it is extremely unlikely that we would know the precise times that individuals became infected during an epidemic, or that other details of the infection mechanism would be observed. Consequently, for many of the standard epidemic models it becomes difficult to write down suitable likelihood functions, since these usually rely on having complete information. Becker (1995) drew attention to these problems within the context of classical inference and described various techniques for dealing with them. These include using approximations to deduce the pattern of the infection process, using simpler models and using martingale techniques for non-likelihood approaches. Another alternative is the use of the EM algorithm (Becker, 1993), where the missing data are treated as parameters to be estimated. This last idea features in the methods that we shall describe in this paper.

To date, almost all of the literature concerning statistical inference for epidemics adopts a classical approach. Our purpose in this paper is to describe a Bayesian approach to inference for both the Reed–Frost epidemic model and the general stochastic epidemic. The Bayesian paradigm is particularly suited to the context of epidemic modelling since the parameters of interest are usually defined in terms of an individual, which naturally leads one to consider the distributions of these parameters over a whole population. Additionally, we shall describe how Markov chain Monte Carlo (MCMC) methods, and in particular the Gibbs sampler,

Address for correspondence: Philip D. O'Neill, Division of Statistics, Department of Mathematical Sciences, University of Liverpool, Liverpool, L69 3BX, UK.
E-mail: P.Oneill@liverpool.ac.uk

 0964–1998/99/162121

can be used to implement the theory. We illustrate our methods by using various data sets, although it should be noted that we do not attempt to analyse these data in great detail.

2. The Reed–Frost epidemic in small populations

2.1. The model

The Reed–Frost model is a discrete time model for an epidemic spreading in a closed population and is defined as follows (see also Bailey (1975), p. 157). Consider a population consisting initially of N susceptible and a infective individuals. For $t = 0, 1, 2, \ldots$ let X_t and Y_t denote respectively the numbers of susceptible and infective individuals in the population at time t.

The model assumes total spatial homogeneity, in that at a given time point each susceptible individual has a probability $1 - q$ of being infected by each infective individual. Here, $q \in (0, 1)$ is known as the avoidance probability. Each of these infections happens independently, so each susceptible individual has the probability q^{Y_t} of avoiding infection within the tth generation. Once infected, the susceptible individual becomes an infective and can then infect others, but only from the following time point.

Thus, given $(X_t = x, Y_t = y)$, X_{t+1} is binomially distributed with parameters x and q^y, and Y_{t+1} is defined to be $X_t - X_{t+1}$. The epidemic continues until no more infective individuals remain in the population. The total number of individuals who contract the disease, excluding those who were infective at time 0, is known as the final size of the epidemic. It is convenient to describe the progress of an epidemic by the vector $(Y_0, Y_1, Y_2, \ldots, Y_m)$ where $m + 1 = \min\{t: Y_t = 0\}$. So, for example, the pattern (1, 2, 1) denotes an epidemic in which one initial infective individual infects two susceptibles, who in turn infect one further susceptible, who then fails to infect any remaining susceptibles.

2.2. Households of size 3

We now consider the case where the available data consist of the final sizes of epidemics in households of three individuals, one of whom is initially infective. Our objective will be to make inferences about the avoidance probability q. For $j = 0, 1, 2$ let n_j denote the number of households in which an epidemic of final size j occurs, so that the data can be written (n_0, n_1, n_2). Next, let n_{21} denote the (unobserved) number of epidemics with pattern of infection given by (1, 1, 1). The full conditional likelihood function for q is thus

$$L(q; n_0, n_1, n_2, n_{21}) = 2^{n_1+n_{21}} q^{2n_0+2n_1+n_{21}} (1-q)^{n_1+2n_2}.$$

In the following, we shall use the notation $\pi(\cdot|\cdot)$ to denote conditional distributions. We assume *a priori* that q has a beta(α, δ) distribution, from which it follows that the posterior distribution is given by

$$\pi(q|n_0, n_1, n_2, n_{21}) \sim \text{beta}(2n_0 + 2n_1 + n_{21} + \alpha, n_1 + 2n_2 + \delta). \tag{2.1}$$

The use of a beta distribution for q has also been considered in a classical inference context; see Bailey (1975), p. 254. Next, by assuming independent behaviour of different households it follows that

$$\pi(n_{21}|n_0, n_1, n_2, q) \sim \text{binomial}\{n_2, 2q/(2q+1)\}. \tag{2.2}$$

Using the two conditional distributions (2.1) and (2.2) we may now employ the Gibbs sampling scheme (see, for example, Smith and Roberts (1993)) to sample from $\pi(q, n_{21}|n_0, n_1, n_2)$. Specifically, we proceed by using the following algorithm.

Table 1. Providence measles data for final size of household epidemics of size 3

Final size of epidemic	*Providence measles data*
0	34
1	25
2 (1, 1, 1)	36
2 (1, 2)	239

Set B, the burn-in time.
Set T, the number of cycles between samples after burn-in.
Set M, the desired sample size.
Set R, the run size, equal to MT.
Define $S(\cdot)$ as the vector for sample output.
Set initial values for $q(-B)$ and $n_{21}(-B)$.
Loop: $i = -B$ to $i = R$;
  sample $q(i+1)$ according to equation (2.1) with $n_{21} = n_{21}(i)$;
  sample $n_{21}(i+1)$ according to equation (2.2) with $q = q(i+1)$;
  if $i > 0$ and $k = i/T$ is an integer then set $S(k) = (q(i+1), n_{21}(i+1))$;
end of loop.

The output from this algorithm will be a sequence $S(1)$, $S(2)$, . . ., $S(M)$, where $S(k)$ is an approximate sample from the joint posterior distribution of q and n_{21}.

This algorithm (and all those described in later sections) was implemented using Fortran running on a mainframe computer. It was found that a burn-in time of $B = 1000$ cycles was adequate, and we set $T = 100$ and $M = 1000$. The actual run time was of the order of a few seconds.

2.3. Results for households of 3

As an example of our methods, we shall use the data set obtained by Wilson and co-workers concerning measles epidemics in Providence, Rhode Island, between 1929 and 1934 (see Bailey (1975), p. 249). Unusually, the data set itself consists of n_0, n_1, n_2 and n_{21}, but we shall proceed by considering only the final size data as described above. This approach also allows for a comparison with the classical inference methods using final size data described in Bailey (1975), p. 251. The data values that we shall use are given in Table 1. Thus $n_0 = 34$, $n_1 = 25$ and $n_2 = 275$.

Fig. 1 shows some typical results for various choices of the prior parameters α and δ (the graphics for the figures in this paper were produced by S-PLUS). In Figs 1(a)–1(c) the prior distributions all have the same mean (namely 0.4), but the values are chosen to reflect respectively a small, medium and a large influence on the posterior distribution (2.1). Fig. 1(d) illustrates the outcome for a different prior, with mean 0.1, and parameters chosen to have a medium influence on the posterior distribution. As would be expected, the general effect of this prior distribution is to cause the posterior distribution to be concentrated around lower values than in Figs 1(a)–1(c). By comparison, the maximum likelihood approach described in Bailey (1975), p. 251, yields the estimate $q = 0.272 \pm 0.018$. It is interesting that, regardless of the prior distribution parameters, the values for q sampled from the posterior distribution are all within an interval of length about 0.1. Thus, crudely speaking, for the

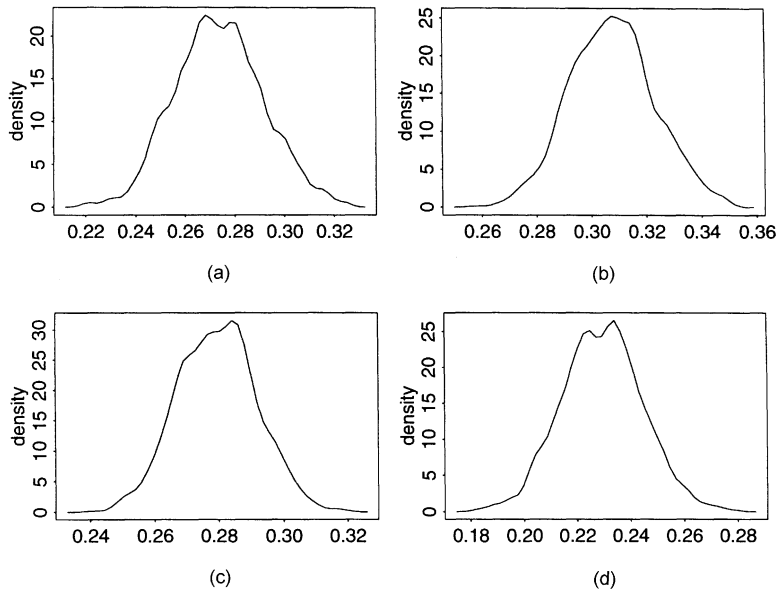

Fig. 1. Posterior distribution for q under different prior specifications: (a) $\alpha = 2$, $\delta = 3$; (b) $\alpha = 100$, $\delta = 150$; (c) $\alpha = 200$, $\delta = 300$; (d) $\alpha = 20$, $\delta = 180$

parameters that we have considered the variance of the posterior distribution seems to be largely insensitive to the choice of prior distribution.

3. General epidemics

We now turn our attention to the most widely studied epidemic model, namely the general stochastic epidemic, defined below. This basic model has several variants, one of which (including births into the population and latent periods) has been considered by Gibson and Renshaw (1998). Specifically, they showed how reversible jump MCMC methods can be employed to perform Bayesian inference given data on births and removals.

The general stochastic epidemic is defined as follows (see, for example, Bailey (1975), p. 88). Consider a population consisting initially of N susceptible and a infective individuals, and denote by X_t and Y_t respectively the numbers of susceptible and infective individuals in the population at time $t \geqslant 0$. The model can be defined in terms of Markov transition rates, so in particular the probabilities of an infection and a removal during a time interval $[t, t + \delta t)$ are respectively $\beta X_t Y_t + o(\delta t)$ and $\gamma Y_t + o(\delta t)$. The epidemic continues until there are no more infective individuals left circulating in the population. From now on, we shall assume that $a = 1$.

In the following, we shall suppose that the observed data consist of a set of removal times, so our approach will involve treating the unobserved infection times as parameters of the model. We adopt the following notation. Let the observed removal times be $\tau_1 = 0$, $\tau_2, \ldots, \tau_n$, these being the times of all the removals during $[0, T]$ where $T > 0$, and write $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_n)$. Let I_1 denote the unobserved time of the first infection, so that with probability 1 $I_1 < 0$, and let $\mathbf{I} = (I_2, I_3, \ldots, I_m)$ denote the remaining unobserved infection times during $(I_1, T]$. If the epidemic is known to have ceased, we must have $m = n$; in general, $n \leqslant m \leqslant N$.

We now define an algorithm for sampling from the desired posterior distribution. In contrast with the Reed–Frost model we shall not be able to use a pure Gibbs sampling approach, but instead we employ a single-component Metropolis–Hastings algorithm. To begin with, we shall describe in detail how to sample from the appropriate conditional distributions. First, note that, conditionally on β, γ and I_1, the density of $(\boldsymbol{\tau}, \mathbf{I})$ is given by

$$f(\boldsymbol{\tau}, \mathbf{I}|\beta, \gamma, I_1) = \prod_{i=1}^{n} \gamma Y_{\tau_j-} \prod_{j=2}^{m} \beta X_{I_j-} Y_{I_j-} \exp\left\{ -\int_{I_1}^{T} (\beta X_t Y_t + \gamma Y_t)\, \mathrm{d}t \right\}, \tag{3.1}$$

where the notation τ_j- denotes the left limit, so for example X_{τ_j-} denotes $\lim_{s\uparrow\uparrow\tau_j-}(X_s)$. We shall suppose that β and γ have, *a priori*, gamma distributions with parameter sets $(\lambda_\beta, \nu_\beta)$ and $(\lambda_\gamma, \nu_\gamma)$ respectively. It then follows from equation (3.1) that, using Γ to denote the gamma distribution,

$$\pi(\beta|\boldsymbol{\tau}, \mathbf{I}, I_1, \gamma) \sim \Gamma\left(\lambda_\beta + \int_{I_1}^{T} X_t Y_t\, \mathrm{d}t,\, m - 1 + \nu_\beta\right),$$

$$\pi(\gamma|\boldsymbol{\tau}, \mathbf{I}, I_1, \beta) \sim \Gamma\left(\lambda_\gamma + \int_{I_1}^{T} Y_t\, \mathrm{d}t,\, n + \nu_\gamma\right).$$

Next, suppose that I_1 has prior density given by $\theta \exp(\theta y)\, I(y < 0)$, where $\theta > 0$, and where $I(\cdot)$ is the indicator function. For $n \geqslant 2$ we must have $I_2 < 0 = \tau_1$, since otherwise the epidemic would cease after the first removal. It then follows from equation (3.1) and the fact that $I_1 < I_2$ that the density of I_1 conditional on $\boldsymbol{\tau}$, $\mathbf{I}$, β and γ is given by

$$g(y|\boldsymbol{\tau}, \mathbf{I}, \beta, \gamma) = \Lambda \exp\{-\Lambda(I_2 - y)\}, \qquad y \in (-\infty, I_2),$$

where $\Lambda = \theta + \gamma + \beta N$.

It now only remains to find a way of sampling from the distribution $\pi(\mathbf{I}|\boldsymbol{\tau}, I_1, \beta, \gamma)$. Although this is problematical to do directly, we can proceed by using a Hastings algorithm with three possible moves, as described below. We shall abbreviate $f(\boldsymbol{\tau}, \mathbf{I}|\beta, \gamma, I_1)$ by $f(\mathbf{I})$, the point here being that $\boldsymbol{\tau}$, β, γ and I_1 are all fixed for the purposes of the algorithm. As before, $\mathbf{I}$ will denote the current set of infection times, and m the size of this set. Further, we shall write $\mathbf{I} - \{s\}$ to denote the set of infection times $\mathbf{I}$ with s excluded, and $\mathbf{I} + \{s\}$ to denote the set of infection times $\mathbf{I}$ with another infection at s.

The three moves of the algorithm are as follows.

(a) Moving an infection time: here one of the existing infection times is chosen uniformly at random, a replacement candidate t is obtained by sampling uniformly on (I_1, T) and the new point is accepted with probability

$$\frac{f(\mathbf{I} - \{s\} + \{t\})}{f(\mathbf{I})} \wedge 1.$$

(b) Removing an infection time: here an infection time is randomly selected, s say, and then removed with probability

$$\frac{f(\mathbf{I} - \{s\})m}{f(\mathbf{I})(T - I_1)} \wedge 1.$$

(c) Adding a new infection time: here a point t is uniformly sampled from (I_1, T) and added to $\mathbf{I}$ with probability

$$\frac{f(\mathbf{I}+\{t\})(T-I_1)}{f(\mathbf{I})(m+1)} \wedge 1.$$

In each case, whenever a proposed move is rejected, the process remains in its current state. Note that, in the case where the epidemic is known to be complete, the only move that is permissible is the first, since the number of infections must always be n.

The computer implementation of this algorithm was achieved in a manner essentially identical with that described in Section 2 for the Reed–Frost case. Of course in this case there is a greater number of variables, which allied to the need to calculate integrals numerically (such as those in equation (3.1), for example) resulted in longer computation times. However, the run times were still only of the order of minutes (e.g. under 10 min for a sample of 1000 taken every 100 cycles after a burn-in time of 10000 cycles).

We now apply our algorithm to two data sets, one of which was obtained by simulating a general stochastic epidemic, and the other coming from a real life disease outbreak.

3.1. Simulated data

We begin by considering some simulated data. Our purpose in applying our algorithm here is primarily practical, since it allows us to see how the algorithm behaves in terms of computation time and convergence characteristics.

A general stochastic epidemic with parameters $\beta = 0.12$, $\gamma = 1$ and $(N, a) = (9, 1)$ was simulated and the following removal times obtained (where time is rescaled so that $t_1 = 0$):

$$t_1 = 0, \qquad t_2 = 1.52292, \qquad t_3 = 1.55004, \qquad t_4 = 1.93064, \qquad t_5 = 2.67492.$$

We shall use these data to consider two scenarios: first, that the epidemic is still in progress (so that we fix T to be less than t_5) and, second, that the epidemic is known to be completed. In both cases, prior parameters were set to be $(\lambda_\beta, \nu_\beta) = (1, 0.1)$, $(\lambda_\gamma, \nu_\gamma) = (0.1, 0.1)$ and $\theta = 0$. Thus the means of the prior distributions for β and γ were 0.1 and 1 respectively, while the prior distribution for I_1 is non-informative.

(a) *Epidemic in progress*: we set $T = 1.7$, so the observed data are $\{t_1, t_2, t_3\}$. Fig. 2 shows posterior distributions for β and γ. The posterior means of β and γ are 0.105 and 0.446 respectively, and their respective variances are 0.007 and 0.185.

(b) *Completed epidemic*: we now set $T = 3$ and assume that the epidemic is completed, so that $m = n$ at all times. Fig. 3 shows posterior distributions for β and γ. The posterior means of β and γ are 0.098 and 0.780 respectively, and their respective variances are 0.004 and 0.203.

Comparing the results for the completed epidemic case with the epidemic in progress case gives us some feel for the effect of extra information on the parameter estimates. In this case the extra information does not appear to affect the posterior distribution of β greatly, although the mean of the posterior distribution of γ has increased considerably. There is some crude intuition to support this, as follows. By time $T = 1.7$ we have merely seen three removals, and thus there have been at least two infections, and possibly as many as nine. However, the extra information that there are only two more removals before completion leads us to conclude that the total number of infections is only 4. Since the rate at which removals occurs is γY, it follows that the estimate of γ must increase to account for the observed removals. Conversely, the XY-term in the infection rate is less sensitive to increases in Y owing to infection, since infections also cause X to decrease.

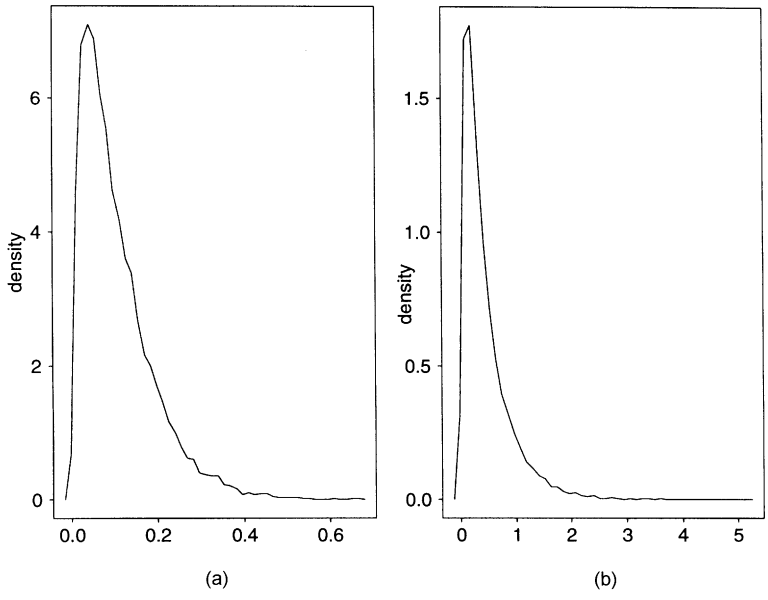

Fig. 2. Marginal posterior distributions of (a) β and (b) γ for the simulated data and the epidemic in progress

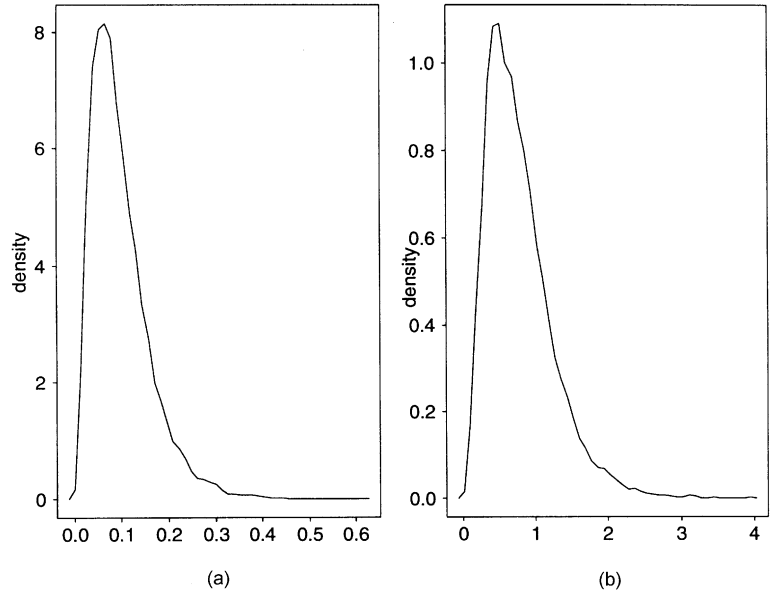

Fig. 3. Marginal posterior distributions of (a) β and (b) γ for the completed epidemic case with simulated data

3.2. Data from smallpox outbreak

We now consider a data set obtained from a smallpox outbreak in a closed community of 120 individuals in Abakaliki, Nigeria (see Bailey (1975), p. 125). The use of the general stochastic epidemic to model smallpox data is not entirely appropriate, since the real life disease has an appreciable latent period. However, these data can be used to illustrate our method, and

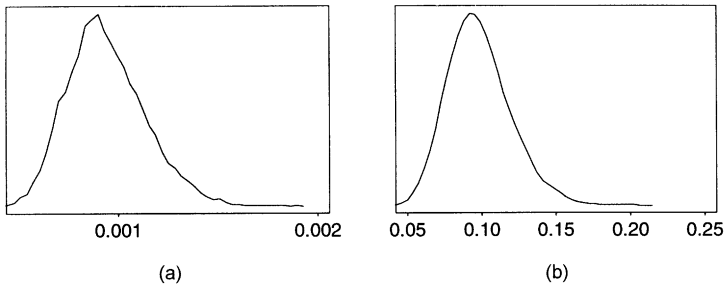

Fig. 4. Marginal posterior distributions of (a) β (weak prior) and (b) γ (weak prior) for the smallpox outbreak data

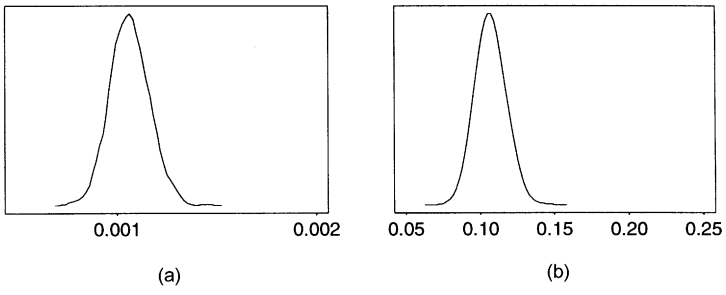

Fig. 5. Marginal posterior distributions of (a) β (medium prior) and (b) γ (medium prior) for the smallpox outbreak data

furthermore we can compare our results with those obtained by using a maximum likelihood approach. With reference to maximum likelihood estimation, the figures given in Bailey (1975), p. 125, are known to be inaccurate; however, Frank Ball (private communication) has obtained maximum likelihood estimates of $\beta = 0.0008254$ and $\gamma = 0.087613$ by using a new method, the details of which will appear elsewhere.

The data themselves consist of the following 29 interremoval times, measured in days:

$$13, 7, 2, 3, 0, 0, 1, 4, 5, 3, 2, 0, 2, 0, 5, 3, 1, 4, 0, 1, 1, 1, 2, 0, 1, 5, 0, 5, 5.$$

A 0 in these times corresponds to the appearance of two cases on the same day. Thus $t_1 = 0$, $t_2 = 13$, $t_3 = 20$ and so on.

We consider two parameter sets for posterior distributions. First, by setting $\lambda_\beta = \nu_\beta = \lambda_\gamma = \nu_\gamma = \theta = 0$ we obtain completely non-informative priors, and thus we expect the posterior distributions for β and γ to be close to the maximum likelihood estimates. This is indeed the case, as seen in Fig. 4. The posterior means of β and γ are 0.0009 and 0.098 respectively, and their respective variances are 3.8×10^{-8} and 4.3×10^{-4}.

Second, we choose more informative priors by setting $(\lambda_\beta, \nu_\beta) = (10^4, 10)$, $(\lambda_\gamma, \nu_\gamma) = (100, 10)$ and $\theta = 0.1$, so the prior means for β, γ and $-I_1$ were 0.001, 0.1 and 10 respectively. Fig. 5 shows the corresponding posterior distributions. The posterior means of β and γ are 0.0011 and 0.107 respectively, and their respective variances are 1×10^{-8} and 9×10^{-5}.

4. Conclusions

We have shown how MCMC methods can be used to carry out Bayesian inference for the two most widely studied epidemic models given partial data. The approach that we have

adopted is extremely flexible and the methods used can clearly be extended to more elaborate models.

Real life epidemics are inherently stochastic in nature, and stochastic models provide an appropriate way of attempting to describe and analyse such phenomena. The statistical analysis of these models is clearly important from a data analysis viewpoint but also necessary in view of the underlying level of stochasticity of the models (for example, with positive probability an epidemic might fail to take off at all). However, the intractability of these models outside the simplest cases has limited their practical appeal, especially as complete data for the epidemic are rarely available. MCMC methodology enables the analysis of more complicated stochastic epidemic models, even in the presence of incomplete data.

As mentioned above, the framework that we have described for simple models can clearly be extended to rather more complex scenarios. Specifically, these include models with latent periods, models with more general infectious periods and multigroup models, as well as situations involving different kinds of incomplete data. Many of these directions will be pursued and reported elsewhere.

Preliminary results for some of these extensions suggest that some models can run into problems of extremely slow convergence when using natural looking corresponding MCMC algorithms. It is therefore necessary to construct the algorithms with extreme care, and considerable further work is necessary here to design algorithms with reliable convergence properties over a wide range of data sets.

Acknowledgement

We are grateful to a referee for spotting an error in one of our original figures.

References

Bailey, N. T. J. (1975) *The Mathematical Theory of Infectious Diseases and Its Applications*, 2nd edn. London: Griffin.

Becker, N. G. (1993) Parametric inference for epidemic models. *Math. Biosci.*, **117**, 239–251.

———(1995) Statistical challenges of epidemic data. In *Epidemic Models: Their Structure and Relation to Data* (ed. D. Mollison), pp. 339–349. Cambridge: Cambridge University Press.

Gibson, G. J. and Renshaw, E. (1998) Estimating parameters in stochastic compartmental models using Markov chain methods. *IMA J. Math. Appl. Med. Biol.*, **15**, 19–40.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.