Mathematical Biosciences 180 (2002) 103–114

Mathematical Biosciences

www.elsevier.com/locate/mbs

A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods

Philip D. O'Neill [*]

School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK

Received 27 June 2001; received in revised form 6 February 2002; accepted 21 February 2002

Abstract

Recent Bayesian methods for the analysis of infectious disease outbreak data using stochastic epidemic models are reviewed. These methods rely on Markov chain Monte Carlo methods. Both temporal and non-temporal data are considered. The methods are illustrated with a number of examples featuring different models and datasets.

Keywords: Bayesian inference; Epidemics; Stochastic epidemic models; Markov chain Monte Carlo; Metropolis–Hastings algorithm

1. Introduction

This paper was written for the conference on Compartmental Models and Disease Transmission, in memory of John Jacquez, held at the University of Michigan, October 2001. The purpose of the paper is to give an introduction and overview of some recent work concerned with methods for performing Bayesian inference for stochastic epidemic models given data on outbreaks of infectious diseases. Important generic ideas are discussed in the present section, with the bulk of the remainder of the paper containing various illustrative examples.

Models and inference for epidemics: Analysing infectious disease data is a non-standard problem, and many approaches have been developed (see [1] for a recent review). In general, inference problems for disease outbreak data are complicated by the facts that (i) the data are

[*] Tel.: +44-115 951 4949; fax: +44-115 951 4951.
E-mail address: pdo@maths.nott.ac.uk (P.D. O'Neill).

PII: S0025-5564(02)00109-8

inherently dependent and (ii) the data are usually incomplete in the sense that the actual process of infection is not observed. However, it is often possible to formulate simple (to define, if not to analyse!) stochastic models which describe the key features of epidemic spread, and these can be used as a convenient starting point for inference. In particular, such models attempt to describe the mechanism by which the observed data are generated. Inference then proceeds by attempting to estimate, either in a classical or Bayesian framework, the parameters of the model. These parameter estimates can in turn be used to provide information concerning quantities of epidemiological interest.

Depending on the application in question, models may incorporate latent periods, variable infectivity, reduced susceptibility following recovery, etc. Similarly, aspects of population heterogeneity such as age structure, varying susceptibility, differential mixing rates between groups of individuals, etc. can be included as appropriate. As with any statistical modelling, there is a balance between models that are too complicated for the data to fully inform, and those which are too simplistic to be regarded seriously as a basis for generating useful inference. In practice it is not always straightforward to achieve this balance via a formal procedure; issues of model adequacy and goodness-of-fit are not especially well-developed in the literature. Situations in which the data essentially consist of repeated independent observations (e.g. different independent outbreaks) are usually easier to assess than those featuring heavily dependent data (e.g. temporal data from a single large outbreak).

Bayesian inference: In classical inference, model parameters are regarded as fixed quantities. The values of the parameters are estimated from data using estimators that are random variables, and whose distributional properties may be known. In Bayesian inference, the model parameters are regarded as random variables, and the main object of interest is the posterior distribution, i.e. the distribution of the parameters given the data. Specifically, the posterior density is defined via Bayes' Theorem as the normalised product of the prior density and the likelihood. The posterior distribution contains all information about the parameters, from which one can obtain point and interval summaries (e.g. mean, mode, credible intervals). Within the present context, the posterior distribution for a parameter provides information concerning parameter uncertainty that might be very hard to obtain using a classical approach. More precisely, in practice it is often straightforward, using the methods we shall describe, to obtain Bayesian credible intervals for parameters, whereas classical confidence intervals may require the development of appropriate theoretical results. In particular, the usual conditions that ensure asymptotic normality of maximum likelihood estimators are often violated. Regarding prior distributions, the choice of prior is largely application-specific, although it is common to use uninformative priors as part of any analysis to provide a kind of baseline. However it is frequently reasonable to use informative priors based on epidemiological beliefs.

Data imputation: As mentioned above, one of the difficulties when dealing with disease outbreak data is that the infection process is unobserved. The reason that this complicates matters is that the likelihood (that is, the probability density or mass function of the data given the model parameters) may become very difficult to evaluate. This problem is especially acute when considering temporal data, since then evaluating the likelihood typically involves integration over all possible infection times, which is rarely analytically possible. Consequently, it is natural to consider the use of data imputation methods. In these methods, unknown quantities that facilitate likelihood evaluation are simply treated as extra model parameters. Two widely used data imputation methods are the EM algorithm and Markov chain Monte Carlo (MCMC) methods. The EM algorithm has been considered for epidemic inference problems (see e.g. [1,2]), although a drawback with this method is that the evaluation of the expectation step can be rather complicated. Conversely, MCMC methods usually allow data imputation in a straightforward manner.

More generally, one of the issues when using data imputation methods is selecting appropriate quantities to impute. For example, suppose that the available data consist only of the number ultimately infected by an outbreak. For some models, the likelihood of such data can be evaluated directly and no imputation is needed. If this is not the case, then in theory the entire (temporal) sample path of the process could be imputed, although in practice dealing with such a large extra quantity of information, about which very little is known, is likely to be problematic. However, alternative non-temporal information, such as who infects who, might be sufficient to enable the likelihood to be evaluated. In general, the choice of imputed variables is facilitated by a good understanding of the probabilistic structure and behaviour of the model under consideration.

Markov chain Monte Carlo: MCMC methods are an established suite of methodologies that enable samples to be drawn from some target density that is only known up to proportionality. The literature on MCMC methods is vast; the reader new to the subject can find a user-friendly introduction in [3]. In the context of Bayesian inference, the target density is the joint posterior density of the model parameters. The methods work by defining a Markov chain whose stationary distribution is equal to the (normalised) target density. The chain is then simulated for a time deemed adequate for convergence to have occurred, and then samples drawn from the simulated chain. These samples are, provided convergence has occurred, samples from the target density of interest. It should be noted that sampling-based methods such as this allow for straightforward exploration not only of the posterior density of interest, but also marginal posterior densities of the model parameters and functions of these parameters. As an example of the latter, a key quantity of interest in epidemic models is the so-called basic reproduction number, denoted R_0. This quantity can often be regarded as a threshold parameter whose value indicates whether an epidemic is likely to take off, or die out quickly. However, most epidemic models are not defined directly in terms of R_0, and so estimating R_0 is often achieved by first estimating the basic model parameters.

Why use MCMC? There are two related reasons why MCMC is an attractive choice of methodology in the present context. First, it permits a huge amount of modelling flexibility. For example, consider a dataset consisting of symptom-appearance times. Evaluation of the likelihood involves integration over all the possible (unknown) infection times. To proceed analytically it becomes necessary to chose a convenient distributional form for the infectious periods, so that the integrals can be explicitly evaluated. Otherwise, the summation and high-dimensional integration can make the problem at best numerically complicated, and at worst intractable. No such modelling restriction on infectious periods is generally necessary for MCMC, in which the missing infection times are included as extra model parameters. Note that MCMC can be used as a tool for exploring the likelihood surface, so that its use extends beyond the purely Bayesian framework. An additional point regarding model flexibility is that there is no restriction that models be Markov, which is necessary when using martingale techniques for inference [1,4].

The second appeal of MCMC is that, in combination with the Bayesian approach, it enables analysis of all of the model parameters, or any function of them. In particular, this includes parameters or functions of parameters for which no classical estimator is known (see, for example,

[5]). Posterior summaries such as means, medians, variances, credible intervals, etc. are all easily obtained for individual parameters, or for joint distributions of parameters.

Scope and outline of paper: In the remainder of this paper we consider examples that illustrate various kinds of datasets, models and MCMC algorithms. Our focus is primarily towards methodology, and thus we do not consider matters such as the quality of actual data, sampling and design issues, laboratory techniques and so on. Similarly, the examples presented do not contain full statistical analyses for specific datasets. However, more details are usually available elsewhere for the datasets that we discuss, as will be indicated. Additionally, we do not consider here those applications of MCMC to infectious disease data analysis in which the transmission process is not explicitly modelled (e.g. [6–8]).

We consider two broad classes of dataset, so that Section 2 address methods for data that do not specifically include time-indexed information, and Section 3 address methods for temporal data. Some concluding comments are given in Section 4.

2. Non-temporal and semi-temporal data

In this section we consider the situation in which the available data do not explicitly contain information concerning the real-time progress of an epidemic. There are two scenarios. First, the data are entirely non-temporal and consist only of the number of cases among a population of known size. Second, the data consist of so-called chain information, which essentially describes the progress of the epidemic in terms of generations of infection: we refer to such data as semi-temporal. These data are usually obtained via expert opinion; for instance, the times between the appearance of symptoms in infected individuals, considered alongside existing knowledge of latent and infectious period lengths, might strongly suggest a particular infection chain. Of course, such data might be viewed with scepticism because they are not clinically verified. Finally, in both scenarios the data might be stratified in some way, for example according to age groups, households, vaccination status, etc.

2.1. Rhode Island measles data

We begin with an analysis of some classic measles data from Providence, Rhode Island, presented in Table 1. These data were considered by Wilson et al. [9], Bailey [10], [11], and Becker [4], and consist of chain data for measles outbreaks in households of size 3. The analysis in [4] considers a number of different models, all of which regard the households as independent, including a model in which the between-individual infection probability can vary between households. We consider this latter model and analyse it from a Bayesian perspective. The motivation

Table 1
Epidemic chain data from outbreaks of measles in households of size 3, Providence, Rhode Island

Chain	Frequency
$\{1\}$	34
$\{1,1\}$	25
$\{1,2\}$	239
$\{1,1,1\}$	36

here is largely pedagogical, since it is possible to illustrate several important features of the MCMC methodology.

Model: The model is defined as follows. The population consists of N households, labelled $1,\dots,N$, each containing n individuals, a of whom are initially infectious, and the rest susceptible. Within household j, an infected individual has a probability $1-q_j$ of being able to independently infect each other individual in the household, where q_j is a realisation of some random variable Q that takes values in [0,1]. This attempt at infection can only be made once, after which the infective can be thought of as playing no further part in the epidemic. Each newly infected individual can then try to infect remaining susceptibles in the same manner, and the process continues until there are no more active infectives left. If $Y_j(t)$ is the number of infectives in the tth generation of infection, where $t = 0, 1, \dots$, and $\tau = \max\{t : Y_j(t) > 0\}$ is the last generation in the outbreak, then $\{Y_j(0) = a, Y_j(1), \dots, Y_j(\tau)\}$ is called the epidemic chain. Finally, each of the q_js are assumed to be independent realisations of Q.

In our application, $n = 3$ and $a = 1$, and so the only possible epidemic chains are $\{1\}$, $\{1,1\}$, $\{1,2\}$ and $\{1,1,1\}$. Note that these correspond respectively to 1, 2, 3 and 3 cases in total within a household.

Inference: The likelihood is given by

$$\pi(n_1, n_{11}, n_{12}, n_{111}|\theta) = (\mathbb{E}[Q^2])^{n_1}(2\mathbb{E}[(1-Q)Q^2])^{n_{11}}(\mathbb{E}[(1-Q)^2])^{n_{12}}(2\mathbb{E}(1-Q)^2Q])^{n_{111}}, \tag{2.1}$$

where θ denotes the parameters of Q. Note that the expectations on the right-hand side of (2.1) are functions of θ.

Following Becker [4], we assume that Q has a beta distribution with parameters α and β, so that the analysis will seek to make inferences about α and β given the data. The likelihood becomes

$$\begin{aligned}\pi(n_1, n_{11}, n_{12}, n_{111}|\alpha,\beta) = &\left(\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\right)^{n_1}\left(\frac{2\beta\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)(\alpha+\beta+2)}\right)^{n_{11}}\\ &\times\left(\frac{\beta(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)}\right)^{n_{12}}\left(\frac{2\alpha\beta(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)(\alpha+\beta+2)}\right)^{n_{111}}.\end{aligned} \tag{2.2}$$

By Bayes' Theorem, the posterior density of α and β is proportional to the product of the likelihood and the prior distribution on (α,β), i.e.

$$\pi(\alpha,\beta|n_1, n_{11}, n_{12}, n_{111}) \propto \pi(n_1, n_{11}, n_{12}, n_{111}|\alpha,\beta)\pi(\alpha,\beta) = h(\alpha,\beta),$$

say. Thus $h(\alpha,\beta)$, when normalised, is the target density of interest.

MCMC algorithm: A Markov chain $\{X_n : n = 0, 1, \dots\}$ with the required target density as its stationary distribution can be defined using a Metropolis–Hastings algorithm (see e.g. [3]), as follows. First, initial values α_0 and β_0 are chosen, so that $X_0 = (\alpha_0, \beta_0)$. These values can be arbitrary, provided $h(\alpha_0,\beta_0) > 0$. Given $X_n = (\alpha_n, \beta_n)$, X_{n+1} is defined as follows. New values for α and β are proposed according to some proposal density $g(\alpha,\beta|\alpha_n,\beta_n)$. With probability

$$p_{\mathrm{acc}}(\alpha,\beta|\alpha_n,\beta_n) = \frac{h(\alpha,\beta)g(\alpha_n,\beta_n|\alpha,\beta)}{h(\alpha_n,\beta_n)g(\alpha,\beta|\alpha_n,\beta_n)} \wedge 1$$

the new values are accepted and $X_{n+1} = (\alpha,\beta)$; otherwise, $X_{n+1} = (\alpha_n,\beta_n)$. The quantity $p_{\mathrm{acc}}(\cdot|\cdot)$ is known as the acceptance probability.

The algorithm just described updates the values of α and β together (sometimes called *block updating*), although it is also possible to update each separately, with a separate proposal and acceptance probability for each. Note also that the choice of the proposal density g is essentially arbitrary, although in practice a careful choice will help the algorithm to move quickly around the parameter space (in this case, $\{(\alpha,\beta):\alpha,\beta>0\}$).

Results: The above algorithm was implemented using independent Gaussian proposal densities for α and β, centred on the current values, each with variance $\sigma^2 = 0.01$ (note that $h(\alpha,\beta)$ is defined as zero if either α or β is negative). Both α and β were assigned gamma-distributed priors with mean 1 and variance 1000, which are esentially uninformative in comparison to the data. Table 2 contains posterior summaries for α and β, based on a sample of 10 000 points from the output of the Markov chain. As expected, the results are in harmony with the maximum likelihood estimates in [4], namely $\hat{\alpha} = 0.264$ and $\hat{\beta} = 1.091$. It is straightforward, and illuminating, to consider posterior distributional information. Fig. 1 contains a scatterplot of α and β based on the sample of 10 000 points, and illustrates clearly the posterior correlation between the two parameters. On the basis of this plot it seems reasonable to suppose that, for example, the data are more informative about $\mathbb{E}[Q] = \alpha/(\alpha+\beta)$ than α or β alone. This is readily investigated by using

Table 2
Posterior mean, median, standard deviation and equal-tailed 95% credible interval for α and β

	α	β
Mean	0.276	1.143
Median	0.266	1.104
Standard deviation	0.071	0.293
95% CI	(0.180, 0.404)	(0.743, 1.679)

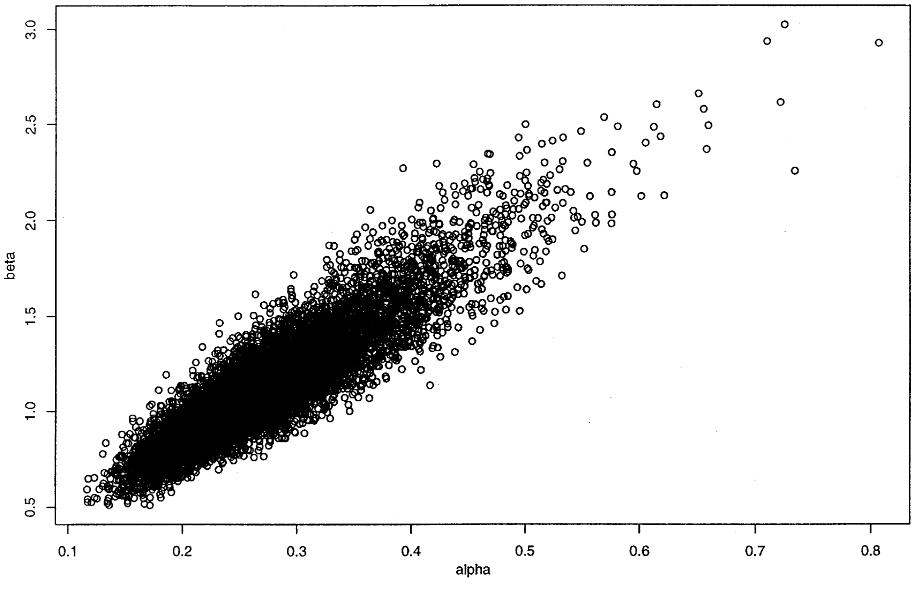

Fig. 1. Pair-wise scatterplot of α and β.

the α and β samples to derive a corresponding sample for $\mathbb{E}[Q]$. From this sample $\mathbb{E}[Q]$ was found to have mean 0.195, and 95% equal-tailed credible interval (0.166, 0.225), the latter indicating the expected increase in precision relative to α and β.

We conclude by considering ways in which the methodology can be generalised.

Missing data: If the data consisted only of the total numbers of cases in each household, so that $n_{111} + n_{12}$ was observed instead of both numbers, then n_{111} (or n_{12}) could be treated as an extra parameter, and updated using some discrete proposal distribution in the same manner as for α and β. In particular, the likelihood expression (2.2) could then be utilised. This approach is described in more detail in [12].

Other distributions for Q: An important aspect of MCMC methods in the context of epidemic modelling is that they often permit far more modelling flexibility than other exisiting methods. In the present example, the use of a Beta distribution for Q is particularly convenient because it yields explicit expressions for the expectations in (2.1). For other choices of distribution this may not be the case, and in particular this can make maximum likelihood estimation far less tractable. However, we can proceed by including each of the q_js as extra model parameters. More specifically, label the N households so that for $j = 1, \dots, N$, each household j has a known epidemic chain y_j and a household probability parameter q_j. The likelihood $\pi(y_1, \dots, y_N | q_1, \dots, q_N) = \prod_{j=1}^{N} \pi(y_j | q_j)$ is straightforward to write down. By Bayes' Theorem we have

$$\pi(\theta, q_1, \dots, q_N | y_1, \dots, y_N) \propto \pi(y_1, \dots, y_N | q_1, \dots, q_N, \theta) \pi(q_1, \dots, q_N | \theta) \pi(\theta),$$

where $\pi(q_1, \dots, q_N | \theta)$ is simply a product of the density function of $Q(\theta)$ evaluated at each of the q_js, and $\pi(\theta)$ is the prior on θ. Note that the posterior density is now augmented to include the extra parameters $q_1, \dots, q_N$. In principle this huge increase in the parameter space does not create a problem for MCMC methods, although in practice careful implementation is necessary to avoid problems of poor algorithm convergence. An MCMC algorithm for this set-up is obtained by simply including updating steps for each of the q_js as well as θ.

2.2. Influenza data in a community of households

We now illustrate how MCMC methods can be used to evaluate data consisting of numbers of cases in a community of households of varying sizes. Such a dataset, taken from outbreaks of influenza in Tecumseh, Michigan, is presented and analysed in [13], and has been considered in a number of subsequent papers. Specifically, the dataset is of the form $\mathscr{D} = \{n_{ij}\}$, where n_{ij} is the number of households containing $i \geqslant 1$ individuals in which $0 \leqslant j \leqslant i$ become infected during the epidemic. Various possible modelling approaches for these data have been considered. First, as in the previous example, it is possible to concentrate simply on within-household dynamics. Second, modelling infection from the community at large in a simple way can be achieved by assuming that every individual escapes infection from outside the household with some fixed probability. In conjunction with the existing within-household process, this gives a two-parameter model, but retains the mathematical convenience of households that act independently of one another. This approach was first considered by Longini and Koopman [14]. A more realistic but analytically less tractable model is to allow two levels of mixing, corresponding to within-household (local) and between-household (global) infections. This approach is described in [15], using a model defined in continuous time.

Longini–Koopman model: An extension of the Longini–Koopman model is considered using MCMC methods in [16], and we now outline the approach used there. The population is assumed to be divided into households. Each individual in the population avoids infection from the community outside its household with probability q_c, and is immune to the disease with probability v. This immunity parameter is included in the analysis to allow for some simple population heterogeneity. The within-household dynamics are identical to those described for the model in the previous section, although now the probability of avoiding infection, q_h, is the same in all households. Finally, all households are assumed to be independent of one another.

The posterior density of (q_c, q_h, v) given $\mathscr{D}$ satisfies

$$\pi(q_c, q_h, v|\mathscr{D}) \propto \pi(q_c, q_h, v) \prod_{i,j} [P(T_i = j)]^{n_{ij}}, \tag{2.3}$$

where T_i is the total number ultimately infected in a household containing i individuals and $\pi(q_c, q_h, v)$ is the prior on (q_c, q_h, v). Calculating the probability mass function of T_i is relatively straightforward for moderately sized households and can be achieved in several ways, as described in [16]. A simple Metropolis–Hastings algorithm for sampling from $\pi(q_c, q_h, v|\mathscr{D})$ is easily defined along the lines described in Section 2.1. The three parameters can either be updated individually, or in blocks. The acceptance probability is calculated using (2.3). The results are detailed in [16], and one notable outcome is that, in the absence of strong prior information, there are strong posterior correlations between the three model parameters. This is largely because the data are insufficient to clearly distinguish between different possible sets of parameter values. If stronger prior information is used then the joint posterior density of the parameters is correspondingly more concentrated.

Two-level mixing model: Consider now a population divided into households. Suppose that each individual in the population avoids infection from a given infective in its own household with probability q_L, and also from a given infective anywhere in the population with probability q_G, with the usual assumptions of independence applying. This model corresponds, in terms of final outcome, to a special case of a continuous-time model described in [15]. Unfortunately, the likelihood $\pi(\mathscr{D}|q_L, q_G)$ is practically intractable unless there are very few households. This is because evaluating this likelihood involves, either explicitly or implicitly, consideration of all possible ways in which the observed infections can have arisen, and the possibility of between-household infections makes this calculation highly involved. This difficulty suggests that some form of data imputation could be of use. One possibility that appears worthy of consideration is to impute an underlying random graph [15] which essentially describes who each individual would succeed in infecting if all other individuals were susceptible. In particular, given this information it is possible to write down a likelihood. The practical challenge in formulating an MCMC algorithm is to update the graph in an efficient manner. This is the subject of current research.

3. Temporal data

In this section we consider the use of MCMC methods to facilitate inference given temporal outbreak data. In practice, at least for human diseases, such data typically consist of case-detection times. The times at which infections actually occurred are invariably unknown, and for most situations this makes calculation of the likelihood infeasible. However, if the infection times are known then the likelihood usually becomes tractable, and consequently the unknown infection times are natural candidates for imputation.

3.1. Smallpox outbreak data

We begin by considering methodology motivated by a well-known dataset compiled from an outbreak of smallpox in Abakaliki, Nigeria and made available by Drs D.M. Thompson and W.H. Foege [17]. The data consist of 29 inter-removal times between detection of cases, measured in days. These data have been considered by a number of authors (e.g. [4,11,12,18]), generally using simplified models, and under the assumption that the disease spread through a closed population consisting of 120 individuals.

Here we describe the methodology used by O'Neill and Becker [19]. As well as illustrating how MCMC methods can be used for datasets of this kind, this example also demonstrates the level of modelling possible using MCMC. In particular, likelihood methods would be highly complex for the model we are about to describe.

Model: The model described below assumes that (i) individuals can have varying susceptibility to infection; (ii) when infected, an individual undergoes a fixed-length latent period, during which time it is not infectious; (iii) an individual's infectious period is assumed to be distributed according to a gamma distribution with unknown parameters; (iv) following its infectious period, an individual acquires immunity (this is known as the removal of the individual). These assumptions are not unreasonable for smallpox, which has both a latent period of appreciable length, and long-term immunity following recovery. The non-random latent period in assumption (ii) is apparently restrictive, but the lack of detail in the data makes it hard to reliably account for variation in both the latent and infectious periods separately.

Denoting the numbers of susceptible, latent, infective and removed individuals at time t by $S(t)$, $L(t)$, $I(t)$ and $R(t)$ respectively the process can be represented by the compartmental diagram

$$S(t) \to L(t) \to I(t) \to R(t).$$

The population is assumed to initially consist of $N-1$ susceptibles and one infective. The data consist of a set of removal times $\boldsymbol{r} = \{r_1, \ldots, r_n\}$, where $r_1 = 0 \leqslant r_2 \leqslant \cdots \leqslant r_n = T$. Individual j, who is removed at time r_j, is infectious from time $i_j < r_j$, and the infectious period $r_j - i_j$ is assumed to be distributed according to a $\mathrm{Gam}(\gamma, \delta)$ distribution. Individual j is assumed to have been initially infected at time l_j, so that $i_j - l_j = c$ is the length of the latent period. It is assumed that c is known. Let κ denote the label of the initial infective, so that $i_\kappa = \min\{i_1, \ldots, i_n\}$, and define $\boldsymbol{i}$ as the set $\{i_j : j \neq \kappa\}$.

A given susceptible, j say, receives an infection intensity $\tilde{u}_j$ from each currently infectious individual, which is to say that j is infected at the first point of a Poisson process with (random) intensity at time $t > i_\kappa$ given by $\tilde{u}_j I(t)$. The intensity $\tilde{u}_j$ is sampled from a $\mathrm{Gam}(\alpha, \beta)$ distribution, and independent of intensities for other susceptibles. The epidemic ceases as soon as there are no more latent or infectious individuals left in the population.

For $j = 1, \ldots, n$, let u_j denote the susceptibility of the jth individual to become infected, so that $u_j = \tilde{u}_k$ for some k. For convenience we also define u_j for $j = n+1, \ldots, N$ so that the set of such

u_js is equal to the set of susceptibilities of those individuals who are never infected. Finally, define $\boldsymbol{u}$ as the set $\{u_1, \dots, u_N\}$.

Inference: As show in [19], the augmented likelihood function is given by

$$\pi(\boldsymbol{r}, \boldsymbol{i} | \alpha, \beta, \gamma, \delta, i_\kappa) = \left(\frac{\alpha^{n-1} \beta^{\alpha(N-1)}}{[\beta + W_T(\boldsymbol{i})])^{\alpha(N-n)}} \right) \left(\prod_{j=1, j \neq \kappa}^{n} \frac{I(l_j -)}{[\beta + W_j(\boldsymbol{i})]^{\alpha+1}} \right) \left(\prod_{j=1}^{n} f(r_i - i_j) \right), \tag{3.1}$$

where

$$W_j(\boldsymbol{i}) = \int_{i_\kappa}^{l_j} I(u) \, \mathrm{d}u, \quad W_T(\boldsymbol{i}) = \int_{i_\kappa}^{T} I(u) \, \mathrm{d}u,$$

f is the density function of a gamma random variable with shape and scale parameters γ and δ, respectively, and $I(\tau -) = \lim_{t \uparrow \tau} I(t)$. Note that in the likelihood in (3.1), the first two terms account for infections, and the third accounts for the infectious periods.

By Bayes' Theorem, the target density is

$$\pi(\alpha, \beta, \gamma, \delta, \boldsymbol{i}, i_\kappa | \boldsymbol{r}) \propto \pi(\boldsymbol{r}, \boldsymbol{i} | \alpha, \beta, \gamma, \delta, i_\kappa) \pi(\alpha, \beta, \gamma, \delta, i_\kappa),$$

where $\pi(\alpha, \beta, \gamma, \delta, i_\kappa)$ is the prior density on $(\alpha, \beta, \gamma, \delta, i_\kappa)$.

MCMC algorithm: A general MCMC algorithm for sampling from the required posterior density is as follows (a specific algorithm is described in detail in [19]). The parameters α, β, γ and δ can be updated using a Metropolis–Hastings step along the lines described in Section 2.1. To update the infection times $\boldsymbol{i}$ and i_κ, a more involved Metropolis–Hastings step is used. More precisely, an infection time is chosen uniformly at random, and a proposed replacement drawn from some specified proposal distribution. Note that it is sensible here to use a proposal distribution that avoids zero-density values (e.g. any time after the individual is removed) and rarely proposes low-density values (e.g. such that the resulting infectious period is very unlikely). The acceptance probability can then be calculated using the target density and proposal density in the usual way. Note also both i_κ and κ can be updated in this procedure.

Related work: The MCMC algorithm described above generalises one described in [12] for a simpler Markov model in which there are no latent periods, and the infectious period has an exponential distribution. However, there it is not assumed that the epidemic has ceased, and thus the number of unknown infections is itself unknown. Accordingly, the mechanism by which the infection times are imputed allows the creation and deletion of infection times. Hawakaya et al. [20] apply similar methodology to a multitype epidemic model in which individuals are grouped according to their susceptibility. Gibson [21] describes related MCMC methods for inference for a spatial epidemic model in continuous time. Gibson and Renshaw [22] address more general problems of inference for stochastic compartmental models, focussing in particular on epidemics.

3.2. Random social structure

There has recently been some interest in models in which the social structure of the population is itself unknown (e.g. [23]). Inference for the simplest example of such is model is considered in

[5], this model being defined as follows. The population consists of N individuals, one of whom is initially infectious and the rest susceptible. Any two individuals have regular social contact with probability p, with independence assumed between all pairs. Thus, these contacts are described by a Bernoulli random graph with parameter p. It is assumed that the epidemic can only propagate among individuals in regular social contact, and thus an epidemic process is defined on a realisation of the graph as follows. Each infectious individual remains so for a period of time having an exponential distribution with mean γ^{-1}. During its infectious period, an infective individual has infectious contacts with each of its neighbours in the graph at the points of a Poisson process of rate β, with independence assumed between different neighbours. Each infectious contact results in the infected individual immediately becoming infectious. The epidemic ceases as soon as there are no more infectives in the population.

The essential ideas of the algorithm in Section 3.1 (e.g. Metropolis–Hasting steps for updating parameter values and unknown infection times) apply to the current model, as described in [5]. However, the realisation of the underlying random graph for social structure also becomes a parameter, and as such it is necessary to find ways of updating this graph. In this particular model, it is not hard to compute the probabilities of edges appearing conditional on the data and other parameter values, thus providing a natural proposal distribution. More generally, efficiently updating conditioned graph structures is a non-trivial problem. Another feature of using graphs as parameters is that their posterior distributions can, in principle, be explored. Exactly how to represent such posterior information is not always clear, although certain summary statistics (e.g. the number of edges) can be used as a first step.

4. Concluding remarks

In this paper we have described some of the key ideas relating to the implementation of Bayesian inference for stochastic epidemics using MCMC methods. Broadly speaking, the methodology is highly flexible and permits consideration of a wide range of models. However, as with any application of MCMC, the existence of an algorithm does not necessarily mean that it is efficient or practical. In the context of epidemic models, missing data and posterior correlation structures can lead to situations where algorithm convergence does not occur in reasonable time (see e.g. [16]). Such difficulties are often made worse when considering large population sizes, containing hundreds or thousands of individuals. In such cases it can sometimes be profitable to consider using approximation results (e.g. central limit theorems for the final size distribution, see e.g. [24]). However, problems can arise in that these results may break down for some sets of parameter values in the posterior density of interest. Developing methods of Bayesian inference for large population models therefore remains an important open problem.

Acknowledgements

The author is pleased to acknowledge helpful comments from two anonymous referees that have improved the content of this paper.

References

[1] N.G. Becker, T. Britton, Statistical studies of infectious disease incidence, J. Roy. Stat. Soc., Series B 61 (Part 2) (1999) 287.

[2] N.G. Becker, Uses of the EM algorithm in the analysis of data on HIV/AIDS and other infectious diseases, Stat. Meth. Med. Res. 6 (1997) 24.

[3] W.R. Gilks, S. Richardson, D.J. Spiegelhalter (Eds.), Markov Chain Monte Carlo in Practice, Chapman and Hall, London, 1995.

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

[5] T. Britton, P.D. O'Neill, Statistical inference for stochastic epidemics in populations with random social structure. Scand. J. Stat., in press.

[6] W.R. Gilks, D. DeAngelis, Projecting the AIDS epidemic in England and Wales: a Bayesian approach, Statistica Applicata 8 (1996) 83.

[7] E.C. Marhsall, A. Frigessi, N.-C. Stenseth, M. Holden, V.S. Ageyev, Plague in Kazakhstan: a Bayesian model for the temporal dynamics of a vector-transmitted infectious disease, J. Am. Stat. Assoc., in press.

[8] L. Knorr-Held, H. Rue, On block updating in Markov random field models for disease mapping, Scand. J. Stat., in press.

[9] E.B. Wilson, C. Bennett, M. Allen, J. Worcester, Measles and scarlet fever in Providence, RI, 1929–34 with respect to age and size of family, Proc. Am. Philos. Soc. 80 (1939) 357.

[10] N.T.J. Bailey, The use of chain-binomials with a variable chance of infection for the analysis of intra-household epidemics, Biometrika 40 (1953) 279.

[11] N.T.J. Bailey, The Mathematical Theory of Infectious Diseases and its Applications, second ed., Griffin, London, 1975.

[12] P.D. O'Neill, G.O. Roberts, Bayesian inference for partially observed stochastic epidemics, J. Roy. Stat. Soc., Series A 162 (Part 1) (1999) 121.

[13] C.L. Addy, I.M. Longini, M. Haber, A generalized stochastic model for the analysis of infectious disease final size data, Biometrics 47 (1991) 961.

[14] I.M. Longini, J.S. Koopman, Household and community transmission parameters from final size distributions of infections in households, Biometrics 38 (1982) 115.

[15] F.G. Ball, D. Mollison, G.-P. Scalia-Tomba, Epidemics with two levels of mixing, Ann. Appl. Prob. 7 (1997) 46.

[16] P.D. O'Neill, D.J. Balding, N.G. Becker, M. Eerola, D. Mollison, Analyses of infectious disease data from household outbreaks by Markov Chain Monte Carlo methods, Appl. Stat. 49 (2000) 517.

[17] N.T.J. Bailey, A.S. Thomas, The estimation of parameters from population data on the general stochastic epidemic, Theor. Populat. Biol. 2 (1971) 53.

[18] N.G. Becker, P. Yip, Analysis of variation in an infection rate, Aust. J. Stat. 25 (1989) 191.

[19] P.D. O'Neill, N.G. Becker, Inference for an epidemic when susceptibility varies, Biostatistics 2 (2001) 99.

[20] Y. Hawakaya, D. Upton, P.S.F. Yip, P.D. O'Neill, Bayesian inference for an epidemic model with several kinds of susceptibles and unknown population size, submitted for publication.

[21] G.J. Gibson, Markov chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology, Appl. Stat. 46 (1997) 215.

[22] G.J. Gibson, E. Renshaw, Estimating parameters in stochastic compartmental models using Markov chain methods, IMA J. Math. Appl. Med. Biol. 15 (1998) 19.

[23] H. Andersson, Epidemic models and social networks, Math. Scientist 24 (1999) 128.

[24] H. Andersson, T. Britton, Stochastic Epidemic Models and Their Statistical Analysis, Springer Lecture Notes in Statistics, New York, 2000.