Analyses of Infectious Disease Data from Household Outbreaks by Markov Chain Monte Carlo Methods

Philip D. O'Neill
Division of Statistics and O.R.
Department of Mathematical Sciences
University of Liverpool
Liverpool L69 3BX
England

PDONeill@liverpool.ac.uk
Tel: (+44) 151 794 4737
Fax: (+44) 151 794 4754

David J. Balding
Department of Applied Statistics
University of Reading
PO Box 240
Reading RG6 6FN
England

D.J.Balding@reading.ac.uk
Tel: (+44) 118 931 8021
Fax: (+44) 118 975 3169

Niels G. Becker
Department of Statistical Science
La Trobe University
Bundoora 3083
Australia

N.Becker@latrobe.edu.au
Tel: (+61) 3 9479 2606
Fax: (+61) 3 9479 2466

Mervi Eerola
National Public Health Institute
Rolf Nevanlinna institute
P.O.BOX 4
00014 University of Helsinki
Finland

Mervi.Eerola@rni.helsinki.fi
Tel: (+358) 9 191 22773
Fax: (+358) 9 191 22779

Denis Mollison
Department of Actuarial Mathematics and Statistics
Heriot-Watt University
Edinburgh EH14 4AS
Scotland

D.Mollison@ma.hw.ac.uk
Tel: (+44) 131 451 3200
Fax: (+44) 131 451 3249

Summary
The analysis of infectious disease data presents challenges arising from the dependence in the data and the fact that only part of the transmission process is observable. These difficulties are usually overcome by making simplifying assumptions. This paper explores the use of Markov chain Monte Carlo (MCMC) methods for the analysis of infectious disease data, with the hope that they will permit analyses to be made under more realistic assumptions. Two important kinds of data sets are considered, containing temporal and non-temporal information respectively, from outbreaks of measles and influenza. Stochastic epidemic models are used to describe the processes that generate the data. MCMC methods are then employed to perform inference in a Bayesian context for the model parameters. The MCMC methods used include standard algorithms, such as the Metropolis-Hastings algorithm and the Gibbs sampler, as well as a new method that involves likelihood approximation. It is found that standard algorithms perform well in some situations, but can exhibit serious convergence difficulties in others. The inferences that we obtain are in broad agreement with estimates obtained by other methods where they are available. However, we are also able to provide inferences for parameters which have not been reported in previous analyses.

Keywords: BAYESIAN STATISTICS; EPIDEMIC DATA; STOCHASTIC EPIDEMIC MODELS; LIKELIHOOD APPROXIMATION; MARKOV CHAIN MONTE CARLO METHODS; METROPOLIS-HASTINGS ALGORITHM; GIBBS SAMPLER; MISSING DATA

1 Introduction

When analysing infectious disease data it is usually desirable to use models which attempt to describe the way in which the data were generated. Such models help the analyses to be focused on the epidemiological quantities of major interest and thereby promote a deeper understanding of the infection process. However, the task of fitting such models to data is hampered by the fact that parts of the infection process are not observed, which has the consequence that the likelihood function involves multiple integrals and is usually intractable. Sometimes only the eventual number of cases is observed, in which case only certain parameters of the transmission model are estimable, and maximum likelihood estimation is difficult because the probability distribution of the eventual size of the outbreak can be tedious to compute even for small groups of individuals. At best we observe the times at which each of the infected individuals are detected. However, the likelihood function is typically very complicated for such data since neither the times of infection nor the times when the infectious periods start are observed, and each unobserved time gives rise to an additional integration in the expression for the likelihood.

There have been substantial advances recently in the development of methods for the

analysis of data from studies with latent variables or missing observations. It is therefore timely to explore the potential these methods have for the analysis of infectious disease data. Typically, the methods rely on some form of data augmentation, see for example Tanner (1996). An example is maximum-likelihood (ML) estimation via the Expectation-Maximisation (EM) algorithm, which uses data augmentation within a likelihood-based analysis of infectious disease data. However, the conditional expectation required for the E-step is usually difficult to compute in these applications. An attractive alternative is to use Markov chain Monte Carlo (MCMC) methods, which are typically straightforward to implement, often even for very complex models, and which have the additional potential of being suitable for models that incorporate heterogeneities.

Our main purpose in this paper is to explore the potential that MCMC methods have for contributing to the analysis of infectious disease data. Specifically, we shall consider two types of data set for outbreaks of a disease in households. The first set consists only of the eventual sizes of outbreaks, but the analysis allows household members to acquire disease from an external source. This is an important type of data set because the diagnoses of disease are laboratory-based and therefore problems with unobserved subclinical infections are avoided. The second data set consists of the times between the detection of measles cases in household outbreaks. These data have previously been analysed using the unrealistic assumption that the infectious period is of constant duration. This assumption is made because it greatly simplifies ML estimation of the parameters. This data set is of interest because it contains the most detail we can typically hope to observe during an infectious disease outbreak and contains some information about the duration of the latent and infectious periods.

The first of our two data sets can be analysed using a relatively straightforward MCMC algorithm, thanks largely to various closed-form formulae for the final size distribution in question. In particular, no data augmentation is necessary. Our purpose here is to provide a simple example of the use of MCMC, and illustrate how various realistic modelling assumptions can readily be incorporated. Our second data set is analysed in two ways, the first using data augmentation to simplify the form of the likelihood, and the second replacing the exact likelihood with a simulation-based approximation. This example illustrates both the powerful nature of the methods in being able to cope with a large increase in the number of model parameters, and the high level of modelling flexibility that can be achieved. It also serves to indicate some of the difficulties, both those associated with MCMC methods applied to large-scale problems and those inherent in attempting to draw inferences from infectious disease data, such as confounding of parameters.

Some preliminary work on the application of MCMC methods for simple epidemic models can be found in O'Neill and Roberts (1999) and Gibson and Renshaw (1998). These papers are both mainly concerned with the situation in which data arise from a single large outbreak of a disease. In contrast, we shall consider data on many small outbreaks across a large number of households.

Number infected	Number of susceptible individuals in the household				
	1	2	3	4	5
0	110	149	72	60	13
1	23	27	23	20	9
2		13	6	16	5
3			7	8	2
4				2	1
5					1
Total	133	189	108	106	31

Table 1: Frequencies of outbreak sizes of influenza in household of various sizes

2 Sizes of Household Influenza Outbreaks

There is a long history of studying outbreaks in households affected by a disease. It is difficult to assess the extent to which disease transmission from outside affects the outbreak size when data are collected only on affected households. It is often true that a susceptible is far more likely to be infected by a given infective household member than by a given infective who is not a household member. However, during an epidemic there are many infectives outside the household and the probability of being infected by at least one of them might not be negligible compared with the probability of being infected by a household member. It is therefore important to consider types of data and analyses that do not ignore disease transmission between households. Our aim is to demonstrate that a Bayesian analysis, based on an existing transmission model for homogeneous individuals, is straightforward by MCMC methods and can easily be extended to estimate the extent of heterogeneity among individuals.

2.1 The Data

Suppose that a random sample of households is selected at a time prior to the epidemic season and every member of these households is tested to see if they are still susceptible to a certain disease. After the epidemic season all those individuals who were susceptible are tested again, to see if they were infected during the season. This is an important type of data set because the diagnoses are verified by laboratory tests, so that subclinical cases are included and case verification is more objective. It is also an important type of study because it can be designed: the number of households and their sizes can be chosen in advance.

Data of this kind were collected on influenza A(H3N2) infections in households as part of the Tecumseh study of respiratory illness (see Monto *et al.* (1985)). The data in Table 1, taken from Addy *et al.* (1991), are a summary in a form suitable for our analysis. Note that there are a number of households with no infections; this information is the key for an analysis that allows for disease transmission from outside the household.

2.2 Modelling Assumptions

Longini and Koopman (1982) propose a model to describe such data, in which disease transmission from outside the household is allowed for by assuming that there is a global force of infection acting equally on all susceptibles. To remain susceptible an individual must escape infection by this global force of infection and must escape infection from any household member infected during the epidemic season. Let q_c denote the probability that a susceptible individual escapes community-acquired infection during the epidemic season and let q_h denote the probability that a susceptible escapes infection when exposed to one infected household member. Then ω_{js}, the probability that exactly j of the s initial susceptibles of a given household are infected during the epidemic season is given as

$$\omega_{js} = \binom{s}{j} \omega_{jj} (q_c q_h^j)^{s-j} \tag{2.1}$$

by Longini and Koopman (1982). Beginning with $\omega_{0s} = q_c^s$, $s = 0, 1, 2, \ldots$, and using the fact that

$$\omega_{jj} = 1 - \sum_{i=0}^{j-1} \omega_{ij}, \tag{2.2}$$

equation (2.1) enables us to compute the ω_{1s}, then the ω_{2s}, and so on. It is also possible, however, to obtain a closed form for ω_{js} in terms of a non-standard family of polynomials called Gontcharoff polynomials, whose definition we now recall (see, for example, Lefèvre and Picard (1990)). Let U be a sequence of real numbers $u_0, u_1, \ldots$. Then the Gontcharoff polynomials associated with U are defined recursively by

$$\begin{cases} G_0(x|U) = 1, \\ G_j(x|U) = \frac{x^j}{j!} - \sum_{i=0}^{j-1} \frac{u_i^{j-i}}{(j-i)!} G_i(x|U) \quad (j = 1, 2, \ldots). \end{cases} \tag{2.3}$$

Following the arguments in Section 5 of Ball and O'Neill (1999), define $H_j = \omega_{jj}/j!$, so that by (2.1) and (2.2) we obtain

$$H_j = \frac{1}{j!} - \sum_{i=0}^{j-1} \frac{1}{(j-i)!} (q_c q_h^i)^{j-i} H_i.$$

By comparison with (2.3) it follows that $H_j = G_j(1|U)$, where U is the sequence with ith term $u_i = q_c q_h^i$. Further,

$$\omega_{js} = \binom{s}{j} (q_c q_h^j)^{s-j} j! G_j(1|U). \tag{2.4}$$

Note that (2.4) is a simplification of the expression for ω_{js} given by Ball *et al.* (1997), equation (3.11), in the case where the within-group epidemic is of Reed-Frost type (see Bailey (1975, Ch. 14)).

So far, we have assumed that the population is homogeneous. One way of checking the validity of this assumption is to fit a more general model that collapses to the

homogeneous case when certain parameter values are specified. We now consider such a model, which incorporates two different kinds of heterogeneity, as follows.

Suppose first that infectives differ in their potential to infect others. In (2.1) this means that q_h depends on the infective. We might model this by letting q_h be a realisation of the random variable Q, selected independently for each infective. More specifically, we shall take $Q = \exp(-Y)$, where Y is the duration of the infectious period when each infective has infectious contacts with a given susceptible at times given by the points of a Poisson process of rate 1. Thus, $q_h = \mathbb{E}[\exp(-Y)] = \phi_Y(1)$, say, where $\phi_Y(\theta) = \mathbb{E}[\exp(-\theta Y)]$ is the moment generating function of Y. In this case, due to the extra dependencies involved we can no longer argue as before to find ω_{js}. However, a closed form is still available, from equation (3.11) of Ball *et al.* (1997), and we find that

$$\omega_{js} = \frac{1}{(s-j)!} \sum_{i=s-j}^{s} \frac{s!}{(s-i)!} \phi_Y(i)^{s-i} q_c^i G_{i-s+j}(0|E^{s-j}U), \tag{2.5}$$

where now U is the sequence with ith term $u_i = \phi_Y(i)$, and where E^iU denotes the sequence $u_i, u_{i+1}, \ldots$. Note that if Y is a constant then (2.5) defines the outbreak size distribution for a Reed-Frost type model, while if Y is negative exponential then (2.5) defines the outbreak size distribution for a general stochastic epidemic model (see Bailey (1975, p. 88)). Note also that since the data do not include temporal information, the temporal scaling of Y can be defined arbitrarily. In particular if we set Y to have mean length one time unit, then $\phi_Y(i) = \exp(-i)$ if Y is constant, and $\phi_Y(i) = (1+i)^{-1}$ if Y is negative exponential.

Additionally, suppose that there is some heterogeneity between susceptibles. A simple way to model this, which ought to enable us to detect the presence of any substantial heterogeneity, is to assume that each susceptible has some probability v of being immune to the disease, perhaps as a result of their cautious behaviour. In the present case we refer to v as the probability of being protected from infection. Thus the number of unprotected susceptibles available at the start of the epidemic has a binomial distribution. So, using $\omega_{js}(v)$ to denote the probability of j infections among s initial susceptibles we have

$$\omega_{js}(v) = \sum_{i=0}^{s-j} \binom{s}{i} v^i (1-v)^{s-i} \omega_{j,s-i}. \tag{2.6}$$

Note that if $v = 0$, the model reverts to the homogeneous susceptible population case.

Finally, of the households that had s *a priori* susceptible members prior to the epidemic season, let there be n_{js} households with j cases at the end of the epidemic season. It follows that the likelihood function is given by

$$L = \prod_{s=1}^{5} \prod_{j=0}^{s} [\omega_{js}(v)]^{n_{js}}, \tag{2.7}$$

so that L can be calculated using equation (2.4) for the completely homogeneous case, and equations (2.5) and (2.6) for the two kinds of heterogeneity.

2.3 Metropolis-Hastings Algorithm for Bayesian Inference

In the general model described in the previous section there are three parameters of interest, namely q_h, q_c and v. Our objective is to make valid and useful inferences about these parameters on the basis of the data and the modelling assumptions. To this end, we compute $\pi(q_h, q_c, v|\,\{n_{js}\})$, the joint posterior of q_h, q_c and v. From Bayes' theorem we have, assuming initial independence between q_h, q_c and v,

$$\pi(q_h, q_c, v|\,\{n_{js}\}) \propto L(q_h, q_c, v)\pi(q_h)\pi(q_c)\pi(v), \tag{2.8}$$

where $\pi(q_h)$, $\pi(q_c)$ and $\pi(v)$ are respectively the prior distributions of q_h, q_c and v, and L is the likelihood function given by (2.7). We shall use MCMC methods, as described below, to sample from $\pi(q_h, q_c, v|\,\{n_{js}\})$. Any desired property of $\pi(q_h, q_c, v|\,\{n_{js}\})$, such as the posterior mean or standard deviation of an individual parameter, is readily approximated by the corresponding property of the sample.

We use a particular type of MCMC method known as a Metropolis-Hastings algorithm (see, for example, Gilks *et al.* (1996)). Suppose that we have a likelihood L and a prior density π. By Bayes' Theorem, the posterior density is $cL\pi$, where c is the normalising constant. Construct a Markov chain $\{Z_n\}$ in the following manner. Given the current state of the chain, $Z_n = x$ say, draw a possible new point (known as a candidate) y from a proposal density $q(y|x)$. Accept the candidate with probability

$$\min\left(1, \frac{L(y)\pi(y)q(x|y)}{L(x)\pi(x)q(y|x)}\right), \tag{2.9}$$

in which case $Z_{n+1} = y$. If however the candidate is rejected, then set $Z_{n+1} = x$. It can be shown under mild conditions (see Gilks *et al.* (1996)) that the stationary distribution of this Markov chain is indeed the (correctly normalised) posterior density. Thus to sample from π we simply run the Markov chain until it is deemed to have converged, and then draw samples from the chain's output.

Although the proposal density is in principle arbitrary, in practice a careful choice can facilitate computation and speed up convergence of the algorithm. If the parameter space is all of $\mathbb{R}^d$, for some d, it is often convenient to set $q(\cdot|y)$ to be a Normal density with mean y. One benefit is that $q(x|y) = q(y|x)$, so that the acceptance probability (2.9) reduces to

$$\min\left(1, \frac{L(y)\pi(y)}{L(x)\pi(x)}\right).$$

In the present example, and those in later sections, informal methods of convergence assessment were adopted, such as visual inspection of the sample chain output, and tracking the estimates of the quantities of interest. This seemed to work well in practice insofar as it generally seemed clear whether or not convergence had occurred.

Returning to the present case, the above algorithm is implemented as follows. The target distribution is $\pi(q_h, q_c, v|\,\{n_{js}\})$. Since each of the three parameters has range

$(0,1)$ we transform each one using the logistic transformation

$$t \to \tilde{t} = \log(t/(1-t)),$$

so that the transformed parameters all have range $\mathbb{R}$, and we can use a Normal proposal density. Note that $t = \exp(\tilde{t})/(1+\exp(\tilde{t})) = g(\tilde{t})$, say, and that the Jacobian of this transformation is given by $J(\tilde{t}) = \exp(\tilde{t})/((1+\exp(\tilde{t}))^2)$. The corresponding transformed version of the RHS of (2.8) is now obtained by replacing q_h, q_c and v by $g(\tilde{q_h})$, $g(\tilde{q_c})$ and $g(\tilde{v})$ respectively, and multiplying by $J(\tilde{q_h})J(\tilde{q_c})J(\tilde{v})$.

2.4 Results and discussion

We found that convergence to the stationary distribution was easily achieved. As a typical example, using a Gaussian proposal density with standard deviation 0.1 for each of the parameters, a 'burn-in' period of 200 steps was ample. Samples were then drawn from the output of the Markov chain at every tenth step.

It is relatively straightforward to find maximum likelihood estimates of parameters. By using uniform prior distributions the posterior density is equivalent to the likelihood, and a simple numerical maximisation technique applied to the MCMC output yields the required estimates. The estimates were found to be virtually identical to the values given in Addy *et al.* (1991) for the case without protection (*i.e.* $v = 0$) considered there. For example, in the Reed-Frost case we obtained $q_c = 0.8677$ and $q_h = 0.8408$ while Addy *et al.* (1991) give an identical q_c, and a q_h value of 0.8406.

Figure 2.1 near here

By way of example, the posterior densities for the Reed-Frost case without protection is shown in Figure 2.1. For the model without protection, we find a simple well-peaked distribution with mode close to the ML estimates of q_h and q_c; estimates of q_h (=0.84) and q_c (=0.87) agree to 2 significant figures for both Reed-Frost and 'general epidemic' models. As can be seen from Figure 2.1, $q_c < 1$ with probability close to 1 and thus there is strong evidence to support the existence of community-acquired infection under the assumptions of the model. Any estimate of q_h made under the assumption that the between-household transmission rate is negligible is vulnerable to the possibility of severe bias.

Figure 2.2 near here

For the model with protection, however, we find that the extra parameter v cannot be well estimated: the joint posterior density of q_h, q_c and v is banana shaped (see Figure 2.2), with high positive correlation between q_h and q_c (0.87), and high negative correlation of v with both q_h (0.85) and q_c (0.91). The mode is at $(q_h, q_c, v) = (0.60, 0.75, 0.47)$ for the 'general epidemic', and at (0.64,0.76,0.45) for the Reed-Frost.

However, the fit of these models is not appreciably better than that of the models without protection, i.e. with $v = 0$.

It may seem remarkable that a data set covering 567 households is insufficient to discriminate between a model in which the modal value of the proportion vaccinated is 47% and one in which it is zero, but closer consideration provides some explanation. This can be given most succinctly in the Reed-Frost case, for which the probabilities of outbreaks of size 0 and 1 with parameters (q_h, q_c, v) are the same as those for $(q_h', q_c', 0)$ where $q_c' = 1-(1-v)(1-q_c) = v+(1-v)q_c$, and $q_h' = (v+(1-v)q_cq_h)/q_c'$. It follows that information classifying outbreaks into sizes 0, 1 and at least 2 is of no use at all for inference on v (in the absence of other information on the values of q_h and q_c). This goes far towards explaining our difficulties in estimating v, since there are only 21 households in our data set with outbreaks of size greater than 2.

It is also relevant to note that the chi-squared goodness-of-fit statistic already shows an acceptable fit for the model without protection (e.g. 14.4 on 13 d.f. for the 'general epidemic'), suggesting that adding any further parameters may be over-fitting the data.

This example has illustrated the use of MCMC to analyse non-temporal final outcome data. Our next example considers a situation where temporal data are available. As we shall see, this necessitates both a more detailed model, and a more sophisticated MCMC implementation.

3 Measles Cases in Household Outbreaks

3.1 The Data

We consider Hope Simpson's data on measles in households of size two around Cirencester (UK), given by Bailey (1975, Ch. 15). There were 264 households with two susceptible children under the age of fifteen which had at least one case of measles. Forty-five households had a single primary case with no further transmission within the household. The time durations between cases in the 219 households with two cases are given in Table 2. Within these households, we do not know the number of primary infectives, which could be one or two. For simplicity we shall initially assume that in the 32 households where the cases occurred within 6 days of each other, both cases are primary infectives. Thus in the remaining 187 households, one of the two infectives is regarded as primary and the cause of the secondary case. However, in Section 3.4 we relax this assumption and treat the proportion of households with two primary infectives as a model parameter.

The analyses of Bailey (1975, Ch. 5) and Becker (1989, Ch. 4), are based on maximum likelihood estimation of disease parameters, such as the mean infectious period, the mean latent period and the transmission rate within households. We apply a Bayesian analysis using a modelling framework that is similar, but with more realis-

Time (days)	0	1	2	3	4	5	6	7	8	9	10
Frequency	5	13	5	4	3	2	4	11	5	25	37
Time (days)	11	12	13	14	15	16	17	18	19	20	21
Frequency	38	26	12	15	6	3	1	3	0	0	1

Table 2: Times between measles cases in 219 two-child households. There were an additional 45 two-child households in which only one child was affected by measles.

tic assumptions.

3.2 Modelling Assumptions

For a given household, let X_i and Y_i denote the durations of the latent and infectious periods, respectively, for individual i, $i = 1, 2$. Capital letters are used for random variables and lower case letters for actual observations on these variables.

When one primary case exerts a force of infection β on a susceptible individual for the entire duration y of the infectious period, the probability of escaping infection is $\exp(-\beta y)$. This is a conditional probability, given the duration $Y = y$, and is the contribution to the likelihood for a household with a single case when the infected individual had an infectious period of duration y. When the infectious periods are not observed, each of the 45 households with a single case makes a contribution

$$\mathrm{E}[\mathrm{e}^{-\beta Y}] = \int_0^\infty \exp(-\beta y)\, f_Y(y)\, \mathrm{d}y$$

to the likelihood function, where f_Y denotes the probability density function (pdf) of the infectious period.

For each of the 32 households with two primary infectives, it is assumed that these infectives were infected simultaneously and so the time between their detection is the difference between two independent realisations of $X + Y$. An observed difference w for one of these households therefore makes a contribution

$$2 \int_0^\infty f_{X+Y}(u) f_{X+Y}(u + w)\, \mathrm{d}u \tag{3.1}$$

to the likelihood. On the other hand, if we had observed both latent periods and both infectious periods the contribution to the likelihood would be

$$f_X(x_1) f_X(x_2) f_Y(y_1) f_Y(y_2)\,. \tag{3.2}$$

Similarly, for a household with one primary case and one secondary case, if we knew the lengths of the latent and infectious periods, and the time t from the start of the infectious period of the primary infective to the infection of the secondary case, then the likelihood could be written

$$f_X(x_1) f_X(x_2) f_Y(y_1) f_Y(y_2)\, \beta \mathrm{e}^{-\beta t}\,. \tag{3.3}$$

The likelihood contribution based only on the time w between the detection of cases is obtained by integrating (3.3) over all variables subject to $t + x_2 + y_2 - y_1 = w$.

Equations (3.1), (3.2), and (3.3) illustrate that failing to observe key times can introduce integrations into the likelihood formula, the integrals being over all possible values for the unobserved times. When multiple integrals arise in this way, the likelihood formula often becomes intractable. We describe below two approaches to avoiding explicit evaluation of these integrals. Firstly, in Section 3.3 we develop a Gibbs sampler algorithm based on a data augmentation approach in which unobserved values are introduced as additional parameters, often called latent variables. Secondly, in Section 3.4, we construct a Metropolis algorithm which employs simulation to approximate the likelihood ratio at each iteration of the algorithm, avoiding both explicit integration and the introduction of latent variables.

Bailey (1975, Ch. 15) and Becker (1989, Ch. 4) simplify the integrations such as (3.1) by assuming a constant infectious period and restricting f_X to a form that enables direct integration. The approaches introduced here permit more flexible modelling assumptions. We shall assume that the latent and infectious periods are drawn from Gamma distributions with parameters $(\theta_{L1}, \theta_{L2})$ and $(\theta_{I1}, \theta_{I2})$, respectively, so that

$$f_X(x) = \frac{1}{\Gamma(\theta_{L2})} \theta_{L1} e^{-x\theta_{L1}} (x\theta_{L1})^{\theta_{L2}-1}, \tag{3.4}$$

and f_Y is defined similarly. The parameter θ_{L1} in (3.4) can be interpreted as a scale parameter, whereas θ_{L2} controls the shape of the distribution: $\theta_{L2} = 1$ gives the exponential distribution, while larger values give distributions which are more symmetric than the exponential.

Our choice of Gamma distributions is by no means a necessary one; other distributions can readily be employed. However, the Gamma distribution is smooth and unimodal which seem appropriate for this application, while the two parameters allow a good deal of flexibility.

3.3 Gibbs sampler algorithm with data augmentation

3.3.1 The augmented data likelihood

The parameters of interest are β, the force of infection, and $\boldsymbol{\theta} = (\theta_{L1}, \theta_{L2}, \theta_{I1}, \theta_{I2})$, the vector of scale and shape parameters for the latent and infectious time distributions. In order to make inferences about β and $\boldsymbol{\theta}$, we compute $\pi(\beta, \boldsymbol{\theta}|\text{data})$, the joint posterior pdf of β and $\boldsymbol{\theta}$.

As discussed above, the likelihood involves many integrals, but we will eliminate explicit integrations by introducing latent variables, such as the lengths of latent and infectious periods, and the infection times. We model the observed times w as if they were continuously distributed, although they are recorded only to the nearest day.

This discretization of the observations is expected to have very little effect on the inferences drawn.

Within-household infections

First, consider the 187 households within which an infection is deemed to have occurred. We set a time-origin in each household by assuming that the cases are detected at times 0 and $w \geq 6$, so that w is taken directly from Table 2. The latent period is sufficiently long that it is reasonable to assume that the first observed case corresponds to the initially infected individual. We introduce latent variables u_1, s, and u_2 (see Figure 3.1) where

$$\begin{aligned} u_1 < 0 &\equiv \text{start time for first infectious period,} \\ s \in (u_1, 0) &\equiv \text{time of within-household infection,} \\ u_2 \in (s, w) &\equiv \text{end time for the secondary infective's latent period.} \end{aligned}$$

Figure 3.1 near here

The augmented likelihood for each household is given by

$$f_Y(-u_1)\beta \exp(-\beta(s-u_1)) f_X(u_2-s) f_Y(w-u_2), \tag{3.5}$$

where f_X and f_Y are the Gamma pdfs given at (3.4). Since each household has its own set of latent variables, we have introduced 3×187 extra parameters into the model. This enormous increase in the dimensionality of the parameter space causes no obstacle in principle for the Gibbs sampler algorithm, although convergence can be slower and more difficult to assess in such large spaces.

Two-primary households

Next, consider the 32 households in which there are deemed to be two primary cases, infected at time $\tau < 0$ and detected at times 0 and w, with $0 \leq w \leq 5$. Since no within-household infection could have been observed, we cannot infer anything about β from these observations. If $v_1 \in (\tau, 0)$ and $v_2 \in (\tau, w)$ denote the ends of the two latent periods, the augmented likelihood for each household can be written

$$f_X(v_1-\tau) f_X(v_2-\tau) f_Y(-v_1) f_Y(w-v_2). \tag{3.6}$$

This introduces 3×32 extra parameters.

No-infection households

Finally, the contribution to the likelihood from a household in which no secondary infection occurs can be evaluated exactly, without the need for latent variables. It is equal to the probability that a single infectious individual fails to infect a single susceptible:

$$\mathbb{E}[\mathrm{e}^{-\beta Y}] = \left(\frac{\theta_{I1}}{\theta_{I1}+\beta}\right)^{\theta_{I2}} = q, \quad \text{say.} \tag{3.7}$$

3.3.2 The Gibbs sampler

Before describing the Gibbs sampler algorithm, we note that a Metropolis-Hastings algorithm following the general procedure described in Section 2.4 can be implemented relatively easily. We did this as follows: all parameters were transformed onto $\mathbb{R}$, the likelihoods multiplied by appropriate Jacobians, and a Metropolis-Hastings algorithm with Gaussian proposal density employed. In practice we found that this approach suffered from severe convergence problems, and additionally exhibited very long run times. Certain blocking strategies were implemented in order to improve matters but these were largely ineffective.

We then turned to the Gibbs sampler, a special case of the Metropolis-Hastings algorithm in which each parameter is updated in turn according to its conditional distribution given the current values of all the other parameters (see Smith and Roberts (1993)). In practice the conditional distributions may be non-standard, but since they are univariate it is often possible to design efficient sampling methods. The attractiveness of the Gibbs sampler for the measles data set is that it enables us to exploit the independence structure inherent in the data. For example, the conditional distributions of each latent variable, given all other parameters, depend only on the 'global' parameters $\boldsymbol{\theta}$ and β and the latent parameters for the same household: the parameters associated with other households are irrelevant.

To fully describe the implementation of the Gibbs sampler it is necessary to consider each of the conditional distributions involved. However, for ease of exposition we focus now on the five global parameters, since they are of most interest, and their conditional distributions help to provide insight into the inter-dependence of the various model parameters. The sampling methods described below are also illustrative of the methods used for the other variables, details of which are given in the Appendix, Section A3.

In describing the sampling schemes we shall often use the techniques of rejection sampling, accounts of which can be found in Morgan(1984, Section 5.3). For a parameter x we adopt the notation $\pi(x|\cdots)$ to denote the density of x conditional upon all other parameters. Parameters for different households are labelled by superscripts in the obvious manner, and we use w_1 and w_2 to denote the data w from households with 1 and 2 primary infectives, respectively. Gamma priors are employed; these are all denoted $\mathrm{Gam}(\lambda, m)$, although in practice the parameters may vary.

Sampling β

¿From (3.5) and (3.7) we find that

$$\pi(\beta|\cdots) \propto \frac{\beta^{186+m} \exp\left\{-\beta[\lambda + \sum_{j=1}^{187}(s^{(j)} - u_1^{(j)})]\right\}}{(\theta_{I1} + \beta)^{45\theta_{I2}}}.$$

We thus need to sample from a density of the form

$$f(x) = \frac{x^a \exp(-bx)}{(x+c)^d},$$

where $a, b, c, d > 0$. In order to do so, we construct a bounding Gamma density via the following Lemma, the proof of which is given in the Appendix, Section A1.

Lemma 3.1

$$f(x) \leq g(x) = \frac{\gamma^{a-b\gamma}}{(\gamma+c)^d} x^{b\gamma} \exp(-bx),$$

where $\gamma = (2b)^{-1}\left(a - bc - d + \sqrt{(a-bc-d)^2 + 4acd}\right)$.

So to sample from $f(x)$, we can use rejection sampling; first draw a sample, Z say, from a $\text{Gam}(b, b\gamma+1)$ distribution, and then accept this with probability $f(Z)/g(Z)$, repeating this process until an acceptance occurs.

Sampling θ_{L1}

¿From (3.5) and (3.6) we have

$$\pi(\theta_{L1}|\cdots) \propto \theta_{L1}^{m-1+251\theta_{L2}} \exp\left\{ -\left[\lambda + \sum_{j=1}^{187}(u_2^{(j)} - s^{(j)}) + \sum_{j=1}^{32}(v_1^{(j)} + v_2^{(j)} - 2\tau^{(j)})\right]\theta_{L1}\right\},$$

yielding that

$$\pi(\theta_{L1}|\cdots) \sim \text{Gam}(\lambda + \sum_{j=1}^{187}(u_2^{(j)} - s^{(j)}) + \sum_{j=1}^{32}(v_1^{(j)} + v_2^{(j)} - 2\tau^{j}), m + 251\theta_{L2}).$$

Sampling θ_{L2}

We find that

$$\pi(\theta_{L2}|\cdots) \propto \frac{\theta_{L2}^{m-1}\exp(B\theta_{L2})}{\Gamma(\theta_{L2})^{251}},$$

where

$$B = \sum_{j=1}^{187}\log(u_2^{(j)} - s^{(j)}) + \sum_{j=1}^{32}\log((v_1^{(j)} - \tau^{(j)})(v_2^{(j)} - \tau^{(j)})) + 251\log(\theta_{L1}) - \lambda.$$

We thus wish to sample from a density of the form

$$h(x) = \frac{x^a \exp(bx)}{\Gamma(x)^c},$$

where $a, c > 0$ and $b \in \mathbb{R}$. Our approach to this is similar to that adopted for the β sampling case above; as before we require a bounding density, and this is provided by the following Lemma, the proof of which is in the Appendix, Section A2.

Lemma 3.2 For $x > 0$ and $\mu > 0$,

$$h(x) \leq l(x) = \delta x^{\gamma\mu}\exp(-\mu x),$$

in which $\delta = \gamma^{a-\gamma\mu}\exp((b+\mu)\gamma)\Gamma(\gamma)^{-c}$ and γ is the unique positive root of the equation

$$\frac{a}{\gamma} + b - c\psi(\gamma) = 0,$$

where ψ denotes the digamma function.

We may thus use rejection sampling to obtain a sample for θ_{L2} in an identical manner to that described for the β case above. The Lemma allows us to choose μ arbitrarily; we set $\mu = \gamma$ which seemed to work fairly well in practice.

Sampling θ_{I1}

It turns out that the conditional distribution of θ_{I1} is identical in form to the conditional distribution of β, and thus we can sample θ_{I1} using the same method. Specifically, from (3.5), (3.6) and (3.7) we obtain that

$$\pi(\theta_{I1}|\cdots) \propto \frac{\theta_{I1}^a \exp(-b\theta_{I1})}{(\theta_{I1}+c)^d},$$

where $a = m - 1 + 483\theta_{I2}$, $b = \sum_{j=1}^{187}(w_1^{(j)} - u_1^{(j)} - u_2^{(j)}) + \sum_{j=1}^{32}(w_2^{(j)} - v_1^{(j)} - v_2^{(j)}) + \lambda$, $c = \beta$ and $d = 45\theta_{I2}$.

Sampling θ_{I2}

The structure of the conditional distribution of θ_{I2} is identical to that of θ_{L2}, so once again we can use our existing method to perform the sampling. Specifically,

$$\pi(\theta_{I2}|\cdots) \propto \frac{\theta_{I2}^{m-1}\exp(B\theta_{I2})}{\Gamma(\theta_{I2})^{438}},$$

where

$$\begin{aligned} B &= 438\log\theta_{I1} + \sum_{j=1}^{187}\log[(-u_1^{(j)})(w_1^{(j)} - u_2^{(j)})] + \sum_{j=1}^{32}\log[(-v_1^{(j)})(w_2^{(j)} - v_2^{(j)})] \\ &\quad + 45\log\left(\frac{\theta_{I1}}{\theta_{I1}+\beta}\right) - \lambda. \end{aligned}$$

3.3.3 Results

In this and subsequent sections μ_L and σ_L denote respectively the mean and standard deviation of the latent period, while μ_I and σ_I are defined similarly for the infection period.

The Gibbs sampler algorithm was implemented using Fortran 90 running under Unix on a mainframe computer. Actual run-time was about 50,000 sweeps per hour (a sweep refers to a single update of the entire parameter set). Prior distributions, where needed, were set as negative exponential with mean 1000 (*i.e.* Gamma with parameters $\lambda = 0.001$ and $m = 1$) so as to be relatively uninformative. Sample

chains of the global parameters moved very slowly around their parameter space. One reason for this is that the conditional densities of the global parameters all involve summations of independent and similarly distributed latent variables. Thus, a law of large numbers effect ensures that the distribution of each global parameter will not change greatly unless there are substantial changes across the latent variables.

The sample chains also exhibited clear correlations, which is expected because of the correlations of the global parameters, and the lack of detail in the data set. It is well-known that such correlations are a cause of slow convergence in the Gibbs sampler (see for example Hills and Smith (1992)); however in the present case any sensible reparameterisation of the global parameters seems likely to drastically increase the complexity of the conditional distributions, with resulting increase in computation time.

More specifically, the β and Y parameter output values were correlated: as β increased, so Y tended to decrease. Intuitively, this can be explained by the data providing little information to distinguish between short infectious periods with high infectivity, and long infectious periods with low infectivity. Furthermore, these sample chains did not appear to converge, but instead displayed rather erratic movement around the sample space. Consequently, no reliable estimates of posterior distributions were obtained for β or for the Y parameters. However, the avoidance probability q, which is derived from β and Y by (3.7), did exhibit convergence. The estimates of posterior mean and standard deviation of q were given by 0.194 and 0.026, respectively, and a 95% equal-tail credibility interval by (0.15, 0.24). In fact, it is straightforward to calculate the posterior distribution of q exactly, and we find that

$$\pi(q|\text{data}) \propto q^{45}(1-q)^{187}\pi(q),$$

so that if q has a Beta prior distribution, it also has a Beta posterior distribution. With a uniform prior (which is essentially comparable with the priors used in our MCMC simulations in the region of interest) we thus find the posterior mean and standard deviation to be 0.197 and 0.026 respectively, and an equal-tail 95% credibility interval (0.15, 0.24), all of which compare favourably with our MCMC estimates, providing some reassurance that the algorithm is functioning correctly.

The sample chains associated with the latent period parameters appeared to exhibit convergence, and we were thus able to obtain posterior estimates. Both μ_L and σ_L had approximately symmetric unimodal posterior densities. The posterior mean and median of μ_L were 10.86 and 10.89 days, respectively, and the 95% equal-tail credibility interval for μ_L was (10.43,11.22). Similarly σ_L had posterior mean and median 2.31 days and 95% equal-tail credibility interval (2.13, 2.51).

Stronger priors

In view of the convergence problems that were encountered, we also considered the use of much stronger prior assumptions. This was found to improve the convergence properties of the algorithm and thus posterior density estimates for all parameters of

	β	μ_L	σ_L	μ_I	σ_I
Weak priors	– –	10.86	2.31	– –	– –
	– –	(10.51,11.15)	(2.17,2.46)	– –	– –
Strong priors	0.39	9.41	1.59	4.53	1.03
	(0.29, 0.51)	(8.91,9.90)	(1.38,1.83)	(3.44,5.63)	(0.81,1.24)

Table 3: Posterior means and equal-tail 90% intervals assuming weak priors (*i.e.* all parameters have prior density Gam(0.001,1)) and strong priors (β is Gam(8,2); θ_{L1}, θ_{I1} are Gam(1,2); θ_{L2}, θ_{I2} are Gam(1,20).) The posterior densities were all roughly symmetric.

interest were obtained. Specifically, β was given prior mean 0.25 and prior standard deviation about 0.2, while $\boldsymbol{\theta}$ was assigned priors such that the infectious and latent periods both had prior mean 10 and prior standard deviation just over 2 days.

The results are summarised in Table 3. For the latent period parameters, the data are sufficiently informative to reduce the effect of the priors, and we find that the posterior estimates are roughly comparable with those obtained in the weak prior case. Note that the prior on β is quite restrictive; for example, values greater than 1 are more than 3 standard deviations from the mean. As a consequence, the posterior credibility interval for β is quite narrow. Restricting β effectively forces the MCMC algorithm to separately distinguish between β and the Y parameter values, in contrast to the weak prior case. Consequently, it is inappropriate to compare results under strong and weak prior assumptions for these parameters. For example, the strong prior results are appreciably different from those obtained using the Monte-Carlo-within-Metropolis algorithm as described in the next Section under weaker priors (see Table 6, Model (b)).

It seems likely that the convergence problems in the weak prior case were caused by the large number of latent variables and the high correlation of model parameters. As an attempt to overcome these difficulties a new algorithm was developed, which we now describe.

3.4 Monte-Carlo-within-Metropolis (MC-w-M) Algorithm

3.4.1 Monte Carlo estimation of the likelihood ratio

The large number of latent variables in the formulation described in Section 3.3 raises concerns about convergence of the Gibbs sampling algorithm. Here, we eliminate the need for latent variables by implementing an 'MC-w-M' algorithm, obtained by replacing the likelihood ratio R in a standard Metropolis algorithm with a simulation-based approximation R^*, defined below at (3.8) and (3.11).

Before describing the algorithm, we introduce two desirable modifications to the modelling assumptions outlined in Section 3.2, which are readily incorporated here. First,

the proportion of 'two-primary' households will be regarded as an additional parameter, Δ, which is estimated from the data. (Previously the two-primary households were assumed to be exactly the 32 households in which cases arose within 6 days of each other.) Secondly, the data show local maxima at 7, 14 and 21 days, suggesting an effect of rounding to the nearest week. It seems difficult to adequately model any such rounding process, and instead we weaken its effect on inference by coarsening the data as shown in the first row of Table 4.

Write $\mathbf{C} \equiv (c_1, \ldots, c_k)$ for the observed frequencies in each time period; for the data of Table 4, we have $k = 13$ and $\sum_{i=1}^{13} c_i = 219$. Let $\mathbf{p}(\Delta, \beta, \boldsymbol{\theta})$ denote the expected proportions given parameter vector $(\Delta, \beta, \boldsymbol{\theta})$, that is

$$\mathbf{p}(\Delta, \beta, \boldsymbol{\theta}) \equiv \mathbb{E}[\mathbf{C}|\Delta, \beta, \boldsymbol{\theta}]/219.$$

As previously discussed, evaluating $\mathbf{p}(\Delta, \beta, \boldsymbol{\theta})$ exactly involves multiple integrations and is generally infeasible, but it is readily approximated by the sample proportion, $\hat{\mathbf{p}}(\Delta, \beta, \boldsymbol{\theta})$, obtained by simulating a large number, n, of households under the model with parameter vector $(\Delta, \beta, \boldsymbol{\theta})$. Similarly, the likelihood ratio,

$$R \equiv \frac{L(\Delta, \beta, \boldsymbol{\theta})}{L(\Delta', \beta', \boldsymbol{\theta'})} = \prod_{i=1}^{k} \left(\frac{p_i(\Delta, \beta, \boldsymbol{\theta})}{p_i(\Delta', \beta', \boldsymbol{\theta'})} \right)^{c_i},$$

is naturally estimated by

$$\hat{R} \equiv \prod_{i=1}^{k} \left(\frac{\hat{p}_i(\Delta, \beta, \boldsymbol{\theta})}{\hat{p}_i(\Delta', \beta', \boldsymbol{\theta'})} \right)^{c_i}, \tag{3.8}$$

(see Diggle & Gratton, 1984).

For notational convenience, we will sometimes write $\mathbf{p}$ and $\mathbf{q}$ for $\mathbf{p}(\Delta, \beta, \boldsymbol{\theta})$ and $\mathbf{p}(\Delta', \beta', \boldsymbol{\theta'})$, with $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ defined similarly. To avoid problems with division by zero, we require that $\hat{q}_i > 0$, for all i, but we are interested in values of n sufficiently large that this requirement has negligible practical effect. If n is also large enough that $(n-1)q_i > c_i$, for all i, then the Dirichlet approximation to the multinomial gives $\mathbb{E}[\hat{R}] \approx \mu_n(\mathbf{p}, \mathbf{q})$, where

$$\mu_n(\mathbf{p}, \mathbf{q}) \equiv \left(\prod_{i=1}^{k} \prod_{j=1}^{c_i} \frac{(n-1)p_i + j - 1}{(n-1)q_i - j} \right) \prod_{i=1}^{219} \frac{n - i - 1}{n + i - 2}. \tag{3.9}$$

This approximation is usually very accurate for the values of n discussed below, so that $\mu_n - R$ approximates the bias of $\hat{R}$.

The detailed balance condition underpinning the Metropolis algorithm requires that

$$\min\left\{1, R\frac{\pi(\Delta, \beta, \boldsymbol{\theta})}{\pi(\Delta', \beta', \boldsymbol{\theta'})}\right\} \pi(\Delta', \beta', \boldsymbol{\theta'}) = \min\left\{1, \frac{1}{R}\frac{\pi(\Delta', \beta', \boldsymbol{\theta'})}{\pi(\Delta, \beta, \boldsymbol{\theta})}\right\} R\pi(\Delta, \beta, \boldsymbol{\theta}),$$

where π denotes the prior pdf. In the application discussed below, uniform priors are employed, and in this case the detailed balance condition becomes

$$\min\{1, R\} = \min\{1, 1/R\}R.$$

	n=4,000		n=8,000		n=16,000	
	$g(\hat{R})$	$g(R^*)$	$g(\hat{R})$	$g(R^*)$	$g(\hat{R})$	$g(R^*)$
$R = 1$	1.00	1.00	1.00	1.00	1.00	1.00
$R = 1.33$	1.31	1.37	1.32	1.34	1.33	1.34
$R = 2.00$	1.90	2.06	1.96	2.01	1.98	2.01
$R = 4.00$	3.69	4.13	3.85	4.03	3.93	4.01

Table 4: Approximate expectations of $g(\hat{R})$ and $g(R^*)$ for various values of the simulation size, n. In row 1, $\mathbf{p} = \mathbf{q}$, with the common value being the observed relative frequencies for measles data: the 13 counts from Table 4, each divided by 219. Rows 2 – 4 correspond to the same $\mathbf{p}$ and different choices of $\mathbf{q}$ obtained by perturbing $\mathbf{p}$; the corresponding value of R is shown in column 1. Each entry in the table is based on 24,000 simulations

It follows that the MC-w-M algorithm will be exactly valid whenever $g(\hat{R}) = R$, where

$$g(\hat{R}) \equiv \frac{\mathbb{E}[\min\{1, \hat{R}\}]}{\mathbb{E}[\min\{1, 1/\hat{R}\}]}. \tag{3.10}$$

If $(\Delta, \beta, \boldsymbol{\theta}) = (\Delta', \beta', \boldsymbol{\theta}')$, then $R = 1$ and $\hat{R}$ has the same distribution as $1/\hat{R}$, so that (3.10) is trivially satisfied. The first entry in each cell of Table 3 shows a simulation-based approximation to $g(\hat{R})$, for $\mathbf{p}$ equal to the vector of sample proportions, and various choices of n and $\mathbf{q}$. The bias in $\hat{R}$ tends to have a 'levelling' effect: $g(\hat{R})$ lies between 1 and R.

Equation (3.9) suggests a bias correction for $\hat{R}$, using an idea similar to one exploited by bootstrap bias corrections. Since R is a smooth function of $(\Delta, \beta, \boldsymbol{\theta})$ and $(\Delta', \beta', \boldsymbol{\theta}')$, it follows that $\mathbb{E}[\hat{R}]/R$, which is (well) approximated by μ_n/R, should be (less well) approximated by $\hat{\mu}_n/\hat{R}$, where $\hat{\mu}_n \equiv \mu_n(\hat{\mathbf{p}}, \hat{\mathbf{q}})$. Hence, R^* should be approximately unbiased for R, where

$$R^* \equiv \hat{R}^2/\hat{\mu}_n \tag{3.11}$$

Table 3 indicates that $g(R^*)$ is closer to R than is $g(\hat{R})$, although there is some tendency to overcorrection.

3.4.2 MC-w-M for the measles data

An MC-w-M algorithm with simulation size $n = 8,000$ was implemented for the data of Table 4 and the model outlined in Section 3.2. To improve convergence a reparameterisation was adopted. For the gamma distribution modelling the length of the latent period, we worked with the expectation, $\mu_L \equiv \theta_{L2}/\theta_{L1}$, and $1/\theta_{L2}$, instead of θ_{L1} and θ_{L2}, and similarly for the infectious period parameters. Further, we rotated the plane spanned by μ_L and μ_I to reduce their correlation. Specifically, we worked

No inf	Time (days)	0	1	2	3	4	5	6–8	9	10	11	12	13–15	≥16
45	Observed	5	13	5	4	3	2	20	25	37	38	26	33	8
44	Fitted (a)	5	9	7	5	3	2	23	25	32	33	28	39	5
45	Fitted (b)	5	10	7	5	3	3	23	23	33	35	29	37	5

Table 5: Row 1 shows the data from Table 2 coarsened to weaken the effect of possible week-rounding and possible inaccurate outliers. Rows 2 and 3 show the mean fitted values (estimated from the same MC-w-M runs as for Table 6) under model (a): fixed length infectious period; and model (b): gamma-distributed infectious period.

	β	μ_L	μ_I	θ_{L2}	θ_{I2}	$\Delta \times 264$
Model (a)	0.51	9.9	3.2	26	– –	34
	(0.3,3.1)	(8.8,10.9)	(0.5,6.0)	(21,35)	– –	(26,43)
Model (b)	2.0	10.7	1.6	54	1.7	32
	(0.6,17)	(10.0,11.1)	(0.2,3.0)	(24,810)	(1.0,30)	(24,42)

Table 6: Posterior medians and equal-tail 90% intervals under model (a): fixed length infectious period; and model (b) gamma-distributed infectious period. Estimates are based on 4,000 outputs from the MC-w-M algorithm ($\equiv 1.6 \times 10^6$ accept/reject decisions), with simulation size $n = 8,000$.

with

$$\begin{aligned} \mu_1' &= (\mu_L - 2\mu_I)/5 \\ \mu_2' &= (2\mu_L + \mu_I)/10. \end{aligned}$$

Improper, uniform prior distributions were assigned to μ_1' and μ_2', while $1/\theta_{L2}$ and $1/\theta_{I2}$ were each assigned Uniform(0,1) priors, so that only Gamma distributions with modes away from zero had *a priori* support. The force of infection parameter, β, was replaced as a basic parameter by the 'escape' probability q, given at (3.7). Both q and Δ, the proportion of 'two-primary' households, were assigned independent Uniform(0,1) priors.

The proposal distributions for all six working parameters were (independent) Uniform(0,1), each centred at the current value but with differing interval lengths, and with reflection at boundaries (0 is a lower boundary for all variables; 1 an upper boundary for all but μ_1' and μ_2', which must satisfy $\mu_2' > \mu_1'$).

3.4.3 Results

Before reporting results for the principal model, we discuss a simpler model similar to that employed in previous studies, in which the infectious period was of constant length. (This corresponds to $\theta_{I1} = \infty$ in our model.) The expected frequencies under the model, shown in row 2 of Table 5, correspond well to the observations: the Pearson χ^2 goodness-of-fit statistic is about 6.8 on 8 degrees of freedom. However, the

parameter estimates (Table 6) are imprecise, particularly the length of the infectious period, μ_I, whose equal-tailed 90% interval is from 0.5 days to 6 days. Further, the posterior distribution for the force of infection, β, obtained by inverting (3.7), is highly skew. The reason is clear from the left panels of Figure 3.2. The upper left panel shows that μ_L and μ_I are highly correlated ($\rho = -0.96$), while the lower left panel shows that $\log(\mu_I)$ and $\log(\beta)$ are almost perfectly correlated ($\rho = -1.00$). Further, they lie on a line of slope -1, indicating that $\beta\mu_I$ is almost constant. Roughly speaking, under this model the data cannot 'choose' between the following two scenarios (and intermediate cases):

$$\begin{array}{lllll} (1) & \mu_L \approx 11, & 0 < \mu_I < 1, & \beta > 3; \\ (2) & \mu_L \approx 9, & 5 < \mu_I < 6, & \beta \approx 1/3. \end{array}$$

Under all scenarios $\theta_{L2} \approx 25$, indicating a distribution for the latent period which is close to normal, with standard deviation about 2 days (90% interval (1.5,2.2)).

Figure 3.2 near here

We turn now to our principal model, in which the infectious period varies from case-to-case according to a Gamma distribution. The expected frequencies (Table 5, row 3) again correspond well to observations: the χ^2 statistic is 5.6 a gain of just over one from the simpler model, at the cost of one extra degree of freedom. However, the posterior distributions of the parameters are substantially altered, as is shown in the right panels of Figure 3.2. In particular, the support for scenario (2) above has almost vanished, while that for scenario (1) has increased. The standard deviation of the latent period is reduced to 1.4 (90% interval (0.4,2.2)), whereas that for the infectious period is about 1.2 (90% interval (0.1,1.8)), which is high compared with its median of 1.6, making the assumption of fixed length infectious period appear untenable. The posterior distribution for β supports larger values than under the simpler model, but remains highly skew.

3.5 Comparisons with previous analyses

We now briefly compare the results obtained by the Gibbs sampler (weak priors) and Monte Carlo-within-Metropolis methods with those obtained in the analyses of Bailey (1975, Ch. 15) and Becker (1989, Ch. 4). Table 7 summarises the findings. The MCMC estimates for μ_L are appreciably higher than the maximum likelihood estimates; specifically, the ML estimates do not lie within the credibility intervals given in Table 6. However, it is reassuring to note that the MCMC values are in harmony with current epidemiological views that the latent period of measles is about ten days; see for example Benenson (1990). The maximum likelihood estimates for μ_I also differ appreciably from the corresponding MCMC values, the former being considerably larger. It is plausible that the restrictive assumption of an infectious

	β	μ_L	σ_L	μ_I	σ_I
Gibbs Sampler	– –	10.9	2.3	– –	– –
MC-w-M, model (a)	0.51	9.9	2.0	3.2	– –
MC-w-M, model (b)	2.0	10.7	1.4	1.6	1.2
MLE (Bailey)	0.26	8.6	1.8	6.7	– –
MLE (Becker)	0.24	8.3	2.4	6.2	– –

Table 7: Posterior median values and Maximum likelihood estimates for the infection rate, and means and standard deviations of the latent and infectious periods.

period of constant length, as employed by both Bailey and Becker, has a bearing on the ML estimates for μ_L and μ_I. However, the difference in the μ_I values is not of great consequence since μ_I cannot be estimated precisely from the data. Our analyses also indicate the presence of significant variation in the infectious period, so that the assumption of a fixed infectious period appears inappropriate.

Finally, it is important to note that our analyses lead to credibility intervals for the parameters of interest. These intervals provide information which is considerably more reliable than that obtained via standard errors as presented in previous analyses (eg Bailey (1975, p.280)). There are two reasons for this : first, such standard errors should be treated with caution since the usual regularity conditions for asymptotic results are violated; and second, the assumptions of a fixed-length infectious period and a fixed value of Δ lead to under-estimation of standard errors because of a failure to incorporate variation in the infectious period length and uncertainty in the value of Δ. For example, Table 5 gives very wide credibility intervals for β. In contrast, the maximum likelihood approach gives an estimate of $\hat{\beta} = 0.256$ with standard error 0.032 (Bailey (1975, p.280)), suggesting a spurious degree of precision.

4 Conclusions

We have considered the application of Markov chain Monte Carlo methods to the analysis of infectious disease data using stochastic epidemic models. In principle, the use of MCMC has a number of important advantages over existing methods. In particular, MCMC can deal with complex models, thus permitting realistic modelling assumptions to be made; it is well-suited to missing data problems; it can often be implemented relatively easily; and it naturally caters for a Bayesian inference framework. Conversely, there are a number of *a priori* obstacles to the use of MCMC for epidemic data analysis. These include the considerable correlation structure inherent in epidemic models; potentially very large numbers of unknown variables; and mathematical difficulties that arise from the models themselves (for example, having to calculate distributions arising from conditioned Markov processes).

Final size data

For the influenza data considered in Section 2, a simple Metropolis-Hastings algorithm was found to be effective. Moreover our approach enabled a reasonable degree of modelling detail. In this case, there was no need to incorporate latent variables, and the analysis was aided considerably by the fact that the likelihood was straightforward to calculate via certain closed-form expressions.

Temporal data

The measles data set considered in Section 3 proved considerably more challenging. The data were temporal, and the modelling approach involved the inclusion of a large number of latent variables. The Metropolis-Hastings and Gibbs sampler algorithms both encountered some convergence problems, which appeared to be particularly due to the high correlations between the parameters of interest. These correlations arise due to a lack of detail in the data. Although it may have been possible to produce results simply by running the simulations for far longer periods of time, this approach was rejected as being practically undesirable. Our Monte-Carlo-within-Metropolis approach seemed to work well in practice, and gave the most complete results for the measles data set. This approach does not appear to have been considered before in the MCMC literature, and is clearly worthy of further investigation and development.

Despite the difficulties encountered with the data set, we nevertheless found that the MCMC methods were able to produce inferences that are in broad agreement with parameter values accepted in the epidemiological literature, and these were obtained under assumptions that are more realistic than those used in previous maximum likelihood analyses. The MCMC methods readily provide credibility intervals for the parameters of interest. In contrast, parameter confidence intervals obtained in previous analyses should be treated with caution, since the assumption of a constant length infectious period invalidates the requisite regularity conditions for standard asymptotic results of maximum likelihood estimation.

Furthermore we observed appreciable differences between the MCMC parameter inferences and the ML parameter estimates for the measles data. It thus appears that MCMC methods have a real contribution to make towards the analysis of infectious disease data.

Future work

In summary, standard MCMC methods can be used to provide useful inferences from infectious disease data, but are not guaranteed to be successful for all problems. There is clearly a great deal of scope for further work; firstly in terms of specific applications, and secondly in terms of improved methodology. Regarding the latter, since we have in this paper deliberately utilised the most common MCMC algorithms, an obvious avenue for further exploration is to consider more elaborate algorithms. In particular it may be fruitful to consider algorithms which take account of the structure of the target density, such as algorithms based on discrete approximations to Langevin diffusions (Roberts and Tweedie (1996)), or Hybrid Monte Carlo methods

(Duane *et al* (1987). In addition to the algorithms themselves, our experiences suggest the need to construct, where possible, model parameterisations that reduce posterior correlations. This strategy, which is generally regarded as desirable in the MCMC literature, is of particular importance in our context due to the correlation structures commonly found in stochastic epidemic models.

Acknowledgement
This project had its origins at the RSS Epidemic Workshop, Sabhal Mor Ostaig, Skye, 31 March-12 April, 1997, funded by EPSRC, and has benefitted considerably from discussions with participants of the Workshop. In particular, we thank Mark Handcock for his part in the early development of this work.

5 Appendix

A1: *Proof of Lemma 3.1*

We first find $\gamma > 0$ such that f is maximised at γ. Differentiating f yields that γ is the positive root of the quadratic

$$a(x+c) - bx(x+c) - dx = 0, \tag{5.1}$$

and it follows that γ is as defined in the statement of the Lemma. Next, $f(x) \leq g(x)$ if and only if

$$\frac{x^{a-b\gamma}}{(x+c)^d} \leq \frac{\gamma^{a-b\gamma}}{(\gamma+c)^d}. \tag{5.2}$$

It is thus sufficient to show that the LHS of (5.2) is maximised over $(0,\infty)$ when $x = \gamma$. By differentiation, the maximum is attained when $(a-b\gamma)(x+c) - dx = 0$. However, this last equation has unique positive root given by $x = \gamma$, since in this case the equation reduces to (5.1).

A2: *Proof of Lemma 3.2*

First define γ as the point at which h is maximised over $(0,\infty)$; differentiating $\log h(x)$ yields the definition of γ given in the statement of the Lemma. Next, note that the Lemma follows if and only if

$$\delta \geq \frac{x^{a-\gamma\mu}\exp((b+\mu)x)}{\Gamma(x)^c}. \tag{5.3}$$

However, the RHS of (5.3) is maximised over $(0,\infty)$ when $(a-\gamma\mu)x^{-1}+b+\mu-c\psi(x) = 0$, which by the definition of γ implies that $x = \gamma$, and the result follows.

A3: *Sampling methods for the non-global parameters*

Sampling u_1

Since $u_1 < 0$, we focus attention on the quantity $-u_1$. From (3.5) we obtain that, for $u_1 < s$,

$$\pi(-u_1|\cdots) \propto (-u_1)^{m-1+\theta_{I2}-1}\mathrm{e}^{-(-u_1)(\lambda+\beta+\theta_{I1})}.$$

Thus $-u_1$ has a $\mathrm{Gam}(\lambda+\beta+\theta_{I1}, m+\theta_{I2}-1)$ distribution, subject to the constraint $u_1 < s$. To obtain a sample from this conditional distribution we could repeatedly sample from the Gamma distribution until a sample satisfied the constraint. However, this approach will be highly time-consuming if the event $u_1 < s$ has low probability according to the Gamma distribution, and so in practice a more efficient method is preferable. An example of such a method is described below in Section A4.

Sampling s

¿From (3.5) we obtain that for $u_1 < s < 0$,

$$\pi(s|\cdots) \propto \exp(-\beta(s-u_1))(u_2-s)^{\theta_{L2}-1}\exp(-\theta_{L1}(u_2-s)).$$

There are now two possibilities. If $\beta \leq \theta_{L1}$ then we find that

$$\pi(s|\cdots) \propto (u_2-s)^{\theta_{L2}-1}\exp(-(\theta_{L1}-\beta)(u_2-s)),$$

so that (u_2-s) has a $\mathrm{Gam}(\theta_{L1}-\beta, \theta_{L2})$ distribution subject to the constraint $\max(0,u_2) < (u_2-s) < (u_2-u_1)$. Sampling from a Gamma distribution over a finite interval I is straightforward; for example, if the density (f, say) has maximum value M over I then we can simply sample Z uniformly over I and accept with probability $f(Z)/M$.

Alternatively, if $\beta > \theta_{L1}$ then we have

$$\pi(s|\cdots) \propto (u_2-s)^{\theta_{L2}-1}\exp((\beta-\theta_{L1})(u_2-s)),$$

again subject to the constraint $\max(0,u_2) < (u_2-s) < (u_2-u_1)$. The situation is similar to the previous case, except that now we are sampling from the density proportional to $f(x) = x^{\theta_{L2}-1}\mathrm{e}^{(\beta-\theta_{L1})x}$ over a finite interval. Such sampling can be carried out fairly efficiently by noting that $f(x) \leq M\mathrm{e}^{(\beta-\theta_{L1})x}$, with M a suitable positive constant, and that this bounding density can be simulated directly by using the inversion method (see for example Morgan (1984, Section 5.2)).

Sampling τ

Since $\tau < 0$ we consider $-\tau$, and note that $-\tau$ must satisfy $-\tau > |\min(v_1,v_2)| > 0$. Define $M = M(m_\tau, \lambda_\tau)$ as the maximum value of the Gamma density $\pi(-\tau)$ over the interval $(|\min(v_1,v_2)|, \infty)$. From (3.6) we find that, for $\tau < \min(v_1,v_2)$,

$$\begin{aligned}
\pi(-\tau|\cdots) &\propto \left(\frac{v_1+v_2}{2}-\tau\right)^{2(\theta_{L2}-1)} \exp\left(-2\theta_{L1}\left(\frac{v_1+v_2}{2}-\tau\right)\right) \frac{[(v_1-\tau)(v_2-\tau)]^{\theta_{L2}-1}}{\left(\frac{v_1+v_2}{2}-\tau\right)^{2(\theta_{L2}-1)}} \\
&\quad \times \frac{(-\tau)^{m-1}\mathrm{e}^{(-\lambda)(-\tau)}}{M} \\
&= \left(\frac{v_1+v_2}{2}-\tau\right)^{2(\theta_{L2}-1)} \exp\left(-2\theta_{L1}\left(\frac{v_1+v_2}{2}-\tau\right)\right) F_1(\tau)F_2(\tau),
\end{aligned}$$

say. By the AM-GM inequality it can easily be shown that $F_1(\tau) \leq 1$, while $F_2(\tau) \leq 1$ from the definition of M.

We can thus sample τ as follows. First sample $\left(\frac{v_1+v_2}{2} - \tau\right)$ from a $\text{Gam}(2\theta_{L2}, 2\theta_{L2}-1)$ distribution, which in turn gives us a sample for τ. This sampling is repeated until $\tau < \min(v_1, v_2)$. Finally, accept the resulting τ value with probability $F_1(\tau)F_2(\tau)$, repeating the whole process until an acceptance occurs.

Sampling u_2, v_1 and v_2

It is straightforward to see from (3.5) and (3.6) that the conditional densities for the remaining parameters u_2, v_1 and v_2 are identical in structure; essentially each involves a product of two Gamma densities. We will therefore only describe the method for the first of these parameters, u_2. Now from (3.5) we obtain that for $s < u_2 < w$,

$$\pi(u_2|\cdots) \propto (u_2 - s)^{\theta_{L2}-1} \exp(-\theta_{L1}(u_2 - s))(w - u_2)^{\theta_{I2}-1} \exp(-\theta_{I1}(w - u_2)).$$

We therefore wish to generate samples from densities of the form

$$f(x) = (x - s)^{m-1}(w - x)^{n-1} \exp(-\lambda(x - s)) \exp(-\mu(w - x)),$$

where $s < x < w$ and $n, m > 0$.

If $\lambda = \mu$ the situation is straightforward, since then f is simply the density of a Beta distribution with parameters (m, n) over the interval (s, w). Now for $p \geq 0$,

$$f(x) \propto \begin{cases} (x - s)^{m-p-1}(w - x)^{n-1}(x - s)^p \exp(-(\lambda - \mu)(x - s)) & \text{if } \lambda > \mu, \\ (x - s)^{m-1}(w - x)^{n-p-1}(w - x)^p \exp(-(\mu - \lambda)(w - x)) & \text{if } \lambda < \mu. \end{cases} \quad (5.4)$$

It follows at once from (5.4) that if $\lambda > \mu$ we can simply sample $(x - s)$ (which immediately specifies x) from a $\text{Beta}(m - p, n)$ distribution over the interval (s, w), and accept this sample with probability proportional to the Gamma density $(x - s)^p \exp(-(\lambda - \mu)(x - s))$. The latter acceptance probability is easy to obtain since we are only considering values of $(x - s)$ in the range $0 < (x - s) < (w - s)$.

Although this sampling scheme works for any $p \geq 0$, in practice we wish to pick p so that the acceptance probability is likely to be high. This amounts to ensuring that the Beta distribution is likely to generate samples in the region where the Gamma density is highest. One way to do this is to make the mean of the Beta distribution, namely $(w - s)(m - p)/(m - p + n)$, equal to the mode of the Gamma density, which is $p/(\lambda - \mu)$. With the additional constraint that $0 < p < m$ we thus find that

$$p = (1/2)\left((m + n + A) - \sqrt{(m + n + A)^2 - 4mA}\right),$$

where $A = (w - s)(\lambda - \mu)$. Finally, the case $\lambda < \mu$ is treated similarly.

A4: *Sampling from a conditional Gamma density*

Our objective here is to sample efficiently from the Gamma density proportional to $f(x) = x^{m-1}\mathrm{e}^{-\lambda x}$, where $m, \lambda > 0$, conditional on $x > y$, where $y \geq 0$. Note that the

Gamma distribution has mean m/λ and standard deviation $\sqrt{m}/\lambda$. We shall use the notation $\mathrm{Exp}(\lambda)$ to denote an Exponential distribution with mean λ^{-1}. We consider various cases, as follows.

(i) $m = 1$. In this case we are simply sampling from the Exponential density $\mathrm{e}^{-\lambda x}$, conditional on $x > y$. But since this density is proportional to $\mathrm{e}^{-\lambda(x+y)}$, it follows that if Z is drawn from the $\mathrm{Exp}(\lambda)$, $Z + y$ will be a sample from the required conditional distribution.

(ii) $m < 1$. There are two methods, as follows. First, if $y \leq m/\lambda$ then it will be reasonably efficient to simply sample repeatedly from $\mathrm{Gam}(\lambda, m)$ until a sample Z satisfies $Z > y$. Second, if $y > m/\lambda$ we use rejection sampling with an exponential bounding density: specifically, if $x > y$, $y^{m-1}\mathrm{e}^{-\lambda x} \geq x^{m-1}\mathrm{e}^{-\lambda x}$. Thus we simply sample Z from $\mathrm{Exp}(\lambda)$, set $X = Z + y$, and accept with probability $(X/y)^{m-1}$. (Note that the choice of m/λ as the boundary between the two sampling methods here is essentially arbitrary; however, it is clear that as the boundary increases, so the first method becomes increasingly inefficient).

(iii) $m > 1$. In this case f is unimodal on $(0, \infty)$. As in (ii) there are two methods. If $y < (\sqrt{m}+m-1)/\lambda$, ie y is less than the mode of f plus one standard deviation, then use simple rejection sampling as in case (ii). Otherwise, set $\mu = (\lambda\sqrt{m})/(\sqrt{m}+m-1)$ and note that $(y^{m-1}\mathrm{e}^{-(\lambda-\mu)y})\mathrm{e}^{-\mu x} \geq x^{m-1}\mathrm{e}^{-\lambda x}$. Then, sample Z from $\mathrm{Exp}(\mu)$, set $X = Z + y$, and accept with probability $(X/y)^{m-1}\mathrm{e}^{-(\lambda-\mu)(x-y)}$.

6 References

Addy, C. L., Longini, I. M. and Haber, M. (1991) A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics* **47**, 961-974.

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

Ball, F. G. and O'Neill, P. D. (1999) The distribution of general final state random variables for stochastic epidemic models. To appear, *J. Appl. Prob.*, June 1999.

Ball, F. G., Mollison, D. and Scalia-Tomba, G. (1997) Epidemics with two levels of mixing. *Ann. Appl. Prob.* **7**, 46-89.

Becker, N. G. (1989) *Analysis of Infectious Disease Data.* London: Chapman and Hall.

Benenson, A. S. (Ed.) (1990) *Control of Communicable Diseases in Man.* 15th Edition. New York : American Public Health Association.

Diggle, P. J. and Gratton, R. J. (1984) Monte Carlo methods of inference for implicit statistical models. *J. Roy. Stat. Soc.* B **46**, 193-227.

Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987) Hybrid Monte Carlo. *Phys. Lett.* B **195**, No. 2, 216-222.

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.

Gilks, W. R., Richardson, S. and Spiegelhalter D. J. (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Hills, S. E. and Smith, A. F. M. (1992) Paramaterization issues in Bayesian inference (with discussion). In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 641-649. Oxford: Oxford University Press.

Lefèvre, C. and Picard, P. (1990) A non-standard family of polynomials and the final size distribution of reed-frost epidemic processes. *Adv. Appl. Prob.* **22**, 25-48.

Longini, I. M. and Koopman, J. S. (1982) Household and community transmission parameters from final distributions of infections in households. *Biometrics* **38**, 115-126.

Monto, A. S., Koopman, J. S., and Longini, I. M. (1985) Tecumseh study of illness. XIII. Influenza infection and disease, 1976-1981. *Am. J. Epid.* **121**, 811-822.

Morgan, B. J. T. (1984) *Elements of Simulation.* London: Chapman and Hall.

O'Neill, P. D. and Roberts, G. O. (1999) Bayesian inference for partially observed stochastic epidemics. *J. Roy. Stat. Soc.* A **162**, 121-129.

Roberts, G. O. and Tweedie, R. L. (1996) Exponential convergence of Langevin diffusions and their discrete approximations. *Bernoulli* **2**, 4, 341-363.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Stat. Soc.* B **55**, 1, 3-23.

Tanner, M. A. (1996) *Methods for the exploration of posterior distributions and likelihood functions*. New York: Springer.

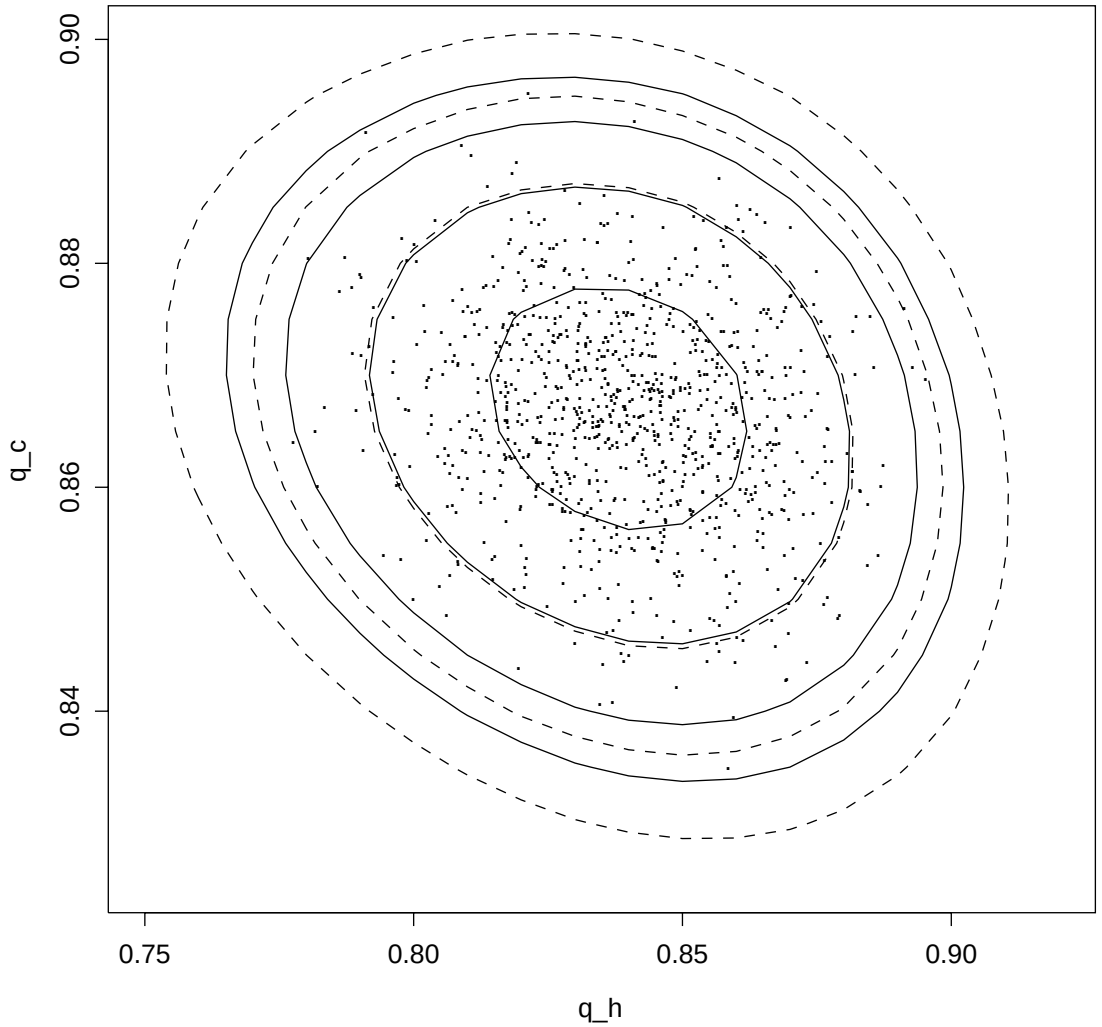

FIG 2.1. Influenza data: Reed-Frost model without protection: MCMC sample values (1000 values, at sampling interval 10). The solid contour lines surround highest posterior density credible intervals at 50, 90, 99 and 99.9 % levels. The dashed contour lines indicate posterior pdf values of 10, 1 and 0.1 % of its maximum.

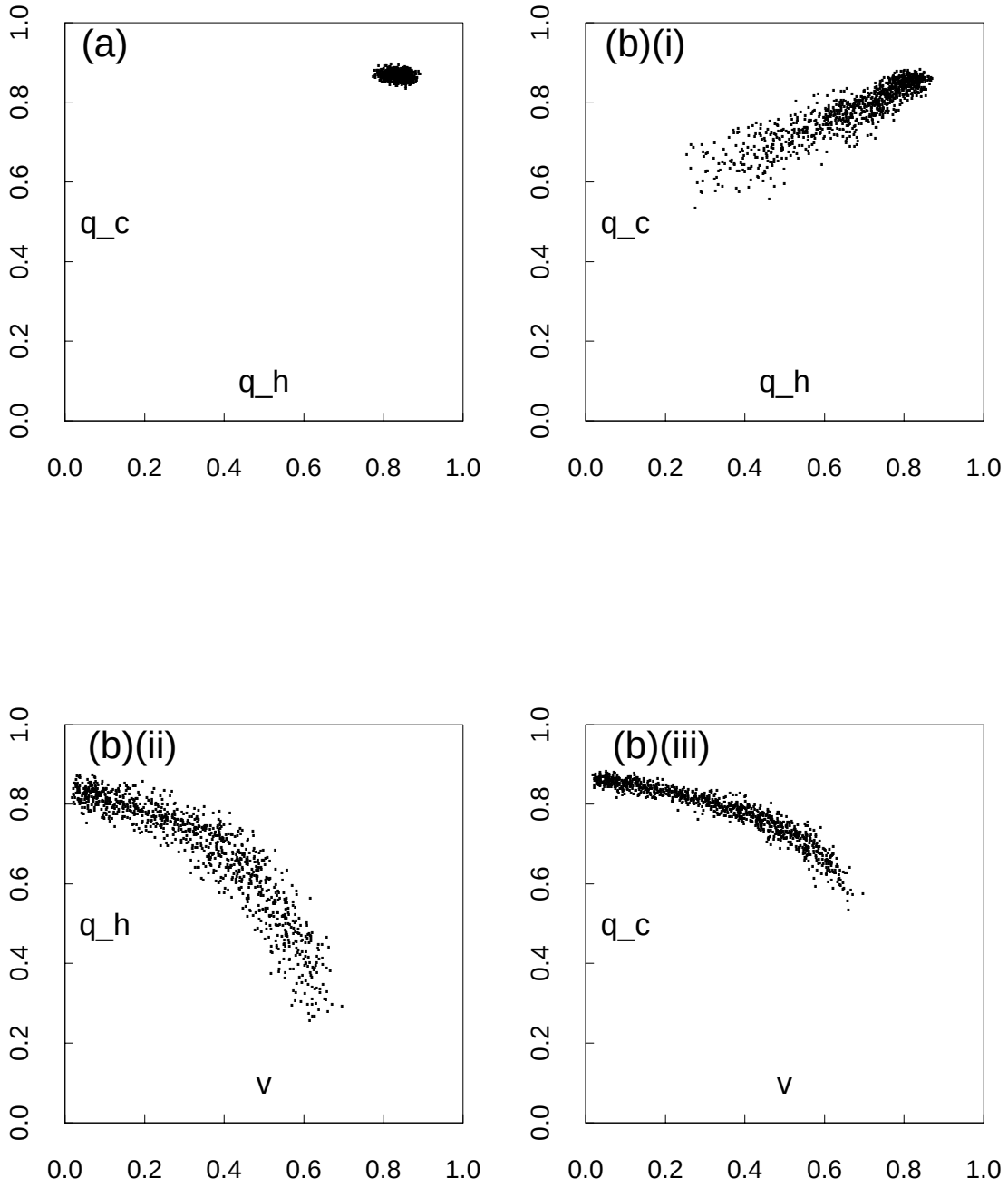

FIG. 2.2. Influenza data: Reed-Frost models (a) without protection (same as in Figure 2.1). (b) with protection (again 1000 MCMC sample values, at sampling interval 10), showing the wide range of credible values for the protection parameter v (roughly 0 to 0.7, with higher values of v corresponding to lower values of both q_h and q_c).

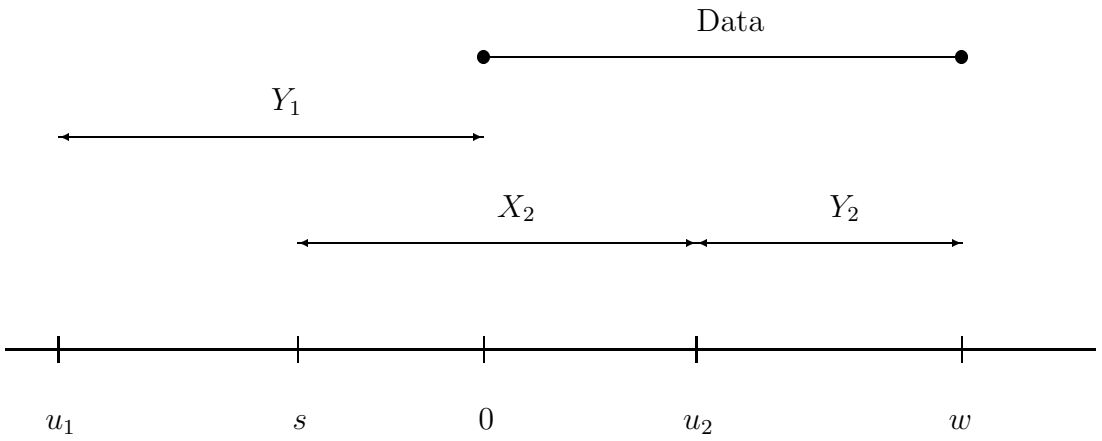

FIG. 3.1. Measles data: latent and infective periods for two individuals. Here Y_1 and Y_2 denote respectively the infectious periods of the primary and secondary infectives, and X_2 denotes the latent period of the secondary infective.

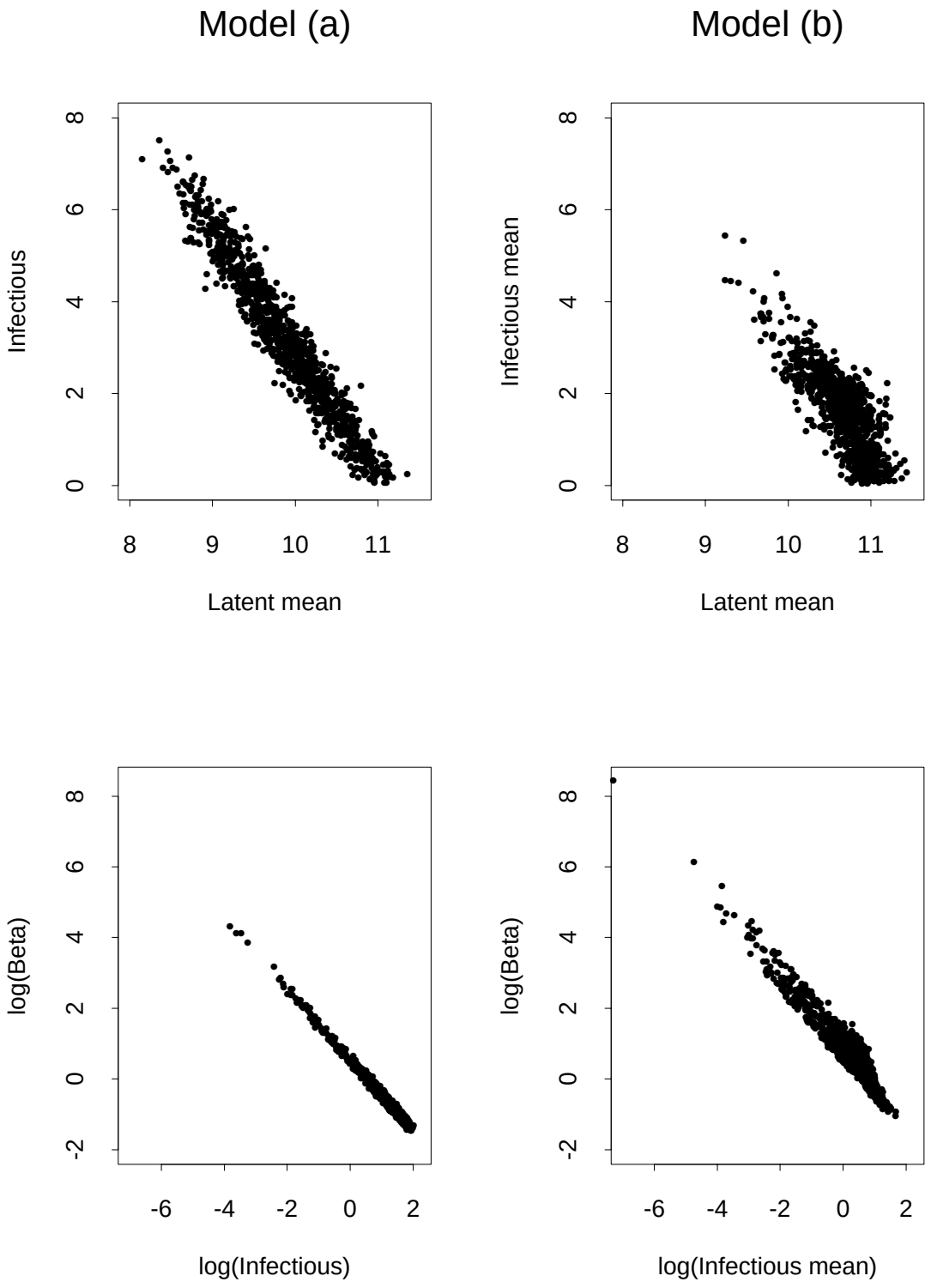

FIG. 3.2. Measles data: scatter plots of 1,000 points drawn from the output of the MC-w-M algorithms under model (a): fixed-length infectious period (left panels); and model (b): gamma-distributed infectious period (right panels). The upper panels show μ_L versus μ_I; the lower panels show $\log(\mu_I)$ versus $\log(\beta)$.