Exact Bayesian Inference and Model Selection for Stochastic Models of Epidemics Among a Community of Households

DAMIAN CLANCY
Department of Mathematical Sciences, University of Liverpool

PHILIP D. O’NEILL
School of Mathematical Sciences, University of Nottingham

ABSTRACT. Much recent methodological progress in the analysis of infectious disease data has been due to Markov chain Monte Carlo (MCMC) methodology. In this paper, it is illustrated that rejection sampling can also be applied to a family of inference problems in the context of epidemic models, avoiding the issues of convergence associated with MCMC methods. Specifically, we consider models for epidemic data arising from a population divided into households. The models allow individuals to be potentially infected both from outside and from within the household. We develop methodology for selection between competing models via the computation of Bayes factors. We also demonstrate how an initial sample can be used to adjust the algorithm and improve efficiency. The data are assumed to consist of the final numbers ultimately infected within a sample of households in some community. The methods are applied to data taken from outbreaks of influenza.

Key words: Bayesian inference, Bayesian model selection, epidemic data, influenza, rejection sampling, stochastic epidemic models

1. Introduction

This paper is concerned with methodology for sample-based Bayesian inference for certain stochastic epidemic models. Specifically, we develop multivariate rejection sampling algorithms to address inference and model-choice questions for the analysis of final-outcome household data on disease outbreaks. The methods are straightforward to use, and practical in terms of real-time implementation. Moreover, they avoid the inherent convergence difficulties with Markov chain Monte Carlo (MCMC) methods. In the following paragraphs we briefly review the relevant background material.

The use of stochastic models to analyse data from infectious disease outbreaks is, in general, a non-standard problem that frequently requires special methodology. This can be due to dependencies in the data (as is the case if, e.g. an epidemic outbreak is observed in real time), missing data (such as the infection process itself being unobserved) or inherent difficulties with calculating likelihoods based on non-linear stochastic models. Recently, considerable progress has been made by the application of MCMC methods (see, e.g. Gibson, 1997; O’Neill et al., 2000; O’Neill, 2002; Becker et al., 2003; Cauchemez et al., 2004; O’Neill & Marks, 2005; and references therein). MCMC methods naturally permit sample-based Bayesian analyses to be performed, and have a very wide range of applicability. However, an inherent drawback with their use is that it is usually only possible to detect non-convergence of the simulated Markov chain. Thus apparent convergence might be misleading, for example, if the Markov chain has actually failed to sufficiently explore the tails of a posterior
density. Sometimes, perfect simulation methods can be employed to ensure convergence has occurred (O’Neill, 2003), but such methods are not widely applicable.

These convergence difficulties can be overcome in certain situations by using rejection sampling (see, e.g. Ripley, 1987), and this is the focus of this paper. Our approach enables us both to perform sample-based Bayesian inference and also to consider questions of model choice via the calculation of Bayes factors. We also show how to use an initial training sample to adjust an algorithm to increase its efficiency. The main challenge is that the setting we consider is typically multivariate, and moreover different models being compared may differ in the dimensionality of their parameter space. We thus extend the standard rejection sampling algorithm, and to our knowledge such an extension has not previously been considered.

The paper is structured as follows. Section 2 contains details of the stochastic epidemic models and data of interest, and section 3 the rejection sampling methodology. Section 4 describes applications to various data sets taken from outbreaks of influenza, and section 5 contains some concluding comments.

2. Epidemic models, data and likelihood

In this section, we describe a basic model for the spread of a disease among a community of households, and then consider variants which allow either for extra uncertainty to be incorporated or for differing individual risk levels.

2.1. Models

Basic model. The following model, described in O’Neill et al. (2000), generalizes a model proposed by Longini & Koopman (1982). For ease of exposition we describe the model in a temporal setting. However, since our objective involves fitting the model to non-temporal data concerning just the final outcome, it will be sufficient for us to parameterize the model in terms of non-temporal quantities, namely transmission probabilities. In particular, the likelihood of the data given the model can be expressed in terms of such probabilities.

Consider a population of individuals partitioned into households of various sizes, so that each individual resides in precisely one household. Initially, all individuals are assumed to be susceptible, meaning that they could potentially become infected by the disease in question. The model assumes that all individuals in the population are subject to a global force of infection that is assumed to come from the community at large. In practice, this usually reflects the assumption that the population of interest is itself just some fraction of a larger community. To remain susceptible throughout the epidemic, an individual must escape infection both from this global force of infection and also from any members of their household who ever become infected. If an individual becomes infected, then they remain so for a period of time known as the infectious period. During this period they are able to potentially infect other members of their household, as described below. At the end of its infectious period, an individual is said to be removed, meaning that it is no longer infectious, and moreover cannot be reinfected. This reflects the assumption, appropriate for many diseases that a period of immunity follows infectiousness.

To model community-acquired infection, it is assumed that the global force of infection acts equally and independently on all individuals in the population. Specifically, if the epidemic occurs during a time interval \([0, T]\), and the force of infection at time \(0 \leq t \leq T\) is \(\lambda(t)\), it follows that the probability that an individual avoids global infection during the epidemic is \(q_c = \exp(- \int_0^T \lambda(t) \, dt)\).
To model infection within the household, we proceed as follows. An infected individual has an infectious period that is distributed according to some specified non-negative random variable $I$. The infectious periods of different infectives are assumed to be mutually independent. During its infectious period, an infective individual has potentially infectious contacts with each other member of its household at times given by the points of a homogeneous Poisson process of rate 1. Each such contact with a susceptible results in the immediate infection of that individual. All the Poisson processes describing contacts between different pairs of individuals are assumed to be mutually independent.

For $k \geq 0$ let $q_h(k)$ denote the probability that a single infective fails to infect $k$-susceptible household members. Then the assumptions above imply that $q_h(k) = E[\exp(-kI)]$, so that in fact the within-household transmission probabilities can be expressed in terms of the moment-generating function of the infectious period random variable $I$. If $I \equiv c$ is non-random, then the within-household epidemic model is of Reed–Frost type, and $q_h(k) = e^{-ck} = q_h^c$, say. This assumption is made in Longini & Koopman (1982). If $I$ is exponentially distributed then the within-household epidemic is a general stochastic epidemic model (see Andersson & Britton, 2000). Note that the parameter vector of $I$, henceforth denoted $\theta_i$, will also be regarded as a parameter of the model.

In order to evaluate a likelihood for the above model, given data on the numbers infected in each household, it is necessary to evaluate $\omega_{js}$, the probability that exactly $j$ of the $s$ initial susceptibles of a given household are ultimately infected. The derivation of the likelihood involves a certain amount of algebra, details of which can be found in Ball et al. (1997), and so here we simply quote the result. We first require the following definition (see, e.g. Lefèvre & 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

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

The desired probability, $\omega_{js}$, is given in equation (2.5) of O’Neill et al. (2000):

$$
\omega_{js} = \frac{1}{(s-j)!} \sum_{i=s-j}^{s} \frac{s!}{(s-i)!} (q_h(i))^{s-i} q_h^c G_{s-i+j}(0 | E^{s-j}U).
$$

in which $U$ is the sequence with $i$th term $u_i = q_h(i)$, and where $E^iU$ denotes the sequence $u_i, u_{i+1}, \ldots$.

**Household variation model.** It is reasonable to suppose that, in reality, the probability of avoiding infection from the community may vary between individuals or households. In the former case, one could assume a random-effects model for the individual avoidance probabilities, so that the $q_c$ value for each individual was sampled, independently of other individuals, from some hyperdistribution $Q$. However, the $q_c$ term in (1), which is the probability that $i$ individuals avoid community-acquired infection (i.e. it is not the probability that a single individual avoids such infection $i$ times) would then be replaced by $(E[Q])^i$, so that in fact the model is essentially as before. Instead we consider household variation, motivated by real-life effects such as household location, environmental factors, dietary factors, etc. All these factors are likely to vary on a household level and could affect an individual’s susceptibility to infection from the community. Modelling such household variation has been considered by a number of other authors; see Becker, 1989, sections 3.4–3.6 and references therein.
Specifically, suppose that the value of \( q_c \) for a given household is a sample from a random variable \( Q \) taking values in \([0, 1]\), \( q_c \) values for distinct households being chosen independently of one another. It follows that the corresponding final outcome probabilities given in (1) will be altered by replacing \( q_i^c \) with its expectation with respect to \( Q \). In general, the required moments of \( Q \) could be included as extra model parameters. However, in practice it might not be possible to separately identify these from the data, and so it is often more pragmatic to adopt a parametric form for \( Q \) with a small number of parameters. We shall denote by \( \theta_Q \) the vector of parameters of the distribution of \( Q \).

\[ \text{Differing levels of susceptibility.} \text{ Longini et al. (1988)} \text{ extend the model of Longini & Koopman (1982) to allow each individual’s probability of avoiding infection (both from the community and from within their own household) to depend upon their own risk level, which may reflect age, sex, pre-immunity via vaccination or previous exposure and so on. Each individual is assigned a risk level } r \in \{1, 2, \ldots, R\}. \text{ The model assumes a within-household epidemic of Reed–Frost type. Thus the probability that an individual of risk level } r \text{ avoids infection from } k \text{-infectious individuals within the same household is } (q_r)^k, \text{ and the probability that the individual avoids infection from the community is } q_c r. \text{ The probabilities corresponding to those in (1) can be computed as described in Longini et al. (1988).} \]

2.2. Data and likelihood

The data are assumed to be of the form \( n = \{n_j\} \), where for \( 0 \leq j \leq s \), \( n_j \) denotes the number of households in which \( j \) of \( s \) initial susceptibles become infected during the epidemic. In the sequel, we use \( \pi \) to denote densities or mass functions, as appropriate. The likelihood of the data for the basic model is

\[ L(q_c, \theta_I) = \pi(n | q_c, \theta_I) \propto \prod_{j,s} \{\omega_j(q_c, \theta_I)\}^{n_j}, \]

where as before \( q_c \) is the community avoidance probability, and \( \theta_I \) the parameter vector associated with the infectious period distribution \( I \). The likelihoods for the household variation model and the varying susceptibility model are similar.

2.3. Bayesian inference and model selection

Our objective is to make inferences about the model parameters given the data. Within the Bayesian framework, our objective is thus to investigate the joint posterior density of the model parameters conditional upon the data. By Bayes’ theorem we have, for the basic model,

\[ \pi(q_c, \theta_I | n) \propto \pi(n | q_c, \theta_I)\pi(q_c, \theta_I), \]

where \( \pi(q_c, \theta_I) \) is the prior density on \((q_c, \theta_I)\), and \( \pi(n | q_c, \theta_I) \) is defined at (2). The posterior densities for the household variation model and varying susceptibility model can be derived similarly.

O’Neill et al. (2000) describe a simple Metropolis–Hastings algorithm for sampling from the target posterior density of the basic model. In the next section, we describe how samples can instead be drawn using an exact rejection sampling method.

We also address the issue of model selection. For instance, we assess whether the data provide real evidence of household variation, in the sense of differing probabilities of infection from the community for different households. Denote by \( M_1 \) our basic model, and by \( M_2 \) the model with household variation. One way to measure the strength of evidence for
one model relative to another is via the Bayes factor (see, e.g. O’Hagan, 1994). The Bayes factor in favour of model $M_1$ is defined as

$$BF_{12} = \frac{\pi(n|M_1)}{\pi(n|M_2)},$$

and denoting by $\pi(M_1), \pi(M_2)$ the prior probabilities of models $M_1, M_2$, respectively, may be conveniently computed by observing that

$$BF_{12} = \frac{\pi(M_1|n)\pi(M_2)}{\pi(M_2|n)\pi(M_1)},$$

i.e. $BF_{12}$ is simply the posterior odds ratio divided by the prior odds ratio. The latter ratio is obviously known, and the former available directly from the rejection sampling algorithm that we shall now describe.

3. Rejection sampling algorithm

Rejection sampling, in its most standard form, is designed to allow sampling from a univariate target distribution with known probability density function $f$, as follows (see, e.g. Ripley, 1987). A candidate value $y$ is first drawn from some proposal probability density $g$, where $g$ satisfies

$$K = \sup_{x \in \mathbb{R}} \left\{ \frac{f(x)}{g(x)} \right\} < \infty.$$  \hspace{1cm} (5)

The candidate $y$ is accepted with probability

$$\pi(y) = \frac{f(y)}{Kg(y)}.$$

Otherwise, a new candidate value $y$ is generated, and the algorithm continues until a candidate point is accepted. The set of accepted points provides a sample from the target density $f$.

The number of trials until a candidate is accepted has a geometric distribution with mean $K$, so the algorithm will work efficiently provided that the distribution $g$ is straightforward to sample from and the value of $K$ is small (or equivalently, the mean acceptance probability $\bar{\pi} = 1/K$ is large).

We now briefly describe the generalization of the above algorithm to a multivariate setting in which the target density is known only up to proportionality, and may be a mixture of components. In the following, we use the term ‘distribution’ to mean a probability measure on $(\mathbb{R}^d, \mathcal{B}^d)$, where $\mathcal{B}^d$ denotes the Borel sets on $\mathbb{R}^d$. For a distribution $F$ we denote its dimensionality by $\dim(F)$, and write $F \ll G$ if $F$ is absolutely continuous with respect to $G$. We also write $\mathcal{M}(F, G; p)$ to denote a mixture distribution with component distributions $F$ and $G$, with respective associated probabilities $p$ and $1-p$. Finally, $\text{Leb}^d$ denotes Lebesgue measure on $\mathbb{R}^d$.

Let $F$ be a target distribution on $\mathbb{R}^d$ from which samples are required. Suppose $F$ is known only up to proportionality, so that $F = kH$ for some $k > 0$, and that $F \ll G$ for some distribution $G$ with $\dim(G) = d$, the latter taking the role of the proposal distribution. Suppose further that the essential supremum with respect to $G$ of the Radon–Nikodym derivative $dH/dG$ is finite, and denoted by $K$. Finally suppose that $G$ is straightforward to sample from, and that $dH/dG$ is a known function on $\mathbb{R}^d$. 

To obtain a sample from the target distribution $F$, candidate points are sampled from the proposal distribution $G$ until one is accepted, the acceptance probability for candidate point $y$ being given by

$$
\pi(y) = \frac{1}{K} \frac{dH}{dG} \bigg|_y.
$$

The set of accepted points is a sample from the target distribution $F$. When considering inference for a single model, $dH/dG$ will simply be the ratio of the (typically multivariate) posterior density, known up to proportionality, and the density of the proposal distribution. Thus the rejection sampling algorithm in this case is entirely analogous to the simple univariate case.

In the sequel we shall frequently be concerned with model-choice issues, and in particular the case of two competing models, giving rise to the situation where the target distribution is a mixture. We now describe how the above algorithm applies to this scenario. Essentially all that is required is to ensure that the proposal distribution is itself of mixture type; then the algorithm proceeds as before.

To be specific, suppose that the target distribution $F$ is of the form $M(F_1, F_2; p_1)$ with $\dim(F_i) = d_i$, $i = 1, 2$. Suppose further that for $i = 1, 2$, $F_i$ has density $f_i$ with respect to $\text{Leb}^{d_i}$, and that $f_i = kh_i$ for some $k > 0$. Then we simply require that the proposal distribution $G$ be of the form $M(G_1, G_2; \hat{p}_1)$, with $\dim(G_i) = d_i$, $i = 1, 2$. Then if $G_i$ has density $g_i$, $i = 1, 2$, the bounding constant $K$ in the rejection sampling algorithm is given by

$$
K = \max \left\{ \sup_{x \in \mathbb{R}^{d_i}} \left\{ \frac{p_i h_1(x)}{\hat{p}_i g_1(x)} \right\}, \sup_{x \in \mathbb{R}^{d_2}} \left\{ \frac{(1 - p_i) h_2(x)}{(1 - \hat{p}_i) g_2(x)} \right\} \right\}.
$$

(6)

In the setting of Bayesian model choice, $F_i$ represents the posterior distribution of model $M_i$ ($i = 1, 2$) and $p_1 = \pi(M_1 | n)$. Samples from the mixture posterior distribution automatically enable estimation of $p_1$, and thus from (4) we may obtain an estimate of $BF_{12}$. For $i = 1, 2$, $f_i$ will be of the form $\pi_i L_i$, where $\pi_i$ denotes the parameter prior density and $L_i$ the likelihood under model $M_i$.

Finally, in practice it is often the case that the competing models are nested in the sense that one can be regarded as a special case of the other. For example, suppose $M_1$ is the basic model (say, with parameters $q_i$ and $\theta_i$ and dimension $d_1$) and $M_2$ the household variation model (say, with parameters $\theta_0$ and $\theta_i$, with dimension $d_2 > d_1$). Then $M_1$ may be nested in $M_2$, for example, by setting the first component of $\theta_0$ to be $q_i$ and the remaining components to zero. Thus the full posterior distribution (say, $F$) can be defined on $\mathbb{R}^{d_2}$. Note that it is not the case that $F \ll \text{Leb}^{d_2}$, as $F$ contains a non-zero mass of dimension $d_1$ corresponding to the region of parameter space where all components but the first of $\theta_0$ are zero. However, this fact causes no problem for the rejection sampling algorithm, provided that $G$ is chosen such that $F \ll G$.

4. Applications

In this section, we apply the rejection sampling algorithm described above to the models of section 2. The data that we consider are taken from outbreaks of influenza in Tecumseh, Michigan (see Monto et al., 1985) and in Seattle, Washington (see Fox & Hall, 1980) and are of the form required for our analysis. These data have been considered by a number of authors, which enables us to compare our findings with theirs.
4.1. Data

The data given in Table 1 are from two epidemics of influenza A (H3N2) that occurred in Tecumseh, Michigan in 1977–8 and 1980–1. Initially, we will follow previous authors (e.g. Addy et al., 1991) in combining the data from the two outbreaks, before treating them separately. More detailed analyses of data obtained from the outbreaks, taking into account other covariates such as age, have been considered elsewhere (see Addy et al., 1991). Our intention here is to illustrate our methodology rather than perform such detailed analyses.

The data given in Table 2 are taken from Longini et al. (1982) and are from two epidemics which occurred in Seattle, Washington: an outbreak of influenza B in 1975–6 and an outbreak of influenza A (H1N1) in 1978–9.

4.2. Algorithm implementation and results

Basic model. We consider first the model without household variation. Two cases are presented, corresponding to different assumptions concerning the within-household disease spread.

Reed–Frost model. For the model of Reed–Frost type, the parameters \( q_h(k) \) are given by \( q_h(k) = q_h^k \) for some \( q_h \in [0, 1] \). Thus the model has two parameters, \( q_c \) and \( q_h \), each with range

<table>
<thead>
<tr>
<th>Susceptibles per household</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1977–78</td>
<td>1980–1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. infected</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 6 7 8</td>
<td>5 6 7 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>79 105 48 44 12 2 1</td>
<td>54 84 60 62 19 6 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Susceptibles per household</th>
<th>Influenza B, 1975–1976</th>
<th>Influenza A (H1N1), 1978–1979</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. infected</td>
<td>0 1 2 3 4 5</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td></td>
<td>0 0 0 0 0 0</td>
<td>0 0 0 0 0 0</td>
</tr>
<tr>
<td>Total</td>
<td>10 20 28 20 9</td>
<td>26 50 17</td>
</tr>
</tbody>
</table>

Taking independent uniform [0, 1] prior distributions for $q_c$ and $q_h$, the joint posterior distribution satisfies

$$
\pi(q_c, q_h | n) \propto \prod_{j,s} \omega_{js}(q_c, q_h)^n_{js}.
$$

In the absence of any information about the shape of the posterior distribution, it is natural to define the proposal distribution $g(q_c, q_h)$ as a product of independent uniform [0, 1] distributions. Condition (5) then becomes

$$
K = \sup_{q_c, q_h \in [0, 1]} \left\{ \prod_{j,s} \omega_{js}(q_c, q_h)^n_{js} \right\} < \infty.
$$

The value of $K$ may be found numerically. We used the `fminsearch` minimization command in the Matlab 6 package, applied to the function $f(q_c, q_h) = -\sum_{j,s} n_{js} \ln(\omega_{js}(q_c, q_h))$.

Using the rejection sampling algorithm to generate 1000 points from the posterior distribution of $(q_c, q_h)$ gave the results shown in Fig. 1. Posterior mean parameter estimates are $\bar{q}_c = 0.87$, $\bar{q}_h = 0.85$. Both estimates are in close agreement with those obtained by O’Neill et al. (2000), although the analysis there did not include household sizes above 5. The mean acceptance probability was $\bar{z} = 0.0012$.

Although the mean acceptance probability $\bar{z}$ is large enough to yield samples rapidly, it is possible to do considerably better by a more careful choice of the proposal distribution $g$. Figure 1 illustrates that the 1000 points all fall within quite a small region within the unit square. Using a product of uniform [0, 1] distributions as the proposal distribution is thus very inefficient, as it will frequently propose points outside the modal region. However, the initial sample of 1000 points suggests an alternative proposal, which more closely reflects the shape of the posterior distribution. Specifically, independent Beta-proposal distributions for $q_c$ and $q_h$ were used, with parameters set so that in each case the mean of the proposal

\[0.76 \quad 0.78 \quad 0.8 \quad 0.82 \quad 0.84 \quad 0.86 \quad 0.88 \quad 0.9 \quad 0.92 \quad 0.94\]

\[0.83 \quad 0.84 \quad 0.85 \quad 0.86 \quad 0.87 \quad 0.88 \quad 0.89 \quad 0.9 \quad 0.91\]

\[q_c \quad 0.87 \quad 0.86 \quad 0.85 \quad 0.84 \quad 0.83 \]

\[0.76 \quad 0.78 \quad 0.8 \quad 0.82 \quad 0.84 \quad 0.86 \quad 0.88 \quad 0.9 \quad 0.92 \quad 0.94\]

\[q_h\]

Fig 1. Scatterplot of $q_c$ and $q_h$, basic model.
distribution was equal to the mean of the corresponding empirical posterior distribution of the first 1000 values generated. Additionally, the variance of the proposal distribution was set to be somewhat larger (a factor of 1.7 was found to be convenient, for each of \( q_c \) and \( q_h \)) than that of the empirical distribution of the first 1000 values generated.

Using these new proposal distributions to generate a further 1000 points, the mean acceptance probability was increased to \( \tilde{x} = 0.54 \). This clearly illustrates that some knowledge of the shape of the posterior distribution can lead to greatly increased algorithm efficiency. Using these improved proposal distributions, we generated a sample of 100,000 points from the posterior distribution. Considering separately the marginal posterior distributions of \( q_c \) and \( q_h \), the 95% equal-tailed credible intervals were found to be [0.84828, 0.88551] for \( q_c \) and [0.81154, 0.88853] for \( q_h \).

**Household variation model.** Consider now the model of Reed–Frost type (i.e. fixed infectious periods) extended to incorporate household variation. Recalling that the \( q_c \) value for each household is to be sampled from a random variable \( Q \), we will suppose that \( Q \sim \text{Beta}(A, B) \), i.e. \( Q \) has density \( f_Q(x) \propto x^{A-1}(1-x)^{B-1} \) for \( x \in [0, 1] \). So as to compute the coefficients \( \omega_j \), it is necessary to calculate expectations of the form \( \mathbb{E}[Q^i(1-Q)^m] \) for non-negative integers \( i \) and \( m \). As \( Q \sim \text{Beta}(A, B) \),

\[
\mathbb{E}[Q^i(1-Q)^m] = \frac{A(A+1)\ldots(A+i-1)B(B+1)\ldots(B+m-1)}{(A+B)(A+B+1)\ldots(A+B+i+m-2)}.
\]

The model has three parameters in total, namely \( q_c \in [0, 1] \) and \( A, B \in [0, \infty) \). However, working with these parameters is problematic, as \( A \) and \( B \) are highly correlated a posteriori, and cannot be estimated individually from the data with much precision. Consequently, we replaced \( A \) and \( B \) with the parameters \( m \) and \( r \) defined by

\[
m = \frac{A}{A+B}, \quad r = \frac{1}{A+B+1},
\]

so that \( m \) is the mean of the distribution from which \( q_c \) values are chosen, and \( r \) is a scale parameter (cf. Lee & Sabavala, 1987).

Note that \( m, r \in [0, 1] \). Again, we took independent uniform \([0, 1]\) prior distributions for each of the parameters \( m, r, q_h \). Initially, independent uniform \([0, 1]\) proposal distributions were used for each of \( m, r, q_h \). The rejection sampling algorithm generated 1000 points from the posterior distribution of \((m, r, q_h)\), giving the results shown in Fig. 2. Posterior mean parameter estimates are \( \bar{m} = 0.84, \bar{r} = 0.19, \bar{q}_h = 0.93 \). Comparing the estimates \( \bar{m}, \bar{q}_h \) with the 95% credible intervals obtained under the basic model for \( q_c \) and \( q_h \), respectively, we see that the model with household variation attributes rather more of the infectious spread to infection from the community, rather than within-household spread. The mean acceptance probability was \( \tilde{x} = 0.00090 \). Thus the increase in complexity in going from a two-parameter to a three-parameter model has resulted in a reduction in mean acceptance probability by a factor of approximately 0.75.

From Fig. 2, it is clear that using uniform \([0, 1]\) proposal distributions leads to an inefficient algorithm. For the parameter \( m \), we can improve efficiency by using a Beta-proposal distribution, with parameters chosen so that the proposal distribution has mean equal to the mean of the corresponding posterior distribution (based on the empirical posterior distribution of the initial sample), and variance somewhat higher (we chose a factor of 3) than that of the empirical posterior distribution. For the parameter \( q_h \), a complication arises when using...
a Beta-proposal distribution. Specifically, the posterior density at $q_h = 1$ is finite but non-zero, but a Beta($u, v$) proposal distribution will have density at 1 either zero or infinite, unless $v = 1$. Hence we take $v = 1$, and then choose $u$ such that the mean $u/(u + v)$ of the proposal distribution is equal to the mean of the empirical posterior distribution of $q_h$. Likewise, the posterior distribution of $r$ has finite, non-zero density at $r = 0$, and so in this case we use a Beta($u, v$) proposal distribution with $u = 1$ and with $v$ chosen such that the mean of the proposal distribution is equal to the mean of the posterior distribution of $r$.

Generating a further 1000 points, the mean acceptance probability was increased to $\bar{z} = 0.092$. The 100-fold improvement in efficiency is not so great as that achieved for the basic model, as the marginal posterior distributions are not so well approximated by Beta-distributions, and there are stronger correlations between parameters. Nevertheless, our choice of proposal distribution has substantially increased the mean acceptance probability, to the extent that it is practical to generate large samples from the posterior distribution. Generating a sample of 100,000 points, and considering the marginal posterior distributions separately, then 95\% equal-tailed credible intervals for $m, r, q_h$ were found to be [0.81102, 0.87177], [0.02296, 0.34094] and [0.84883, 0.99681] respectively.

Returning to the fact that the posterior density of $q_h = 1$ is non-zero, this means that, under the current model, it is possible that all infections were caused by contact from the community, with no within-household spread of infection at all. This conclusion is unlikely to be an accurate reflection of reality for this particular data set, but is suggestive of some degree of over-parameterization. Note that the three-parameter model collapses to the two-parameter model in the case $r = 0$, and that the mode of our posterior distribution occurs close to $r = 0$. 

Furthermore, the basic two-parameter model was found by O’Neill et al. (2000) to provide an adequate fit to the data.

To test formally whether the data provide evidence of household variation, we can use rejection sampling to compute a sample-based estimate of the appropriate Bayes factor, as described in section 3. Denoting by $M_1$ the basic model, and by $M_2$ the model with household variation, and setting $\pi(M_1)=0.5$, from (4) the required Bayes factor is

$$
BF_{12} = \frac{\pi(M_1 | n)}{1 - \pi(M_1 | n)}.
$$

As before, we take $q_c, q_h$ to have independent uniform [0, 1] prior distributions under $M_1$, and $m, r, q_h$ to have independent uniform [0, 1] prior distributions under $M_2$. As model $M_1$ corresponds to the nested submodel of $M_2$ with $r=0$, then the full prior distribution can be regarded as a distribution on [0, 1]$^3$ with $m, q_h$ each having independent uniform [0, 1] prior distributions, while the prior distribution of $r$ is a mixture of $r=0$ with probability 0.5, $r \sim \text{uniform}[0,1]$ with probability 0.5 (and $r$ taken to be, a priori, independent of $m, q_h$). For simplicity, we can take the proposal distribution $g(m, r, q_h)$ equal to the prior distribution. Expression (6) then reduces to

$$
K = \sup_{m, r, q_h \in [0, 1]} \left\{ \prod_{j, s} \omega_j(m, r, q_h)^{a_j} \right\}.
$$

Generating a sample of 1000 points, there were 653 points with $r=0$ compared with 347 points with $r>0$, giving an estimated Bayes factor of 653/347 = 1.88 in favour of the model without household variation ($r=0$). This is very close to 1, indicating no strong evidence one way or the other. As there is no strong evidence for household variation, it seems sensible to accept the simpler model $M_1$ as adequate for these data. We could of course greatly improve the efficiency of the algorithm by using a proposal distribution which more closely reflects the shape of the posterior distribution, but since it does not take long to produce a sample of 1000 points, even using the naive proposal distribution above, this seems unnecessary.

**Exponential infectious periods** Returning to our basic model without household variation, if the infectious periods $I$ are exponentially distributed, rather than fixed, then $q_h(k) = q_h/(k - (k - 1)q_h)$ for some $q_h \in [0, 1]$. Note that for both the Reed–Frost and the exponential infectious period case, $q_h(1) = q_h$, and so $q_h$ has the same interpretation in both models, namely the probability of a susceptible avoiding infection from a single infective. It is therefore reasonable to compare this parameter directly between the two models.

As for the Reed–Frost case, we assign independent uniform [0, 1] prior distributions for $q_c$ and $q_h$. Taking independent uniform [0, 1] proposal distributions for $q_h$ and $q_c$ and generating 1000 points from the joint posterior distribution of $(q_c, q_h)$, posterior mean parameter estimates were found to be $\bar{q}_c = 0.87$, $\bar{q}_h = 0.85$, with a mean acceptance probability $\bar{z} = 0.0013$. The posterior distribution appears to have about the same amount of spread in the $q_c$ direction as for the Reed–Frost model, but with greater spread in the $q_h$ direction. This seems likely to be attributable to the extra variability in the infectious period distribution compared with the Reed–Frost model.

Taking Beta-proposal distributions for $q_c$ and $q_h$ chosen to reflect the shape of the posterior distribution and generating a further 1000 points, the mean acceptance probability was increased to $\bar{z} = 0.53$. Continuing to generate a sample of size 100,000, 95% credible intervals for $q_c$ and $q_h$ were found to be [0.84744, 0.88476] and [0.80350, 0.88878] respectively.
To compare the Reed–Frost model $M_1$ with the model $M_2$ with exponentially distributed infectious periods, we took $\pi(M_j)=0.5$, with the parameters $q_c, q_h$ having independent uniform $[0, 1]$ priors under each of $M_1, M_2$, and generated a sample of 1000 points from the full posterior distribution. Taking the proposal distribution for rejection sampling to be equal to the prior distribution, expression (6) in this case becomes

$$K = \max \left\{ \sup_{q_c, q_h \in [0, 1]} \left\{ \prod_{j, s} \left( \omega_{js}^{(1)}(q_c, q_h) \right)^{n_{js}} \right\}, \sup_{q_c, q_h \in [0, 1]} \left\{ \prod_{j, s} \left( \omega_{js}^{(2)}(q_c, q_h) \right)^{n_{js}} \right\} \right\},$$

where for $i = 1, 2$, $\omega_{js}^{(i)}$ is the probability $\omega_{js}$ computed under model $M_i$.

In the resulting sample, 433 of the sample points corresponded to model $M_1$ and 567 to model $M_2$, giving an estimated Bayes factor of $567/433 = 1.31$ in favour of the model with exponentially distributed infectious periods. Thus, there is no real evidence to favour either model over the other. As the Reed–Frost type model is simpler, hereafter we assume fixed infectious periods.

**Comparisons between different outbreaks.** The data of Table 1 are from two separate outbreaks. Previous authors (Longini *et al.* 1988) have concluded that these two outbreaks were sufficiently similar that it is reasonable to combine the data. We now investigate this formally using Bayes factors.

Generating samples of 100,000 points for each outbreak, for the 1977–8 outbreak we find $\tilde{q}_c = 0.85$ (95% CI [0.81137–0.91748]), $\tilde{q}_h = 0.87$ (95% CI [0.81838–0.87670]), whereas for the 1980–1 outbreak, $\tilde{q}_c = 0.88$ (95% CI [0.85649–0.90460]), $\tilde{q}_h = 0.83$ (95% CI [0.77584–0.88781]). Scatter plots of $(q_c, q_h)$ for the two outbreaks have similar shapes, but slightly displaced in both $q_h$ and $q_c$ directions. To test whether it is sensible to combine data from the two outbreaks, we will compare a four-parameter model, in which each outbreak has its own values of $q_c$ and $q_h$, with the nested two-parameter submodel in which the outbreaks share a common $q_c$ value and a common $q_h$ value. Taking all parameters to have uniform $[0, 1]$ prior distributions, independently of one another, taking the two-parameter model to have prior probability 0.5, and taking a proposal distribution equal to the prior distribution, then generating a sample of 1000 points gives an estimated Bayes factor of 39 in favour of the two-parameter submodel (the two-parameter model was chosen in 975 of 1000 cases). This provides confirmation that it is indeed reasonable to combine data from the two outbreaks.

In addition to influenza data from Tecumseh, Longini *et al.* (1982) provide data from two outbreaks of influenza in Seattle, an influenza B outbreak in 1975–6 and an influenza A (H1N1) outbreak in 1978–9. The Seattle data are given in Table 2. In comparing the Tecumseh 1977–8 and 1980–1 outbreaks of influenza A (H3N2), we have concluded that it is reasonable to use common values for both $q_h$ and $q_c$ for the two outbreaks, in line with previous authors’ conclusions (Longini *et al.*, 1988). Biologically, it may be of interest to treat the $q_c$ values from distinct outbreaks as distinct, and to test only for a common $q_h$ value, as the value of $q_c$ may depend upon the size of the outbreak and the structure of households, whereas the value of $q_h$ may be more closely linked to properties of the infectious organism itself. Thus, we compare the full four-parameter model with a nested three-parameter submodel.

In the case of the Seattle data of Table 2, comparing the influenza B outbreak of 1975–6 with the influenza A (H1N1) outbreak of 1978–9, Fig. 3 shows a scatterplot of 1000 sampled points from the posterior distributions for each of these two outbreaks (as usual, taking independent uniform $[0, 1]$ prior distributions for all parameters). We see from Fig. 3 a distinct separation in the $q_c$ direction between the two outbreaks, but considerable overlap in the $q_h$ direction. For the influenza B outbreak, we find $\tilde{q}_c = 0.83, \tilde{q}_h = 0.87$; for the influenza A
Fig. 3. Scatterplots of $q_c$ and $q_h$ for outbreaks of two different strains of influenza in Seattle, Washington. Circles represent parameter values corresponding to the influenza B outbreak, crosses represent parameter values corresponding to the influenza A (H1N1) outbreak.

(H1N1) outbreak, $\bar{q}_c = 0.54$, $\bar{q}_h = 0.69$. Allowing distinct $q_c$ values for the two outbreaks and testing for a common $q_h$ value, our estimated Bayes factor value is 1.05 in favour of different $q_h$ values for the two outbreaks. Thus there is no real evidence of differences between the two influenza strains in terms of their within-household infectiousness.

Prior sensitivity. So far, we have always taken independent uniform $[0, 1]$ prior distributions for all parameters in our models, to represent prior ignorance. The value of the Bayes factor for comparison between two models $M_1$, $M_2$ depends upon the within-model prior distributions used. In the case of comparison between the two Seattle outbreaks, one could argue that although we have no prior knowledge of $q_c$ or $q_h$ for either outbreak, we would at least expect the $q_h$ values for two different strains of influenza to be reasonably similar to one another. Thus denoting by $q_h^{(1)}$, $q_h^{(2)}$ the within-household avoidance probabilities for influenza B and influenza A (H1N1), respectively, we require a prior distribution on $(q_h^{(1)}, q_h^{(2)})$ such that the marginal distributions are both uniform on $[0, 1]$, with a reasonably strong positive correlation between $q_h^{(1)}$ and $q_h^{(2)}$. Such a prior distribution may be constructed using the method of Minhajuddin et al. (2004), as follows. Fix three parameters $A, B > 0$, $N \in \mathbb{Z}^+$. Generate a random variable $K$ from the Beta-binomial distribution with parameters $(N, A, B)$. Generate values of $q_h^{(1)}$, $q_h^{(2)}$ which are independent and identically distributed, given $K$, from the Beta-distribution with parameters $(A + K, B + N - K)$. Minhajuddin et al. (2004) show that the random variables $q_h^{(1)}$, $q_h^{(2)}$ generated by the above algorithm are exchangeable random variables, each having Beta-distribution with parameters $(A, B)$, with correlation $N/(N + A + B)$. Taking $A = B = 1$ provides a prior distribution such that, a priori, $q_h^{(1)} \sim \text{uniform}[0, 1]$, $q_h^{(2)} \sim \text{uniform}[0, 1]$ and the correlation between $q_h^{(1)}$ and $q_h^{(2)}$ is $N/(N + 2)$. 

Taking $N=18$, so that the prior correlation between $q_h^{(1)}$ and $q_h^{(2)}$ is 0.9, we generated a sample of size 1000 from the appropriate posterior distribution, with prior probability of 0.5 for each of the two models ($M_1: q_h^{(1)} \neq q_h^{(2)}$ versus $M_2: q_h^{(1)} = q_h^{(2)}$) and independent uniform [0,1] prior distributions for $q_c^{(1)}$, $q_c^{(2)}$. For simplicity, we used the prior distribution as the proposal distribution, leading to cancellation between the prior and proposal densities in (6). The estimated Bayes factor was 1.62 in favour of different $q_h$ values for the two outbreaks. Although this is larger than the Bayes factor of 1.05 computed under a priori independence between $q_h$ values, it nevertheless remains close to 1, leading us to conclude that there is no real evidence of differences in $q_h$ values for the two influenza outbreaks under consideration. Similarly, setting $N=198$, so that the prior correlation between $q_h^{(1)}$ and $q_h^{(2)}$ is 0.99, yielded a Bayes factor of 1.51, leading to the same conclusion.

Differing levels of susceptibility. In the above analyses we have assumed that all individuals are equally susceptible. Longini et al. (1988) consider the effect of age and pre-epidemic season antibody level upon susceptibility, using the model with differing susceptibility levels described in section 2.1. Table 4 of Longini et al. (1988) shows data from the combined 1977–8 and 1980–1 outbreaks of influenza A (H3N2) in Tecumseh, categorizing each individual’s pre-season antibody level as ‘low’ or ‘higher’. We can fit a model with two risk levels to this data set, with parameters $(q_c, q_h)$ corresponding to individuals with low pre-season antibody levels and $(q_c, q_h)$ corresponding to individuals with higher pre-season antibody levels. Note that Table 4 of Longini et al. (1988) only includes households up to size 5, and also combines data in some categories. For instance, in households which initially consisted of five susceptible individuals, all with the low level of pre-season antibodies, that only three households avoided infection altogether whereas the remaining three households did not. Consequently the likelihood function includes terms such as $(a(0,0), (5,0))^3(1-a(0,0), (5,0))^3$, in the obvious notation.

Generating a sample of size 1000 from this model (with independent uniform [0,1] prior distributions for the four parameters), a scatter plot of the sampled points $(q_c, q_h)$ is shown in Fig. 4. We see a distinct separation between the sampled parameter values for the two different antibody levels, with the points corresponding to higher antibody levels having consistently higher values for both $q_c$ and $q_h$, as one would expect. For low antibody levels, we have $\bar{q}_c = 0.83$, $\bar{q}_h = 0.72$; for higher antibody levels, $\bar{q}_c = 0.91$, $\bar{q}_h = 0.95$. Note that the sampled points corresponding to higher antibody levels go all the way up to $q_h^{(2)} = 1$. This suggests the possibility that individuals with higher antibody level may be completely immune to within-household transmission, although biologically it seems unlikely that individuals could be truly immune to within-household transmission while remaining susceptible to infection from the community. To compare the above four-parameter model with the nested two-parameter submodel such that $(q_c, q_h) = (q_c, q_h)$, we took prior probabilities of 0.5 for each model and generated a sample of 1000 points. None of the sampled points had common parameter values for the two susceptibility levels, indicating a Bayes factor >1000 in favour of different parameter values for the two susceptibility levels.

5. Discussion

5.1. Computational issues

Here, we mention a few computational issues relating to the implementation of the rejection sampling algorithm. First, although proposals that more closely resemble the target posterior density lead to more efficient sampling, in practice, there is a trade-off between this and
the time taken to sample from such a proposal. In addition, caution is needed at boundary points, as indicated in our use of Beta-proposal distributions that require non-zero density at the point 1. Secondly, in some cases we found that it was essential to evaluate the \( \omega_j \) probabilities algebraically (using Maple) and then import the resulting algebraic expressions into Matlab. This is because of well-known numerical instabilities in the recursive equations for the \( \omega_j \) terms; see e.g. Addy et al., 1991. Finally, reparameterization is sometimes necessary, as illustrated by our parameterization of the Beta-distribution used in the household variation model analysis.

5.2. General remarks
We have illustrated that methods of rejection sampling can be successfully applied to generate exact samples from posterior distributions of interest for household epidemic models. Such methods make an important contribution to Bayesian inference for stochastic epidemic models because, unlike MCMC methods, the samples obtained are known to have the correct distribution. This in turn enables a tractable Bayesian analysis. Moreover, at least for the examples we have considered, the algorithms run very quickly, which is of considerable practical importance. In particular, this enables thorough sensitivity analyses (e.g. using a range of different prior distributions or prior parameter choices, using slight variants of the models, etc.) to be performed relatively quickly, whereas in an MCMC setting such analyses might be prohibitively time-consuming. We have also illustrated that the methods can be used to straightforwardly calculate Bayes factors. Again, in an MCMC setting such analyses are typically performed using transdimensional methods, which are notoriously awkward to implement in a reliably efficient manner.
The key practical challenges of the methods described are determining a good proposal distribution $g$, and finding the bounding factor $K$. The latter can clearly be problematic for multimodal posterior densities. However, the problem of locating global maxima of complicated functions has been widely studied, and so in practice this problem is far from insurmountable.

Regarding the proposal density, an efficient choice should resemble the target density of interest as closely as possible. For the models we considered, using naive uniform proposal distributions worked sufficiently well to enable generation of a ‘training’ sample of 1000 points, which was then used to guide the construction of better proposals. For more complicated models, this may not be practical, as generating the initial training sample using a bad proposal may take an unfeasibly long time. However, in this case, MCMC methods could be used to generate the training sample, from which a more efficient proposal could be designed.

The applications of our methods which we have presented go some way towards illustrating what can and cannot be ascertained from data consisting of final outcome. For example, we found that models with exponential infectious periods are hard to formally distinguish from those with fixed infectious periods. This seems likely to be a consequence of the fact that we are using household data, so that the final outcome distributions contributing to the likelihoods are based on populations of just a few individuals. In particular, such distributions do not vary greatly between different infectious period choices when the mean is fixed. It is possible that studies featuring larger household sizes or with some temporal data (e.g. Cauchemez et al., 2004), would be better able to discriminate between different infectious periods. Similarly, we found little evidence to favour models featuring household variation over those which did not. Again it seems plausible that studies which collect additional temporal data might provide a better means of discriminating between models. Finally, we note that since our methods are not generally time intensive, it would be possible to conduct fairly detailed studies to ascertain how the Bayes factors changed over a variety of different data sets and study designs, yielding useful information about the best way to collect future data.

References


Received August 2005, in final form April 2006.

Damian Clancy, Department of Mathematical Sciences, University of Liverpool, Liverpool L69 7ZL, UK.

E-mail: d.clancy@liv.ac.uk
Dear Author,

During the preparation of your manuscript for publication, the questions listed below have arisen. The numbers pertain to the numbers in the margin of the proof. Please attend to these matters and return the form with this proof.

Many thanks for your assistance.

<table>
<thead>
<tr>
<th>Refs.</th>
<th>Query</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Au: Please provide the manufacturer information for Matlab 6: company name, city, state (if USA) and country.</td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td>Au: Please provide the manufacturer information for maple: Compny name, city, State (if USA) and country.</td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>Au: Please provide the publisher name in full and also provide the city and state name of publisher in Fox &amp; Hall (1980).</td>
<td></td>
</tr>
<tr>
<td>Q4</td>
<td>Au: Please provide the location of publisher in O’Hagan (1994)</td>
<td></td>
</tr>
</tbody>
</table>