Bayesian inference for stochastic multitype epidemics in structured populations using sample data

PHILIP D. O’NEILL

School of Mathematical Sciences, University of Nottingham, Nottingham, UK.

SUMMARY

This paper is concerned with the development of new methods for Bayesian statistical inference for structured-population stochastic epidemic models, given data in the form of a sample from a population with known structure. Specifically, the data are assumed to consist of final-outcome information, so that it is known whether or not each individual in the sample ever became a clinical case during the epidemic outbreak. The objective is to make inference for the infection rate parameters in the underlying model of disease transmission. The principal challenge is that the required likelihood of the data is intractable in all but the simplest cases. Demiris and O’Neill (2005b) used data augmentation methods involving a certain random graph in a Markov chain Monte Carlo (MCMC) setting to address this situation in the special case where the sample is the same as the entire population. Here we take an approach relying on broadly similar principles, but for which the implementation details are markedly different. Specifically, to cover the general case of data in a sample we use an alternative data augmentation scheme, and employ non-centering methods. The methods are illustrated using data from an influenza outbreak.
Keywords: Bayesian inference; Epidemics; Markov chain Monte Carlo Methods; Stochastic epidemic models.

1 Introduction

This paper is concerned with the problem of inferring information about infection rates in a disease transmission model, given data on a sample of an at-risk population. Specifically, we suppose that the data contain (i) individual-level information that specifies which individuals in the sample ever became clinical cases of the disease in question during an epidemic outbreak; and (ii) information on the mixing structure of the population structure within the sample. A typical example of the latter, for human diseases, is that the data specify which household each individual in the sample lives in.

We start by briefly recalling some background to the problem. Stochastic epidemic models that feature population structure, such as households, have been widely studied during the past decade or so (see e.g. Becker and Dietz, 1995; Ball et al., 1997; Ball and Lyne, 2001; Li et al., 2002; Demiris and O’Neill, 2005a, 2005b and references therein). Such models are motivated by concerns of realism, providing a natural refinement over a homogeneously mixing population. Furthermore, many real-life outbreak control measures are based around structures within populations (e.g. school closures). Structured population epidemic models are not solely appropriate for human diseases: for example, for diseases among farm animals kept in pens it can be important to consider within-pen and between-pen transmission separately (e.g. Höhle et al., 2005).

We shall focus here on two-level mixing epidemic models, defined formally in
the next section. These models, introduced in Ball et al. (1997), assume that a population is partitioned into groups, not necessarily of equal sizes. In the simplest case, within-group and between-group transmission of the disease are governed by two separate model parameters corresponding to infection rates. For data of the kind we shall consider, the inference problem is then to try and estimate these two parameters. This problem is complicated because the fates of different groups in the model are dependent on one another, and so the likelihood of a particular final outcome is not simply a product over the probabilities of individual group outcomes.

Broadly speaking, there are two approaches to solving this kind of inference problem. In the first approach, an approximation is employed in which the fates of different groups are regarded as being independent. Such an approximation is not too unreasonable in practice, at least when the population is sufficiently large or the outbreak sufficiently small. This approach underlies the statistical analyses in Ball et al. (1997), Britton and Becker (2000), and Demiris and O’Neill (2005a).

The second approach, described in Demiris and O’Neill (2005b), uses a data augmentation method. Specifically, a certain random directed graph, which describes potentially infectious contacts between individuals, is imputed within a Markov chain Monte Carlo (MCMC) algorithm. Knowledge of this graph is sufficient to provide a tractable likelihood, from which the inference problem can be readily addressed. This approach is very flexible, does not rely on any kind of independence approximation, and has various other benefits as described in detail in Demiris and O’Neill (2005b). However, the method also assumes that the observed sample comprises the entire at-risk population, and it is the relaxation of this assumption that is the principal motivation for this paper. More precisely,
here we shall develop the basic idea of the data augmentation method in three ways. First, an alternative data augmentation scheme is used, which in turn leads to a simpler MCMC algorithm than that given in O’Neill and Demiris (2005b). This is of practical importance since it means that the coding is simpler, and the resulting algorithms tend to be faster. Second, we extend the methods to allow the observed sample to be any part of the at-risk population. As well as the obvious benefit in terms of realism, this also facilitates exploration of the effects that observing a sample actually has on inference. Third, we make use of the ideas of non-centered parameterisations (Neal and Roberts, 2005) in order to improve the performance of our algorithms. This again is of considerable importance in practice in reducing the time taken for algorithms to run.

The paper is structured as follows. The underlying stochastic epidemic model is described in Section 2, along with a threshold parameter of interest. Section 3 describes the form of the data, the standard data augmentation scheme and Bayesian framework. The MCMC algorithms are given in Section 4, and here also the non-centering approach is described. Section 5 contains applications of the methods, and Section 6 has some concluding remarks.

2 Multitype epidemic models with two levels of mixing

In this section we describe the epidemic model of interest, a threshold parameter for the model, and an alternative representation of the model that will be used in the sequel.
2.1 Multitype two-level-mixing model

The following model of a continuous-time epidemic process is defined in Ball and Lyne (2001). Consider a closed population of $N$ individuals, labelled $1,\ldots,N$, that is partitioned into groups, where the sizes of different groups need not be the same. Each individual in the population is assumed to be one of a possible $k$ types, where type refers to the value of a categorial covariate such as age, sex, vaccination status, etc. For $j = 1,\ldots,k$, let $N_j$ denote the total number of individuals in the population of type $j$. For convenience, suppose that the possible types are labelled $1,\ldots,k$, and for $i = 1,\ldots,N$ denote by $\tau(i)$ the type of individual $i$.

For each time $t \geq 0$, each individual in the population will be in one of three states, namely susceptible, infective, or removed. A susceptible individual may potentially contract the disease in question. An infective individual has become infected, and can transmit the disease to others. A removed individual is no longer infectious, and plays no part in further disease spread. In practice this could occur either because of actual immunity (induced by antibodies), or by isolation following the appearance of symptoms. If a type $j$ individual, $i$ say, becomes infective, then they remain so for a random time $I(i)$ whose distribution is the same as some specified non-negative random variable $I_j$. The infectious periods of different individuals are assumed to be mutually independent.

The epidemic is initiated at time $t = 0$ by a small number of individuals becoming infective. During its infectious period, an individual, $i$ say, has (global) contacts with the population of type $j$ individuals at times given by the points of a homogeneous Poisson process of rate $\lambda_{\tau(i)j}^G$. These contacts are potentially infectious, meaning that they are adequate for transmission of the pathogen in question from
an infective individual to another individual. Note that in the sequel, the word ‘contact’ will always refer to contacts of this kind. The individual contacted on each such occasion is chosen uniformly at random from the $N_j$ type $j$ individuals in the population, with different choices being mutually independent. Note that this description is equivalent to $i$ having contacts with each type $j$ individual in the population according to a Poisson process of rate $\lambda^G_{\tau(i)j}/N_j$, independently of all other contact processes. However, we use the former description because it is more directly related to the data imputation methods used in the sequel.

In addition to and independently of the global contact process, the infective individual $i$ also makes (local) contacts with individuals in its own group according to a Poisson process of rate $\lambda^L_{\tau(i)j}N^L_{ij}$, where $N^L_{ij}$ denotes the number of type $j$ individuals in the same group as $i$. The individual actually contacted on each occasion is chosen independently and uniformly at random from the $N^L_{ij}$ individuals in question.

If either a global or local contact is made between an infective individual and a susceptible, the susceptible immediately becomes infective. Conversely, contacts that an infective individual makes with either infective or removed individuals have no effect. Note that under the assumptions of the model, an infective individual can contact themselves. This fact has no bearing on the progress of the epidemic, and moreover it leads to simpler notation than explicitly excluding self-contacts. At the end of its infectious period, an infective individual becomes removed, and plays no further part in the epidemic. The epidemic ceases as soon as there are no infectives present in the population.

The above model does not include a latent period, i.e. a period of time after a
susceptible individual has been contacted by an infective, but before they themselves become infective. However, including a latent period makes no difference to the distribution of the final outcome of the epidemic, i.e. to the random subset of individuals who ever become infected (see, for example, Ball et al., 1997). In particular, this means that the methods described in this paper can be applied to epidemic models that include a latent period.

2.2 Threshold parameter

Threshold parameters are of considerable importance in both the theory of stochastic epidemic models (Andersson and Britton, 2000, p.6), and epidemiology (Farrington et al., 2001). Typically, for a stochastic epidemic model, there is a real-valued parameter \( R \) such that epidemics in an infinite population of susceptibles infect only finitely many individuals, almost surely, if and only if \( R \leq 1 \). Such parameters can usually be thought of as the average number of new infections caused by a single infective unit in a large susceptible population, where a ‘unit’ is typically an individual or a group, as appropriate.

A threshold parameter for the multitype two-level mixing model can be obtained by allowing the population to become large globally, i.e. by allowing the number of groups to tend to infinity. The details are given in Ball and Lyne (2001), and are briefly as follows. Define the \( k \times k \) matrix \( M := (m_{ij}) \) where \( m_{ij} \) is the average number of global contacts to type \( j \) individuals made by individuals in a group in which the first infected individual is of type \( i \). The threshold parameter, \( R_\ast \), is then defined as the maximal eigenvalue of \( M \). It follows that, in a large population, epidemics are extremely unlikely to take off if \( R_\ast \leq 1 \). Calculation of \( R_\ast \) in practice involves computing \( M \); explicit details of how to do this are given
2.3 A non-temporal representation of the epidemic model

As described above, we shall be concerned with final-outcome data, i.e. knowledge of which individuals in a sample ever become infected during the epidemic. In order to determine the final outcome of a realisation of the epidemic model it is sufficient to record, for each individual in the population, the set of individuals that they themselves would contact if they became infective, ignoring the times of such contacts. Given such information, and initial conditions specifying which individuals are initially infective, it is then a simple matter to determine which individuals are ultimately infected and which are not. Specifically, individual $j$ is ultimately infected if and only if there is an ordered sequence of individuals $i_1, i_2, \ldots, i_l = j$ such that $i_1$ is an initial infective and $i_p$ contacts $i_{p+1}$ for $p = 1, \ldots, l - 1$. This representation of the epidemic can be equivalently described by a directed random graph in which vertices are individuals and an edge from individual $i$ to $j$ occurs if and only if $i$ has a contact with $j$ whilst $i$ is infective (see, for example, Andersson and Britton, Chapter 7, or Demiris and O’Neill, 2005b, for full details). We now describe how this representation can be used in the inference problem of interest.

3 Data and data augmentation

The most natural likelihood, namely the probability that the observed data occur under the model for a given set of parameters, is analytically and numerically intractable for all but the simplest data sets. This is essentially because the
required calculation involves summing over all possible ways in which the data can have arisen, which is highly non-trivial due to the inherent dependencies between different groups (see e.g. O’Neill and Demiris, 2005b). Therefore we shall adopt a form of data augmentation in which, for each individual, we keep track of two kinds of information, namely (i) the total number of contacts they have, and (ii) which individuals the contacts are with. Both quantities will be required for both local and global contacts, and for each of the $k$ types of individual with whom contacts can be made. Note that both of (i) and (ii) appear explicitly in the definition of the epidemic, and that the information is sufficient to describe the progress of the epidemic in terms of the non-temporal representation described above. Moreover, the assumption that the contact processes of different individuals are independent leads to a tractable augmented likelihood, as described later. We now give the details, starting with notation.

### 3.1 Final outcome data and inference

It is assumed that the data consist of a sample of groups from the population. Specifically, we assume that within each group we know the number of initially susceptible individuals of each type, and how many of these ever became infected during the epidemic. Thus the data are of the form $n = \{n(s_1, \ldots, s_k; i_1, \ldots, i_k)\}$, where $n(s_1, \ldots, s_k; i_1, \ldots, i_k)$ denotes the number of groups in the sample containing $s_j$ initially susceptible individuals of type $j$, of whom $i_j$ ever become infected, where $j = 1, \ldots, k$.

We also assume that the group structure of the entire population is known, meaning we know the total number of groups of size $j$ for $j \geq 1$, and the composition of types of individual within each group. Note that we therefore assume that the
total number of individuals of type \( j \), \( N_j \), is known for each \( j = 1, \ldots, k \). In practice, group structure information might come from a mixture of census information (for human diseases) and extrapolation of the observed group structure in the sample, assuming that the sample is representative.

Our objective is to conduct Bayesian statistical inference for the two infection rate matrices \( \Lambda^L := (\lambda^L_{ij}) \) and \( \Lambda^G := (\lambda^G_{ij}) \), given the data \( n \). Let \( \Lambda := (\Lambda^L, \Lambda^G) \). By Bayes’ Theorem, the posterior density of interest, \( \pi(\Lambda|n) \), satisfies

\[
\pi(\Lambda|n) \propto \pi(n|\Lambda)\pi(\Lambda),
\]

where \( \pi(n|\Lambda) \) denotes the likelihood and \( \pi(\Lambda) \) denotes the joint prior density of \( \Lambda^L \) and \( \Lambda^G \). As noted above, the likelihood is analytically and numerically intractable in all but the most trivial cases. Before describing a data augmentation scheme to overcome this problem, we make two important observations.

First, final outcome data contain no explicit temporal information. In particular, there is no information regarding the mean length of the infectious period. We therefore fix the infectious period distribution in advance of data analysis. This implicitly creates a time-scale, with respect to which the values of \( \Lambda^G \) and \( \Lambda^L \) (but not \( R_* \)) should be interpreted.

Second, if there are more than \( k = 2 \) types of individual, then the model parameters will not usually all be identifiable. Specifically, the \( \Lambda^L \) matrix is identifiable with sufficiently rich data, but \( \Lambda^G \) is not (see e.g. Demiris and O’Neill, 2005b, and Britton, 1998). Although this issue has no bearing in principle on the Bayesian analysis, it is both pragmatic and computationally preferable to consider models with additional restraints on \( \Lambda^G \), as described later.
3.2 Data augmentation scheme

It is convenient to partition the population of \( N \) individuals into three subsets, namely (i) \( A \), the individuals in the observed sample who ever become infective; (ii) \( B \), the individuals who are not in the observed sample; and (iii) \( C \), the individuals in the observed sample who never become infective. In what follows, most of our attention will focus on \( A \) and \( B \), the point being that there is no need to keep track of contacts from individuals in \( C \) since they themselves are known to have avoided infection. For \( j = 1, \ldots, k \), denote by \( A_j \) and \( B_j \) respectively the sets of individuals in \( A \) and \( B \) respectively who are of type \( j \). Similarly denote by \( N^A_j \), \( N^B_j \) and \( N^C_j \) the numbers of type \( j \) individuals in \( A \), \( B \) and \( C \), respectively, so that \( N_j = N^A_j + N^B_j + N^C_j \). For individual \( i \), define \( N_{ij}^{A,L} \), \( N_{ij}^{B,L} \) and \( N_{ij}^{C,L} \) as the numbers of type \( j \) individuals in \( i \)'s group who are in \( A \), \( B \) and \( C \) respectively. Thus \( N_{ij}^L = N_{ij}^{A,L} + N_{ij}^{B,L} + N_{ij}^{C,L} \), although since the observed sample only contains whole groups then \( N_{ij}^{B,L} = 0 \) if \( i \in A \), and \( N_{ij}^{B,L} = N_{ij}^L \) if \( i \in B \). Finally, note that all of the sets and quantities defined here are assumed to be known.

For each individual \( i \in A \), and \( j = 1, \ldots, k \), denote by \( x_{ij}^{A,L} \) the number of local contacts made by \( i \) with type \( j \) individuals in \( A \) in \( i \)'s group. It follows from the definition of the epidemic that \( x_{ij}^{A,L} \) has a Poisson distribution with (random) mean \( I(i) \lambda_{r(i)j}^{L} N_{ij}^{A,L} \), and each contact is with an individual chosen independently and uniformly at random from the \( N_{ij}^{A,L} \) individuals in question (possibly zero). Let the set of such contacted individuals be denoted \( c_{ij}^{A,L} \), where this set is allowed to contain repeated elements since individuals might be contacted more than once. (In fact, in reality it is sufficient to ignore repeated contacts, but for ease of exposition we do not do so).
Similarly, for $i \in A$ define $x_{ij}^{AG}$ and $x_{ij}^{ABG}$ to be the number of global contacts made by $i$ with type $j$ individuals in $A$, and in $B$, respectively. For $i \in B$, define equivalent quantities for the numbers of contacts made locally to individuals in $B$, globally to individuals in $B$, and globally to individuals in $A$ by $x_{ij}^{BL}$, $x_{ij}^{BG}$ and $x_{ij}^{BAG}$, respectively. Note that all of these quantities have Poisson distributions, and moreover are (conditional upon the $I(i)$ values in question) mutually independent. Define corresponding sets of contacted individuals in the natural way, so that $c_{ij}^{AG}$ is the set of individuals contacted by the $x_{ij}^{AG}$ global contacts made by $i \in A$ to type $j$ individuals in $A$, etc. Note that we define contact numbers and sets for all individuals in $B$, regardless of whether or not such individuals are actually infected in a given realisation.

Define

$$E_A = \left\{ x_{ij}^{AL}, x_{ij}^{AG}, x_{ij}^{ABG}, c_{ij}^{AL}, c_{ij}^{AG}, c_{ij}^{ABG} : i \in A, 1 \leq j \leq k \right\},$$

$$E_B = \left\{ x_{ij}^{BL}, x_{ij}^{BG}, x_{ij}^{BAG}, c_{ij}^{BL}, c_{ij}^{BG}, c_{ij}^{BAG} : i \in B, 1 \leq j \leq k \right\},$$

so that $E_A$ ($E_B$) contains information on all the contacts made by individuals in $A$ ($B$) to individuals in $A \cup B$. Denote by $K$ the (unobserved) set of individuals who are initially infective. Thus $E_A$, $E_B$ and $K$ are latent variables, or ‘missing data’, with which we can form an augmented likelihood as described below. Note that there is no need to explicitly account for contacts from individuals in $A \cup B$ to those in $C$. This is because (i) individuals in $A$, or ever-infected individuals in $B$ must have zero contacts with individuals in $C$, while (ii) individuals in $B$ who are never infected can never infect individuals in $C$ so their contacts with $C$ are irrelevant.

For simplicity, we henceforth assume that the infectious periods $I(i)$, for $i =$
1, \ldots, N, are known. It is not necessary to do this: as described in O’Neill and Demiris (2005b), the infectious periods can also be included as latent variables, and need not be fixed. However, in practice it is often realistic to assume fixed infectious periods, and in any case inference for the infection rates does not vary greatly between different standard choices of infectious period distribution with the same mean (O’Neill et al., 2000; Demiris and O’Neill, 2005b).

3.3 Augmented posterior distribution

From Bayes’ Theorem we have

$$\pi(\Lambda | n) \propto \pi(n | E_A, E_B, K, \Lambda) \pi(E_A, E_B, K, \Lambda) = \pi(n | E_A, E_B, K, \Lambda) \pi(E_A | \Lambda) \pi(E_B | \Lambda) \pi(\Lambda) \pi(K) \tag{3.1}$$

where $\pi(K)$ denotes the prior mass function of $K$, which is assumed to be a priori independent of $\Lambda$, and where we have used the fact that $E_A$ and $E_B$ are independent given $\Lambda$. Let $\Gamma(a, b)$ denote a Gamma random variable with probability density function $f(x) \propto x^{a-1} \exp(-bx)$ for $x > 0$. We assume that, for $i, j = 1, \ldots, k$, $\lambda_{L_{ij}} \sim \Gamma(m_{L_{ij}}, \nu_{L_{ij}})$ and $\lambda_{G_{ij}} \sim \Gamma(m_{G_{ij}}, \nu_{G_{ij}})$ a priori.

The various terms in the augmented posterior distribution refined via (3.1) can be evaluated as follows. First, let $\chi_A = \chi_A(E_A, E_B, K) \subseteq A$ denote the (random) set of individuals in $A$ who ever become infected given $E_A, E_B$ and $K$. Similarly, define $\chi_B = \chi_B(E_A, E_B, K) \subseteq B$ to denote the set of individuals in $B$ who ever become infected. Then

$$\pi(n | E_A, E_B, K, \Lambda) = \mathbb{1}_{\{\chi_A = A\}} P(C \text{ not infected} | E_A, E_B, K, \Lambda)$$

$$= \mathbb{1}_{\{\chi_A = A\}} \prod_{i,j=1}^k \exp \left\{ -(T_{A_i}^A + T_{i}^B) \lambda_{L_{ij}} N_j^C / N_j - \lambda_{G_{ij}} \sum_{l \in A_l} I(l) N_l^C \right\} \tag{3.2}$$
where \( \mathbb{1}_D \) denotes the indicator function of the event \( D \), \( T_i^A = \sum_{j \in A_i} I(j) \) and \( T_i^B = \sum_{j \in \mathcal{D} \cap B_i} I(j) \). Note that \( T_i^A \) is the sum of all infectious periods of type \( i \) individuals in \( A \); such a sum is usually called a severity. The indicator function in (3.2) ensures that the observed data \( n \) actually arise, i.e. that all of the individuals in \( A \) do become infected, according to \( E_A \), \( E_B \) and \( K \), and is straightforward to evaluate. The product in (3.2) arises from the independent Poisson processes that describe infection processes in the underlying epidemic model. There are two parts: individuals in \( C \) must avoid global contact from individuals in \( A \) and infective individuals in \( B \), and also avoid local contact from individuals in \( A \).

Next,

\[
\pi(E_A|\Lambda) = \prod_{i \in A} \prod_{j=1}^k \phi_{ij}^{A,L} \phi_{ij}^{A,G} \phi_{ij}^{AB,G},
\]

where

\[
\phi_{ij}^{A,L} = \exp \left\{ -I(i)N_{ij}^{A,L}\lambda_{r(i)j}^L \right\} \left( I(i)\lambda_{r(i)j}^L \right)_j^{x_{ij}^{A,L}} / x_{ij}^{A,L}!,
\]

\[
\phi_{ij}^{A,G} = \exp \left\{ -I(i)\lambda_{r(i)j}^G N_j^A / N_j \right\} \left( I(i)\lambda_{r(i)j}^G / N_j \right)_j^{x_{ij}^{A,G}} / x_{ij}^{A,G}!,
\]

\[
\phi_{ij}^{AB,G} = \exp \left\{ -I(i)\lambda_{r(i)j}^G N_j^B / N_j \right\} \left( I(i)\lambda_{r(i)j}^G / N_j \right)_j^{x_{ij}^{AB,G}} / x_{ij}^{AB,G}!.
\]

Thus, for example, \( \phi_{ij}^{A,L} \) is the probability that individual \( i \) makes \( x_{ij}^{A,L} \) local contacts with type \( j \) individuals, as listed in \( c_{ij}^{A,L} \), each such contact being chosen uniformly at random from the \( N_{ij}^{A,L} \) type \( j \) individuals in \( i \)'s group. Similarly,

\[
\pi(E_B|\Lambda) = \prod_{i \in B} \prod_{j=1}^k \phi_{ij}^{B,L} \phi_{ij}^{B,G} \phi_{ij}^{BA,G},
\]

where

\[
\phi_{ij}^{B,L} = \exp \left\{ -I(i)N_{ij}^{B,L}\lambda_{r(i)j}^L \right\} \left( I(i)\lambda_{r(i)j}^L \right)_j^{x_{ij}^{B,L}} / x_{ij}^{B,L}!,
\]

\[
\phi_{ij}^{B,G} = \exp \left\{ -I(i)\lambda_{r(i)j}^G N_j^B / N_j \right\} \left( I(i)\lambda_{r(i)j}^G / N_j \right)_j^{x_{ij}^{B,G}} / x_{ij}^{B,G}!,
\]

\[
\phi_{ij}^{BA,G} = \exp \left\{ -I(i)\lambda_{r(i)j}^G N_j^A / N_j \right\} \left( I(i)\lambda_{r(i)j}^G / N_j \right)_j^{x_{ij}^{BA,G}} / x_{ij}^{BA,G}!.
\]
We make two remarks regarding the data augmentation scheme described above. First, it may appear inefficient to keep track of the contacts made by all of the individuals in $B$, since it is sufficient to know contacts from individuals in $\chi_B$ in order to write down a likelihood. However, in practice it is far more complicated to work only with $\chi_B$, the difficulty being that updating $\chi_B$ in an MCMC algorithm would require generating additional contact numbers and sets as part of the proposal mechanism. Second, in a similar vein it is sufficient simply to know whether or not a given individual $i$ contacts a given individual $j$, as opposed to keeping track of possibly-repeated contacts as we do here. Such an approach is taken in O’Neill and Demiris (2005b), but the augmentation scheme used in the current paper leads to a simpler MCMC algorithm, as now described.

4 Markov chain Monte Carlo algorithms

In order to obtain samples from the posterior distribution defined at (3.1), we now describe two different Metropolis-Hastings algorithms (see e.g. Gilks et al., 1996). The first utilises the data augmentation scheme described in the previous Section. This algorithm is straightforward to implement and works well when the unobserved population is not too large. The second algorithm, which is slightly more involved, is suitable for situations where the unobserved population is large, and thus complements the first algorithm. In this case, non-centering methods are used to reduce certain dependencies and thus improve algorithm performance, as discussed in detail below.
4.1 Standard algorithm

Here we define an MCMC algorithm by specifying the methods used to update the various relevant parameters. The algorithm itself proceeds in the conventional way, i.e. given initial conditions, the Markov chain is simulated for a large number of iterations, after which samples are taken in order to approximate samples from the posterior distribution of interest (see e.g. Gilks et al., 1996). We use the notation \( \tilde{\theta} \) to denote a proposed value of a parameter \( \theta \).

Infection rates Since \( \lambda^G_{ij} \) and \( \lambda^L_{ij} \) have Gamma prior distributions, it follows from (3.1) - (3.4) that they will have Gamma-distributed full conditional distributions. Thus these infection rate parameters can be updated using a Gibbs step. Specifically, for \( i, j = 1, \ldots, k \),

\[
\lambda^L_{ij} | \ldots \sim \Gamma \left( \sum_{l \in A_i} x_{ij}^{A,L} + \sum_{l \in B_i} x_{ij}^{B,L} + m^L_{ij}, \sum_{l \in A_i} I(l)(N_{ij}^{A,L} + N_{ij}^{C,L}) + \sum_{l \in B_i} I(l) N_{ij}^{B,L} + \nu^L_{ij} \right),
\]

and

\[
\lambda^G_{ij} | \ldots \sim \Gamma \left( \sum_{l \in A_i} \left(x_{ij}^{A,G} + x_{ij}^{AB,G}\right) + \sum_{l \in B_i} \left(x_{ij}^{B,G} + x_{ij}^{BA,G}\right) + m^G_{ij}, \right.
\]

\[
T^A_i + T^B_i + \sum_{l \in B_i \setminus \chi_B} I(l)(N_{ij}^A + N_{ij}^B) / N_j + \nu^G_{ij} \bigg) \bigg),
\]

where e.g. \( \lambda^L_{ij} | \ldots \) denotes the distribution of \( \lambda^L_{ij} \) given all other model parameters and \( E_A, E_B \) and \( K \).

Numbers of contacts and contact sets Let Po(\( \mu \)) denote a Poisson random variable with mean \( \mu \). As previously described, for \( i \in A \) and \( j = 1, \ldots, k \), the underlying
epidemic model has

\[ x_{ij}^{A,L} \sim \text{Po}(I(i)N_{ij}^{A,L}\lambda_{r(i)j}^{L}), \quad (4.3) \]
\[ x_{ij}^{A,G} \sim \text{Po}(I(i)N_{ij}^{A}\lambda_{r(i)j}^{G}/N_j), \quad (4.4) \]
\[ x_{ij}^{AB,G} \sim \text{Po}(I(i)N_{ij}^{B}\lambda_{r(i)j}^{G}/N_j), \quad (4.5) \]

while for \( i \in B \) and \( j = 1, \ldots, k \),

\[ x_{ij}^{B,L} \sim \text{Po}(I(i)N_{ij}^{B,L}\lambda_{r(i)j}^{L}), \quad (4.6) \]
\[ x_{ij}^{B,G} \sim \text{Po}(I(i)N_{ij}^{B}\lambda_{r(i)j}^{G}/N_j), \quad (4.7) \]
\[ x_{ij}^{BA,G} \sim \text{Po}(I(i)N_{ij}^{A}\lambda_{r(i)j}^{G}/N_j). \quad (4.8) \]

To update the contacts made by individuals in \( A \) we proceed as follows. First pick an individual \((i, \text{ say})\) in \( A \) uniformly at random. For each type \( j = 1, \ldots, k \) propose \( \tilde{x}_{ij}^{A,L} \) according to (4.3). If \( \tilde{x}_{ij}^{A,L} \geq 1 \), propose \( \tilde{c}_{ij}^{A,L} \) by \( \tilde{x}_{ij}^{A,L} \) independent uniformly distributed draws from the set of \( N_{ij}^{A,L} \) type \( j \) individuals in \( i \)'s group, otherwise set \( \tilde{c}_{ij}^{A,L} = \emptyset \). Finally, the proposed new values are accepted if they are compatible with the data, i.e. the acceptance probability equals \( 1_{\{\chi_A = A\}} \); otherwise, the original values are restored.

Global contacts made by \( i \) within \( A \), as described by \( x_{ij}^{A,G} \) and \( c_{ij}^{A,G} \), are updated in a similar manner using (4.4). Note that when proposing updates for \( i \)'s contacts within \( A \), both of the severities \( T_i^{A} \) and \( T_i^{B} \) in (3.2) are unchanged if the proposed values are accepted (in fact, \( T_i^{A} \) is always constant, assuming infectious periods are fixed). In contrast, \( T_i^{B} \) can change whenever contacts to individuals in \( B \) are updated, since \( \chi_B \) may change. Therefore, to update \( x_{ij}^{AB,G} \) and \( c_{ij}^{AB,G} \) we proceed as before using (4.5) and draws from the \( N_j^{B} \) relevant individuals as necessary.
From (3.2), the proposed new values are accepted with probability

$$\mathbb{I}_{\{\tilde{\chi}_A = A\}} \exp \left( \sum_{i,j=1}^{k} (T_i^B - \tilde{T}_i^B) \lambda_{ij}^G N_j^C / N_j \right) \land 1,$$  \hspace{1cm} (4.9)

where $\tilde{\chi}_A$ and $\tilde{T}_i^B$ are respectively the proposed new values of $\chi_A$ and $T_i^B$, and $\land$ denotes minimum.

The contacts for individuals in $B$ are updated in a similar manner using (4.6) - (4.7) and appropriately sampled sets of contacted individuals. As before, the acceptance probability at (4.9) is required for updates of contacts made to individuals in $B$ (i.e. $x_{ij}^{B,L}$, $x_{ij}^{B,G}$ and the contact sets). Conversely, updating $x_{ij}^{BAG}$ and $c_{ij}^{BAG}$ only requires checking that $\tilde{\chi}_A = A$, since $\chi_B$ is unchanged.

**Initial infectives** The set of initial infectives $K$ enters the posterior density at (3.1) only through the prior density $\pi(K)$ and (3.2). Thus if a new set of initial infectives, $\tilde{K}$ say, is proposed from a distribution with probability mass function $q(\cdot|K)$, then $\tilde{K}$ is accepted with probability

$$\mathbb{I}_{\{\tilde{\chi}_A = A\}} \frac{\pi(\tilde{K}) q(\tilde{K}|K)}{\pi(K) q(K|\tilde{K})} \exp \left( \sum_{i,j=1}^{k} (T_i^B - \tilde{T}_i^B) \lambda_{ij}^G N_j^C / N_j \right) \land 1.$$

Under the assumption that there is a single initial infective who is *a priori* equally likely to be any individual in the population, and proposing $\tilde{K}$ uniformly at random independently of $K$, the acceptance probability simplifies to (4.9).

### 4.2 Non-centered algorithm

The standard algorithm defined above is straightforward to implement in practice. However, it can suffer from poor performance in some settings due to inherent correlations in the parameterisation that is used. Specifically, consider a pair of
parameters $x$ and corresponding $\lambda$ (e.g. $x_{ij}^{A,L}$ and $\lambda_{r(i)j}$, etc.). Such pairs are directly dependent on each other, as shown in (4.1) - (4.8), which in practice means that the updating scheme described above will tend to propose values that are ‘close’ to the current ones. Such correlations can in turn lead to poor mixing when they appear in data augmentation schemes such as that we have used, and in particular the problem becomes worse as the amount of missing data increases (see e.g. Neal and Roberts, 2005). In our setting, this means that the performance of the standard algorithm will deteriorate as the unobserved population $B$ becomes larger.

In order to overcome these difficulties, we adopt methods of non-centering (see Neal and Roberts, 2005 and references therein). The idea is a simple one: instead of using $(x, \lambda)$ in the model definition, where $x \sim \text{Po}(c\lambda)$ for some known constant $c > 0$ (i.e. as in (4.3) - (4.8)) we use $(d, \lambda)$ where $d$ is a Uniform$(0,1)$ random variable, and define $x$ deterministically as $x = F^{-1}_{\lambda}(d)$, where $F_{\lambda}$ is the distribution function of $x$. The point now is that in this parameterisation of the model, $d$ and $\lambda$ are independent quantities. Furthermore, when one of the $\lambda$ parameters is updated, the $x$ variables (and contact lists $c$, as appropriate) for all individuals are updated simultaneously, thus allowing much faster movement around the parameter space.

Specifically, define $d_{ij}^{A,L}$, $d_{ij}^{A,G}$, $d_{ij}^{AB,G}$, $d_{ij}^{B,L}$, $d_{ij}^{B,G}$ and $d_{ij}^{BA,G}$ to be independent Uniform$(0,1)$ ‘decision variables’ associated with the $x$’s in the original parameterisation. So for example, for $x = 0, 1, \ldots$, $x_{ij}^{A,L} = x$ if and only if $F_{ij}(x - 1) < d_{ij}^{A,L} \leq F_{ij}(x)$, where $F_{ij}$ is the distribution function of a Po$(I(i)N_{ij}^{A,L}\lambda_{r(i)j}^{L})$ random variable.
Next, instead of the contact sets $c_{ij}^{A,L}$, $c_{ij}^{A,G}$ etc. previously defined, it is more convenient to employ ordered lists of contacts. This is essentially because updating the infection rate parameters will lead to wholesale updating of the contacts made by individuals, and it is easiest to do this by having an ordered list of potential contacts already available, as opposed to proposing new lists for every $\lambda$ update. Specifically, for $i \in A$ and $j = 1, \ldots, k$ define $b_{ij}^{A,L}$ to be an ordered sequence of individuals, each randomly and independently selected from the $N_{ij}^{A,L}$ type $j$ individuals in $i$’s group. The sequence represents the (possibly repeated) contacts made, in order, by individual $i$, although only the first $x_{ij}^{A,L}$ contacts are actually realised. The sequence is infinite in length, although of course in practice it is necessary and sufficient to choose some suitably large fixed length. Define $b_{ij}^{A,G}$, $b_{ij}^{AB,G}$ etc. in the obvious manner. Note that, under the model definition, these sequences are mutually independent and also independent of the decision variables and infection rate parameters.

Denote by $d$ the set of all the decision variables and by $b$ the set of the contact sequences $b_{ij}^{A,L}$, $b_{ij}^{A,G}$ etc. It follows that

$$\pi(\Lambda|n) \propto \pi(n|b, d, K, \Lambda)\pi(b)\pi(d)\pi(\Lambda)\pi(K),$$

(4.10)

where both $\pi(b)$ and $\pi(d)$ are constants, since $b$ and $d$ are both uniformly distributed under the model. Furthermore,

$$\pi(n|b, d, K, \Lambda) = 1_{\{\chi_{A}=A\}} \prod_{i,j=1}^{k} \exp \left\{ -(T_{i}^{A} + T_{i}^{B})\lambda_{ij}^{G}N_{j}^{C}/N_{j} - \lambda_{ij}^{L}\sum_{l \in A_{i}} I(l)N_{lj}^{C} \right\},$$

(4.11)

where now $\chi_{A} = \chi_{A}(b, d, K, \Lambda)$, so that in contrast to the original parameterisation, $\chi_{A}$ now depends on $\Lambda$. The same is true for $\chi_{B}$, and the severities $T_{i}^{B}$, $i = 1, \ldots, k$. Note that in order to evaluate $\chi_{A}$, it is necessary to compute all
Infection rates

The infection rates are updated using Gaussian proposal distributions centered on the current values. Specifically, for $i, j = 1, \ldots, k$, propose $\tilde{\lambda}_L^{ij} \sim N(\lambda_L^{ij}, (\sigma_L^{ij})^2)$, where $N(\mu, \sigma^2)$ denotes a Gaussian distribution with mean $\mu$ and variance $\sigma^2$. It follows from (4.10) and (4.11) that the new value is accepted with probability

$$
\mathbb{I}\{\tilde{\chi}_A = A\} \frac{\pi(\tilde{\Lambda})}{\pi(\Lambda)} \exp \left( - \sum_{i,j=1}^{k} \left\{ (\tilde{T}_i^B - T_i^B)\lambda_G^{ij} N_j^C / N_j + (\tilde{\lambda}_L^{ij} - \lambda_L^{ij}) \sum_{l \in A_i} I(l) N_l^C \right\} \right) \wedge 1.
$$

Similarly, we propose $\tilde{\lambda}_G^{ij} \sim N(\lambda_G^{ij}, (\sigma_G^{ij})^2)$ and accept with probability

$$
\mathbb{I}\{\tilde{\chi}_A = A\} \frac{\pi(\tilde{\Lambda})}{\pi(\Lambda)} \exp \left( - \sum_{i,j=1}^{k} \left\{ (\tilde{\lambda}_G^{ij} - \lambda_G^{ij}) T_i^A + (\tilde{\lambda}_G^{ij} \tilde{T}_i^B - \lambda_G^{ij} T_i^B) \right\} N_j^C / N_j \right) \wedge 1.
$$

Decision variables

For $i \in A \cup B$ and $j = 1, \ldots, k$, $d_{ij}$ is updated by proposing a new value $\tilde{d}_{ij} \sim U(0, 1)$ and accepting with probability

$$
\mathbb{I}\{\tilde{\chi}_A = A\} \exp \left( \sum_{i,j=1}^{k} (T_i^B - \tilde{T}_i^B) \lambda_G^{ij} N_j^C / N_j \right) \wedge 1. \quad (4.12)
$$

Contact lists

The sequences in $b$ are updated by independently choosing each contact uniformly at random from the appropriate set of individuals. The acceptance probability is given by (4.12).

Initial infectives

Finally, $K$ is updated by proposing $\tilde{K}$ from a proposal distribution with mass function $q(\cdot|K)$ and accepting with probability

$$
\mathbb{I}\{\tilde{\chi}_A = A\} \frac{\pi(\tilde{K}) q(K|\tilde{K})}{\pi(K) q(K|K)} \exp \left( \sum_{i,j=1}^{k} (T_i^B - \tilde{T}_i^B) \lambda_G^{ij} N_j^C / N_j \right) \wedge 1.
$$
5 Application to influenza data

In this section we demonstrate the methods using a data set on an outbreak of influenza. In addition to illustrating the estimation methodology, we also examine the effect on inference of the size of the unobserved population.

5.1 Influenza data and model

We consider a data set taken from the Tecumseh study of illness (see Monto et al., 1985) which is given in Table 3 of Haber et al. (1988). The data consist of a sampled survey of households from Tecumseh, Michigan, constituting roughly 10% of the entire population of the town. For each household in the sample, the number of adults (age 18+) and children (age 0-17) were recorded, and also how many of each became cases of influenza A(H3N2) during the 1977-78 influenza season. In total there were 289 households containing 491 adults of whom 62 were cases, and 180 children of whom 63 were cases. The households in the data set contain five or fewer individuals. More details of the data can be found in Haber et al. (1988).

Let $\alpha$ denote the fraction of the population observed, so that in reality $\alpha \approx 0.1$. We shall analyse the data assuming that $\alpha = 1$, $\alpha = 0.5$ and $\alpha = 0.1$ in order to see the effect that a sampled data set has on the estimation of the model parameters. Note that here we assume that the $B$ population itself gets larger as $\alpha$ decreases, i.e. the observed population $A \cup C$ remains unchanged.

The data provide details of two types of individuals, namely adults and children, so that a full model would contain eight infection rate parameters. As mentioned
previously, identifiability issues make it pragmatic to adopt a simpler model. Here, following Haber *et al.* (1988), we consider age as a risk factor for infection. Specifically, denoting \( j = 1 \) as adults and \( j = 2 \) as children, define \( \lambda^L_j := \lambda^L_{1j} = \lambda^L_{2j} \) and \( \lambda^G_j := \lambda^G_{1j} = \lambda^G_{2j} \), so there are now just four infection rate parameters. The model thus assumes that children and adults are equally infectious, but need not be equally susceptible.

5.2 Algorithm implementation

The standard algorithm was found to be practically adequate for \( \alpha = 1 \) and \( \alpha = 0.5 \). For \( \alpha = 0.1 \) the non-centered algorithm was used instead, since the standard algorithm mixed rather poorly.

Prior distributions were assigned such that for \( j = 1, 2 \), \( \lambda^L_j \sim \Gamma(1, 10^{-6}) \) and \( \lambda^G_j \sim \Gamma(1, 10^{-6}) \), corresponding to vague prior assumptions. It was assumed that the set of initial infectives \( K \) consisted of one individual from the unobserved \( B \) population (unless \( \alpha = 1 \), in which case the individual resided in the \( A \) population), with this individual being *a priori* equally likely to be any of those in question. The infectious period was set as 4.1 days for all individuals, in keeping with similar analyses (e.g. Addy *et al.*, 1991; Demiris and O’Neill, 2005b).

For convenience, it was assumed that the observed population \( A \cup C \) was representative, in the sense that the \( B \) population had a similar household structure. For example, with \( \alpha = 0.5 \) the \( B \) population consisted of an identical copy of the observed population as regards numbers of households, and of adults and children in each household.
Table 1: Posterior means and standard deviations for the infection rate and threshold parameters assuming different values for the observed fraction of the population $\alpha$.

### 5.3 Results

Table 1 contains posterior means and standard deviations for the infection rate parameters and the threshold parameter $R_s$. The estimates are reasonably precise, as illustrated by the posterior standard deviations in Table 1, while each infection rate had a unimodal marginal posterior density (not shown). Point estimation is not greatly affected by the value of $\alpha$, which is to be expected since the actual observed data remain the same. Since the infectious period is 4.1 days, it follows that a typical individual (adult or child) makes, roughly, an average of (i) 0.4 infectious contacts in total with both adults and children in the population at large, (ii) 0.1 infectious contacts with each adult household member and (iii) 0.25 infectious contacts with each child household member. Thus children are around two and a half times more ‘at risk’ from an infected household member than adults. It is also instructive to consider probabilities of infection, as opposed to infection rates, and this is discussed further below.

A key observation from Table 1 is that the precision of estimates improves as $\alpha$ decreases. At first sight this seems highly counterintuitive, since as $\alpha$ decreases,
less and less is actually observed. The situation can be explained as follows. First, it is well-known that epidemics in large populations that actually take off follow a deterministic path, with relatively small random fluctuations, see for example Andersson and Britton (1999), Chapter 5. Now as $\alpha \downarrow 0$, the probability of the observed population $A \cup C$ containing the vast majority of all actual cases of infection tends to zero, since $A \cup C$ is assumed to be randomly chosen (and in particular, not chosen to increase the likelihood of observing cases). It follows that observing a reasonable number of cases in $A$ implies that the epidemic must have taken off in the entire population, so the large-population behaviour then gives increasingly precise estimation as $\alpha$ decreases. For example, $95\%$ equal-tailed posterior credible intervals for $R_\ast$ were found to be $(0.99, 1.56)$, $(1.04, 1.45)$ and $(1.11, 1.32)$ for $\alpha = 1$, 0.5 and 0.1, respectively. Note also that the epidemic having taken off means that $P(R_\ast \leq 1| n)$ decreases as $\alpha$ decreases; we obtained values of 0.03, 0.008 and $< 10^{-5}$ for $\alpha = 1$, 0.5 and 0.1, respectively, which illustrates this effect. The phenomena described in this paragraph also occur when using approximate methods of inference for two-level mixing models, see Demiris and O’Neill (2005a).

Haber et al. (1988) presents an analysis of the data using a slightly simpler model in which the fates of different households are assumed to be independent of one another. Instead, each individual in the population independently avoids infection from outside their household with some fixed probability. This independence assumption can be thought of as an approximation to our model when $\alpha$ is very small, since then the fates of randomly sampled households are approximately independent. The within-household dynamics of the Haber model are the same as for our model with a a fixed-length infectious period. The Haber
model is defined in terms of transmission probabilities rather than infection rates, specifically the probabilities $Q_j$ and $B_j$ that a type $j$ individual avoids infection from a single infective within the household, and from global infection, respectively. In our model these probabilities are given by

$$Q_j = 1 - \exp(-4.1\lambda_j^L)$$

and

$$B_j = 1 - \exp[ -4.1\lambda_j^G \sum_i (T_i^A + T_i^B)/N_j ], \quad j = 1, 2.$$  

There is reasonable agreement between the maximum likelihood estimates reported in Haber et al. (1988) for these quantities (namely $\hat{Q}_1 = 0.087, \hat{Q}_2 = 0.213, \hat{B}_1 = 0.104, \hat{B}_2 = 0.272$) and the corresponding posterior means from our analysis (namely 0.092, 0.224, 0.105 and 0.280, respectively) with $\alpha = 0.1$.

In addition to being able to obtain point estimates and associated measures of uncertainty such as those given in Haber et al. (1988), our approach readily provides information on functions of model parameters (such as $R_\ast$, already discussed above), the relationship between parameters a posteriori, and inference on the quantities used for data augmentation. For example, when $\alpha = 0.1$ the posterior correlation coefficients for pairs of infection rate parameters are given in Table 2. None of the correlation coefficients of distinct infection rates are greater than 0.5 in magnitude, which broadly indicates that the data are sufficient to enable each infection rate to be estimated separately. The pairs $(\lambda_j^L, \lambda_j^G), \quad j = 1, 2,$ have moderate negative correlation, which is to be expected because individuals of type $j$ will be infected either locally or globally according to the rates $\lambda_j^L$ and $\lambda_j^G$, respectively. The other values in Table 2 are less straightforward to interpret, but the negative correlations between the pairs $(\lambda_1^L, \lambda_2^L)$ and $(\lambda_1^G, \lambda_2^G)$ could be accounted for by considering the order in which individuals become infected. For instance, in a household with precisely one adult and one child infected, it is highly likely that only one was infected globally, so that an increase in $\lambda_1^G$ suggests
Table 2: Posterior correlation coefficients for the infection rate parameters assuming $\alpha = 0.1$.

<table>
<thead>
<tr>
<th></th>
<th>$\lambda^L_1$</th>
<th>$\lambda^L_2$</th>
<th>$\lambda^G_1$</th>
<th>$\lambda^G_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda^L_1$</td>
<td>1</td>
<td>-0.21</td>
<td>-0.45</td>
<td>0.12</td>
</tr>
<tr>
<td>$\lambda^L_2$</td>
<td>1</td>
<td>0.12</td>
<td>-0.46</td>
<td></td>
</tr>
<tr>
<td>$\lambda^G_1$</td>
<td>1</td>
<td></td>
<td>-0.34</td>
<td></td>
</tr>
<tr>
<td>$\lambda^G_2$</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

a decrease in $\lambda^G_2$. A similar argument applies to the local infection rates, i.e. the adult infecting the child suggests an increase in $\lambda^L_2$ and a decrease in $\lambda^L_1$, and vice versa.

Regarding inference for quantities used in data augmentation, suppose that $\alpha = 0.5$, so that the observed and unobserved populations are of the same size. The number of infected individuals in the unobserved $B$ population, $|\chi(B)|$, was found to have posterior mean and standard deviation 130 and 8.5, respectively, as compared to 125 infected individuals observed in $A$. As well as being of interest in its own right, such information gives a crude indication that the model fit is reasonable. More precisely, the epidemic in $B$ can, very approximately, be regarded as a prediction of a future outbreak based on the inference obtained from observing $A \cup C$. If the model was poorly fitting, then the observed data in $A \cup C$ would be relatively unlikely even under the most plausible estimated model parameters, while we would expect that the epidemic in $B$ would be a typical realisation based on such parameters. In particular, the epidemic in $B$ would look rather different to that in $A \cup C$. Thus a similar outbreak in $B$ to that observed in $A \cup C$ suggests that the observed data do lie within the range of typical outcomes of the underlying epidemic model. Obviously our arguments here are heuristic and overlook
the inherent dependencies between the $A$, $B$ and $C$ populations, but nevertheless the results are reassuring.

6 Concluding comments

We have extended the basic ideas contained in Demiris and O’Neill (2005b) to allow for sample-based data, and moreover developed new and more efficient MCMC algorithms. The algorithms seem to work quite well in practice and are reasonably fast, with typical run-times for the influenza data set ranging from a few hours (the standard algorithm with $\alpha$ large) to several days (the non-centered algorithm for small $\alpha$). The non-centered algorithm itself benefits from appropriate tuning of the proposal variances used to update the infection rate parameters. For larger data sets than we have considered here, methods of adaptive MCMC are then likely to be useful in order to perform such tuning automatically (see e.g. Roberts and Rosenthal, 2006).

There are several advantages to using the kind of methods described in this paper. Primarily, the methods permit very detailed data analyses to be conducted, yielding far more than just point estimates alone for the key model parameters. Additionally, the methods are highly flexible, and can be tailored both to different models and to data collected under different conditions to those described here. An example of such a data set, as described in detail in Park et al. (2004), consists of final outcome data on the spread of equine influenza in a population of Thoroughbred horses during 2003. In this setting, the groups in the two-level mixing model correspond to different yards where the horses were kept. All horses were vaccinated against influenza, but varying individual responses to the vaccine
meant that it was not known precisely which horses were susceptible, and which immune. Our methods can be easily adapted to analyse such data.

REFERENCES


