Simulation-based Bayesian inference for epidemic models

Trevelyan J. McKinley*

Disease Dynamics Unit, Department of Veterinary Medicine, University of Cambridge, Cambridge, UK.

Joshua V. Ross

School of Mathematical Sciences, The University of Adelaide, Adelaide, Australia.

Rob Deardon

Department of Mathematics and Statistics, University of Guelph, Guelph, Canada.

Alex R. Cook

Saw Swee Hock School of Public Health, Department of Statistics and Applied Probability, and Duke-NUS Graduate Medical School Singapore, National University of Singapore, Singapore.

Abstract

A powerful and flexible method for fitting dynamic models to missing and censored data is to use the Bayesian paradigm via data-augmented Markov chain Monte Carlo (DA-MCMC). This samples from the joint posterior for the parameters and missing data, but requires high memory overheads for large-scale systems. In addition, designing efficient proposal distributions for the missing data is typically challenging. Pseudo-marginal methods instead integrate across the missing data using a Monte Carlo estimate for the likelihood, generated from multiple independent simulations from the model. These techniques can avoid the high memory requirements of DA-MCMC, and under certain conditions produce the exact marginal posterior distribution for parameters. A novel method is presented for implementing importance sampling for dynamic epidemic models, by conditioning the simulations on sets of validity criteria (based

*Corresponding author. Tel.: (+44) 1223 337685.
Email address: tjm44@cam.ac.uk (Trevelyan J. McKinley)

on the model structure) as well as the observed data. The flexibility of these techniques is illustrated using both removal time and final size data from an outbreak of smallpox. It is shown that these approaches can circumvent the need for reversible-jump MCMC, and can allow inference in situations where DA-MCMC is impossible due to computationally infeasible likelihoods.
Keywords: Bayesian inference, epidemic models, Markov chain Monte Carlo, pseudo-marginal methods, smallpox

---

1. Introduction

Mathematical models of infectious disease dynamics are useful tools to help explore the biological mechanisms of disease spread and to provide predictive information to guide the implementation of control policies and interventions (see e.g. Bailey, 1975; Keeling and Rohani, 2008). A common way to model epidemic systems is to consider that individuals progress through different epidemiological states over time. A simple example for a single epidemic of a disease such as influenza is an $\mathcal{SIR}$ model, in which individuals are classified as either susceptible to infection ($\mathcal{S}$), infected and infectious ($\mathcal{I}$), or removed ($\mathcal{R}$; corresponding to recovered and immune, or dead). A functional form is then chosen to describe the movements of individuals between states, governed by a set of epidemiological parameters. Due to the inherently stochastic nature of infectious disease outbreaks, we eschew deterministic approximations in favor of fully stochastic models, in which state transitions are governed by sets of probability equations. Hence, multiple realizations of the system will result in a distribution of outcomes, even for a fixed set of parameter values (i.e. with no parameter uncertainty). Therefore the observed data are one realization of a stochastic process, the dynamics of which we are attempting to explore using the chosen model.

To ensure that the outputs from the model can be interpreted robustly, it is vital to account for *parameter* uncertainty, as well as stochasticity arising from the model dynamics. Various techniques exist in order to fit dynamic

models to data (see e.g. Bailey, 1975; Weirman and Marchette, 2004; Ionides et al., 2006; Cook et al., 2007; Höhle and Feldmann, 2007; Yang et al., 2007; Keeling and Ross, 2008; Jewell et al., 2009; Chis Ster et al., 2009; Deardon et al., 2010; Wong et al., 2013), many of which use a likelihood function to quantify the propensity of a given model and set of parameters to explain the observed data. However, the likelihood function can be difficult to calculate in practice, particularly when data are missing or incomplete. Although techniques exist to generate maximum likelihood estimates of dynamic temporal epidemic systems when data are missing/censored (e.g. Ionides et al., 2006), since it is often useful to supplement case time-series data with other forms of information—on the incubation period, say—here we use the Bayesian paradigm.

Readers unfamiliar with the Bayesian framework are referred to many excellent texts available, such as those by Gilks et al. (1996) and Gelman et al. (2004). This framework treats all parameters and variables as random, and the aim is to estimate the *posterior distribution* for the unknown parameters, $\boldsymbol{\theta}$, given the observed data, $\boldsymbol{D}$, written as $f(\boldsymbol{\theta}|\boldsymbol{D}) \propto f(\boldsymbol{D}|\boldsymbol{\theta})\, f(\boldsymbol{\theta})$, up to some normalizing constant, where $f(\boldsymbol{\theta})$ represents our *prior* knowledge about the parameters, and $f(\boldsymbol{D}|\boldsymbol{\theta})$ is the likelihood. The normalizing constant is often difficult to evaluate, and so we resort to numerical estimation methods such as Markov chain Monte Carlo (MCMC; e.g. Gilks et al., 1996) or Sequential Monte Carlo (SMC; e.g. Doucet et al., 2001). The techniques discussed in this paper relate directly to the former, and in particular are linked to the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970).

The Bayesian framework offers a natural environment to parameterize epidemic systems, since missing/censored data can simply be included as extra parameters in the model. One implementation of this approach is through data-augmented MCMC (DA-MCMC; Gibson and Renshaw, 1998; O'Neill and Roberts, 1999), which, particularly when coupled with reversible-jump (RJ) methodology (Green, 1995), is perhaps the most flexible computational technique currently available for fitting dynamic epidemic models to data. However, implementation of DA-MCMC can be challenging, particularly in defining effi-

cient proposal distributions for the missing data. For large amounts of missing data, it may be necessary to update each missing value, or subsets of the missing values, in turn. Furthermore, it may also be necessary to track the full history of each augmented variable. This can lead to large memory requirements for high-dimensional problems and highly autocorrelated chains. A recent paper by Andrieu et al. (2010) uses SMC methods to build efficient high-dimensional proposals for use in MCMC. Known as particle MCMC, this method has the potential to be widely applicable for inference in many epidemiological problems. However, in this paper we focus on an alternative method, based on using information from multiple repeated simulations instead of direct evaluation of the likelihood function. This idea goes back at least to Diggle and Gratton (1984), who approximate the log-likelihood through simulation, and use this to develop a numerical approximation routine for performing maximum likelihood calculations.

A general technique—based on these ideas—that is growing in popularity in various scientific fields is Approximate Bayesian Computation (ABC). For a given parameter value, multiple simulations from the model are produced and the proportion that 'match' the observed data are used to provide an estimate of the likelihood. This basic idea can be incorporated into rejection sampling (e.g. Tavaré et al., 1997; Beaumont et al., 2002), MCMC (e.g. Marjoram et al., 2003; Wilkinson, 2010) or SMC routines (e.g. Sisson et al., 2007; Toni et al., 2009; Beaumont et al., 2009; Erhardt and Smith, 2012). In practice the requirement to match the observed and simulated data exactly is relaxed, and instead some metric, $\rho(\cdot)$, is defined that characterizes the distance between the observed and simulated data sets. Simulations then 'match' if $\rho(\cdot)$ is less than some tolerance ϵ. This introduces three areas of approximation: the choice of metric, tolerance and the number of simulations used to produce the approximate Monte Carlo estimate. In McKinley et al. (2009), ABC techniques were employed to produce approximate posterior estimates for the parameters of a temporal epidemic model, both with and without missing data. The authors showed that it was possible to produce simple metrics that provided accurate estimates of the

true posterior (relative to the gold-standard of DA-MCMC) in the case where there is negligible missing data. However, they showed that the accuracy of the approximation begins to break down when the amount of missing data increases. Although these techniques are potentially useful to provide estimates of parameter uncertainty in complex models for which it is difficult to calculate a likelihood, it is not always clear how to define a metric sensibly, or decide on a suitable value for the tolerance. Questions also remain as to the impact of these choices on what the approximate posterior distribution actually represents (see e.g. Wilkinson, 2010), although a recent paper by Fearnhead and Prangle (2012) made some exciting developments in terms of re-casting ABC as an inferential framework in its own right, as opposed to simply approximating the true posterior. The reader is also encouraged to see Tanaka et al. (2006), Blum and Tran (2010) and Neal (2010) for other applications of ABC in epidemic modeling. Nonetheless, some of these complexities motivate the interest here to explore alternative simulation methods.

Pseudo-marginal approaches (see e.g. O'Neill et al., 2000; Beaumont, 2003; Andrieu and Roberts, 2009) are based on importance sampling. O'Neill et al. (2000) employ a so-called Monte Carlo within Metropolis (MCWM) algorithm to analyze epidemiological models based on household outbreak data. Beaumont (2003) introduces a similar algorithm called grouped-independence Metropolis-Hastings (GIMH) to analyze genealogical data. The convergence properties of both MCWM and GIMH are explored more theoretically in Andrieu and Roberts (2009), where the general moniker of 'pseudo-marginal approaches' is applied to cover both cases. Although they are based on a similar central concept, GIMH can be shown to produce an *exact* marginal posterior for the parameters, despite the use of a Monte Carlo (MC) estimate for the likelihood (Beaumont, 2003; Andrieu and Roberts, 2009). MCWM produces an approximation, though we show in Section 4 that this approximation is good for the sorts of applications discussed here. Similar techniques have been implemented with some success particularly in the field of statistical genetics (see e.g. O'Ryan et al., 1998; Berthier et al., 2002). A related method, using a slightly different

implementation of importance sampling in an SMC framework, was developed by Cauchemez et al. (2008) for making inference for a dynamic epidemic model based on a large-scale sentinel influenza data set. An alternative, more general, importance sampling technique would be population Monte Carlo (e.g. Cappé et al., 2004; Celeux et al., 2006), which employs adaptive importance sampling, where the samples at each generation depend on those in previous generations. However, the challenge of generating importance estimates for missing data in dynamic models still remains.

We introduce the MCWM and GIMH algorithms in Section 2. In Section 3 we discuss the formulation of general stochastic epidemic models and how we might implement these in pseudo-marginal routines. In Section 4 we introduce a data set from an outbreak of smallpox in Abakaliki, Nigeria, in 1967, and show that it is possible to generate simulation algorithms that match the data exactly for a range of compartmental epidemic models, assumptions and forms of the data (including when data are missing). We also show that in a range of cases pseudo-marginal routines can provide efficient alternatives to DA-/RJ-MCMC. We also show that in some scenarios simple changes to a simulation algorithm can bypass the need to use RJ-MCMC to account for changes in the dimensionality of the system, and that in other situations computationally feasible importance ratios can be generated when computationally infeasible likelihood functions cannot. We conclude with a discussion in Section 5.

2. Model fitting algorithms

For a large enough number of iterations, N_{iter}, the Metropolis-Hastings (M-H) algorithm generates a Markov chain that will converge to the correct posterior distribution, $f(\boldsymbol{\theta}|\boldsymbol{D})$, regardless (theoretically) of the starting point of the chain. (Here $\boldsymbol{\theta}$ represents a vector of parameters, and the notation $\boldsymbol{\theta}^{(i)}$ represents the values of the parameters at the i^{th} iteration.) The algorithm begins by proposing initial values for the chain, $\boldsymbol{\theta}^{(0)}$, and then proposing new values at each iteration from some (multidimensional) proposal distribution $q_\theta(\cdot|\boldsymbol{\theta}^{(i)})$.

Algorithm 1: Monte Carlo within Metropolis algorithm (MCWM)

Require: N_{iter}, N_{sim}, $\boldsymbol{\theta}^{(0)}$.
Set: $i = 0$
while $i < N_{\text{iter}}$ **do**
  $\boldsymbol{\theta}' \sim q_\theta\left(\cdot|\boldsymbol{\theta}^{(i)}\right)$
  Generate $\hat{f}\left(\boldsymbol{D}|\boldsymbol{\theta}^{(i)}\right)$ and $\hat{f}\left(\boldsymbol{D}|\boldsymbol{\theta}'\right)$ from equation (1)
  $u \sim U(0,1)$
  $\alpha = \min\left[1, \frac{\hat{f}(\boldsymbol{D}|\boldsymbol{\theta}')}{\hat{f}\left(\boldsymbol{D}|\boldsymbol{\theta}^{(i)}\right)} \times \frac{\pi(\boldsymbol{\theta}')}{\pi\left(\boldsymbol{\theta}^{(i)}\right)} \times \frac{q_\theta\left(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}'\right)}{q_\theta\left(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(i)}\right)}\right]$
  if $u < \alpha$ **then**
    $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}'$
  else
    $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)}$
  end if
  $i = i + 1$
end while

These candidate values are probabilistically accepted or rejected based on a ratio of posterior and proposal distributions evaluated at the current and proposed values (Metropolis et al., 1953; Hastings, 1970).

The form of $q_\theta(\cdot|\boldsymbol{\theta}^{(i)})$ is arbitrary, but affects the convergence and mixing properties of the chain, and an ideal sampler will have an independent proposal density that is close to the true posterior. In practice this can be difficult to achieve, and hence various adaptive proposal mechanisms exist that 'learn' how to sample more efficiently as the chain progresses (e.g. Haario et al., 2001; Roberts and Rosenthal, 2009). Once convergence has been reached, the initial draws (the burn-in) are discarded and the chain continues to be run until the required number of samples have been generated. Each iteration of the chain produces a random (but correlated) sample from the posterior.

In general, MCMC is good at dealing with high-dimensional problems—a further reason why it is a particularly useful framework for dealing with missing data problems. In DA-MCMC the parameter vector, $\boldsymbol{\theta}$, is augmented to include the missing data, $\boldsymbol{z}$, before using MCMC to explore the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{z}$. The marginal posterior for $\boldsymbol{\theta}$, $f(\boldsymbol{\theta}|\boldsymbol{D}) = \int_{\mathcal{Z}} f(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{D})\, d\boldsymbol{z}$,

Algorithm 2: Grouped independence Metropolis-Hastings algorithm (GIMH)

Require: N_{iter}, N_{sim}, $\boldsymbol{\theta}^{(0)}$.

Set: $i = 0$ and generate $\hat{f}\left(\boldsymbol{D}|\boldsymbol{\theta}^{(0)}\right)$ from equation (1)

while $i < N_{\text{iter}}$ **do**
  $\boldsymbol{\theta}' \sim q_\theta\left(\cdot|\boldsymbol{\theta}^{(i)}\right)$
  Generate $\hat{f}\left(\boldsymbol{D}|\boldsymbol{\theta}'\right)$ from (1)
  $u \sim U(0,1)$
  $\alpha = \min\left[1, \frac{\hat{f}(\boldsymbol{D}|\boldsymbol{\theta}')}{\hat{f}\left(\boldsymbol{D}|\boldsymbol{\theta}^{(i)}\right)} \times \frac{\pi(\boldsymbol{\theta}')}{\pi\left(\boldsymbol{\theta}^{(i)}\right)} \times \frac{q_\theta\left(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}'\right)}{q_\theta\left(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(i)}\right)}\right]$
  if $u < \alpha$ **then**
    $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}'$
    $\hat{f}\left(\boldsymbol{D}|\boldsymbol{\theta}^{(i+1)}\right) = \hat{f}\left(\boldsymbol{D}|\boldsymbol{\theta}'\right)$
  else
    $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)}$
    $\hat{f}\left(\boldsymbol{D}|\boldsymbol{\theta}^{(i+1)}\right) = \hat{f}\left(\boldsymbol{D}|\boldsymbol{\theta}^{(i)}\right)$
  end if
  $i = i + 1$
end while

can be obtained by integrating over the missing data, which is trivial to obtain from an MCMC run.

Given data $\boldsymbol{D}$ and parameters $\boldsymbol{\theta}$, O'Neill et al. (2000) propose a method to estimate the likelihood ratio in a M-H algorithm as $\hat{f}(\boldsymbol{D}|\boldsymbol{\theta}')/\hat{f}\left(\boldsymbol{D}|\boldsymbol{\theta}^{(i)}\right)$, where $\hat{f}(\boldsymbol{D}|\boldsymbol{\theta})$ is an MC estimate of $f(\boldsymbol{D}|\boldsymbol{\theta})$. Specifically this approximation is done by using importance sampling, where

$$\hat{f}(\boldsymbol{D}|\boldsymbol{\theta}) = \frac{1}{N_{\text{sim}}} \sum_{k=1}^{N_{\text{sim}}} \frac{f\left(\boldsymbol{D}, z^{(k)}|\boldsymbol{\theta}\right)}{q_{D,Z}\left(\boldsymbol{D}, z^{(k)}|\boldsymbol{\theta}\right)}, \tag{1}$$

with $\boldsymbol{Z}$ a set of auxiliary random variables (in this case representing the unobserved events and event times), $q_{D,Z}(\cdot)$ an importance-sampling distribution and $z^{(k)}$ the k^{th} random sample from $q_{D,Z}(\cdot)$. Whereas in DA-MCMC the auxiliary variables are integrated out of the joint posterior, in the pseudo-marginal approaches they are integrated out in the likelihood. The MCWM algorithm utilizing this estimate is shown in Algorithm 1.

In MCWM the MC estimates of the likelihood are re-simulated at each iteration of the chain. This leads to a biased estimate of the marginal pos-

terior $f(\boldsymbol{\theta}|\boldsymbol{D})$—though the bias should decrease as N_{sim} increases (Andrieu and Roberts, 2009). For their specific model, O'Neill et al. (2000) suggest a correction to help minimize this bias.

The GIMH algorithm, proposed by Beaumont (2003), is given in Algorithm 2. Here the auxiliary variables Z are *reused* at each iteration of the chain (rather than re-simulated). Beaumont (2003) provides an elegant proof that Algorithm 2 will produce samples from the correct marginal $f(\boldsymbol{\theta}|\boldsymbol{D})$ as $N_{\text{iter}} \to \infty$, regardless of the value of N_{sim}; though he noticed that increasing the number of simulations increases the acceptance rate of the chain. Qualitatively at least, a similar pattern was remarked upon in McKinley et al. (2009) for a related Approximate Bayesian Computation method (ABC-MCMC) when the simulations are reused.

3. General compartmental epidemic models

As previously discussed, epidemic systems can be characterized by allowing individuals in the population to move between a series of discrete epidemiological states, where transitions between states are governed by probability statements. As an example consider the $\mathcal{SIR}$ model discussed earlier. Many variations of this basic system exist, but here we assume that we have a closed population of $N_{\text{pop}} > 1$ individuals, that individuals who become infected move through states $\mathcal{S} \to \mathcal{I} \to \mathcal{R}$ in that order. The transition probabilities in this case are:

$$\begin{aligned} P(SI) &= \beta S^{(t)} I^{(t)} dt + o(dt), \text{ and} \\ P(IR) &= \gamma I^{(t)} dt + o(dt), \end{aligned} \tag{2}$$

where the notation SI corresponds to the movement of a single individual from state $\mathcal{S}$ to state $\mathcal{I}$ in the time period $(t, t+dt)$—where $dt \ll 1$—and likewise for IR. Here β is the transmission parameter, γ^{-1} is the mean infectious period and the model results in exponentially-distributed inter-event times (see e.g. Keeling and Ross, 2008). $S^{(t)}$ and $I^{(t)}$ represent the numbers of susceptibles and infectives at time t. If an epidemic begins at time $t^{(1)}$, then at some subsequent time point $T_{\max} > t^{(1)}$, N_E events will have occurred, where $N_E = N_{SI} + N_{IR}$

is the sum of the number of SI and IR events in $[t^{(1)}, T_{\max})$. If the epidemic dies out before $T_{\max}$, then the final epidemic size will be N_F (where $N_{SI} = N_{IR} = N_F$ and $N_E = 2N_F$).

A key inference problem for these systems is that the epidemic process operates in continuous time but available data are almost always discrete—being snapshots of the system—and hence are likely to form (partial) time-series counts of individuals in some, but not necessarily all, of the epidemic states. This makes evaluation of the likelihood difficult unless we introduce latent variables to account for the unobserved (continuous) events. It is possible in some cases to approximate this process using a discrete-time model, but here we will assume that we wish to fit a continuous-time model to discrete-sampled data. In addition there are often missing data, such as missing counts at different time points, or, often, we observe events of one type but not another (such as removals but not infections). We will use the random variable X to denote the type of event, such that

$$X = \begin{cases} 0 & \text{if event is an } SI \text{ event, and} \\ 1 & \text{if event is an } IR \text{ event.} \end{cases} \tag{3}$$

T denotes the corresponding *continuous* event times, and Y denotes *discrete-time* observations. Let $\boldsymbol{D} = \left\{\left(y^{(1)}, x_d^{(1)}\right), \ldots, \left(y^{(N_O)}, x_d^{(N_O)}\right)\right\}$ represent N_O observations, and $\boldsymbol{z} = \left\{\left(t^{(1)}, x_z^{(1)}\right), \ldots, \left(t^{(N_E)}, x_z^{(N_E)}\right)\right\}$ correspond to N_E unobserved events, where $N_E \geq N_O$. (Note that observations $\boldsymbol{y}$ and $\boldsymbol{x}_d$ can be obtained directly from time-series count data and visa-versa.) As noted in various studies (e.g. Gibson and Renshaw, 1998; O'Neill and Roberts, 1999; Ross et al., 2006) the likelihood function based on the discrete-time events, $f(\boldsymbol{D}|\beta, \gamma)$, is often infeasible to evaluate directly. However, $f(\boldsymbol{z}|\beta, \gamma)$ has a more efficient

closed form, given as:

$$f(\boldsymbol{z} \mid \beta, \gamma) = N_{\text{pop}} f\left(t^{(1)}\right) \left\{ \prod_{j=2}^{N_E} \left[\left(\beta S^{(j-1)} I^{(j-1)}\right)^{1-x_z^{(j)}} \left(\gamma I^{(j-1)}\right)^{x_z^{(j)}} \right] \times \exp\left[-\left(\beta S^{(j-1)} I^{(j-1)} + \gamma I^{(j-1)}\right)\left(t^{(j)} - t^{(j-1)}\right)\right] \right\} \times \exp\left[-\left(\beta S^{(N_E)} I^{(N_E)} + \gamma I^{(N_E)}\right)\left(T_{\max} - t^{(N_E)}\right)\right]. \quad (4)$$

(A more general form for non-exponential infectious period distributions is provided in Appendix A.) For a given vector $\boldsymbol{z}$, the counts $S^{(j)}$ and $I^{(j)}$ can be calculated as required. The marginal posterior of interest is then

$$f(\beta, \gamma | \boldsymbol{D}) \propto \int_{\boldsymbol{\mathcal{Z}}} f(\boldsymbol{D}|\boldsymbol{z}, \beta, \gamma) f(\boldsymbol{z}|\beta, \gamma) f(\beta, \gamma) d\boldsymbol{z}, \quad (5)$$

where $\mathcal{Z}$ is the multidimensional parameter space of all possible latent event times and orderings, and $f(\boldsymbol{D}|\boldsymbol{z}, \beta, \gamma)$ is the probability that we observe data $\boldsymbol{D}$ given $\boldsymbol{z}$, β and γ. Samples from this marginal posterior can be generated using DA- and/or RJ-MCMC, in which candidate values for the missing data $\boldsymbol{z}$ are proposed at each iteration of the MCMC chain, alongside candidate values for the parameters. For a given set of N_E ordered events, in order to produce a non-zero value for $f(\boldsymbol{z}|\beta, \gamma)$, $\boldsymbol{z}$ must constitute a valid epidemic based on the model (2). For this to be true, each event j must adhere to certain conditions given the history of the events so far and the limiting states of the system. Let $C_I^{(j)}$ and $C_R^{(j)}$ represent the *cumulative* number of SI and IR events, and $S^{(j)}$ and $I^{(j)}$ the numbers of susceptible and infective individuals just after the j^{th} event (occurring at time $t^{(j)}$). If the first event is an infection event at time $t^{(1)}$, then for $j > 1$,

$$S^{(j-1)} > 0,\ I^{(j-1)} > 0 \qquad \text{if event } j \text{ is an } SI \text{ event } (x_z^{(j)} = 0), \quad (6)$$

$$I^{(j-1)} > 0 \qquad \text{if event } j \text{ is an } IR \text{ event } (x_z^{(j)} = 1), \quad (7)$$

$$C_R^{(j)} < C_I^{(j)} \leq N_{\text{pop}} \qquad \text{for all } j < N_E, \text{ and} \quad (8)$$

$$C_R^{(j)} \leq C_I^{(j)} \leq N_{\text{pop}} \qquad \text{for } j = N_E. \quad (9)$$

These follow directly from the model specification (2) and hold true for any

fitting mechanism that requires events to be proposed in some way. Note that these conditions can be modified to deal with different models, for example the introduction of a latent (i.e. infected but not infectious) class $\mathcal{E}$—a so-called $\mathcal{SEIR}$ model (see section 4.1.3).

One way that MCWM or GIMH could be implemented for these systems would be to generate N_{sim} realizations of an epidemic, $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(N_{\text{sim}})}\}$, by first simulating a time of initial infection from some prior distribution, and then applying Gillespie's algorithm (Gillespie, 1977). In this case $f\left(\boldsymbol{D}, \boldsymbol{z}^{(k)}|\boldsymbol{\theta}\right) = f\left(\boldsymbol{D}|\boldsymbol{z}^{(k)}, \boldsymbol{\theta}\right) f(\boldsymbol{z}^{(k)}|\boldsymbol{\theta})$, and the importance estimate (1) reduces to

$$\hat{f}(\boldsymbol{D}|\boldsymbol{\theta}) = \frac{1}{N_{\text{sim}}} \sum_{k=1}^{N_{\text{sim}}} f\left(\boldsymbol{D}|\boldsymbol{z}^{(k)}, \boldsymbol{\theta}\right), \tag{10}$$

since $f(\boldsymbol{z}^{(k)}|\boldsymbol{\theta}) = q_Z\left(\boldsymbol{z}^{(k)}|\boldsymbol{\theta}\right)$. Here $f\left(\boldsymbol{D}|\boldsymbol{z}^{(k)}, \boldsymbol{\theta}\right) = 1$ if $\boldsymbol{z}^{(k)}$ is consistent with $\boldsymbol{D}$, and 0 otherwise. This approach is equivalent to the ABC-MCMC routine implemented by McKinley et al. (2009) in the case where the simulations match the data exactly. Simulating in this way ensures that for any realization of the model, $\boldsymbol{z}^{(k)}$, the ordering conditions are automatically adhered to. The main challenge is that in highly stochastic and/or high dimensional systems the probability of matching the simulations to the data is very low. In the subsequent sections we show how alternative simulation mechanisms can be developed for different epidemic models, in which the model structure, the observed data and the ordering criteria are exploited to ensure (at least for the models presented here) that $f\left(\boldsymbol{D}|\boldsymbol{z}, \boldsymbol{\theta}\right) = 1$. (Of course if the observation process is not deterministic, then this requirement can be relaxed—see discussion in Section 5.)

4. Applications

All routines were run on a Dell XPS 15Z laptop with an Intel(R) Core(TM) i7-2640M CPU @ 2.80Ghz × 4 processors running Ubuntu 12.04. The algorithms were coded in C using the GNU Scientific Library. To protect against precision issues when calculating anti-logs, we used multiple precision arithmetic provided by the GNU Multiple Precision Arithmetic Library (http://gmplib.org/) and

the GNU MPFR library (http://www.mpfr.org/). All plots and results were generated using the R statistical language (R Development Core Team, 2011) with the `coda` (Plummer et al., 2010) package. All software and packages used are open-source and free to download.

4.1. Outbreak of smallpox in Abakaliki, Nigeria in 1967

These data consist of a set of 29 inter-removal times from an epidemic of smallpox in a closed population of $N_{\text{pop}} = 120$ individuals, and have been studied by various authors, for example Bailey (1975); Becker (1983); O'Neill and Roberts (1999); Gibson and Renshaw (2001); O'Neill and Becker (2001); Fearnhead and Meligkotsidou (2004) and Boys and Giles (2007). O'Neill and Roberts (1999) assume that the model follows a density-dependent $\mathcal{SIR}$ structure; in reality there is an appreciable latent period (as fitted in Becker, 1983, Gibson and Renshaw, 2001 and O'Neill and Becker, 2001). For comparison we use the $\mathcal{SIR}$ here, fitting to both removal time as well as final size data only (see e.g. Ball, 1986; Becker, 1989; Rida, 1991 and Demiris and O'Neill, 2005a,b, 2006). An alternative, more detailed, version of this data set is available that allows more complex, and arguably more epidemiologically correct models to be fitted (see Eichner and Dietz, 2003). Here we use the simpler form in order to allow direct comparison of the parameter estimates to those obtained from previous Bayesian fitting methods.

4.1.1. $\mathcal{SIR}$ model for removal data

Consider that the observed data consist of a set of N_{IR} discrete removal times, and denote these as $\boldsymbol{D}_R = \{(y_R^{(r)}, 1); r = 1, \ldots, N_{IR}\}$, where without loss-of-generality $y_R^{(1)} = 0$. We consider two scenarios: the first when the epidemic is known to end on the N_{IR}^{th} removal, and the second when there is the possibility that the epidemic is still ongoing at some time point $T_{\max} > y_R^{(N_{IR})}$. We assume that the data are observed at daily time intervals, such that a removal observed at time $y_R^{(r)}$ will have occurred in the period $\left(y_R^{(r)} - 1, y_R^{(r)}\right]$. We also follow O'Neill and Roberts (1999) and place an exponential prior on the time between the first (unobserved) infection and the first observed removal, such

that $y_R^{(1)} - t^{(1)} \sim \text{Exp}(\theta)$.

Final size known

Here $N_F = N_{IR}$. In this case the ordering criteria (6)–(9) can be simplified to:

$$C_I^{(j)} \leq N_F \qquad \text{for all } j, \tag{11}$$

$$C_I^{(j)} > C_R^{(j)} \qquad \text{if } C_R^{(j)} < N_F, \text{ and} \tag{12}$$

$$C_I^{(j)} = C_R^{(j)} \qquad \text{if } C_R^{(j)} = N_F. \tag{13}$$

The first criterion is obvious, since we know that the final epidemic size is N_F, and if $N_F \leq N_{\text{pop}}$ in a closed population then $S^{(j-1)} \geq 0$ for all j. The second and third criteria follow from the fact that if $C_I^{(j)} = C_R^{(j)}$ then $I^{(j)} = 0$, and therefore the epidemic is over. Hence $C_I^{(j)} = C_R^{(j)}$ if and only if $C_I^{(j)} = N_F$, otherwise $C_I^{(j)} > C_R^{(j)}$. These criteria make it possible to generate stochastic simulations of the missing events and event times, $\boldsymbol{z}$, such that $f\left(\boldsymbol{D}_R|\boldsymbol{z}, \beta, \gamma, \theta\right) = 1$.

Let $\boldsymbol{t}_R$ be a vector of length N_F recording the subset of $\boldsymbol{t}$ corresponding to the removal times. The algorithm is initialized as follows: firstly, generate a set of continuous removal times using uniform order statistics based on the observed data: i.e. if there are $N_R^{(0)}$ removals in the time period $(-1, 0)$, then simulate $N_R^{(0)}$ events from a $U(-1, 0)$ distribution and sort into ascending order. Assign $t_R^{(j)}$ to these values for $j = 1, \ldots, N_R^{(0)}$. Repeat these steps for $j = N_R^{(0)} + 1, \ldots, N_R^{(0)} + N_R^{(1)}$ and so on to end up with an ordered set of simulated removal times.

Then, conditional on $t_R^{(1)}$, the initial infection time, $t^{(1)}$, is sampled such that $t^{(1)} < t_R^{(1)}$ and $t_R^{(1)} - t^{(1)} \sim \text{Exp}(\theta)$ (see Figure 1: Step 1). Letting j denote the current event, and r the *next* removal event, set $j = 1$, $r = 1$ and $z^{(1)} = \left(t^{(1)}, 0\right)$. If $t^{(j)}$ is the current event time, and $t_R^{(r)}$ is the next removal time, we generate a probability that an infection event occurs in the interval $\left(t^{(j)}, t_R^{(r)}\right)$, subject to a series of constraints to ensure that the simulations are valid and that the corresponding time-series counts match the observed data. If we have already had N_F infection events, then there is a zero probability of having any more from criterion (11). Otherwise, if $C_I^{(j)} = r$ and $r < N_F$, then we must have *at*

least one infection event in the interval due to criterion (12). These conditions ensure that criterion (13) is also matched. Otherwise we generate a non-zero probability based on the model structure and the current infection rate. Hence

$$p^{(j)} = \begin{cases} 0 & \text{if } C_I^{(j)} = N_F, \\ 1 - \exp\left[-\lambda_{SI}^{(j)}\left(t_R^{(r)} - t^{(j)}\right)\right] & \text{if } \left(C_I^{(j)} < N_F\right) \cap \left(C_I^{(j)} > r\right), \\ 1 & \text{otherwise,} \end{cases} \quad (14)$$

where $\lambda_{SI}^{(j)} = \beta S^{(j)} I^{(j)}$ is the rate of infection events directly after the j^{th} event. Based on $p^{(j)}$, we then randomly sample whether an infection event occurs. If not, then we set $t^{(j+1)} = t_R^{(r)}$ and $z^{(j+1)} = \left(t^{(j+1)}, 1\right)$ and increment r by one before continuing. If an infection event does occur, then we sample the event time, $t^{(j+1)}$, from a truncated distribution constrained in the interval $\left(t^{(j)}, t_R^{(r)}\right)$. Here we choose a truncated exponential distribution with probability density function

$$q_{t_R^{(r)} - t^{(j)}}\left(t; \lambda_{SI}^{(j)}\right) = \frac{\lambda_{SI}^{(j)} e^{-\lambda_{SI}^{(j)} t}}{1 - e^{-\lambda_{SI}^{(j)}\left(t_R^{(r)} - t^{(j)}\right)}} \qquad 0 < t < t_R^{(r)} - t^{(j)},\ \lambda_{SI}^{(j)} > 0, \quad (15)$$

and set $z^{(j+1)} = \left(t^{(j+1)}, 0\right)$. Finally we increment j by one and continue until the final removal (Figure 1: Steps 2 and 3).

In practice, it is not necessary to simulate all removal times in advance, as it would be feasible to generate only the $N_R^{(y)}$ times required in each discrete period $(y-1, y)$ in turn. It is also not necessary to record the entire history of the epidemic in order to calculate the importance estimate, and so to improve computational efficiency the log-importance contribution can be updated recursively as the simulation progresses. Algorithm C.1 (Supp. Mat.) provides pseudo-code for an efficient implementation of this routine, and discussion regarding the importance contributions from each simulated event is given in Appendix C.1.

Final size unknown

In the situation in which the epidemic is still ongoing at time $T_{\max}$, the number of infection events is known only to be greater or equal to the observed number

of removals at $T_{\max}$ (i.e. $N_{SI} \geq N_{IR}$). Gibson and Renshaw (1998) and O'Neill and Roberts (1999) deal with this by introducing reversible-jump steps to allow the number of unobserved infections to vary. In our approach we simply modify our ordering criteria as follows:

$$C_I^{(j)} \leq N_{\text{pop}} \qquad \text{for all } j, \tag{16}$$

$$C_I^{(j)} > C_R^{(j)} \qquad \text{if } C_R^{(j)} < N_{IR}, \text{ and} \tag{17}$$

$$C_I^{(j)} \geq C_R^{(j)} \qquad \text{if } C_R^{(j)} = N_{IR}. \tag{18}$$

The first condition follows from the fact that the maximum number of possible infections in a closed community of a fixed size N_{pop} is N_{pop}. The second and third are a direct result of the model specification as before, allowing for the fact that the epidemic could still be ongoing when $C_I^{(j)} = N_{IR}$. Therefore there are two ways that the simulation can end: firstly, if $C_I^{(j)} = N_{IR}$ at the time of the N_{IR}^{th} removal (in which case $\lambda_{SI}^{(j)} = \lambda_{IR}^{(j)} = 0$), and secondly if the epidemic is still going at time $T_{\max}$. In the latter case, at the time of the N_{IR}^{th} removal $\left(\text{i.e. when } t^{(j)} = t_R^{(N_{IR})}\right)$, $C_I^{(j)} > N_{IR}$, and so we must continue to simulate potential infection events in $\left(t^{(j)}, T_{\max}\right)$ until no more occur (Algorithm C.2, Supp. Mat.).

4.1.2. $\mathcal{SIR}$ model results for removal data

In slight contrast to our approach, O'Neill and Roberts (1999) assume that $\boldsymbol{y}_R$ constitute the *exact* removal times and implement an MCMC algorithm in which the parameters $t^{(1)}$, β and γ are updated using Gibbs sampling steps, and then the unobserved infection times are updated using a Metropolis-Hastings step in which an event is either moved, added or removed. In the first instance we focus on the case where the epidemic is known to have finished at the final removal time (Algorithm C.1). In the algorithm of O'Neill and Roberts (1999), the probability of adding or removing an infection time is therefore zero, so the only valid proposal for the M-H step is to move an existing infection time. An alternative Bayesian approach was proposed by Fearnhead and Meligkotsidou (2004), who develop an exact filtering algorithm to fit the same model.

Let $G(\rho, \tau)$ represent a gamma distribution with shape parameter ρ and scale parameter τ. Following O'Neill and Roberts (1999), we set the prior for the time between initial infection and first removal, θ, to 0.1, and use independent priors for β and γ of $G(10, 10^{-4})$ and $G(10, 10^{-2})$ (with prior means of 0.001 and 0.1) respectively. We use $N_{\text{sim}} = 512$ repeated simulations and run the algorithms for 10,000 iterations burn-in plus a further 90,000 updates. We use an adaptive proposal scheme on the log-scale for the parameters (see Roberts and Rosenthal, 2009). Table 1 compares the mean and posterior variance of β and γ obtained from both pseudo-marginal approaches against the maximum likelihood estimates obtained by Frank Ball (as reported in O'Neill and Roberts, 1999), the MCMC algorithm of O'Neill and Roberts (1999) and the exact filtering algorithm of Fearnhead and Meligkotsidou (2004). The means in all cases are fairly similar (though the Bayesian estimates are on the whole slightly higher than the ML estimates; except in the case of MCWM, which are very similar). The posterior variances obtained from the GIMH and exact filtering algorithms are approximately four times larger than those obtained from the DA-MCMC approach. Fearnhead and Meligkotsidou (2004) suggest that with respect to their method this difference may be due to the fact that MCMC algorithms are known to struggle with exploring heavy-tailed posteriors. This explanation is also feasible for the pseudo-marginal approaches, since the independent simulations used to produce the MC estimate of the likelihood removes some of the correlation structure that is inherent in the proposal mechanisms used in the MCMC algorithm, which might prevent it from efficiently exploring the tails of the posterior distributions. As expected, since MCWM is an approximate method, the posterior variances are higher than the other approaches. The GIMH algorithm using the informative prior took ≈ 26 minutes to run 100,000 iterations.

We also fit this model using uninformative priors for β, γ and $t^{(1)}$ [uniform on $(0, \infty)$ for the two former variables and on $(-\infty, 0)$ for the latter]. In this case it is no longer sensible to draw initial infection times during the simulations from the prior. Instead we simulate the initial infection time such that $t_R^{(1)} - t^{(1)} \sim$

Table 1: Posterior means and variances for β and γ for an $\mathcal{SIR}$ model fitted to the Abakaliki smallpox data set obtained using GIMH and MCWM, compared to estimates obtained from other methods.

Final size	**Priors**	**Method**	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$
Known	Informative	GIMH	9.4×10^{-4} (3.9×10^{-8})	0.098 (4.0×10^{-4})
		MCWM	8.5×10^{-4} (4.5×10^{-8})	0.089 (6.4×10^{-4})
		O'Neill and Roberts (1999)	1.1×10^{-3} (1.0×10^{-8})	0.11 (9.0×10^{-5})
		Fearnhead and Meligkotsidou (2004)	9.4×10^{-4} (3.6×10^{-8})	0.098 (4.0×10^{-4})
		Ball	8.3×10^{-4}	0.088
	Uninformative	GIMH	9.5×10^{-4} (6.6×10^{-8})	0.10 (8.3×10^{-4})
		MCWM	8.0×10^{-4} (6.6×10^{-8})	0.087 (1.2×10^{-3})
		O'Neill and Roberts (1999)	9.0×10^{-4} (3.8×10^{-8})	0.098 (4.3×10^{-4})
Unknown	Informative	GIMH	9.1×10^{-4} (3.6×10^{-8})	0.083 (4.6×10^{-4})
		MCWM	8.3×10^{-4} (3.7×10^{-8})	0.073 (6.1×10^{-4})

Exp(γ), and change the log-importance contribution accordingly. The results from this model are also shown in Table 1, and we can see that our estimates of the posterior variances are approximately 1.5–1.7 times larger than those for the informative prior. As an exercise, we also fit a model assuming N_F is unknown (Algorithm C.2, Supp. Mat.), with results similar to the case where $N_F = 30$ (although the posterior means are slightly lower). Note that posterior propriety is not trivial to establish for these priors, although impropriety would usually manifest in poor samples from the posterior, which is not the case here. An alternative would simply be to choose a proper prior distribution with very large variance (e.g. uniform on a large but finite support).

4.1.3. $\mathcal{SEIR}$ model for removal time data when final epidemic size is known

Full details of the model specification, likelihood and ordering criteria are given in Appendix C.2. When the final epidemic size is known, the ordering

criteria are:

$$S^{(j-1)} > 0,\ I^{(j-1)} > 0 \qquad \text{if event } j \text{ is an } SE \text{ event,} \tag{19}$$

$$E^{(j-1)} > 0 \qquad \text{if event } j \text{ is an } EI \text{ event,} \tag{20}$$

$$I^{(j-1)} > 0 \qquad \text{if event } j \text{ is an } IR \text{ event,} \tag{21}$$

$$C_R^{(j)} \leq C_I^{(j)} \leq N_F \qquad \text{for all } j < N_E, \tag{22}$$

$$C_I^{(j)} \leq C_E^{(j)} \leq N_F \qquad \text{for all } j < N_E, \tag{23}$$

$$C_R^{(j)} < C_E^{(j)} \leq N_F \qquad \text{for all } j < N_E, \text{ and} \tag{24}$$

$$C_R^{(j)} = C_I^{(j)} = C_E^{(j)} = N_F \qquad \text{for } j = N_E. \tag{25}$$

To simulate from this model, we first generate the continuous removal times and the initial infection time in the same manner as for the $\mathcal{SIR}$ model. We then set event indicator $j = 1$ and the indicator for the *next* removal $r = 1$, and generate a probability of a non-removal event occurring in $\left(t^{(j)}, t_R^{(r)}\right)$ as

$$p^{(j)} = \begin{cases} 1 - \exp\left[-\lambda_{EI}^{(j)}\left(t_R^{(r)} - t^{(j)}\right)\right] & \text{if } \left(C_E^{(j)} = N_F\right) \cap \left(C_I^{(j)} > r-1\right), \\ 1 - \exp\left[-\left(\lambda_{SE}^{(j)} + \lambda_{EI}^{(j)}\right)\left(t_R^{(r)} - t^{(j)}\right)\right] & \text{if } \left(C_E^{(j)} < N_F\right) \cap \left(C_I^{(j)} > r-1\right) \\ & \quad \cap \left[\left(C_E^{(j)} > r\right) \cup \left(r = N_F\right)\right], \text{ and} \\ 1 & \text{otherwise.} \end{cases} \tag{26}$$

We also generate a conditional probability that a simulated non-removal event is an SE event as

$$p_{SE}^{(j)} = \begin{cases} 0 & \text{if } \left(C_E^{(j)} = N_F\right), \text{ and} \\ \frac{\lambda_{SE}^{(j)}}{\lambda_{SE}^{(j)} + \lambda_{EI}^{(j)}} & \text{otherwise.} \end{cases} \tag{27}$$

Correspondingly, the conditional probability that a non-removal event is an EI event is given by $p_{EI}^{(j)} = 1 - p_{SE}^{(j)}$. These follow from conditions (19)–(25): if $C_E^{(j)} = N_F$, then there can be no further SE events [condition (25)]; however, there could be an EI event. Here this will occur with probability 1 if $C_I^{(j)} = r-1$ [condition (22)], or probability $1 - \exp\left[-\lambda_{EI}^{(j)}\left(t_R^{(r)} - t^{(j)}\right)\right]$ otherwise (note that if $C_I^{(j)} = N_F$, then $\lambda_{EI}^{(j)} = 0$ and so no further EI events can occur [condition (25)]).

If $C_E^{(j)} < N_F$, then we could have SE or EI events occurring. If, in addition to $C_E^{(j)} < N_F$, we have that $C_E^{(j)} = r$ and $r < N_F$, then an SE event must occur before the next removal [condition (24)]. As before, if $C_I^{(j)} = r - 1$, then an EI event must also occur [condition (22)]. If neither of these latter conditions are violated, then a non-removal event occurs with probability $1 - \exp\left[-\left(\lambda_{SE}^{(j)} + \lambda_{EI}^{(j)}\right)\left(t_R^{(r)} - t^{(j)}\right)\right]$. Conditional on a non-removal event occurring, the type of event is chosen by the conditional probability $p_{SE}^{(j)} = \lambda_{SE}^{(j)} / \left(\lambda_{SE}^{(j)} + \lambda_{EI}^{(j)}\right)$ and the event time is sampled from a truncated distribution in a similar manner to before. Appendix C.2 gives more details of the importance ratio calculations and pseudo-code is provided in Algorithm C.3 (Supp. Mat.).

4.1.4. $\mathcal{SEIR}$ model results for removal data

Table 2 provides results from an $\mathcal{SEIR}$ model fitted using GIMH, MCWM and DA-MCMC. To illustrate the sampling properties, the trace plots for these models are shown in Figure S1. The MCWM algorithm has better mixing properties than either the GIMH or DA-MCMC for this example, though we must state that the DA-MCMC algorithm used does not employ more sophisticated methods that improve mixing, such as partial non-centering (Papaspiliopoulos et al., 2003; Kypraios, 2007; Jewell et al., 2009). Nevertheless, both pseudo-marginal methods perform well, albeit at the cost of more uncertainty in the approximate posteriors for the MCWM routine (characterised by the variability in the values of the importance estimates accepted—see Figure S1).

Table 2: Posterior means and variances for β, δ and γ for an $\mathcal{SEIR}$ model fitted to the Abakaliki smallpox data set obtained using GIMH, MCWM and DA-MCMC.

Final size	**Priors**	**Method**	$\boldsymbol{\beta}$	$\boldsymbol{\delta}$	$\boldsymbol{\gamma}$
Known	Informative	GIMH	1.1×10^{-3} (6.0×10^{-8})	0.14 (9.1×10^{-4})	0.12 (6.2×10^{-4})
		MCWM	1.1×10^{-3} (1.0×10^{-7})	0.15 (2.1×10^{-3})	0.11 (1.2×10^{-3})
		DA-MCMC	1.1×10^{-3} (5.1×10^{-8})	0.13 (9.1×10^{-4})	0.11 (5.6×10^{-4})

4.1.5. $\mathcal{SIR}$ model for final size data

This simulation approach can also be applied for inference when only the final epidemic size is known (see e.g. Ball, 1986; Becker, 1989; Rida, 1991 and Demiris and O'Neill, 2005a,b, 2006). As a simple example consider the Abakaliki data with all of the temporal information removed, leaving the size of the initial susceptible population ($N_{\text{pop}} = 120$) and the final epidemic size ($N_F = 30$). For consistency with Demiris and O'Neill (2006), we use an $\mathcal{SIR}$ model with frequency-dependent transmission [i.e. $P(\mathcal{S} \to \mathcal{I}) = \beta S^{(t)} I^{(t)} N_{\text{pop}}^{-1} dt + o(dt)$]. Since the population is closed (i.e. no births, deaths or migrations occur), N_{pop} is constant, and so this is the same model as before only with β re-scaled by a factor of $1/N_{\text{pop}}$. We are interested in producing a posterior for the basic reproduction number R_0 (defined as the average number of secondary infections produced from a single primary infection introduced into a fully susceptible population). However, since there is no temporal information in the data, it is not possible to estimate the length of the infectious period; nonetheless it is possible to make inference about R_0 under different choices for the infectious period distribution. To mirror Demiris and O'Neill (2006) we choose three options, such that in each case the mean length is 4.1. These are: i) constant, ii) a gamma distribution with variance 8.405, and iii) an exponential distribution with variance 4.1^2.

Demiris and O'Neill (2006) use a set of triangular equations for the final size probabilities, derived by Ball (1986), that can be calculated recursively. They use multiple precision arithmetic to enable accurate calculation of Ball's result, and implement this within a Bayesian MCMC algorithm to estimate the posterior distributions for $R_0 = 4.1\beta$ given each infectious period distribution. To do this in a pseudo-marginal framework, we generate an importance sample estimate of the final size likelihood by repeatedly simulating a set of continuous-time epidemics, each constrained to have a final size of N_F, in a similar manner

to before, hence

$$\hat{f}(N_F|\beta) = \frac{1}{N_{\text{sim}}} \sum_{k=1}^{N_{\text{sim}}} \frac{f\left(N_F|\boldsymbol{z}^{(k)},\beta\right) f\left(\boldsymbol{z}^{(k)}|\beta\right)}{q_Z\left(\boldsymbol{z}^{(k)}|\beta\right)}, \tag{28}$$

where $f\left(N_F|\boldsymbol{z}^{(k)},\beta\right) = 1$ if $\boldsymbol{z}^{(k)}$ are consistent with N_F and 0 otherwise. Each simulation is initialized with an index infection event at time $t_1 = 0$.

4.1.6. Fixed infectious period of length T_I

For a given value of β and T_I, set $j = 1$, $r = 1$, $C_I^{(j)} = 1$, $t_I^{(1)} = 0$ and $t_R^{(1)} = T_I$. The probability an SI event occurs in $\left[t_I^{(j)}, t_R^{(r)}\right)$ is given by:

$$p^{(j)} = \begin{cases} 0 & \text{if } C_I^{(j)} = N_F, \\ 1 - \exp\left[-\lambda_{SI}^{(j)}\left(t_R^{(r)} - t_I^{(j)}\right)\right] & \text{if } \left(C_I^{(j)} > r\right) \cap \left(C_I^{(j)} < N_F\right), \text{ and} \\ 1 & \text{otherwise,} \end{cases} \tag{29}$$

where $\lambda_{SI}^{(j)} = \beta S^{(j)} I^{(j)} / N_{\text{pop}}$. If an infection event occurs then sample the next infection time, $t_I^{(j+1)}$, from a truncated distribution in the period $\left[t_I^{(j)}, t_R^{(r)}\right)$, and set a new removal time $t_R^{(C_I^{(j)}+1)} = t_I^{(j+1)} + T_I$, before updating the states of the system. If a removal event occurs then increment r by one and update the states. Finally increment j by one and continue until $r = N_F + 1$. Pseudo-code and importance contributions are given in Algorithm C.6 and Appendix C.3.3.

4.1.7. Gamma infectious period, with shape ρ and scale τ

In this case we set $j = 1$, $r = 1$, $C_I^{(j)} = 1$ and $t_I^{(1)} = 0$, and sample the first removal time $t_R^{(r)} \sim G(\rho, \tau)$. If an infection event occurs [with probability $p^{(j)}$, given by (29)], then sample the next infection time, $t_I^{(j+1)}$, from a truncated distribution in the period $\left[t_I^{(j)}, t_R^{(r)}\right)$, before simulating a new removal time $t^* = t_I^{(j+1)} + t'$ where $t' \sim G(\rho, \tau)$. An added complexity is that t^* needs to be added to the vector $\boldsymbol{t}_R$, and this vector sorted into ascending order (resetting $t_R^{(r)}$ if necessary), before updating the states. If a removal event occurs then increment r by one and update the states. Finally, increment j by one and continue until $r = N_F + 1$. Pseudo-code and importance contributions are given in Algorithm C.5 and Appendix C.3.2.

4.1.8. Exponential infectious period

This can be done in exactly the same manner as in Section 4.1.7 by setting the shape parameter $\rho = 1$. However, it is also possible to avoid having to simulate event times at all. In this case set $j = 1$, $C_I^{(j)} = 1$ and $C_R^{(j)} = 0$, and sample whether an infection event occurs before the next removal with probability

$$p^{(j)} = \begin{cases} 0 & \text{if } C_I^{(j)} = N_F, \\ \frac{\lambda_{SI}^{(j)}}{\lambda_{SI}^{(j)} + \lambda_{IR}^{(j)}} & \text{if } \left(C_I^{(j)} > C_R^{(j)} + 1\right) \cap \left(C_I^{(j)} < N_F\right), \text{ and} \\ 1 & \text{otherwise,} \end{cases} \tag{30}$$

where $\lambda_{SI}^{(j)} = \beta S^{(j)} I^{(j)} / N_{\text{pop}}$ and $\lambda_{IR}^{(j)} = \gamma I^{(j)}$. States are updated as before and the algorithm continued until $C_I^{(j)} = N_F$. Pseudo-code and importance contributions are given in Algorithm C.4 and Appendix C.3.1.

4.1.9. Results for final size data

We used 10,000 iterations burn-in with a further 90,000 updates. A $G(0.0001, 100^2)$ prior distribution was used for β (i.e. mean=1 and variance=100^2). The GIMH model with the exponential infectious period took ≈ 13 minutes to run 120,000 iterations, compared to ≈ 12 minutes for the fixed and ≈ 13 minutes for the gamma infectious periods. It can be seen from Table 3 that our estimates are consistent with those of Demiris and O'Neill (2006), though our posterior means are slightly lower. A nice property of the simulation algorithm employed here is that the complicated removal process cancels out in the importance ratio (see Appendix A).

5. Discussion

A significant challenge for inference in epidemic systems is dealing with missing and censored data. In order to generate a likelihood it is typically necessary to infer the missing information as part of the fitting process. This problem becomes more challenging as the size and complexity of the system increase. The Bayesian framework offers a natural environment in which to attempt to

Table 3: Posterior means and standard deviations for R_0 for an $\mathcal{SIR}$ model fitted to the Abakaliki smallpox final size data obtained using GIMH and MCWM, compared to estimates obtained from other methods.

Method	**Infectious period**	$\boldsymbol{R_0}$
GIMH	Fixed	1.14 (0.21)
	Gamma	1.15 (0.26)
	Exponential	1.17 (0.31)
MCWM	Fixed	1.14 (0.21)
	Gamma	1.16 (0.27)
	Exponential	1.18 (0.32)
Demiris and O'Neill (2006)	Fixed	1.18 (0.21)
	Gamma	1.22 (0.27)
	Exponential	1.26 (0.34)
Becker (1989)		1.10
Rida (1991)		1.11

tackle these issues, since missing data can simply be included as extra parameters in the model. As such, the method of DA-/RJ-MCMC provides the current gold-standard fitting mechanism for epidemic systems, allowing the additional uncertainty due to the missing data to be implicitly captured in the marginal posteriors for the parameters. It can also be used to facilitate the evaluation of infeasible likelihood functions via the introduction of latent variables.

Nonetheless, for complex systems with large amounts of missing data, the complexity and computational overheads of DA-MCMC algorithms can be prohibitive. A major challenge is designing efficient proposal distributions for the missing data and parameters, such that the acceptance rate of the chain is reasonable whilst allowing good mixing of the chain and controlling for excessive autocorrelation. Here we employ pseudo-marginal methodology, using importance sampling to generate an MC estimate of the likelihood that can be used in place of the true value in MCMC routines. These methods have various useful properties: firstly, they update the parameters and all of the missing data at the same time. If implemented successfully this allows the chain to move efficiently around the parameter space. Of course there are various parallels between GIMH and DA-MCMC, since the latter using independence sampling is equivalent to GIMH using a single simulation to generate the importance estimate

(since the augmented data need not be stored—see Beaumont, 2003). However, this would typically lead to low acceptance rates, and hence in DA-MCMC some form of conditional update scheme is usually used instead. Pseudo-marginal algorithms alleviate this problem by using multiple repeated simulations to produce the importance estimate of the likelihood—essentially reducing the MC error and potentially improving the efficiency of the chain. In addition, GIMH will produce the *exact* posterior for the parameters in probability, despite an approximation to the likelihood being used.

Developing simulation algorithms that have a high probability of matching the observed data is key if pseudo-marginal routines are to be implemented successfully for epidemic systems. We provide various examples of how this can be done for a range of model structures and data types. By constraining the simulations based on the observed data we use the model to define efficient proposal distributions for the unobserved events, improving both the acceptance rate and mixing of the chain. The algorithms we propose here match the simulations to the data exactly. One useful extension, not discussed thus far, is that in many cases there may also be a stochastic observation process above the epidemic process. We anticipate that this could be included simply by requiring that the simulated time-series counts are *equal to or greater than* the observed data, before adjusting the likelihood calculation based on the probability distribution for the observation process (e.g. binomial).

By producing *independent* simulations of the unobserved data, some of the autocorrelation and memory overheads—due to storing and simulating conditional on previous values of the augmented data—can be reduced. One deficiency of the GIMH method is that since the MC estimates are re-used at each iteration of the chain, if an uncharacteristically large estimate is produced at one iteration (i.e. from the upper tail of the sampling distribution of the importance estimate), then sometimes the chain can become stuck. A simulation study (results not shown) suggests that this may preclude the use of GIMH for systems where the variance of the importance sampling distributions for different parameters are large, unless a large enough number of repeats can be

generated. Of key importance here is to generate simulations from a model that match as closely as possible to the real model, so that elements of the likelihood/importance estimate ratio cancel. Nonetheless, it is worth noting that the MCWM algorithm overcomes these problems through re-sampling at each step, albeit at the cost of increased simulations and some potential bias in the posterior, although these routines are natural candidates for parallelization, which may go some way towards alleviating the requirement for more simulations. We have illustrated the efficacy of these methods on a well-studied data set from an outbreak of smallpox. These routines are very flexible, and we have shown how simple adaptations to the simulation algorithm for fitting to removal data can allow the dimensionality of the system to be changed, and as such the pseudo-marginal routines can be used without requiring reversible-jump methodology. Also, further small changes in the simulation algorithm allows the model to be fitted to final size data.

To conclude, pseudo-marginal methods provide an exciting variation on traditional DA approaches to inference. Whilst the choice of method will vary according to the specific application, it is clear that the methods are very flexible, and show some advantages over traditional DA-/RJ-MCMC in terms of exploring the parameter space for the sorts of systems described here, albeit potentially at the cost of the outcome being approximate in the case of MCWM. Future work will focus on extending these approaches to more complex systems.

Acknowledgements

T. J. M. was in part supported by Department for the Environment, Food and Rural Affairs/Higher Education Funding Council of England, grant number VT0105 and BBSRC grant [BB/I012192/1]. J. V. R was in part supported by Australian Research Council's Discovery Projects funding scheme [project number DP110102893]. R. D. was in part supported by Natural Sciences and Engineering Research Council (NSERC) of Canada's Discovery Grants Program. A. R. C. was in part supported by National Medical Research Council [NMRC/HINIR/005/2009] and NUS Initiative to Improve Health in Asia. The

authors would like to thank Andrew Conlan and Theo Kypraios for useful discussions.

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010), "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society, Series B (Methodological)*, 72, 269–342.

Andrieu, C. and Roberts, G. O. (2009), "The pseudo-marginal approach for efficient Monte Carlo simulation," *The Annals of Statistics*, 37, 697–725.

Bailey, N. T. (1975), *The Mathematical Theory of Infectious Diseases*, Charles Griffin and Company Ltd., London and High Wycombe, 2nd ed.

Ball, F. (1986), "A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models," *Advances in Applied Probability*, 18, 289–310.

Beaumont, M., Zhang, W., and Balding, D. (2002), "Approximate Bayesian Computation in population genetics," *Genetics*, 162, 2025–2035.

Beaumont, M. A. (2003), "Estimation of population growth and decline in genetically monitored populations," *Genetics*, 164, 1139–1160.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009), "Adaptive approximate Bayesian computation," *Biometrika*, 96, 983–990.

Becker, N. G. (1983), "Analysis of data from a single epidemic," *Australian Journal of Statistics*, 25, 191–197.

— (1989), *Analysis of Infectious Disease Data*, Chapman and Hall, CRC.

Berthier, P., Beaumont, M. A., Cornuet, J.-M., and Luikart, G. (2002), "Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach," *Genetics*, 160, 741–751.

Blum, M. G. B. and Tran, V. C. (2010), "HIV with contact-tracing: a case study in Approximate Bayesian Computation," *Biostatistics*, 11, 644–660.

Boys, R. J. and Giles, P. R. (2007), "Bayesian inference for stochastic epidemic models with time-inhomogeneous removal rates," *Journal of Mathematical Biology*, 55, 223–247.

Cappé, O., Guilin, A., Marin, J.-M., and Robert, C. P. (2004), "Population Monte Carlo," *Journal of Computational and Graphical Statistics*, 13 (4), 907–929.

Cauchemez, S., Valleron, A.-J., Boëlle, P.-Y., Flahault, A., and Ferguson, N. M. (2008), "Estimating the impact of school closure on influenza transmission from Sentinel data," *Nature*, 452, 750–755.

Celeux, G., Marin, J.-M., and Robert, C. P. (2006), "Iterated importance sampling in missing data problems," *Computational Statistics and Data Analysis*, 50, 3386–3404.

Chis Ster, I., Singh, B. K., and Ferguson, N. M. (2009), "Epidemiological inference for partially observed epidemics: The example of the 2001 foot and mouth disease epidemic in Great Britain," *Epidemics*, 1, 21–34.

Cook, A., Otten, W., Marion, G., Gibson, G., and Gilligan, C. (2007), "Estimation of multiple transmission rates for epidemics in heterogeneous populations," *Proceedings of the National Academy of Sciences USA*, 104, 20392–20397.

Deardon, R., Brooks, S. P., Grenfell, B. T., Keeling, M. J., Tildesley, M. J., Savill, N. J., Shaw, D. J., and Woolhouse, M. E. (2010), "Inference for individual level models of infectious diseases in large populations," *Statistica Sinica*, 20, 239–261.

Demiris, N. and O'Neill, P. D. (2005a), "Bayesian inference for epidemics with two levels of mixing," *Scandinavian Journal of Statistics*, 32, 265–280.

— (2005b), "Bayesian inference for stochastic multitype epidemics in structured populations via random graphs," *Journal of the Royal Statistical Society. Series B (Methodological)*, 67, 731–745.

— (2006), "Computation of final outcome probabilities for the generalised stochastic epidemic," *Statistics and Computing*, 16, 309–317.

Diggle, P. J. and Gratton, R. J. (1984), "Monte Carlo methods of inference for implicit statistical models (with discussion)," *Journal of the Royal Statistical Society, Series B (Methodological)*, 46, 193–227.

Doucet, A., Freitas, N. D., and Gordon, N. (eds.) (2001), *Sequential Monte Carlo Methods in Practice*, Springer.

Eichner, M. and Dietz, K. (2003), "Transmission potential of smallpox: Estimates based on detailed data from an outbreak," *American Journal of Epidemiology*, 158, 110–117.

Erhardt, R. J. and Smith, R. L. (2012), "Approximate Bayesian computing for spatial extremes," *Computational Statistics and Data Analysis*, 56 (6), 1468–1481.

Fearnhead, P. and Meligkotsidou, L. (2004), "Exact filtering for partially-observed continuous time models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 66, 771–789.

Fearnhead, P. and Prangle, D. (2012), "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation," *Journal of the Royal Statistical Society. Series B (Methodological)*, 74, 419–474.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman and Hall/CRC, 2nd ed.

Gibson, G. J. and Renshaw, E. (1998), "Estimating parameters in stochastic compartmental models using Markov chain methods," *IMA Journal of Mathematics Applied in Medicine and Biology*, 15, 19–40.

— (2001), "Likelihood estimation for stochastic compartmental models using Markov chain methods," *Statistics and Computing*, 11, 347–358.

Gilks, W., Richardson, S., and Spiegelhalter, D. (eds.) (1996), *Markov Chain Monte Carlo In Practice*, Chapman and Hall.

Gillespie, D. T. (1977), "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, 81, 2340–2361.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Haario, H., Saksman, E., and Tamminen, J. (2001), "An adaptive Metropolis algorithm," *Bernoulli*, 7, 223–242.

Hastings, W. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Höhle, M. and Feldmann, U. (2007), "RLadyBug—An R package for stochastic epidemic models," *Computational Statistics and Data Analysis*, 52 (2), 680–686.

Ionides, E., Bretó, C., and King, A. (2006), "Inference for nonlinear dynamical systems," *Proceedings of the National Academy of Sciences USA*, 103, 18438–18443.

Jewell, C. P., Kypraios, T., Neal, P., and Roberts, G. O. (2009), "Bayesian analysis for emerging infectious diseases," *Bayesian Analysis*, 4, 465–496.

Keeling, M. J. and Rohani, P. (2008), *Modeling Infectious Diseases in Humans and Animals*, Princeton University Press.

Keeling, M. J. and Ross, J. V. (2008), "On methods for studying stochastic disease dynamics," *Journal of the Royal Society Interface*, 5, 171–181.

Kypraios, T. (2007), "Efficient Bayesian inference for partially observed stochastic epidemics and a new class of semi-parametric time series models," PhD Thesis, Lancaster University.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003), "Markov chain Monte Carlo without likelihoods," *Proceedings of the National Academy of Sciences USA*, 100, 15324–15328.

McKinley, T. J., Cook, A. R., and Deardon, R. (2009), "Inference in epidemic models without likelihoods," *The International Journal of Biostatistics*, 5.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), "Equations of state calculations by fast computing machine," *Journal of Chemical Physics*, 21, 1087–1091.

Neal, P. (2010), "Efficient likelihood-free Bayesian Computation for household epidemics," *Statistics and Computing*.

O'Neill, P., Balding, D., Becker, N., Eerola, M., and Mollison, D. (2000), "Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods," *Applied Statistics*, 49, 517–542.

O'Neill, P. D. and Becker, N. G. (2001), "Inference for an epidemic when susceptibility varies," *Biostatistics*, 2, 99–108.

O'Neill, P. D. and Roberts, G. O. (1999), "Bayesian inference for partially observed stochastic epidemics," *Journal of the Royal Statistical Society. Series A (General)*, 162, 121–129.

O'Ryan, C., Harley, E. H., Bruford, M. W., Beaumont, M., Wayne, R. K., and Cherry, M. I. (1998), "Microsatellite analysis of genetic diversity in fragmented South African buffalo populations," *Animal Conservation*, 1, 85–94.

Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (1998), "Non-centered parameterizations for hierarchical models and data augmentation," in J. M. Bernardo and M. J. Bayarri and J. O. Berger and A. P. Dawid and D. Heckerman and A. F. M. Smith and M. West (Eds.), Bayesian Statistics 7, Oxford University Press, 307–326.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2010), *coda: Output analysis and diagnostics for MCMC*, r package version 0.14-2.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rida, W. N. (1991), "Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model," *Journal of the Royal Statistical Society. Series B (Methodological)*, 53, 269–283.

Roberts, G. O. and Rosenthal, J. S. (2009), "Examples of adaptive MCMC," *Journal of Computational and Graphical Statistics*, 18, 349–367.

Ross, J., Taimre, T., and Pollett, P. (2006), "On parameter estimation in population models," *Theoretical Population Biology*, 70, 498–510.

Sisson, S., Fan, Y., and Tanaka, M. M. (2007), "Sequential Monte Carlo without likelihoods," *Proceedings of the National Academy of Sciences USA*, 104, 1760–1765.

Tanaka, M. M., Francis, A. R., Luciani, F., and Sisson, S. (2006), "Using Approximate Bayesian Computation to estimate tuberculosis transmission parameters from genotype data," *Genetics*, 173, 1511–1520.

Tavaré, S., Balding, D. J., Griffiths, R., and Donnelly, P. (1997), "Inferring coalescence times from DNA sequence data," *Genetics*, 145, 505–518.

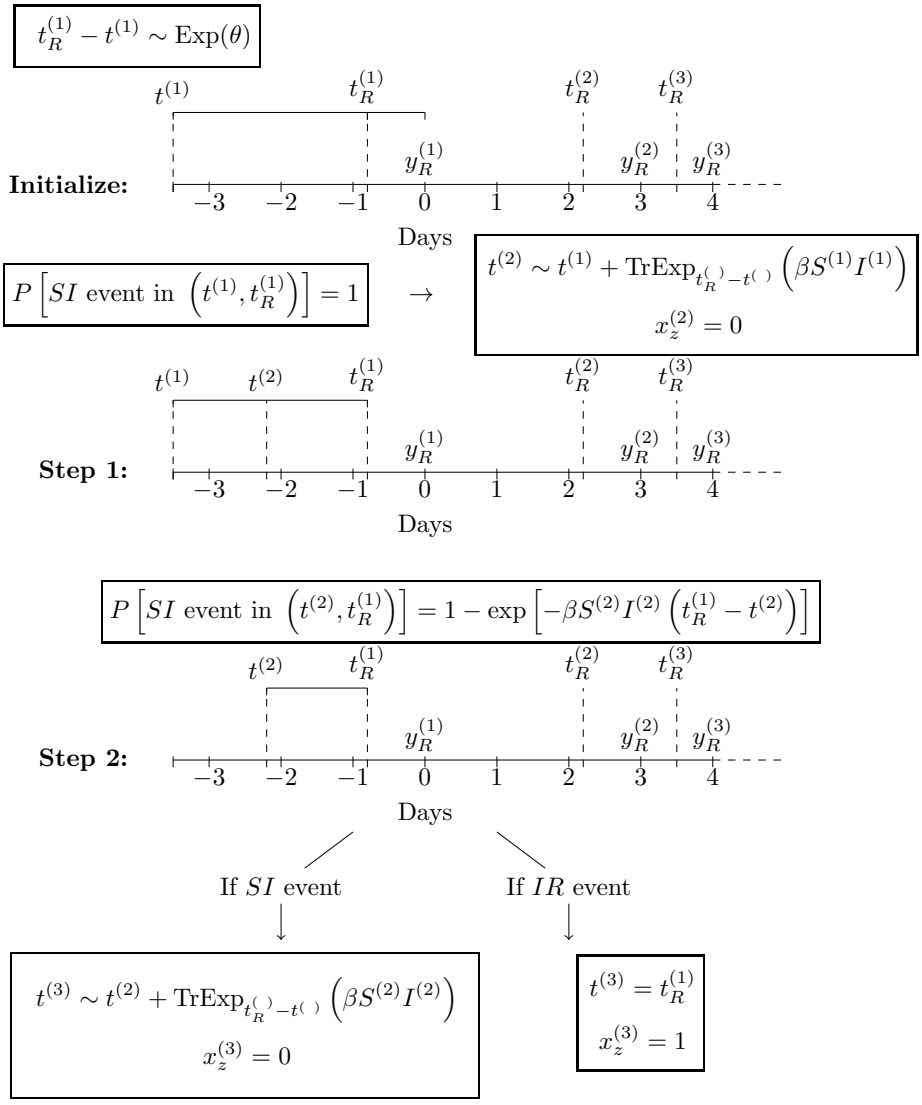
Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Strumpf, M. P. (2009), "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," *Journal of the Royal Society Interface*, 6, 187–202.

Weirman, J. C. and Marchette, D. J. (2004), "Modeling computer virus prevalence with a susceptible-infected-susceptible model with reintroduction," *Computational Statistics and Data Analysis*, 45 (1), 3–23.

Wilkinson, R. D. (2010), "Approximate Bayesian computation (ABC) gives exact results under the assumption of model error," *in submission*.

Wong, H., Shao, Q., and Ip, W. (2013), "Modeling respiratory illnesses with change point: A lesson from the SARS epidemic in Hong Kong," *Computational Statistics and Data Analysis*, 57 (1), 589–599.

Yang, Y., Longini Jr., I. M., and Halloran, E. (2007), "A data-augmentation method for infectious disease incidence data from close contact groups," *Computational Statistics and Data Analysis*, 51 (12), 6582–6595.



Step 3: continue ...

Figure 1: Schematic of constrained simulation algorithm for discrete removal time data. The y_R values are the *observed* discrete removal times, and the t_R values the *simulated* continuous removal times. The t values in this case correspond to the simulated infection times. In the initialisation step all removal times, $t_R^{(j)}$, $j = 1, \ldots, N_F$, are simulated from sets of uniform order statistics conditional on the observed removal counts $\boldsymbol{N_R}$. Then an initial infection time is simulated conditional $t_R^{(1)}$. In Step 1 we force an infection event to occur in $\left(t^{(1)}, t_R^{(1)}\right)$ with probability 1, in order to ensure the epidemic does not die out. This time of this event, $t^{(2)}$, is simulated from a truncated exponential distribution. In Step 2 we simulate whether a further infection event occurs in $\left(t^{(2)}, t_R^{(1)}\right)$, which then determines how the simulation progresses at further stages.