Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 42: 378 –396, 2015
doi: 10.1111/sjos.12111

Forward Simulation Markov Chain Monte Carlo with Applications to Stochastic Epidemic Models

PETER NEAL and CHIEN LIN TERRY HUANG

Department of Mathematics and Statistics, Lancaster University

ABSTRACT. For many stochastic models, it is difficult to make inference about the model parameters because it is impossible to write down a tractable likelihood given the observed data. A common solution is data augmentation in a Markov chain Monte Carlo (MCMC) framework. However, there are statistical problems where this approach has proved infeasible but where simulation from the model is straightforward leading to the popularity of the approximate Bayesian computation algorithm. We introduce a forward simulation MCMC (fsMCMC) algorithm, which is primarily based upon simulation from the model. The fsMCMC algorithm formulates the simulation of the process explicitly as a data augmentation problem. By exploiting non-centred parameterizations, an efficient MCMC updating schema for the parameters and augmented data is introduced, whilst maintaining straightforward simulation from the model. The fsMCMC algorithm is successfully applied to two distinct epidemic models including a birth–death–mutation model that has only previously been analysed using approximate Bayesian computation methods.

Key words: approximate Bayesian computation, birth–death–mutation model, importance sampling, Markov chain Monte Carlo, non-centred parameterization, SIR and SIS epidemic models

1. Introduction

Often when studying a stochastic process, the observed data $\mathbf{x}^*$ are insufficient for straightforward estimation of the parameters θ underlying the process. In particular, for likelihood-based inference, frequentist or Bayesian, it is often impossible to calculate the likelihood, $L(\theta; \mathbf{x}^*) = \pi(\mathbf{x}^*|\theta)$. A common solution underpinning both the expectation–maximization and Markov chain Monte Carlo (MCMC) algorithms is data augmentation, in that, the data are augmented by additional information, $\mathbf{y}$, such that $L(\theta; \mathbf{x}^*, \mathbf{y}) = \pi(\mathbf{x}^*, \mathbf{y}|\theta)$ is tractable. Then estimation of θ can proceed by iterating between the following two steps with the details being algorithm dependent.

1. Update θ given $\mathbf{x}^*$ and $\mathbf{y}$.
2. Update $\mathbf{y}$ given $\mathbf{x}^*$ and θ.

There are many situations where no straightforward augmented data $\mathbf{y}$ exist. In such circumstances, an alternative approach is required, and in the case where simulation from the model is straightforward, approximate Bayesian computation (ABC) (Tavaré *et al.* (1997), Beaumont *et al.* (2002)) gives a useful alternative for obtaining samples from the (approximate) posterior distribution of the parameters θ given the observed data $\mathbf{x}^*$. The ABC algorithm naturally follows from an exact Bayesian computation (EBC) algorithm (see, for example, White *et al.* (2014)), which, whilst obtaining samples from the posterior distribution using simulation, is often far too inefficient for practical purposes. The ABC and EBC algorithms in their simplest forms are rejection algorithms as follows.

ABC/EBC algorithms
For $i = 1, 2, \ldots, N$.

1. Sample θ_i from $\pi(\cdot)$, the prior distribution on θ.
2. Simulate a realization $\mathbf{x}(\theta_i)$ from the model with parameters θ_i.
3. • *ABC*: Choose summary statistics $T(\mathbf{x})$, a distance metric $\rho(\cdot, \cdot)$ and a precision $\epsilon \geq 0$. If $\rho(T(\mathbf{x}(\theta_i)), T(\mathbf{x}^*)) \leq \epsilon$, accept the simulation and set $\chi_i = 1$. Otherwise reject the simulation and set $\chi_i = 0$.

   • *EBC*: If $\mathbf{x}(\theta_i) = \mathbf{x}^*$ accept the simulation and set $\chi_i = 1$. Otherwise reject the simulation and set $\chi_i = 0$.
4. Store $(\theta_1, \chi_1), (\theta_2, \chi_2), \ldots, (\theta_N, \chi_N)$.

For the EBC algorithm, the sample $\{\theta_i; \chi_i = 1\}$ is an independent and identically distributed sample from $\pi(\theta|\mathbf{x}^*)$. That is, the EBC algorithm is a data augmentation algorithm that provides an unbiased 0/1 estimate of the likelihood. However, the probability that $\mathbf{x}(\theta_i) = \mathbf{x}^*$ is often negligible or even 0 in the case of continuous data, hence, the use of the ABC algorithm. The choice of $T(\cdot)$, $\rho(\cdot, \cdot)$ and $\epsilon \geq 0$ is the source of much discussion, see, for example, Fearnhead & Prangle (2012) for an overview. There are many improvements that can be made to the aforementioned ABC algorithm, for example, in the choice of θ_i in step 1 using either MCMC–ABC (Marjoram *et al.* (2003)) or sequential Monte Carlo (SMC)–ABC (Sisson *et al.* (2007)) and local-linear regression in step 3 when using $\rho(T(\mathbf{x}(\theta_i)), T(\mathbf{x}^*)) \leq \epsilon$ (Beaumont *et al.* (2002)). However, these improvements do not address the fundamental problem with the ABC algorithm that it produces a sample from an approximate posterior distribution where the level of the approximation is in practice impossible to quantify. Therefore, wherever it is practical, the MCMC algorithm is preferable to the ABC algorithm. This provides the statistical motivation for this paper, to explore what makes the ABC algorithm successful and a useful statistical tool, namely straightforward simulation, and seek to incorporate this into an MCMC framework to produce an effective MCMC algorithm.

The ABC and EBC algorithms can be viewed as data augmentation algorithms, in that, in simulating from a stochastic process, we are generating data $\mathbf{y}$ of how the process evolves. Typically, the simulated data that we are interested in, $\mathbf{x}$, corresponding to the observed data, $\mathbf{x}^*$, are a subset of $\mathbf{y}$. For example, for an epidemic model, $\mathbf{x}$ could denote which individuals are infectious at time $T(> 0)$, say, whilst $\mathbf{y}$ consists of the progression of the epidemic up to time T. The augmented data $\mathbf{y}$ are not chosen to construct a tractable likelihood in the usual manner but to give a realization of the stochastic process. However, for the EBC algorithm, $\mathbb{E}_{\mathbf{y}}[\chi(\theta; \mathbf{y})|\theta] = \pi(\mathbf{x}^*|\theta)$, where in writing $\chi(\theta; \mathbf{y})$, we are explicit of χ's dependence on θ and $\mathbf{y}$. This provides the starting point for using simulation in an MCMC context. Specifically, we focus on two features of the simulation process. Firstly, on how to improve upon *naive* simulation from the model by using importance sampling to direct the simulation towards the observed data. The key consequence of this is replacing $\chi(\theta; \mathbf{y})$ by a probability $P(\theta; \mathbf{y})$ with $P(\theta; \mathbf{y})$ being an unbiased estimator of $L(\theta; \mathbf{x}^*) = \pi(\mathbf{x}^*|\theta)$ with a smaller variance than $\chi(\theta; \mathbf{y})$. Secondly, we seek to construct non-centred parameterizations (Papaspoliopoulos *et al.* (2003)) for the simulation process. The details of the non-centred parameterization are problem specific and are discussed in detail in Sections 3 and 4 with the key requirement being that θ and $\mathbf{y}$ are *a priori* independent. The non-centred parameterization enables us to iterate between updating θ and $\mathbf{y}$ with implementation of the simulation process straightforward given the updated parameters and augmented data. This gives us an efficient way to make small changes to the underlying random variables (seeds) $\mathbf{Y}$ used in the simulation process. Specifically, rather

than at each iteration simulating the stochastic process with a completely new set of random variables $\mathbf{Y}$ which is the case, as far as we are aware, for all ABC algorithms in the literature, we propose a new set of random variables $\mathbf{y}'$, which can depend upon the current $\mathbf{y}$. The idea of using the same random variables for different parameter values is the basis of the coupled ABC algorithm in Neal (2012), although in that paper, each iteration uses a completely new set of random variables $\mathbf{y}$. Also Andrieu *et al.* ((Andrieu *et al.*, 2012)) made suggestions in terms of how to choose the underlying random variables $\mathbf{Y}$ in an ABC context, noting that we are not restricted to making independent and identically distributed draws from $\mathbf{Y}$.

The forward simulation MCMC (fsMCMC) algorithm introduced in Section 2 is successfully applied to three disparate epidemic examples, two of which, final size data for a homogeneously mixing SIR epidemic model (Section 3) and the initial (branching) stages of a mutating SIR epidemic model observed at a single point in time (Section 4) appear in the main text. The third example, a single snapshot (from the quasi-stationary distribution) of a multiple strain SIS epidemic, is presented in the Supplementary information. Thus, all the data sets considered are cross-sectional for which large-scale data augmentation is required to obtain a tractable likelihood. The latter two models are Markov processes for which the Gillespie algorithm (Gillespie (1976)) can be used to simulate from. As noted in Neal (2012), Section 5, it is straightforward to construct non-centred parameterizations for the Gillespie algorithm, and hence, implement the forward simulation MCMC algorithm. The methodology is by no means restricted to epidemic models and should be applicable to a wide range of population-based stochastic models, especially Markov models.

The remainder of the paper is structured as follows. In Section 2, we give a brief overview to the fsMCMC algorithm. We highlight two scenarios for which the algorithms can be useful, and these are studied in detail in Sections 3 and 4. Also in Section 2, we outline an importance sampling version of the EBC algorithm (isEBC), which can in some circumstances offer a simple, efficient alternative to the fsMCMC algorithm. Throughout, we seek to keep the simulation process as straightforward as possible both in terms of implementation and computational burden because the fsMCMC algorithm offers most benefits where the time taken per simulation in the fsMCMC algorithm is similar to that required for *naive* simulation. In Section 3, the fsMCMC and isEBC algorithms are applied to estimating the infection rate λ of a homogeneously mixing epidemic model from final size data. This forms a useful test of the methodology as estimation of λ is available via alternative methods, see Demiris & O' Neill (2006) and Neal (2012). We demonstrate that both the fsMCMC and isEBC algorithms substantially outperform the EBC algorithm. In Section 4 and the Supporting information, we consider partially observed cross-sectional epidemic data. This is where the methodology of the paper is most useful, and we illustrate the methodology with a birth–death–mutation (BDM) model (Section 4), and a slightly different example, a multiple strain SIS epidemic model, is considered in the Supporting information. MCMC has not previously been applied successfully to either data set, with the BDM data set being extensively analysed using ABC, see Tanaka *et al.* (2006), Sisson *et al.* (2007), Fearnhead & Prangle (2012) and Del Moral *et al.* (2012). For the BDM model, the fsMCMC algorithm is shown to perform well with a similar computational burden to the ABC algorithms used in Tanaka *et al.* (2006) and Fearnhead & Prangle (2012), requiring more time per iteration but fewer iterations. Furthermore, we are able to obtain samples from the posterior distribution of the parameters of the BDM model rather than an approximate posterior distribution. The multiple strain SIS epidemic model considered in the Supporting information, despite its biological limitations, offers further interesting insight into the implementation of the isEBC and fsMCMC algorithms for cross-sectional epidemic data. Finally, in Section 5, we make concluding remarks concerning possible directions of future research for the fsMCMC algorithm, in particular in optimizing the performance of the fsMCMC algorithm.

2. Forward simulation MCMC

In this section, we give a brief overview to the fsMCMC algorithm and the importance sampling exact Bayesian computation (isEBC) algorithm. The details for implementing each algorithm efficiently are problem specific and discussed in Sections 3 and 4.

Let $\mathbf{x}^*$ denote the observed data that we assume arise from a realization of a stochastic parametric model, $\mathcal{M}$. Typically, $\mathbf{x}^*$ will denote a partial observation of a stochastic process $\mathcal{G}$, and we assume this to be the case throughout this paper. Let θ denote the parameters of $\mathcal{M}$ with $\pi(\theta)$ and $\pi(\theta|\mathbf{x}^*)$ denoting the prior and posterior probability density functions of θ, respectively. Throughout, we assume that the likelihood $\pi(\mathbf{x}^*|\theta)$ is not tractable, and hence, that it is necessary to use a data augmentation technique such as ABC or MCMC. Let $\mathbf{X}(\theta)$ denote a random vector that generates a realization of the data from $\mathcal{M}$ with parameters θ. In Neal (2012), it is assumed that there exists a deterministic function $h(\cdot;\cdot)$ and a random vector $\mathbf{Y}$ with probability density function $\pi_{\mathbf{Y}}(\mathbf{y})$, *a priori* independent of θ such that a realization of $\mathbf{X}(\theta) = h(\theta;\mathbf{y})$. It is straightforward to construct an EBC algorithm by simulating θ and $\mathbf{y}$ from $\pi_\theta(\cdot)$ and $\pi_{\mathbf{Y}}(\cdot)$, respectively, and setting $\chi(\theta;\mathbf{y}) = 1$ if $\mathbf{X}(\theta) = \mathbf{x}^*$ (the simulated data agree with the observed data) and $\chi(\theta;\mathbf{y}) = 0$ otherwise. We focus on discrete data; otherwise, $\chi(\theta;\mathbf{y})$ is almost surely 0, which are often the case for epidemic models and many other population models, where typically data consist of the number of individuals in different categories, such as, how many individuals are infectious at a given point in time.

In principle it is straightforward to adapt the aforementioned EBC algorithm to give an MCMC algorithm as follows. Given $(\theta,\mathbf{y})$, we propose updates $(\theta',\mathbf{y}')$ according to some transition kernel $q(\theta',\mathbf{y}'|\theta,\mathbf{y})$ with the proposed move accepted with probability

$$\min\left\{1, \frac{\chi(\theta';\mathbf{y}')q(\theta,\mathbf{y}|\theta',\mathbf{y}')}{\chi(\theta;\mathbf{y})q(\theta',\mathbf{y}'|\theta,\mathbf{y})}\right\}. \tag{2.1}$$

Note that $\chi(\theta;\mathbf{y})$ is a 0/1 indicator, and hence is discontinuous over the joint $(\Theta,\mathbf{Y})$ space. Consequently, great care needs to be taken in the choice of $q(\cdot|\cdot)$; otherwise, the MCMC algorithm will have a prohibitively low acceptance rate. For example, it has been shown in (Neal *et al.*, 2012) that random walk Metropolis algorithms perform poorly for discontinuous target densities. However, (2.1) forms a central basis for this paper by replacing $\chi(\theta;\mathbf{y})$ by a probability (likelihood) $P(\theta;\mathbf{y})$ such that $\mathbb{E}_{\mathbf{Y}}[\chi(\theta;\mathbf{Y})] = \mathbb{E}_{\mathbf{Y}}[P(\theta;\mathbf{Y})] = \pi(\mathbf{x}^*|\theta)$. (It suffices for the MCMC that $\mathbb{E}_{\mathbf{Y}}[P(\theta;\mathbf{Y})] \propto \pi(\mathbf{x}^*|\theta)$.)

There are two scenarios considered in this paper in which simulation of the stochastic processes can be performed with $\chi(\theta;\mathbf{y})$ being replaced by $P(\theta;\mathbf{y})$. The first is where we can *bias* the simulation process, so that $\mathbf{X}(\theta) = \mathbf{x}^*$, or at least that there is a relatively high probability that $\mathbf{X}(\theta) = \mathbf{x}^*$. We can take account of the biasing of the simulation process by computing the probability of the bias we introduce happening by chance, and this can be viewed as an example of importance sampling. This is illustrated in Section 3 where we ensure that the simulation of the homogeneously mixing epidemic infects the correct number of individuals. The second scenario that is more important between the two from a practical perspective is where the observed data form a partial observation of the stochastic process at a given point in time (or times). In this case, we simulate the stochastic process without biasing (or in the examples in Section 4 and the Supporting information with limited biasing) and then compute the probability of the observed data arising as a partial observation of the stochastic process. For example, in Section 4, it is assumed that detailed data are available for $n(= 473)$ individuals selected uniformly, at random, from an infectious population of $N(= 10000)$ individuals. The construction of $\mathbf{X}(\theta)$ not only is still deterministic given θ and $\mathbf{y}$ but also includes the importance sampling probability weight $P(\theta;\mathbf{y})$, which takes into account any biasing of the simulation

process and/or the sampling process to obtain the observed data. The sampling probability is the probability of obtaining a sample of n individuals with the observed characteristics from a population of N individuals. Note that if $\mathbf{X}(\theta) \neq \mathbf{x}^*$, then $P(\theta; \mathbf{y}) = 0$.

The structure of the fsMCMC algorithm is given in the succeeding text with the details being problem specific. However, in all cases, we found it useful to follow the standard structure of data augmentation MCMC algorithms in alternating between updating the parameters (θ) and augmented data ($\mathbf{y}$).

Forward simulation MCMC (fsMCMC) algorithm

1. Choose initial values for θ and $\mathbf{y}$. Construct $\mathbf{X}(\theta)$ using θ and $\mathbf{y}$ and compute $P(\theta; \mathbf{y})$. If $P(\theta; \mathbf{y}) = 0$, reinitialize with new θ and $\mathbf{y}$ values.
2. For $i = 1, 2, \ldots, N$.
   (a) Propose a new value θ' from $q_\theta(\cdot|\theta)$. Construct $\mathbf{X}(\theta)$ using θ' and $\mathbf{y}$ and compute $P(\theta'; \mathbf{y})$. Accept θ' with probability $1 \wedge \frac{P(\theta'; \mathbf{y}) q_\theta(\theta|\theta')}{P(\theta; \mathbf{y}) q_\theta(\theta'|\theta)}$ and set $\theta_i = \theta'$. Otherwise, reject θ' and set $\theta_i = \theta$.
   (b) Propose a new value $\mathbf{y}'$ from $q_y(\cdot|\mathbf{y})$. Construct $\mathbf{X}(\theta)$ using θ and $\mathbf{y}'$ and compute $P(\theta; \mathbf{y}')$. Accept $\mathbf{y}'$ with probability $1 \wedge \frac{P(\theta; \mathbf{y}') q_y(\mathbf{y}|\mathbf{y}')}{P(\theta; \mathbf{y}) q_y(\mathbf{y}'|\mathbf{y})}$ and set $\mathbf{y} = \mathbf{y}'$. Otherwise, leave $\mathbf{y}$ unchanged.
3. Discard the first B iterations as burn-in and store $\theta_{B+1}, \theta_{B+2}, \ldots, \theta_N$ as a sample from $\pi(\theta|\mathbf{x}^*)$.

Whilst the main motivation for introducing $P(\theta; \mathbf{y})$ is for construction of the fsMCMC algorithm, a useful byproduct is the following isEBC algorithm.

Importance sampling EBC (isEBC) algorithm

1. Sample θ from $\pi(\theta)$, the prior on θ and $\mathbf{y}$ from $\pi_Y(\mathbf{y})$.
2. Construct $\mathbf{X}(\theta)$ using θ and $\mathbf{y}$ and compute $P(\theta; \mathbf{y})$.
3. Store $(\theta, P = P(\theta; \mathbf{y}))$.

Variations on the ABC (EBC) algorithm such as MCMC–ABC, Marjoram *et al.* ((Marjoram *et al.*, 2003)), and grouped independence Metropolis–Hastings (GIMH), Beaumont ((Beaumont, 2003)), can similarly be obtained by allowing θ to be drawn from a proposal distribution depending upon the current value of θ and independent realizations of $\mathbf{Y}$ at each iteration. In the case of the GIMH, this involves repeating the simulation process multiple times for a given θ.

We briefly discuss the pros and cons of the isEBC and fsMCMC algorithms before applying them to the epidemic examples in Sections 3 and 4. The isEBC algorithm is easy to implement, and because it generates fresh simulations at each iteration, there are no questions concerning convergence or mixing of the algorithm. Also because each iteration is independent, it is trivial to parallelize. The first major drawback to the isEBC is that we require a proper prior distribution, and the efficiency of the algorithm is severely reduced by having a diffuse prior. This can to some extent be circumvented by proposing θ values from an alternative distribution to the prior using importance sampling (cf. Fearnhead & Prangle(2012)). A second major drawback to the isEBC algorithm is that estimation of $\mathbb{E}[g(\theta)|\mathbf{x}^*]$ by $\sum_{j=1}^N g(\theta_j) P_j / \sum_{j=1}^N P_j$ can often be dominated by one or a few simulations as $P(\theta; \mathbf{y})$ is usually heavy tailed. This is a consequence of *throwing away* simulations and starting afresh at each iteration. The aforementioned drawbacks can be partially alleviated by using (variants of) the GIMH algorithm ((Beaumont 2003), Andrieu & Roberts ((Andrieu & Roberts, 2009))) instead. However, if $P(\theta; \mathbf{y})$ is heavy tailed,

the GIMH is liable to become stuck (see Lee *et al.*(for suggested improvements to reduce problems with the GIMH becoming stuck), and it is costly, in terms of computer time, to perform multiple simulations for each value of θ. The fsMCMC algorithm seeks to efficiently explore the joint space of $(\Theta, \mathbf{Y})$. As mentioned earlier, in principle, the fsMCMC algorithm can be applied in the case where there is no biasing with $P(\theta; \mathbf{Y})$ being an indicator random variable for whether or not the simulation generates $\mathbf{x}^*$. However, in practice, the conditioning is important in ensuring efficient mixing of the fsMCMC algorithm as MCMC algorithms, especially random walk Metropolis algorithms, often perform poorly for discontinuous likelihood surfaces, see, for example, Neal *et al.* (2012). Another downside of the fsMCMC algorithm are those commonly seen with MCMC algorithms involving large-scale data augmentation with large serial autocorrelation between successive iterations of the algorithm. There is also the question of how to choose $q_\theta(\cdot|\theta)$ and $q_y(\cdot|\mathbf{y})$ for an efficient MCMC algorithm that we address to some extent, but there is a scope for further investigation. Finally, each iteration requires multiple (typically two or three) simulations, which means that more computational time is often required per iteration than for the isEBC algorithm.

3. Conditioned simulation of homogeneously mixing epidemics

The homogeneously mixing SIR epidemic forms a useful benchmark for comparing the isEBC and fsMCMC algorithms with the EBC algorithm for the following reasons. Firstly, simulation of homogeneously mixing epidemics is trivial and the implementation of the isEBC and fsMCMC algorithms straightforward. Secondly, using the multiple precision procedures in Demiris & O'Neill (2006), it is possible to compute $\pi(\mathbf{x}^*|\theta)$ exactly, and hence, using numerical integration to compute $\pi(\theta|\mathbf{x}^*)$ to desired accuracy. This provides a benchmark for estimates obtained using the various algorithms. Thirdly, it is possible to construct a coupled isEBC (cisEBC) algorithm in the spirit of the coupled ABC algorithm introduced in Neal (2012), thus further improving our analysis. All the algorithms are compared fitting the generalized stochastic epidemic model to final epidemic size data. The data used are the Abakiliki smallpox data set (see, Bailey (1975), page 125) with 30 individuals infected out of population of 120 individuals. These final size data have previously been analysed in Demiris & O' Neill (2006), Neal (2012) and McKinley *et al.* (2014) as final size data, and many authors have studied the full temporal data, see, for example O' Neill & Becker (2001), Neal & Roberts (2005) and McKinley *et al.* (2014). The work of McKinley *et al.* (2014) is particularly relevant as they also consider importance sampling for epidemic models with their main focus on temporal data. Our approach is slightly different being based upon the Sellke construction (Sellke (1983)) of the epidemic process that is particularly amenable to a non-centred parameterization. Moreover, McKinley *et al.* (2014) requires a separate algorithm for each different infectious period distribution, whereas our approach has a single algorithm applicable for all infectious period distributions with trivial amendments.

The generalized stochastic epidemic model in a closed population of size n is defined as follows. Suppose that there are a initial infectives and $n - a$ initial susceptibles. Throughout, we assume that $a = 1$, although it is trivial to adapt the arguments to $a > 1$. The infectives have independent and identically distributed infectious periods according to an arbitrary, but specified, non-negative random variable I. Whilst infectious, infectives make infectious contacts at the points of a homogeneous Poisson point process with rate λ with the individual contacted being chosen uniformly at random from the whole population (including the infective). At the end of their infectious period, an individual recovers from the disease and is immune to further infection and subsequently plays no further role in the epidemic process. An infectious contact with a susceptible results in the susceptible individual becoming an infective, whilst an

infectious contact with a non-susceptible has no affect on the recipient of the contact. Thus an individual, i say, with infectious period I_i makes $\text{Po}(\lambda I_i)$ infectious contacts during its infectious period. The probability that an infectious contact at time t is with a susceptible is S_t/n, where S_t denotes the total number of susceptibles at time t.

From a statistical perspective, we are interested in the estimation of λ and the parameters of I under the assumption that I belongs to a specified probability distribution. Because we focus upon final size data, the only information is m, the total number of individuals infected during the course of the epidemic. Therefore, we have only one data point, and it is only possible to make meaningful inference about one parameter, which in our case will be λ. Consequently, we assume that I is known and without loss of generality that $\mathbb{E}[I] = 1$.

Under the assumption that $\mathbb{E}[I] = 1$, λ denotes the basic reproduction number, R_0, of the epidemic model. Consequently, the aforementioned epidemic construction can be applied in situations where the infection rate is non-constant during the course of the infectious period. In particular, suppose that β_t denotes the infection rate t time units after infection and T_i is the infectious period of individual i, then we can set $\lambda I_i = \int_0^{T_i} \beta_t \, dt$, the total amount of infectivity generated by individual i, where I_i now denotes the relative infectiousness of individual i. Then the total number of infectious contacts made by individual i is $\text{Po}(\int_0^{T_i} \beta_t \, dt) = \text{Po}(\lambda I_i)$ as before, and the aforementioned simulation process can be used by drawing the I_i's from the appropriate probability distribution. However, now λ represents the basic reproduction number rather than the infection rate.

In Neal (2012), ABC and coupled ABC algorithms were applied to the generalized stochastic epidemic model. It was found that it was convenient to use the equivalent Sellke construction (Sellke (1983)) of the epidemic process, see Neal (2012) for a full description. The Sellke construction replaces the homogeneous Poisson infectious point process by an infectious threshold T for each individual. Let $T_1, T_2, \ldots, T_n$ be independent and identically distributed according to $T \sim \text{Exp}(1/\text{n})$. An individual i with infectious threshold T_i becomes infected once the total amount of infectious pressure in the population exceeds T_i, and an infective individual j contributes λI_j units of infectious pressure. As observed in Sellke (1983), Section 2, and Neal (2012), Section 3, it is convenient to study the ordered infectious thresholds, $T_{(1)} (= T_1 = 0) < T_{(2)} < \ldots < T_{(n)}$ with the individuals relabeled according to the order of their infectious thresholds. For $i = 1, 2, \ldots, n$, let $(\tilde{T}_i, \tilde{I}_i) = (T_{(i)}, I_{(i)})$, where $\tilde{I}_i \overset{D}{=} I$. Furthermore, letting $L_1, L_2, \ldots, L_{n-1}$ be independent random variables with $L_j \sim \text{Exp}((\text{n} - \text{j})/\text{n})$, we can construct the ordered thresholds by setting $\tilde{T}_1 = 0$ and for $2 \leq i \leq N$, $\tilde{T}_i = \sum_{j=1}^{i-1} L_j$. Thus, given $\lambda > 0$, $L_1, L_2, \ldots, L_n$ and $\tilde{I}_1, \tilde{I}_2, \ldots \tilde{I}_n$, the final size of the epidemic is M, where

$$\begin{aligned} M &= \min \left\{ m; \tilde{T}_{m+1} > \lambda \sum_{j=1}^{m} \tilde{I}_j \right\}, \\ &= \min \left\{ m; \sum_{j=1}^{m} L_j / \sum_{j=1}^{m} \tilde{I}_j > \lambda \right\}. \end{aligned} \tag{3.1}$$

As noted in Neal (2012), Section 3, we can simulate $L_1, L_2, \ldots, L_n$ and $\tilde{I}_1, \tilde{I}_2, \ldots \tilde{I}_n$ and consider epidemics for all values of $\lambda \in \mathbb{R}^+$ simultaneously. This is termed coupled ABC in Neal (2012) and, for example,

$$\lambda \in \left[\max_{1 \leq k \leq m-1} \left\{ \sum_{j=1}^{k} L_j / \sum_{j=1}^{k} \tilde{I}_j \right\}, \sum_{j=1}^{m} L_j / \sum_{j=1}^{m} \tilde{I}_j \right), \tag{3.2}$$

will produce an epidemic with final size m.

Whilst $L_1, L_2, \ldots, L_n$ are already *a priori* independent of λ, it will be convenient to express $L_j = -\{n/(n-j)\}\log(U_j)$ where $U_j \sim U(0,1)$ $(j = 1, 2, \ldots, n)$. From (3.1), for an epidemic of size m, we require that for all $1 \leq k \leq m-1$, $\sum_{j=1}^{k} L_j \leq \lambda \sum_{j=1}^{k} \tilde{I}_j$ and $\sum_{j=1}^{m} L_j > \lambda \sum_{j=1}^{m} \tilde{I}_j$. This is achieved for given $(\lambda, \mathbf{U}, \tilde{\mathbf{I}})$ as follows.

SIR final size simulation

1. Fix $(\lambda, \mathbf{U}, \tilde{\mathbf{I}})$, and set $P_0(\lambda, \mathbf{U}, \tilde{\mathbf{I}}) = 1$. (In the notation of Section 2, $\mathbf{Y} = (\mathbf{U}, \tilde{\mathbf{I}})$.)
2. For $k = 1, 2, \ldots, m-1$, L_k is drawn from $\text{Exp}((n-k)/n)$ conditioned to be less than or equal to $\lambda \sum_{j=1}^{k} \tilde{I}_j - \sum_{j=1}^{k-1} L_j$. Thus, we set, using inversion of the cumulative distribution function (CDF),

$$L_k = -\frac{n}{n-k}\log\left(1 - U_k\left\{1 - \exp\left(-\frac{n-k}{n}\left\{\lambda\sum_{j=1}^{k}\tilde{I}_j - \sum_{j=1}^{k-1} L_j\right\}\right)\right\}\right),$$

   and we set

$$P_k(\lambda, \mathbf{U}, \tilde{\mathbf{I}}) = \left\{1 - \exp\left(-\frac{n-k}{n}\left\{\lambda\sum_{j=1}^{k}\tilde{I}_j - \sum_{j=1}^{k-1} L_j\right\}\right)\right\} \times P_{k-1}(\lambda, \mathbf{U}, \tilde{\mathbf{I}}),$$

   the probability of the imposed conditions up to and including the k^{th} infection occurring by chance.
3. L_m is drawn from $\text{Exp}((n-m)/n)$ conditioned to be greater than $\lambda \sum_{j=1}^{m} \tilde{I}_j - \sum_{j=1}^{m-1} L_j$. Thus, we set, using inversion of the CDF,

$$L_m = \lambda\sum_{j=1}^{m}\tilde{I}_j - \sum_{j=1}^{m-1} L_j - \frac{n}{n-m}\log(1-U_m),$$

   and we set

$$P(\lambda, \mathbf{U}, \tilde{\mathbf{I}}) = \exp\left(-\frac{n-m}{n}\left\{\lambda\sum_{j=1}^{m}\tilde{I}_j - \sum_{j=1}^{m-1} L_j\right\}\right) \times P_{m-1}(\lambda, \mathbf{U}, \tilde{\mathbf{I}}),$$

   which takes into account the probability of the imposed conditions of the epidemic infecting exactly m individuals occurring by chance.

For the isEBC algorithm, we generate new $(\lambda, \mathbf{U}, \tilde{\mathbf{I}})$ at each iteration with λ drawn from $\pi(\cdot)$, the components of $\mathbf{U}$ and $\tilde{\mathbf{I}}$ being independent and identically distributed according to $U(0,1)$ and $\tilde{I}_1$, respectively. We follow Neal (2012) in assuming $U(0,5)$ prior on λ and Demiris & O' Neill (2006) and Neal (2012) in considering $\tilde{I}_1 \equiv 1$, $\tilde{I}_1 \sim \text{Exp}(1)$ and $\tilde{I}_1 \sim \text{Gamma}(2,2)$. For the fsMCMC algorithm, we update λ, $\mathbf{U}$ and $\tilde{\mathbf{I}}$ one at a time sequentially. We propose $\lambda' \sim N(\lambda, 0.3^2)$, and for both $\mathbf{U}'$ and $\tilde{\mathbf{I}}'$, we propose to update eight components drawn independently from $U(0,1)$ or $\tilde{I}_1$ as appropriate. (Note that if $\tilde{I}_1 \equiv 1$, updating of $\tilde{\mathbf{I}}$ is omitted.) That is, we separately update $\mathbf{U}$ and $\tilde{\mathbf{I}}$, so that each iteration of the fsMCMC involves three simulations, updating λ, $\mathbf{U}$ and $\tilde{\mathbf{I}}$ in turn. The proposal variance for λ and the number of components of $\mathbf{U}$ and $\tilde{\mathbf{I}}$ to update are chosen to optimize the performance of the fsMCMC algorithm details given in the succeeding text.

Constructing a cisEBC algorithm in the spirit of the coupled ABC algorithm of Neal (2012) is even easier to implement as the only conditional event is L_m. That is, provided $\max_{1\leq k\leq m-1} \sum_{j=1}^{k} L_j / \sum_{j=1}^{k} \tilde{I}_j \leq \sum_{j=1}^{m} L_j / \sum_{j=1}^{m} \tilde{I}_j$, there exists a range of λ values, given by (3.2), such that $(\lambda, \mathbf{L}, \tilde{\mathbf{I}})$ yields an epidemic of size m.

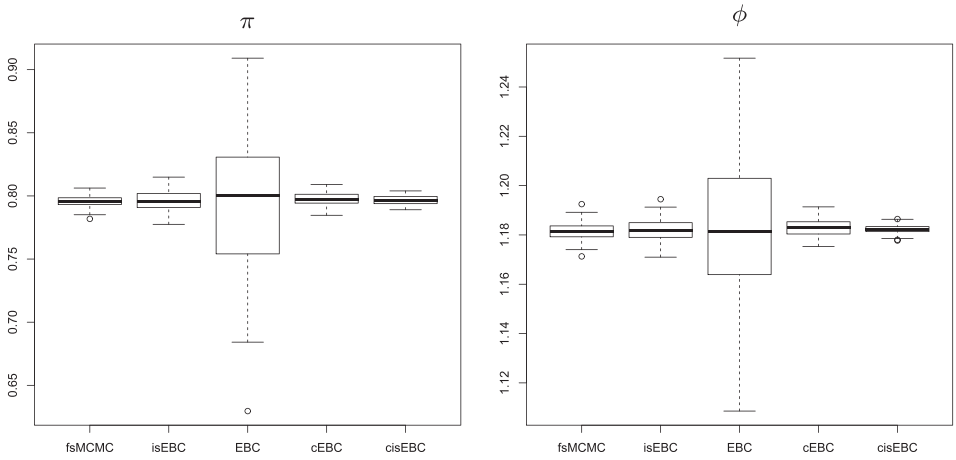

Fig. 1. Boxplots of the estimates of (a) $\pi = \mathbb{P}(\lambda > 1|\mathbf{x}^*)$ and (b) $\phi = \mathbb{E}[\lambda|\mathbf{x}^*]$ based on 100 samples of size 100,000 using the fsMCMC, isEBC, EBC, cEBC and cisEBC algorithms.

cisEBC algorithm

1. Simulate a realization from $(\mathbf{U}, \tilde{\mathbf{I}})$.
2. For $k = 1, 2, \ldots, m-1$, L_k is drawn from $\text{Exp}((n-k)/n)$ with $L_k = -\frac{n}{n-k}\log(1-U_k)$.
3. L_m is drawn from $\text{Exp}((n-m)/n)$ conditioned to be greater than

$$A_m = \max_{1\le k\le m-1}\left\{\sum_{j=1}^{k} L_j / \sum_{j=1}^{k}\tilde{I}_j\right\}\sum_{j=1}^{m}\tilde{I}_j - \sum_{j=1}^{m-1} L_j.$$

   Using inversion of the CDF, we set $L_m = A_m - \frac{n}{n-m}\log(1-U_m)$ and $P(\mathbf{U}, \tilde{\mathbf{I}}) = \exp(-\frac{n-m}{n}A_m)$.
4. Then $[\max_{1\le k\le m-1}\left\{\sum_{j=1}^{k} L_j / \sum_{j=1}^{k}\tilde{I}_j\right\}, \sum_{j=1}^{m} L_j / \sum_{j=1}^{m}\tilde{I}_j)$ is a set of λ values from $\pi(\lambda|\mathbf{x}^*)$ with weight $P(\mathbf{U}, \tilde{\mathbf{I}})$.

In order to compare the performance of the algorithms, we study the estimation of $\phi = \mathbb{E}[\lambda|\mathbf{x}^*]$, the posterior mean of $\lambda(= R_0)$, and $\pi = \mathbb{P}(\lambda > 1|\mathbf{x}^*)$, the posterior probability that the epidemic is supercritical ($R_0 > 1$). We choose ϕ and π as they are quantities of epidemiological interest and are the expectations of a continuous and a discontinuous function of λ, respectively. Using multiple precision arithmetic (Demiris & O' Neill (2006)), $\mathbb{P}(X = 30|\lambda)$ can be computed exactly, with ϕ and π then computed to the desired accuracy using numerical integration for each infectious period distribution. For each algorithm and each infectious period distribution, we generated 100 samples of size $N = 100,000$ and estimated ϕ and π. For the isEBC and EBC algorithms, we use the consistent estimator of $\mathbb{E}[g(\lambda)|\mathbf{x}^*]$ given by $\sum_{j=1}^{N} g(\lambda_j)P_j / \sum_{j=1}^{N} P_j$. For the cisEBC and coupled EBC (cEBC) algorithms, a consistent estimator of $\mathbb{E}[g(\lambda)|\mathbf{x}^*]$ is given by

$$\frac{\frac{1}{N}\sum_{i=1}^{N}\{P_i \int_{a_i^L}^{a_i^H} g(\lambda)\pi(\lambda)\,d\lambda\}}{\frac{1}{N}\sum_{i=1}^{N}\{P_i \int_{a_i^L}^{a_i^H} \pi(\lambda)\,d\lambda\}}, \tag{3.3}$$

where $\mathcal{A}_i = [a_i^L, a_i^H)$ and for the cEBC algorithm, P_i is a $0-1$ indicator. For a $U[0,5]$ prior on λ and $g(x) = x$ or $g(x) = 1_{\{x>1\}}$, (3.3) is easy to compute. In Fig. 1, boxplots

of the estimates of π and ϕ using each of the five algorithms are given for the case $I \equiv 1$. Similar plots were observed for the cases $I \sim \text{Exp}(1)$ and $I \sim \text{Gamma}(2, 2)$, and hence, omitted. The plots show that the standard EBC algorithm performs far worse than the other algorithms with the cisEBC algorithm clearly performing best. The other three algorithms have similar performances, although the fsMCMC algorithm does outperform the isEBC algorithm. In the coupled case, an fsMCMC algorithm, updating a proportion of the components of $\mathbf{U}$ at each iteration, was found to be less efficient than the cisEBC algorithm. This gives an example where the simple isEBC algorithm is preferable, not only in ease of implementation but also in performance, to the fsMCMC algorithm.

4. Cross-sectional epidemic data

We study the second scenario outlined in Section 2, where the epidemic is simulated with limited *biasing* and the observation of the epidemic process is a partial observation of a cross-sectional snapshot of the epidemic. Thus, the main focus in computing $P(\theta; \mathbf{y})$ is the probability of observing the given cross-sectional snapshot of the epidemic given the simulated epidemic process, and this is usually relatively easy to compute. In particular, we consider a BDM model that models the initial (branching) stages of a mutating SIR epidemic model. The model is applied to an outbreak of tuberculosis in San Francisco in the early 1990s, Small *et al.* (1994), and it is assumed that the data are a snapshot of the epidemic corresponding to when the total number of infectives reaches a given fixed size for the first time. In contrast to the homogeneously mixing example studied in Section 3, the BDM model has not previously analysed using MCMC because of the problems of applying standard MCMC algorithms. However, the BDM model has been extensively analysed using ABC, for example, MCMC–ABC (Tanaka *et al.* (2006)), SMC–ABC (Sisson *et al.* (2007), semi-automatic ABC (Fearnhead & Prangle (2012) and adaptive SMC–ABC (Del Moral *et al.*(2012). We show that MCMC can be efficiently applied to this model with computation times comparable with ABC. A second example is given in the Supporting information involving the spread of a multiple strain SIS epidemic model with interactions between the different strains of the disease. Both the BDM model and the multiple strain SIS model are Markov models, and the construction of the epidemic process has a similar structure in both cases, although slightly different constructions are useful in the two cases.

We analyse the San Francisco data using both the isEBC and the fsMCMC algorithms in order to perform exact Bayesian inference for the parameters of the model. The isEBC algorithm is shown to perform poorly for this model, and we highlight why this should be the case. The fsMCMC algorithm is very effective and is successfully applied to obtain a sample from the posterior distribution. We outline the data and the model before describing a non-centred simulation procedure for the BDM model. This is followed by analysis of the BDM model using the isEBC and the fsMCMC algorithms.

The data consist of the genotypes of 473 bacteria samples sampled from individuals infected with tuberculosis in San Francisco during an observational period in 1991–2. The data are clustered by genotype and summarized in Table 1. Let N_t denote the total number of tuberculosis cases at time t. The data are assumed to be a random sample taken at time T, where $T = \min\{t; N_t = K\}$ for some $K \in \mathbb{N}$. Thus, we have a cross-sectional study. In Tanaka *et al.* (2006) and Fearnhead & Prangle (2012), K is taken to be equal to 10,000, although Tanaka

Table 1. *Observed cluster size distribution of tuberculosis bacteria genotype data, Small et al. (1994)*

Cluster size	1	2	3	4	5	8	10	15	23	30
Number of clusters	282	20	13	4	2	1	1	1	1	1

et al. (2006) noted that analysis is insensitive to reasonable choices of K, and unless otherwise stated, we fix $K = 10,000$.

Let Z_t^i denote the total number of tuberculosis cases of type i at event time t, and let $\mathbf{Z}_t = (Z_t^1, Z_t^2, \ldots, Z_t^{s_t})$, where s_t denotes the total number of genotypes present in the population at event time t. Then $\mathbf{Z}_t$ is modelled using a continuous time Markov process with $\mathbf{Z}_0 = (1)$, a single introductory infectious case. There are three types of event: birth (infection), death (recovery) and mutation. Let α, δ and ϑ denote the birth, death and mutation rates, respectively. Then given $N_t (= \sum_{i=1}^{s_t} Z_t^i) = n$, the time until the next event is exponentially distributed with rate $n(\alpha + \delta + \vartheta)$. The probability that the next event is a birth, a death or a mutation is $a = \alpha/(\alpha + \delta + \vartheta)$, $d = \delta/(\alpha + \delta + \vartheta)$ or $q = \vartheta/(\alpha + \delta + \vartheta)$, respectively, with the selected individual equally likely to be any member of the population. Hence, the probability that the individual belongs to genotype i is Z_t^i/n. In the event of a birth, the offspring inherits the same genotype as the parent. In the event of a mutation, a completely new genotype emerges. Therefore, the model assumes that there is no difference in the way that the different genotypes behave, and we only record the genotypes with at least one member in $\mathbf{Z}_t$ for both mathematical and computational conveniences. Also the structure of $\mathbf{Z}_t$ does not change if a singleton (an individual with a unique genotype) mutates. We can exploit this observation in noting that if $N_0 = 1$ and $N_T = K > 1$, then there exists $0 < s < T$ such that $\mathbf{Z}_s = (2)$ because a birth must occur before the first death with any mutations of the singleton becomes irrelevant.

The cross-sectional nature of the data means that we cannot make meaningful inference on $(\alpha, \delta, \vartheta)$ on the basis of the data alone, and we follow Fearnhead & Prangle (2012) in reparameterizing the model in terms of (a, d, q). In Tanaka *et al.* (2006), $\pi(\vartheta)$, the marginal prior on ϑ is taken to be $N(0.198, 0.06735^2)$ on the basis of previous analysis of tuberculosis mutation. Then in Tanaka *et al.* (2006) and Fearnhead & Prangle (2012), MCMC–ABC is used to analyse the data with priors $\pi(\vartheta, \alpha, \delta) \propto \pi(\vartheta) 1_{\{\alpha > \delta\}}$ and $\pi(\vartheta, a, d) \propto \pi(\vartheta) 1_{\{a > d\}} 1_{\{a+d<1\}}$, respectively. The constraint that the birth rate is greater than the death rate is consistent with the tuberculosis epidemic growing to an infectious population of $K = 10,000$ individuals. For the isEBC algorithm, we require a proper prior distribution and therefore use a uniform prior on (a, d, q) conditioned upon $a \geq 0.5$, $\pi(a, d, q) \propto 1_{\{a \geq 0.5\}} 1_{\{a+d+q=1\}}$. The condition is motivated not only by wanting a vague but proper prior with $a > d$ but also by noting that from the perspective of a single genotype, death and mutation are equivalent. The presence of a cluster of size 30 suggests that there are approximately 634 $(= (30/473) \times 10,000)$ individuals of that genotype in the population, which is unlikely if $a < d + q$ (each genotype will almost surely go extinct). This is further supported by Fearnhead & Prangle (2012), Fig. 3, which shows no accepted values with $a < 0.5$. Furthermore in test runs, we found that simulating the BDM model with a high mutation rate $q > 0.5$ was very time consuming and had very little posterior support. For the fsMCMC algorithm, we consider both the Tanaka prior and the uniform prior. We analyse the data using the isEBC and fsMCMC algorithms for $K = 10,000$ and using the fsMCMC algorithms for K unknown, with prior $\pi(K = k) \propto 1/k$ $(k = 5000, 5001, \ldots, 20,000)$.

In both Tanaka *et al.* (2006) and Fearnhead & Prangle (2012), the iterations of the MCMC–ABC proceed as follows. Proposed parameter values $(\alpha, \delta, \vartheta)$ or (a, d, q) are generated. A realization of a tuberculosis epidemic is simulated with $N_0 = 1$ using the chosen parameters until either $N_T = 0$ (the simulation is rejected) or $N_T = 10,000$. Given that $N_T = 10,000$, a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_g)$ of size 473 is taken from the simulated population $\mathbf{z} = (z_1, z_2, \ldots, z_s)$, where g and s denote the total number of distinct genotypes in the sample and population, respectively, and x_i and z_j denote the total number of individuals of genotype i in the sample and genotype j in the population, respectively. (Note that genotype i in the random sample does not necessarily correspond to genotype i in the population.) Summary statistics, T, of the sample are computed. If T is sufficiently close to T^*, the summary statistics

of the original data, the parameter values are accepted with an appropriate probability. Otherwise the proposed parameter values are rejected. In Tanaka *et al.* (2006), the summary statistics are g, the total number of different genotypes in the sample and $H = 1 - \sum_{i=1}^{g}(x_i/473)^2$, where x_i is the total size of cluster i in the sample. The motivation for these summary statistics, which are related to Rényi entropies (Rényi(1961)) of the sample, is given with reference to Ewens(1972). In Fearnhead & Prangle (2012), a wide range of summary statistics are considered as part of the semi-automatic ABC algorithm, see Fearnhead & Prangle (2012) for details. Note that $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_{326}^*)$, and for convenience in using the isEBC and the fsMCMC algorithms, we order the clusters so that for $1 \leq i < j \leq 326$, $x_i^* \geq x_j^*$.

The non-centred simulation algorithm for the BDM process splits into two parts: simulation of the BDM until time T, $N_T = K$ and the sampling of $\mathbf{x}^*$ from $\mathbf{Z}_T = \mathbf{z}$. The proof that the algorithm gives an unbiased estimator of $P(\mathbf{X} = \mathbf{x}^*|\theta)$ (up to a constant of proportionality) is given after the algorithm. Let $\mathbf{U} = (U_1, U_2, \ldots)$, $\mathbf{W} = (W_1, W_2, \ldots)$ and $\mathbf{V} = (V_1, V_2, \ldots V_{326})$ be random vectors consisting of independent $U(0,1)$ random variables. The vectors $\mathbf{U}$ and $\mathbf{W}$ are used for the simulation of BDM process. The number of components of $\mathbf{U}$ and $\mathbf{W}$ that are required for the non-centred BDM procedure is random, and we discuss how this is dealt with in the implementing of the isEBC and fsMCMC algorithms. The vector $\mathbf{V}$ is used for sampling $\mathbf{x}^*$ from $\mathbf{Z}_T$.

Non-centred BDM procedure

1. Fix $(a, d, q) = (\alpha/(\alpha+\delta+\vartheta), \delta/(\alpha+\delta+\vartheta), \vartheta/(\alpha+\delta+\vartheta))$ and $(\mathbf{U}, \mathbf{W}, \mathbf{V})$. Set $\tilde{P} = 1$.
2. Simulate the evolution of the tuberculosis using the parameters (a, d, q) and random vectors starting from $N_0 = 1$ until $N_T = K$ as follows with t_i denoting the time at which the i^{th} event (birth, death or mutation) occurs and $t_0 = 0$. (Note that we do not compute t_i as it is not needed for our analysis.)

   (a) If $N_{t_i} = 1$, set $N_{t_{i+1}} = 2$, $\mathbf{Z}_{t_{i+1}} = (2)$ and $\tilde{P} = \tilde{P} \times a/(a+d)$.
   We condition upon a birth occurring before a death to stop the population going extinct.

   (b) If $1 < N_{t_i} < K$, choose an individual from genotype I, where I satisfies $\sum_{j=1}^{I-1} Z_{t_i}^j < [N_{t_i} U_i] + 1 \leq \sum_{j=1}^{I} Z_{t_i}^j$.
   If $W_i \leq a$, the individual gives birth; if $a < W_i \leq a+d$, the individual dies; and if $W_i > a+d$, the individual mutates. We update $\mathbf{Z}_{t_{i+1}}$ accordingly.

   (c) If $N_{t_i} = K$, stop the simulation and set $\mathbf{z} = \mathbf{Z}_{t_i}$.

3. If $s(= s_T) < 326$, set $\tilde{P} = 0$; otherwise, reorder $\mathbf{z}$ such that for all $i < j$, $z_i \geq z_j$ and sample $\mathbf{x}^*$ from $\mathbf{z}$ as follows.
   For $k = 1, 2, \ldots, 326$, we sample x_k^* individuals from the same genotype.

   - Let $\chi_i^k = 1$ if genotype i has not been sampled in the first $k-1$ genotypes to be sampled and $z_i \geq x_k^*$. Otherwise, let $\chi_i^k = 0$.
   - Choose an individual from the $A_k = \sum_{i=1}^{s} \chi_i^k z_i$ individuals who belong to genotypes that have not previously been chosen and have at least x_k^* members. The individual chosen belongs to genotype I, where $\sum_{i=1}^{I-1} \chi_i^k z_i < [A_k V_k] + 1 \leq \sum_{i=1}^{I} \chi_i^k z_i$.
     Set $\tilde{P} = \tilde{P} \times A_k/M_k$, where $M_k = (K - \sum_{l=1}^{k-1} x_l^*)$.
   - If $x_k^* > 1$, choose $x_k^* - 1$ other individuals of genotype I and set $\tilde{P} = \tilde{P} \times \prod_{j=1}^{x_k^*-1} \frac{z_I - j}{M_k - j}$.

4. Then $\tilde{P}$ is the probability that whenever the population reaches $N_t = 1$, we condition upon a birth occurring before a death, and that in sampling from $\mathbf{z}$, the first x_1^* individuals are of the same genotype, the next x_2^* individuals are of the same genotype distinct from the genotype of the first x_1^* individuals and so on. To account for the fact that the individuals in $\mathbf{x}^*$ could be sampled in any order, we can compute $P(= P(a, d, q, \mathbf{U}, \mathbf{W}, \mathbf{V})) = L\tilde{P}$, where

$$L = \frac{473!}{(30!)^1(23!)^1(15!)^1(10!)^1(8!)^1(5!)^2(4!)^4(3!)^{13}(2!)^{20}(1!)^{282}1!1!1!1!1!2!4!13!20!282!},$$
$$= 4.033 \times 10^{357}.$$

However, because L is a constant across all simulations, $\tilde{P}$ suffices. Moreover, for numerical stability, it is more convenient to compute and record $\log \tilde{P}$.

The key behind the non-centred BDM procedure is to express the likelihood as

$$P(\mathbf{X} = \mathbf{x}^*|\theta) = \sum_{\mathbf{z}} P(\mathbf{X} = \mathbf{x}^*|\mathbf{Z}_T = \mathbf{z})P(\mathbf{Z}_T = \mathbf{z}|\theta),$$
$$= \sum_{\mathbf{z}\neq(0)} P(\mathbf{X} = \mathbf{x}^*|\mathbf{Z}_T = \mathbf{z})P(\mathbf{Z}_T = \mathbf{z}|\theta), \tag{4.1}$$

where the second line of (4.1) follows because $P(\mathbf{X} = \mathbf{x}^*|\mathbf{Z}_T = (0)) = 0$. We show how step 2 of the BDM procedure gives an unbiased estimate of $P(\mathbf{Z}_T = \mathbf{z}|\theta)$ ($\mathbf{z} \neq \mathbf{0}$) by constructing $\mathbf{Z}_T$ in a deterministic manner given $(\mathbf{U}, \mathbf{W}, \theta)$. Let $\tilde{\mathbf{W}} = (\tilde{W}_1, \tilde{W}_2, \ldots)$ be a random vector of independent $U(0, 1)$ random variables and suppose that step 2(a) in the BDM procedure is replaced by

(a′) If $N_{t_i} = 1$, implement the following. If $\tilde{W}_i \leq a$, the individual gives birth and set $\mathbf{Z}_{t_i+1} = (2)$; if $a < \tilde{W}_i \leq a + d$, the individual dies and set $\mathbf{Z}_{t_i+1} = (0)$; and if $\tilde{W}_i > a + d$, the individual mutates and set $\mathbf{Z}_{t_i+1} = (1)$. If $\mathbf{Z}_{t_i+1} = (0)$, the BDM process has died out and the procedure terminates with $\tilde{P} = 0$.

Note that in (a′), $\tilde{W}_i$ has the same role as W_i in step 2(b) of the BDM procedure in determining which event takes place at time t_i. Because there is only one genotype when $N_{t_i} = 1$, we do not need to choose the genotype of the individual in (a′). Let $H'_T(\theta, \mathbf{U}, \mathbf{W}, \tilde{\mathbf{W}})$ denote the simulation generated by the BDM procedure with (a′). Then

$$E\left[1_{\{H'_T(\theta,\mathbf{U},\mathbf{W},\tilde{\mathbf{W}})=\mathbf{z}\}}\right] = P(\mathbf{Z}_T = \mathbf{z}|\theta), \tag{4.2}$$

and hence, $1_{\{H'_T(\theta,\mathbf{U},\mathbf{W},\tilde{\mathbf{W}})=\mathbf{z}\}}$ is an unbiased estimator of $P(\mathbf{Z}_T = \mathbf{z}|\theta)$. For $\mathbf{z} \neq (0)$, step 2(a) of the BDM procedures averages over $\tilde{\mathbf{W}}$ to give $\tilde{P} = E_{\tilde{\mathbf{W}}}[1_{\{H'_T(\theta,\mathbf{U},\mathbf{W},\tilde{\mathbf{W}})=\mathbf{z}\}}]$ as an unbiased estimator of $P(\mathbf{Z}_T = \mathbf{z}|\theta)$ at the end of step 2. Then given $\mathbf{Z}_T = \mathbf{z}$, step 3 gives an unbiased estimate of $P(\mathbf{X} = \mathbf{x}^*|\mathbf{Z}_T = \mathbf{z})$, and hence, $E[P](= LE[\tilde{P}]) = P(\mathbf{X} = \mathbf{x}^*|\theta)$ as required. In principle, step 3 of the BDM procedure could be replaced by computing exactly the probability of observing $\mathbf{x}^*$ given $\mathbf{z}$. However, this is not practical given the total number of ways that $\mathbf{x}^*$ can arise. The reordering of $\mathbf{z}$ is crucial for the successful implementation of the fsMCMC algorithm as it is the relative size of each genotype that is important, not their order, and without reordering the acceptance rate is significantly lower.

For the isEBC algorithm, we generate new parameters $\theta = (a, d, q)$ and random vectors $\mathbf{U}$, $\mathbf{W}$ and $\mathbf{V}$ at each iteration. Because we do not need to store $\mathbf{U}$ and $\mathbf{W}$ for future iterations, we simulate the components of the vectors as required in using the non-centred BDM

procedure. We ran 10 batches of 10^6 iterations of the isEBC algorithm. We found that $\tilde{P}$, and consequently P, is heavy tailed, and in all the batches, there were a few dominant simulations. For example, in three cases, $\max_j\{\tilde{P}_j\}/\sum_l \tilde{P}_l$ is in excess of 0.99 and in all cases exceeds 0.40. The situation does not improve when we combine the batches with the best simulation (highest value of $\tilde{P}$) accounting for 88.9 per cent of the weighting across the 10^7 simulations. The failure of the isEBC is primarily down to the lack of conditioning in the simulation part of the BDM procedure. Therefore, we are highly unlikely to generate data sets $\mathbf{z}$ from which $\mathbf{x}^*$ is likely to have arisen. However, it is difficult to see how to run a conditioned simulation of $\mathbf{z}$ to increase the chance of sampling $\mathbf{x}$ without abandoning the simple Gillespie-type algorithm (Gillespie (1976), which is used to generate $\mathbf{z}$ (subject to the minor conditioning at $N_t = 1$), and is extremely fast to implement. The dominance of a few simulations means that the GIMH is not a practical solution to this problem unless a large number of simulations are carried out for each set of parameters. Hence, we turn our attention to the fsMCMC algorithm.

For the fsMCMC algorithm, we considered four scenarios corresponding to each combination of the Tanaka and uniform priors with $K = 10,000$ and K unknown. In all cases, we ran the algorithm for 1.1×10^6 iterations discarding the first 1×10^5 iterations as burn-in. Because of substantial serial correlation in the MCMC output, we thinned the MCMC output retaining every 100 iterations giving a sample of size 10,000 from the posterior distribution of the parameters. The BDM fsMCMC algorithm is outlined in the succeeding text.

BDM fsMCMC algorithm

1. Choose initial values for $\theta = (\alpha, \delta, \vartheta)$ or $\theta = (a, d, q)$ as appropriate, K and $\mathbf{y} = (\mathbf{u}, \mathbf{w}, \mathbf{v})$ drawn.

   Run the non-centred BDM procedure and compute $P(\theta, \mathbf{y}, K)$.

   We choose $\theta = (0.75, 0.25, 0.2)$ for the Tanaka prior and $\theta = (0.65, 0.2, 0.15)$ for the uniform prior. These values were seen as reasonable starting points on the basis of the results presented Tanaka *et al.* (2006) and Fearnhead & Prangle (2012) and giving $P(\theta, \mathbf{y}, K) > 0$. Similar performance was seen with other initial choices of θ. We initialize with $K = 10,000$. The components of $\mathbf{y}$ are independent draws from $U(0, 1)$. Note that the required length of $\mathbf{u}$ and $\mathbf{w}$ is unknown. We found that it sufficed to fix the lengths of $\mathbf{u}$ and $\mathbf{w}$ at 100,000. Furthermore, if extra $\mathbf{u}$ and $\mathbf{w}$ terms are required, these can simply be obtained by making independent draws from $U(0, 1)$.
2. For $i = 1, 2, \ldots, 1.1 \times 10^6$. Update the parameters and augmented data using the proposal distributions given in the succeeding text by running the non-centred BDM procedure, computing $P(\theta', \mathbf{y}', K')$ and accept or reject the new proposed values accordingly.

   (a) θ': Propose a new value θ' from a multivariate Gaussian with mean θ and variance–covariance matrix $0.025^2 I$, where I is the identity matrix. (For the uniform prior, we only propose new values for a and d with $q = 1 - a - d$.)

   The scaling of the random walk Metropolis update was found using pilot runs and, although it performs well, could be improved especially for the Tanaka prior by choosing a proposal variance Σ, which more closely resembles the dependence between the parameters in the posterior distribution using adaptive MCMC. The acceptance rates ranged between 10 and 13 per cent.

   (b) $(\mathbf{u}', \mathbf{w}')$: Fix $G \geq 1$. Partition $\mathbf{u}$ and $\mathbf{w}$ into blocks of length G. Construct $\mathbf{u}'$ ($\mathbf{w}'$) by proposing to update one element of $\mathbf{u}$ ($\mathbf{w}$) in each block chosen uniformly at random

from $U(0, 1)$. The remaining $G-1$ elements in each block $\mathbf{u}'$ and $\mathbf{w}'$ are left unchanged from $\mathbf{u}$ and $\mathbf{w}$. Note that we update different elements in $\mathbf{u}$ and $\mathbf{w}$.

We found that taking $G = 50$ worked well, that is, updating 2 per cent of the augmented data, 2000 values, at each iteration. Typically between 300 and 1000 of the augmented values were used in the simulation and gave an acceptance rate of approximately 25 per cent.

(c) $\mathbf{v}'$: Uniformly at random, select five elements of $\mathbf{v}$ and propose replacements from $U(0, 1)$.

Note that this step is very quick as it is not necessary to re-run the simulation of the BDM process.

(d) K': If K is a parameter in the model, draw $K' = K + U(-500, 500)$.

This is random walk Metropolis update and in both cases resulted in an acceptance rate over 50 per cent suggesting that a uniform proposal with a larger range would have been more efficient, even though the algorithm performs very well for updating K.

3. Discard the first 1×10^5 iterations as burn-in and store every 100^{th} iteration after the burn-in to obtain a sample of size 10,000 from $\pi(\theta|\mathbf{x}^*)$.

The reordering of $\mathbf{z}$ for the BDM procedure is particularly useful for steps 2(a) and 2(b) as a small difference early in the construction of two BDM processes can lead to different genotypes being the most populous, which can have a dramatic affect $P(\theta', \mathbf{y}', K')$ without reordering. The fsMCMC algorithm takes just over two times as long per iteration as the isEBC algorithm for fixed $K = 10000$ and approximately six times as long per iteration for unknown K. The increased time for unknown K is due to the time taken to run the BDM algorithm increasing approximately quadratically in K. In generating 1.1×10^6 iterations, this is less than half the number used in either Tanaka *et al.* (2006) and Fearnhead & Prangle (2012) for the ABC algorithm. The mixing of the thinned chains is good with estimated effective independent and identically distributed (iid) sample size ranging from 500 to 2000 for all parameters with particularly good performance for the uniform prior.

The output from the four implementations of the BDM fsMCMC algorithm is summarized in Table 2 with parameter estimates provided by Dennis Prangle using the semi-automatic ABC algorithm in Fearnhead & Prangle (2012). A plot of a sample of d against a for each run of the algorithm is given in Fig. 2. We observe consistent results across the four cases in terms of estimation of parameter means and standard deviations. We observe more parameter uncertainty for a, d and q with the Tanaka prior than the uniform prior

Table 2. *Estimated posterior means and standard deviations of the parameters of the BDM model applied to the San Francisco tuberculosis data for the four runs of the fsMCMC algorithm and the semi-automatic ABC algorithm, Fearnhead & Prangle (2012)*

MCMC run	$E[a\|\mathbf{x}^*]$ $(sd(a\|\mathbf{x}^*))$	$E[d\|\mathbf{x}^*]$ $(sd(d\|\mathbf{x}^*))$	$E[q\|\mathbf{x}^*]$ $(sd(q\|\mathbf{x}^*))$	$E[K\|\mathbf{x}^*]$ $(sd(K\|\mathbf{x}^*))$
Tanaka prior, K fixed	0.694 (0.043)	0.102 (0.073)	0.204 (0.033)	10,000 (—)
Uniform prior, K fixed	0.708 (0.035)	0.075 (0.059)	0.217 (0.028)	10,000 (—)
Tanaka prior, K unknown	0.691 (0.054)	0.138 (0.091)	0.172 (0.040)	15,707 (3110)
uniform prior, K unknown	0.715 (0.044)	0.091 (0.072)	0.194 (0.034)	14,972 (3417)
Semi-automatic ABC	0.702 (0.041)	0.090 (0.070)	0.208 (0.033)	10,000 (—)

BDM, birth–death–mutation; fsMCMC, forward simulation Markov chain Monte Carlo; ABC, approximate Bayesian computation.

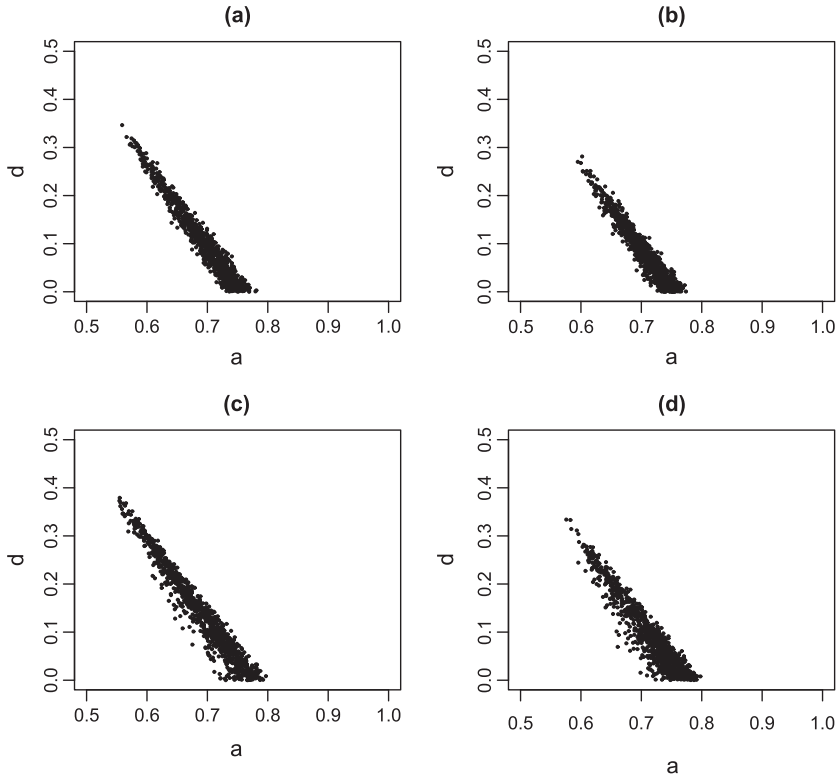

Fig. 2. Plots of d against a for sample of size 1000 (every $\mathbf{10}^{\text{th}}$ stored value) from $\pi(a, d|\mathbf{x}^*)$; (a) Tanaka prior and $K = \mathbf{10,000}$; (b) uniform prior and $K = \mathbf{10,000}$; (c) Tanaka prior and K, unknown; (d) uniform prior and K, unknown.

and with unknown K as opposed to $K = 10,000$. The results obtained are similar to those obtained in Fearnhead & Prangle (2012) using semi-automatic ABC, in terms of both the means and standard deviations of parameters. Thus, our analysis provides support for using the semi-automatic ABC, although the fsMCMC algorithm is clearly preferable for this model.

The analysis shows that the results are robust to the choice of K supporting the use of $K = 10,000$ as noted in Tanaka *et al.* (2006). For unknown K, there is posterior support for a wide range of K values with the samples for K ranging over $(5090, 19999)$ and $(5109, 20000)$ for the Tanaka and uniform priors, respectively, with standard deviations for K in excess of 3000 in both cases. The results for K, including the noted high acceptance rate for the random walk Metropolis proposal, are not surprising as the structure of a BDM population at $K = 10,000$ and $K = 20,000$ is usually not dissimilar in terms of the proportion of the population in the most populous genotype or the proportion of the population who are singletons and it is quantities such as these, which play a key role in the chances of observing the random sample $\mathbf{x}^*$.

Finally, there are limitations to the BDM model being applied to tuberculosis. Firstly, we assume a Markov model with constant infection, recovery and mutation rates. It would be more appropriate for these quantities to be time varying. However, estimating general time-varying parameters is a much harder problem and would probably require more data than are

available through a cross-sectional snapshot. By expanding the parameter and state space and using the method of stages approach (Barbour (1976)), it is straightforward to allow for non-exponential infectious periods with the infection and mutation rate varying between stages. For example, we could assume that the infectious period of an individual consists of two successive stages ($i = 1, 2$) with the length of stage i being distributed according to $\text{Exp}(\delta_i)$, and whilst in stage i, an individual infects at rate α_i and mutates at rate θ_i. The Gillespie algorithm can still be used to simulate the process, but the simulation is more time consuming as we need to identify the total number of individuals in each stage of each genotype rather than than just the total number of individuals of each genotype. The method of stages can be used to construct more realistic Markov processes to model other population processes. Secondly, we have assumed that the population is closed, in that, there is a single introductory case of tuberculosis into San Francisco from which the entire epidemic emanates. However, it could be that the observed data come from outbreaks attributable to multiple introductory cases. If the total number of introductory cases is small, then the findings are likely to be similar to those presented here; otherwise, it will be necessary to incorporate immigration of tuberculosis cases into the population. However, this can be performed only if the relative rate of introductory cases is known.

5. Conclusions

We have introduced a simulation-based MCMC algorithm that has been successfully applied to two different epidemic scenarios. The key features are using a non-centred parameterization for ease of constructing the simulations and importance sampling (pseudo-likelihood) to improve the efficiency of the algorithm. The BDM in Section 4 is a Markovian process simulated using the Gillespie algorithm (Gillespie (1976)). Generally, a non-centred parameterization is straightforward to implement for the Gillespie algorithm, and thus, the fsMCMC algorithm can easily be applied to other Markov population processes such as the Lotka–Volterra predator–prey model (Boys *et al.* (2008), White *et al.* (2014)). We have focussed upon analysing cross-sectional data that involve large-scale data augmentation and limited observed data, because this is where the fsMCMC algorithm is particularly effective.

Throughout the paper, we have sought to optimize the performance of the fsMCMC through the scaling of the proposal variance for the random walk Metropolis updates for the parameters θ and the choice of G, the proportion $1/G$ of components of $\mathbf{y}$ to update at each iteration. This has been performed fairly successfully on the basis of pilot runs to tune the proposal variance and G, but it would be useful to develop generic theory to choose these. For example, the optimal scaling of the random walk Metropolis algorithm in Section 4 was found to correspond to an acceptance rate between 10 and 13 per cent, and for the multiple strain SIS epidemic model in the Supporting information, the acceptance rate was even lower. These acceptance rates are less than the 23.4 per cent suggested in Roberts *et al.* (1997), and this is due to us having a noisy (unbiased) estimate of the likelihood depending upon the augmented data $\mathbf{y}$. This optimal scaling issue is a topic of ongoing investigation. There are alternatives for updating $\mathbf{y}$, in particular, different independence sampler updating schemes. For example, changes at the start of the BDM or multiple strain SIS process will generally have more impact on the evolution of the process than changes later on in the process, and this could be incorporated into updating schemes for $\mathbf{y}$.

Whilst we have presented simulation as a data augmentation tool for obtaining a tractable likelihood, it would be interesting to explore the use of both non-centred parameterizations and importance sampling in an ABC framework as suggested by Andrieu *et al.* (2012).

Acknowledgements

The authors would like to thank the anonymous referees for their helpful comments. The authors would like to thank Dennis Prangle for sharing his parameter estimates for the BDM model. The first author was partly supported by the Engineering and Physical Sciences Research Council under grant EP/J008443/1.

References

Andrieu, C., Doucet, A. & Lee, A. (2012). Discussion of constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74**, 451–452.

Andrieu, C. & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37**, 697–725.

Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications*, (Second edition)., Griffin, London.

Barbour, A. D. (1976). Networks of queues and the method of stages. *Adv. in Appl. Probab.* **8**, 584–591.

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

Beaumont, M., Zhang, W. & Balding, D. (2002). Approximate Bayesian computation in population genetics. *Genet.* **162**, 2025–2035.

Boys, R. J., Wilkinson, D.J. & Kirkwood, T. B. L. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statist. Comput.* **18**, 125–135.

Del Moral, P., Doucet, A. & Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian compuation. *Statist. Comput.* **22**, 1009–1020.

Demiris, N. & O'Neill, P. (2006). Computation of final outcome probabilities for the generalised stochastic epidemic. *Statist. Comput.* **16**, 309–317.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Population Biol.* **3**, 87–112.

Fearnhead, P. & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74**, 419–474.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434.

Kurtz, T. (1971). Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *J. Appl. Probab.* **8**, 344–356.

Lee, A., Andrieu, C. & Doucet, A. (2012). Discussion of constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74**, 449–450.

Marjoram, P., Molitor, J., Plagnol, V. & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100**, 15324–15328.

McKinley, T. J., Ross, J. V., Deardon, R. & Cook, A. R. (2014). Simulation-based Bayesian inference for epidemic models. *Comput. Statis. Data Anal.* **71**, 434–447.

Neal, P. (2012). Efficient likelihood-free Bayesian computation for household epidemics. *Statist. Comput.* **22**, 1239–1256.

Neal, P. J. & Roberts, G. O. (2005). A case study in non-centering for data augmentation: stochastic epidemics. *Statist. Comput.* **15**, 315–327.

Neal, P., Roberts, G. O. & Yuen, W. K. (2012). Optimal scaling of random walk Metropolis algorithms with discontinuous target densities. *Ann. Appl. Probab.* **22**, 1880–1927.

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

Papaspoliopoulos, O., Roberts, G. O. & Sköld, M. (2003). Non-centered parameterisations for hierarchical models and data augmentation. In *Bayesian Statistics 7* (eds Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M. & West, M.), Oxford University Press: Oxford; 307–326.

Rényi, A. (1961). On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960*. University of California Press: Berkeley, California, 547–561.

Roberts, G. O., Gelman, A. & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7**, 110–120.

Sellke, T. (1983). On the asymptotic distribution of the size of a stochastic epidemic. *J. Appl. Probab.* **20**, 390–394.

Sisson, S. A., Fan, Y. & Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **104**, 1760–1765.

Small, P. M., Hopewell, P. C., Singh, S. P., Paz, A., Parsonnet, J., Ruston, D. C., Schecter, G. F., Daley, C. L. & Schoolnik, G. K. (1994). The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N. Engl. J. of Med.* **330**, 1703–1709.

Tanaka, M. M., Francis, A. R., Luciani, F. & Sisson, S. A. (2006). Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genet.* **173**, 1511–1520.

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

White, S., Kypraios, T. & Preston S. (2014). Fast approximate Bayesian computation for discretely observed Markov models using a factorised posterior distribution. *To appear in Statist. Comput.* Available at: http://link.springer.com/article/10.1007/s11222-013-9432-2.

Received February 2013, in final form June 2014

Peter Neal, Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YW, UK.
E-mail: p.neal@lancaster.ac.uk

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's website.