The International Journal of Biostatistics

Volume 5, Issue 1

2009

Article 24

Inference in Epidemic Models without Likelihoods

Trevelyan McKinley, *University of Cambridge*

Alex R. Cook, *National University of Singapore*

Robert Deardon, *University of Guelph*

Recommended Citation:

McKinley, Trevelyan; Cook, Alex R.; and Deardon, Robert (2009) "Inference in Epidemic Models without Likelihoods," *The International Journal of Biostatistics*: Vol. 5: Iss. 1, Article 24.

DOI: 10.2202/1557-4679.1171

Inference in Epidemic Models without Likelihoods

Trevelyan McKinley, Alex R. Cook, and Robert Deardon

Abstract

Likelihood-based inference for epidemic models can be challenging, in part due to difficulties in evaluating the likelihood. The problem is particularly acute in models of large-scale outbreaks, and unobserved or partially observed data further complicates this process. Here we investigate the performance of Markov Chain Monte Carlo and Sequential Monte Carlo algorithms for parameter inference, where the routines are based on approximate likelihoods generated from model simulations. We compare our results to a gold-standard data-augmented MCMC for both complete and incomplete data. We illustrate our techniques using simulated epidemics as well as data from a recent outbreak of Ebola Haemorrhagic Fever in the Democratic Republic of Congo and discuss situations in which we think simulation-based inference may be preferable to likelihood-based inference.

KEYWORDS: Approximate Bayesian Computation, inference, Markov Chain Monte Carlo, infectious disease modelling, Sequential Monte Carlo

Author Notes: T.J.M. was supported by the Department for the Environment, Food and Rural Affairs/Higher Education Funding Council of England, grant number VT0105; A.R.C. by the Biotechnology and Biological Sciences Research Council, grant number BB/C007263/1; and R.D. by the Natural Sciences and Engineering Research Council of Canada Discovery Grants Program. The authors would like to thank Gavin Gibson, Olivier Restif and Andrew Conlan for helpful discussions. All models were coded in C++ and all data were analysed in the R statistical package (R Development Core Team, 2008).

Erratum

On Page 4, in the second full paragraph, equation (2.2) should appear as follows:

$$\begin{aligned} f(\mathbf{t} \mid \beta, \delta, \gamma) = & \left\{ \prod_{i=1}^k \left[\left(\frac{\beta}{N} S_{t_{(i-1)}} I_{t_{(i-1)}} \right)^{\eta_{1(i)}} \left(\delta E_{t_{(i-1)}} \right)^{\eta_{2(i)}} \left(\gamma I_{t_{(i-1)}} \right)^{1-\eta_{1(i)}-\eta_{2(i)}} \right] \right. \\ & \times \exp \left[- \left(\frac{\beta}{N} S_{t_{(i-1)}} I_{t_{(i-1)}} + \delta E_{t_{(i-1)}} + \gamma I_{t_{(i-1)}} \right) (t_{(i)} - t_{(i-1)}) \right] \left. \right\} \\ & \times \exp \left[- \left(\frac{\beta}{N} S_{t_{(k)}} I_{t_{(k)}} + \delta E_{t_{(i-1)}} + \gamma I_{t_{(k)}} \right) (t - t_{(k)}) \right]. \end{aligned} \quad (2.2)$$

The authors apologise for this typographic error and thank Simon Frost for bringing it to their attention.

1 Introduction

Mathematical models of disease transmission provide useful insights into the behaviour of epidemic systems (Anderson and May, 1991). They also provide quantitative predictions for the future path of a disease outbreak (Riley et al., 2003), allowing the impact of potential strategies for disease control to be assessed (Haydon et al., 2003; Tildesley et al., 2006; Cook et al., 2008a). In this paper we focus on stochastic models, in which epidemiological events (such as infection and recovery) are governed by a set of probability statements. Traditional deterministic formulations are useful as approximations to fully stochastic models when the population size is large (Becker, 1989), but the latter provide a much more natural framework for studying most epidemic systems. In particular they quantify the *uncertainty* in a dynamical system, and are straightforward to extend to more complex scenarios: such as individual-level or spatially-explicit models.

Simulation-based models (e.g. Bailey, 1975; Keeling and Rohani, 2008) can be used to produce Monte Carlo estimates for various aspects of the system, conditional on a given set of parameter values; however the uncertainty generated is related only to the model structure and does not reflect *parameter* uncertainty per se. In these models the effect of parameter uncertainty is often assessed using some form of sensitivity analysis.

Various alternative methods have been proposed to estimate parameters for epidemiological systems, the more sophisticated of which rely on the calculation of a likelihood function (e.g. Becker, 1976; Andersson and Britton, 2000; Fearnhead and Meligkotsidou, 2004; Ionides et al., 2006). Likelihood-based inference for epidemic models poses many challenges, not least because available epidemic data are often censored or incomplete. In this case the unobserved data must often be inferred from the observed data. Alternatively, approximations can be made in order to match the model structure to the form of the data (e.g. using discrete-time models).

A popular and particularly effective solution is to use the Bayesian paradigm to estimate parameters, using numerical techniques such as Markov Chain Monte Carlo (MCMC) algorithms (e.g. Gibson and Renshaw, 1998; O'Neill and Roberts, 1999, Streftaris and Gibson, 2004; Lekone and Finkenstädt, 2006, Cook et al., 2007), or Sequential Monte Carlo (SMC) routines (e.g. Doucet et al., 2001; Liu, 2001). Bayesian methodology provides a useful means to infer unobserved or missing data, since it treats all unknown parameters and data alike as random variables, for which full posterior distributions can be estimated. However, as the population gets larger and/or the process gets more intricate, the likelihood can become mathematically or computationally

intractable (e.g. Deardon et al., 2009). Some approximation techniques have been considered for fitting these sorts of models (e.g. Filipe and Gibson, 2001; Krishnarajah et al., 2005; Diggle, 2006; Ross et al., 2006, Cauchemez and Ferguson, 2008; Deardon et al., 2009) with some success, but these are usually specific to particular models or data, thus limiting their general application.

The method of Ionides et al. (2006) provides a more generic approximation method for estimating parameters in nonlinear dynamical systems by using iterated filtering to produce maximum likelihood estimates in the case of time-series epidemic data. The authors use a Sequential Monte Carlo framework in which the current state of the system is estimated through a series of conditional distributions based on observed data at previous time points. They then introduce a novel method for estimating the *static* parameters of the system by treating them as time-varying parameters and honing in on the “true” values as they iterate through the sequence. Here we are interested in exploring an alternative approach in which the explicit calculation of the full joint likelihood is replaced with Monte Carlo approximations based on simulating from the underlying model, an idea dating back at least to Diggle and Gratton (1984). These ideas have recently been extended to the Bayesian framework, in which *approximate* posterior distributions for the unknown parameters can be generated in a similar vein (Tavaré et al., 1997).

There are various potential advantages of this system for modelling epidemic processes. Firstly, since it is based on simulation models it provides a natural method for imputing missing data. Secondly, the simulated data are matched to the observed data through the use of metrics, usually based on some form of summary (or sufficient) statistic. This allows key features of the epidemic to be used to drive the model fit, even though information on individual events or event times may be incomplete or unobserved. Thirdly, the use of a Bayesian framework allows prior information about the parameters to be easily incorporated into the model. This is particularly useful in epidemic systems where there is often a high degree of correlation between parameters, but for which prior information about some of the sub-processes (such as the length of the infectious period) are sometimes known (e.g. through experimental studies or historical data).

Another major motivation for using these approaches for inference in epidemic models is that in many situations simulation algorithms are much faster to program and perform (Gillespie, 1977) than repeated calculation of the likelihood function. Speed is particularly important in the face of an ongoing epidemic, where model predictions would need to be regularly updated and refined as new data emerge, and having a reliable but quick model-fitting algorithm is essential. The ability to combine elements of model and parameter

uncertainty into model predictions is also an attractive property.

These Approximate Bayesian Computation (ABC) techniques have already been employed with considerable success in various fields, for example: population genetics (Beaumont et al., 2002; Tallmon et al., 2004; Marjoram et al., 2003), systematics (Tavaré et al., 1997; Pritchard et al., 1999; Tavaré et al., 2002), molecular epidemiology (Tanaka et al., 2006; Sisson et al., 2007a) and stereology (Bortot et al., 2007).

In this paper we will deal with the case where we have discrete temporal count data—which is a familiar scenario in many epidemic situations—and begin by introducing the model formulation for these types of systems. We then discuss parameter inference techniques for these models within a Bayesian framework, initially through the gold-standard of data-augmented MCMC (DA-MCMC; Gibson and Renshaw, 1998), in which the unobserved event times are inferred as part of the model fitting process. We then introduce various ABC methods and explore the performance of some simple metrics and tolerances for capturing the epidemic dynamics by directly comparing the utility of the approximate posteriors to the exact posteriors obtained through DA-MCMC. In particular we examine the utility of these simple metrics in situations where there are large amounts of missing or incomplete data. We illustrate these techniques using simulated data as well as data from a 1995 outbreak of Ebola Haemorrhagic Fever in the Democratic Republic of Congo, and conclude with a discussion of the advantages and disadvantages of ABC methods, and how simulation-based approaches may be extended to more complex, epidemic-specific models.

2 Temporal $\mathcal{SEIR}$ model

A popular way to model epidemic systems is to classify individuals according to their epidemiological status over time. Various simple state-space models exist for epidemic systems (see, for example, Bailey, 1975 or Renshaw, 1993), but here we consider one form only, the $\mathcal{SEIR}$ model, since this relates directly to the Ebola scenario discussed in Section 4.1.

We consider four stages, so at any point in time an individual can be classified as either susceptible to infection ($\mathcal{S}$), infected but not infectious ($\mathcal{E}$), infected and infectious ($\mathcal{I}$), or removed from the population ($\mathcal{R}$). We also assume that individuals can only move through the states in that order. Assuming a homogeneous closed population of size N (i.e. with negligible births, deaths or migration), and homogeneous mixing within the population, the rate of transition of an individual from $\mathcal{S} \rightarrow \mathcal{E}$ can be modelled as being pro-

portional to the proportion of infective individuals in the population (scaled here by the population size—known as frequency dependence). The rate of transition from $\mathcal{E} \rightarrow \mathcal{I}$ and $\mathcal{I} \rightarrow \mathcal{R}$ are constant. These assumptions are not necessary for either likelihood- or simulation-based methodologies but are convenient for their exposition, since exponentially-distributed inter-event times are produced—allowing *event* transitions to be modelled without having to condition on specific individuals in the population.

If S_t , E_t , I_t and R_t are random variables representing the number of susceptible, latent, infective and removed individuals at time t , and dt is a vanishingly small time period such that only one event can occur in $[t, t + dt)$, then we can write the probabilities of events within the population as:

$$\begin{aligned} P(\mathcal{S} \rightarrow \mathcal{E}) &= \beta N^{-1} S_t I_t dt + o(dt), \\ P(\mathcal{E} \rightarrow \mathcal{I}) &= \delta E_t dt + o(dt), \\ P(\mathcal{I} \rightarrow \mathcal{R}) &= \gamma I_t dt + o(dt). \end{aligned} \quad (2.1)$$

The notation $\mathcal{S} \rightarrow \mathcal{E}$ corresponds to the movement of a single individual from state $\mathcal{S}$ to state $\mathcal{E}$ (likewise for the other possible transitions). Here β is the transmission parameter, δ^{-1} is the mean incubation period and γ^{-1} is the mean infectious period.

If, in a closed population of N individuals, we were able to observe a series of exact continuous (ordered) event times, $t_{(0)} = 0 < t_{(1)} < \dots < t_{(k)} < t$, then the likelihood function for model (2.1) would be given by:

$$\begin{aligned} f(\mathbf{t} \mid \beta, \delta, \gamma) &= \\ &\left\{ \prod_{i=1}^k \left[\left(\frac{\beta}{N} S_{t_{(i-1)}} I_{t_{(i-1)}} \right)^{\eta_{1(i)}} \left(\delta E_{t_{(i-1)}} \right)^{1-\eta_{2(i)}} \left(\gamma I_{t_{(i-1)}} \right)^{1-\eta_{1(i)}-\eta_{2(i)}} \right] \right. \\ &\quad \times \exp \left[- \left(\frac{\beta}{N} S_{t_{(i-1)}} I_{t_{(i-1)}} + \delta E_{t_{(i-1)}} + \gamma I_{t_{(i-1)}} \right) (t_{(i)} - t_{(i-1)}) \right] \Big\} \\ &\quad \times \exp \left[- \left(\frac{\beta}{N} S_{t_{(k)}} I_{t_{(k)}} + \delta E_{t_{(i-1)}} + \gamma I_{t_{(k)}} \right) (t - t_{(k)}) \right], \end{aligned} \quad (2.2)$$

where t is the current point of the epidemic and $\eta_{(i)} = (\eta_{1(i)}, \eta_{2(i)})$ are a set of binary indicator variables such that $\eta_{1(i)} = 1$ if event $t_{(i)}$ is an infection (e.g. a move from $\mathcal{S} \rightarrow \mathcal{E}$), and 0 otherwise, and $\eta_{2(i)} = 1$ if event $t_{(i)}$ is a move from $\mathcal{E} \rightarrow \mathcal{I}$, and 0 otherwise. $S_{t_{(i)}}$, $E_{t_{(i)}}$ and $I_{t_{(i)}}$ are counts of susceptible, latent and infective individuals at time $t_{(i)}$ —i.e. in the time period $[t_{(i)}, t_{(i+1)})$. As previously stated, since the background population is homogeneous and the model Markovian, it is enough to condition on the time to next event rather than on specific individuals.

3 Bayesian analysis and MCMC

In a Bayesian framework we consider all parameters (θ) of the model to be random variables, and given a data set ($\mathbf{D}$) we are interested in estimating the *posterior* distribution for the parameters conditional upon the data, $f(\theta|\mathbf{D})$. Using Bayes' theorem we can write the joint posterior as a function of the likelihood, $f(\mathbf{D}|\theta)$, multiplied by the *prior* distribution(s) for the unknown parameter(s), $\pi(\theta)$, up to some normalising constant. Due to analytical complexity it is often necessary to evaluate the posterior numerically, using techniques such as MCMC or SMC. We assume that the reader is familiar with the Bayesian paradigm, MCMC methods and SMC methods; if not the reader is referred to excellent texts by Gelman et al. (2004), Gamerman and Lopes (2006), Doucet et al. (2001) and Liu (2001).

3.1 Data-augmented MCMC

In practice epidemic data are usually only *partially* observed (for example new infections/removals are often only available at discrete time intervals). One approach for dealing with this problem is to modify the model to match the data, using a discrete version of the continuous-time transmission model (e.g. Lekone and Finkenstädt, 2006). There are two reasons we prefer not to do this. One is that, since the underlying process is continuous, a more satisfying alternative is to treat the data as a censored realisation of the continuous time model, as one would in the analogous situation in survival analysis. The other is that data can be recorded at different time granularity for the same disease (e.g. 12-hourly [Moser et al., 1979], daily [Chowell et al., 2007], or weekly [Vynnycky et al., 2007]), making comparison between different outbreaks difficult otherwise.

Let $\mathbf{t}$ be the set of unobserved event times and $\mathcal{T} = \{\mathbf{t} : f(\mathbf{D}|\mathbf{t}) = 1\}$ be the space of event times that are consistent with the censored data $\mathbf{D}$. If $\mathbf{t}$ were known, the likelihood function can be written $f(\mathbf{D}, \mathbf{t}|\theta) = f(\mathbf{D}|\mathbf{t}, \theta)f(\mathbf{t}|\theta) = f(\mathbf{t}|\theta)$. As $\mathbf{t}$ is not known, the likelihood can be found only by integrating the unobserved event times of the (highly complex) space that is consistent with the data: $f(\mathbf{D}|\theta) = \int_{\mathcal{T}} f(\mathbf{D}, \mathbf{t}|\theta) d\mathbf{t} = \int_{\mathcal{T}} f(\mathbf{t}|\theta) d\mathbf{t}$.

The problem of calculating this integral can be tackled using data-augmented MCMC following Gibson and Renshaw (1998). Under this approach, the parameter vector is expanded from θ to $(\theta, \mathbf{t})$, MCMC is used to explore the joint distribution of parameters and event times, and the resulting marginal distribution of the parameters θ integrating over the event times $\mathbf{t}$ is the desired posterior distribution. The approach is well established, see O'Neill and

Roberts (1999); O'Neill et al. (2000); Streftaris and Gibson (2004); Höhle et al. (2005); Cook et al. (2007) and Cook et al. (2008a). A schematic representing this approach is depicted in Figure 1a. Essentially the algorithm is inferring the trajectories of the censored/missing data *conditional* on them remaining consistent with the observed data.

So far we have discussed the use of DA to deal with the issue of *censoring*, however the same process can be applied when there are missing data. In this case the missing data are simply augmented to the parameter vector and estimated as part of the MCMC algorithm. This can be done in continuous- or discrete-time (see e.g. Lekone and Finkenstädt, 2006 for an example of the latter).

The DA-MCMC routines used in this paper are straight-forward applications of the Metropolis–Hastings algorithm, although these generally involve searching over a very high dimensional space (since usually $\dim(\boldsymbol{\theta}, \boldsymbol{t}) \gg \dim \boldsymbol{\theta}$). The usual MCMC considerations apply; after initialisation, the chain needs to be monitored to ensure convergence to the stationary distribution. This can be especially difficult for DA-MCMC approaches, since there are so many dimensions in the expanded parameter space $(\boldsymbol{\theta}, \boldsymbol{t})$. Consequently, there are always doubts that the estimated posterior really does correspond to the true posterior; for instance, results from an early application of DA-MCMC to an epidemic of smallpox by O'Neill and Roberts (1999) differ from those found using an alternative DA method by Fearnhead and Meligkotsidou (2004), which the latter put down to the poor performance of MCMC when estimating heavy-tailed distributions. If, however, we satisfy ourselves that the chain has indeed converged then the initial samples (the *burn-in*) are discarded and the chain run until the required number of posterior samples has been obtained.

Two important aspects are the choice of prior and proposal distributions. The former characterises the degree of prior knowledge about the parameters, though a misinformed prior can lead to a posterior inconsistent with the data. A carefully chosen proposal plays an important role in controlling the acceptance rate and mixing of the chain. For some models, it is possible to use Gibbs's samplers (see e.g. Gamerman and Lopes, 2006) for parameter proposals, but to minimise differences between approaches and facilitate the fairest comparison possible, in all the examples in this paper we will use random-walk Gaussian proposals on the log-scale to ensure their positivity. We do not go into detail here about MCMC algorithms, for more detail the reader is referred to O'Neill (2002).

3.2 Approximate Bayesian Computation

ABC utilises the Bayesian paradigm in a simple and elegant manner. The theory behind the approach is that if $\mathcal{M}$ represents the model of interest, then the observed data $\mathbf{D}$ are simply one realisation from $\mathcal{M}$, conditional on its (unknown) parameters, $\boldsymbol{\theta}$. For a given set of candidate parameters, $\boldsymbol{\theta}'$, drawn from the prior, we can simulate a data set $\mathbf{D}'$ from $\mathcal{M}$. It follows that if $\mathbf{D}' = \mathbf{D}$ then $\boldsymbol{\theta}'$ is a draw from the posterior.

Early ABC routines (Tavaré et al., 1997) embedded these simulation steps into simple acceptance-rejection algorithms, in order to generate Monte Carlo estimates for the parameters. This involves matching the simulated to observed data at each step in the algorithm. However, in practice requiring an exact match ($\mathbf{D}' = \mathbf{D}$) tends to lead to very low acceptance rates (zero if the data are continuous). An alternative is to use a metric $\rho(\mathbf{D}', \mathbf{D})$ that measures *proximity* between $\mathbf{D}'$ and $\mathbf{D}$ (Plagnol and Tavaré, 2002). Matching to this will generate draws from the *approximate* posterior $f(\boldsymbol{\theta} \mid \rho(\mathbf{D}', \mathbf{D}) \leq \epsilon)$. In order to make our approximation worthwhile, we require a good metric to determine how ‘close’ the simulated and observed data sets are, and a tolerance that balances the accuracy of our approximation (which increases as $\epsilon \rightarrow 0$) with computationally manageable acceptance rates (which increase as $\epsilon \rightarrow \infty$). In practice it is often difficult to compare the full data set directly, in which case we can compare sufficient, or *approximately* sufficient statistics (Marjoram et al., 2003).

One issue with using acceptance-rejection algorithms in which samples are drawn from the prior is that they can be very inefficient if the prior and posterior are quite different, for instance if the data contain a lot of information. Two elegant extensions of these approaches that deal with this issue were developed by Marjoram et al. (2003) and Sisson et al. (2007a). The former embed the simulation steps into a standard Metropolis–Hastings MCMC routine (henceforth, ABC-MCMC), whilst the latter do the same for Sequential Monte Carlo routines (ABC-SMC). Both Toni et al. (2009) and (independently) Sisson et al. (2007b) (in an update of Sisson et al., 2007a on their website) produce more stable ABC-SMC routines, based on a more robust weighting scheme. A further interesting method was proposed by Beaumont et al. (2002), who use an acceptance-rejection algorithm but improve the accuracy of the resulting Monte Carlo approximations by weighting the contributions of each sample through kernel-weighted estimates of the corresponding metric values. They also reduce dimensionality of the summary statistics using local-linear regression.

In the next two sections we introduce the ABC algorithms of Marjoram

et al. (2003) and Toni et al. (2009). The SMC algorithms produced using the updated form of Sisson et al. (2007a) and Toni et al. (2009) are very similar, and we choose to report the slightly simpler version of Toni et al. (2009) here (see also Beaumont, 2008). We also introduce a further extension based on conditional simulation algorithms (Section 3.5.1).

3.3 ABC-MCMC

The basic ABC-MCMC routine is very similar to the standard Metropolis–Hasting’s algorithm. Over a series of iterations, parameters are sampled often from the vicinity of the values at the current iteration, with proposals being accepted with some probability, otherwise the current values are retained at the next iteration. Where it differs is in the calculation of the acceptance probability. Here we present a version in which this probability is the proportion of simulations that “match” the data.

Let $\mathbf{D}$ denote the matrix of observed data and $\boldsymbol{\theta}$ the vector of unknown parameters. Marjoram et al. (2003) then propose a variant of the following algorithm:

- A1. Set $i = 0$ and choose initial values $\boldsymbol{\theta}^{(0)}$.
- A2. Sample a candidate value, $\boldsymbol{\theta}'$, from a proposal distribution $q(\cdot)$.
- A3. Simulate R data sets, $\mathbf{D}'_1, \dots, \mathbf{D}'_R$, from model $\mathcal{M} \mid \boldsymbol{\theta}'$, and calculate

$$\hat{f}(\mathbf{D} \mid \boldsymbol{\theta}') = \frac{1}{R} \sum_{r=1}^R \mathcal{I}[\rho(\mathbf{D}'_r, \mathbf{D}) < \epsilon],$$

where $\rho(\mathbf{D}'_r, \mathbf{D})$ is a metric measuring ‘proximity’ of the simulated and observed data, and $\mathcal{I}(X) = 1$ if X is true, and 0 otherwise.

- A4. Calculate the acceptance probability α as:

$$\alpha = \min \left(1, \frac{\hat{f}(\mathbf{D} \mid \boldsymbol{\theta}') \pi(\boldsymbol{\theta}') q(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}')}{\hat{f}(\mathbf{D} \mid \boldsymbol{\theta}^{(i)}) \pi(\boldsymbol{\theta}^{(i)}) q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{(i)})} \right).$$

- A5. Sample u from a $U(0, 1)$ distribution and set $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}'$ if $u < \alpha$ and $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)}$ otherwise.
- A6. Set $i = i + 1$, go to A2 and repeat for the required number of samples.

This produces Monte Carlo estimates of the approximate posterior. See Section 4 for discussion on the choice of proposal distribution(s), $q(\cdot)$.

3.4 ABC-SMC

The idea underlying SMC methods for parameter estimation is that the posterior is not sampled directly but instead through a series of proxy distributions that converge to the posterior. Let us define a sequence of K distributions, $f_1, \dots, f_K$, where $f_1 > f_2 > \dots > f_K$, and f_K is the posterior distribution of interest, $f(\boldsymbol{\theta} \mid \mathbf{D})$. At the outset a large number of *particles* are randomly drawn from f_1 and then subjected to a series of sequential importance sampling steps, where at each stage the particles are weighted assuming that they are generated from the corresponding sequential distribution. One issue with SMC methods is the potential for particle degradation, in which only a very small number of particles with non-zero weight remain after a few stages of the algorithm. This can occur if the initial sampling distribution, f_1 , is very overdispersed compared to the posterior. To deal with this and to make the process more efficient, it is common to introduce a resampling stage, whereby if the sample degeneracy gets too high then particles can be resampled proportionately to their weight. A recent review of SMC methods can be found in Cappé et al. (2007), and other examples of these approaches can be found in Doucet et al. (2001); Liu (2001); Moral et al. (2006); Sisson et al. (2007a); Toni et al. (2009); Buckland et al. (2007) and Jégat et al. (2008).

SMC methods have a number of potential advantages over MCMC and simple acceptance-rejection sampling methods, which are explained succinctly in Sisson et al. (2007a) and Toni et al. (2009). Principally these include the fact that SMC methods produce uncorrelated samples, avoid getting stuck in areas of low probability, and help improve sampling by eliminating particles that do not represent the posterior well in favour of those that do. In an ABC setting the aim is to draw samples from the approximate posterior, $f(\boldsymbol{\theta} \mid \rho(\mathbf{D}', \mathbf{D}) < \epsilon)$, and so a natural way to specify the series of intermediate distributions is to implement decreasing metric tolerances (ϵ) at each stage. This also provides a good way of evaluating the tolerance level for the approximation, without having to re-run the entire chain (as would be necessary if using MCMC-type methods). This is an important point and will be returned to in the next section.

Following Sisson et al. (2007a), we can define a sequence of distributions $f_1, \dots, f_K$ as

$$f_k(\boldsymbol{\theta} \mid \rho(\mathbf{D}', \mathbf{D}) \leq \epsilon_k) = \frac{\pi(\boldsymbol{\theta})}{R_k} \sum_{r=1}^{R_k} \mathcal{I}[\rho(\mathbf{D}'_r, \mathbf{D}) \leq \epsilon_k].$$

where $\pi(\boldsymbol{\theta})$ is the prior as before, $\mathbf{D}'_1, \dots, \mathbf{D}'_{R_k}$ are R_k simulated data sets and $\{\epsilon_k\}$ are a strictly decreasing sequence of tolerances. If we have a fixed number

of repeats and tolerances at each stage, so $R_k = R$ and $\epsilon_k = \epsilon$, then we recover the approximate likelihood in step A3 of Algorithm A.

Let N denote the number of required particles and R denote the number of repeated simulations used to generate the approximate likelihoods. Toni et al. (2009) suggest the following algorithm:

- B1. Initialise $\epsilon_1, \dots, \epsilon_K$. Set population indicator $k = 1$.
- B2. Set particle indicator $i = 1$.
- B3. If $k = 1$, sample θ'' independently from $\pi(\theta)$.
If $k > 1$, sample θ' from the previous population $\{\theta_{t-1}^{(i)}\}$ with weights $\{W_{t-1}^{(i)}\}$, and perturb the particle to $\theta'' \sim Q_k(\theta \mid \theta')$ according to a Markov transition kernel Q_k .
- B4. If $\pi(\theta'') = 0$, return to B3.
- B5. Generate R data sets $\mathbf{D}_r'' \sim f(\mathbf{D} \mid \theta'')$.
If $\hat{f}(\mathbf{D} \mid \theta'') = (1/R) \sum_{r=1}^R \mathcal{I}(\rho(\mathbf{D}_r'', \mathbf{D}) < \epsilon_k) = 0$, then go to B3.
- B6. Set

$$\theta_t^{(i)} = \theta'' \quad \text{and} \quad W_k^{(i)} = \begin{cases} \hat{f}(\mathbf{D} \mid \theta_k^{(i)}) & \text{if } k = 1, \\ \frac{\hat{f}(\mathbf{D} \mid \theta_k^{(i)}) \pi(\theta_k^{(i)})}{\sum_{j=1}^N W_{k-1}^{(j)} Q_k(\theta_k^{(i)} \mid \theta_{k-1}^{(j)})} & \text{if } k > 1. \end{cases}$$

If $i < N$, increment $i = i + 1$ and go to step B3.

- B7. Normalise the weights so that $\sum_{i=1}^N W_k^{(i)} = 1$.
- B8. If $k < K$, increment $k = k + 1$ and go to B2.

In a similar manner to the proposal distribution for Metropolis–Hastings algorithms, the perturbation kernel, $Q_k(\cdot)$, effects the efficiency of the sampling mechanism. As with the ABC- and DA-MCMC proposal distributions we choose to use random-walk perturbation kernels here (see Section 4 for more details). In Section 3.5 we discuss various metric choices for the type of time-series data used in this paper, and in Section 4.1 we explore the accuracy and efficiency of the chains for different numbers of repeated simulations (R).

3.5 Choice of metric for temporal count data

In the first instance assume that fully observed discrete time series counts for each stage in the infection process are available, hence S_j , E_j , I_j and R_j correspond to the number of susceptible, exposed, infective and removed individuals at time $j = 0, \dots, T$, with $\mathbf{D} = (\mathbf{E}, \mathbf{I}, \mathbf{R})$ the matrix of observed data. (We are considering a closed population and choose to use only the latent, infective and removal curves, since $S_j + E_j + I_j + R_j = N \forall j$.)

It seems natural for time-series data to develop metrics based on differences between observed and simulated counts. In this case we could produce metrics based on aggregating over each of the available epidemic curves (latent, infective and removal), however each curve provides different information about the parameters in the model and how they interact. It seems more intuitive therefore to produce a series of metrics that match each time-series separately, and then to accept the simulation if each of the metrics lies within a pre-specified tolerance level. (Note that we need not use the same metric/tolerance for each curve, though we choose to here for simplicity.) If $\rho(\cdot)$ represents our metric of choice, then this approach would correspond to

$$\rho(\mathbf{D}'_r, \mathbf{D}) < \epsilon = \rho(\mathbf{E}'_r, \mathbf{E}) < \epsilon \text{ and } \rho(\mathbf{I}'_r, \mathbf{I}) < \epsilon \text{ and } \rho(\mathbf{R}'_r, \mathbf{R}) < \epsilon. \quad (3.1)$$

There are various ways to specify “differences” between two sets of counts. Treating each curve independently, one option would be to place an envelope around the data such that for a simulated epidemic curve to “match” the observed epidemic curve it must not deviate beyond the envelope at any point. A schematic diagram of this type of approach is given in Figure 1b. Here the solid points represent the observed data and the dashed lines the envelope (derived using an equal tolerance both above and below the data across the whole epidemic period). For an ABC-MCMC algorithm, the metric and tolerance are both fixed at the outset. For an ABC-SMC algorithm—depicted in Figure 1c—the tolerance is reduced sequentially as the algorithm progresses.

One potential issue here is that this metric might be quite sensitive to spurious single time-point deviations between the simulated and observed epidemics (which might be expected to be reasonably common in large-scale stochastic systems). An alternative is to match to *summary* statistics ($\mathbf{S}$), instead of each individual data point. An intuitive summary measure could be to produce a sum-of-squared differences between observed and simulated counts; or perhaps a sum-of-absolute differences. For the remainder of this paper we choose to use one based on a chi-squared goodness-of-fit criterion. This is very similar to the sums-of-squares approach, but with the contribution at each time point scaled by the observed data. This adjusts the contribution of each point along the

epidemic curve to reflect the fact that the variation changes as the epidemic progresses. For time-series counts $\mathbf{X} = (X_1, \dots, X_T)$ and $\mathbf{Y} = (Y_1, \dots, Y_T)$, this metric is given by:

$$\rho_{\chi}(\mathbf{X}, \mathbf{Y}) = \sum_{j=1}^T \frac{(X_j - Y_j)^2}{Y_j}. \quad (3.2)$$

The choice of metric is likely to be highly problem-specific, and it is not the purpose of this manuscript to optimise the choice of metric for this exact system. Nonetheless we tried out various metric choices based on the above discussion and got broadly similar degrees-of-accuracy when applied to the simulated epidemic data described in Section 4.1 (see Appendix A), with arguably the chi-squared and envelope metrics performing better given similar run times.

3.5.1 Composite ABC

An alternative, for situations in which we have discrete temporal count data and the model is such that the data satisfy the Markov property, is to decompose the likelihood into a product (two examples of such situations are described in Cook et al., 2008a). Here the contribution of each term is the likelihood of changing from one observation to the next. In such situations, it may be even more beneficial (in terms of improving acceptance rates) to think about constructing a *composite* likelihood approximation over the discrete time periods, because the likelihood per sequence of two observations is much higher than the product of likelihoods over all sequences. This motivates the following approximate composite likelihood (CABC) scheme.

Using the same notation as above we can replace steps A3 and B5 (above) with:

A3b/B5b. For each day, $j = 1, \dots, T$, generate R simulations from $\{\mathcal{M}|\boldsymbol{\theta}', \mathbf{D}_{j-1}\}$ in the period $[j-1, j)$. Then calculate

$$\hat{f}_j(\mathbf{D}_j | \boldsymbol{\theta}') = \frac{1}{R} \sum_{r=1}^R \mathcal{I}[\rho(\mathbf{D}'_{jr}, \mathbf{D}_j) < \epsilon],$$

where $\mathbf{D}_j$ is the observed data and $\mathbf{D}'_{jr}$ is the r^{th} set of simulated data in $[j-1, j)$. $\rho(\mathbf{D}'_{jr}, \mathbf{D}_j)$ is a metric measuring ‘proximity’ between the simulated and observed data. The full approximate likelihood is then given by:

$$\hat{f}(\mathbf{D} | \boldsymbol{\theta}') = \prod_{j=1}^T \hat{f}_j(\mathbf{D}_j | \boldsymbol{\theta}').$$

We can then specify the neighbourhood metric to be:

$$\rho(\mathbf{D}'_{jr}, \mathbf{D}_j) < \epsilon = |E'_{jr} - E_j| < \epsilon \text{ and } |I'_{jr} - I_j| < \epsilon \text{ and } |R'_{jr} - R_j| < \epsilon. \quad (3.3)$$

This approach is depicted in Figure 1d: the simulations within each observation window have initial conditions given by the data at the last observation time.

4 Application

In this section we will explore the utility of these various ABC approaches compared to a gold-standard of data-augmented MCMC, which avails an exact draw from the posterior. We apply the techniques to both simulated data and data from an outbreak of Ebola Haemorrhagic fever in the Democratic Republic of Congo in 1995. In order to assess the behaviour of the approximation techniques to different tolerances and to different degrees of missing data, we use simulated epidemics based on parameter values consistent with the 1995 outbreak.

4.1 An outbreak of Ebola Haemorrhagic Fever in the Democratic Republic of Congo (DRC)

Ebola Haemorrhagic Fever (EHF) is a highly pathogenic infectious disease caused by the Ebola virus. After large outbreaks in the DRC (Anon, 1978a) and Sudan (Anon, 1978b) in the late 1970's the disease re-emerged to cause large epidemics in the DRC in 1995 (Anon, 1995) and Uganda in 2000–2001 (Okware et al., 2002). Further large outbreaks have since been reported in the Republic of Congo between 2002–2003 (Formenty et al., 2003). Here we will use data from the 1995 outbreak in the DRC, which resulted in 316 cases with an 81% mortality rate (Khan et al., 1999).

This outbreak has been well-documented and since we are interested in the comparative performance of competing algorithms we will not go into detail about the epidemiology of the outbreak or the justification of the model. For these discussions, and the data, the reader is referred to Khan et al. (1999), Chowell et al. (2004) and Lekone and Finkenstädt (2006).

The data consist of two incomplete time-series: one for date-of-onset of clinical signs and the other for date-of-death. The former contains 291 cases and the latter 236 deaths, implying that there are at least 25 missing dates-of-onset and 80 missing dates-of-removal. In the case of Ebola it is reasonable

Figure 1: Schematic diagram of different model fitting algorithms: a) data-augmentation using MCMC, b) Approximate Bayesian Computation, c) sequential Monte Carlo based on decreasing tolerance levels, and d) ABC using composite simulations

to assume that the date-of-onset corresponds closely to the start of infectivity, since transmission is usually through contact with infected fluids secreted from the body. Since mortality is so high, it is also reasonable to assume date-of-death corresponds closely to the removal date. The net result is that we have partial infective and removal *incidence* curves, but because of the missing values we cannot directly reconstruct the actual numbers of infectives and removals at any point in time. In addition there is also a latent period associated with the infection process, for which no information exists in the data, other than that which can be inferred from the relative onset and removal counts. So any likelihood-based fitting mechanism will need to infer not only the missing dates-of-onset and death, but also the whole latent infection curve.

To this end, Lekone and Finkenstädt (2006) developed a stochastic discrete-time likelihood-based $\mathcal{SEIR}$ model in which the unobserved (discrete) event times are inferred using data-augmented MCMC. Here we extend their model to continuous-time, and develop a DA-MCMC algorithm in which all unobserved event times are inferred, as described in section 3.1.

The underlying model is described by (2.1), but with an additional term modelling the effect of control interventions. Usual response strategies for Ebola include quarantine of infected patients, increased surveillance, training of hospital staff in the use of better preventive measures, and the rapid cremation or burial of the dead (Chowell et al., 2004; WHO, 2008). Chowell et al. (2004) developed a deterministic model, which they fit using least-squares, and assume that the effect of control interventions is such that the transmission rate is constant up to the point of intervention, before decaying exponentially over time to some reduced rate. Lekone and Finkenstädt (2006) use the same basic form, but highlight that there is an identifiability issue with using two transmission rates in this system, so they simplify to allow the transmission parameter to decay to zero instead. Following this latter approach,

$$\beta_t = \begin{cases} \beta & t < t_{\text{int}} \\ \beta e^{-q(t-t_{\text{int}})} & t \geq t_{\text{int}}, \end{cases} \quad (4.1)$$

where t_{int} is the time point at which intervention measures are introduced.

For that reason, and the fact that they fit to daily counts of infection but weekly counts of removals, the parameter estimates of Chowell et al. (2004) are not directly comparable to the approaches explored in the next section; nonetheless the implications of their results will be covered in the discussion.

We choose to start our simulations from the date at which the first case was recorded (6th January). To facilitate between-model comparisons, we compute the posterior of the derived statistic, R_0 (Anderson and May, 1991), which represents the ‘average’ number of new cases per ‘average’ infected individual

in an otherwise fully susceptible population. In this case we let R_0 equal the transmission rate per unit time multiplied by the expected length of the infectious period (β/γ).

4.2 Simulation study

We fit initially to data simulated from model (2.1), with control interventions modelled as in (4.1), using (specified) parameter values that are representative of the those from the true outbreak (see Lekone and Finkenstädt, 2006): $S_0 = 5,364,500$, $I_0 = 1$, $\beta = 0.2$, $\delta = 0.2$, $\gamma = 0.143$, $q = 0.2$ and $t_{\text{int}} = 130$. The simulated epidemic ran for 166 days and resulted in 596 infections. The data were aggregated into daily counts of the numbers of latent, infective and removed individuals such that $\mathbf{D} = (\mathbf{E}, \mathbf{I}, \mathbf{R})$.

The simulated data were censored to varying degrees in order to compare the performance of the routines for data of varying quality: in the first instance we assume that we have observed counts for each stage of the disease cycle, deemed *complete* in the sense that we have counts for the latent, infective *and* removal curves. A continuous time DA-MCMC model can then be fitted in which the censored latent, infective and removal times are inferred from the data. We then employ a series of ABC routines, using metric (3.1), with $\rho(\cdot) = \rho_{\chi}(\cdot)$, given in (3.2), as well as the composite-likelihood approach (Section 3.5.1). For the ABC-MCMC algorithms we run multiple chains with different tolerances (ϵ) as well as varying the number of repeated simulations contributing to the Monte Carlo approximation of the likelihood (R). For the ABC-SMC routines the tolerance levels are varied as a direct component of the model fitting, and so we only need to repeat the full algorithm in order to compare the effect of changing R .

For the purposes of this comparison we used log-normal random-walk proposal distributions with variance of 0.2 for most of the MCMC-type routines—e.g. $\log(\theta') \sim N(\log(\theta), \sigma_q^2)$. We used uniform perturbation kernels for the SMC-type routines, such that $Q_k(\theta' | \theta) = \theta + \tau_q$, where $\tau_q \sim U(-\sigma_q, \sigma_q)$, with $\sigma_q = 0.05$. For the ABC routines we simulate from (2.1) using Gillespie's algorithm (Gillespie, 1977). It is worth noting that some of the DA-MCMC routines required additional tuning of the proposal distributions to aid mixing and convergence of the chain—but these situations are highlighted in the text.

We complete the specification of the model by choosing gamma priors for each of the unknown parameters, such that $\pi(\theta) = G(a, b)$, where $\theta = \beta, \delta, \gamma$ and q and $G(a, b)$ is a gamma density with mean ab and variance ab^2 . We use the same priors as Lekone and Finkenstädt (2006), centred around the true mean such that $\{(a_\theta, b_\theta); \theta = \beta, \delta, \gamma, q\} = \{(2, 0.1), (2, 0.1), (2, 0.07), (2, 0.1)\}$.

Summary results for the ABC- and DA-MCMC routines are shown in Table 1a. These posterior summaries are derived from 100 000 iterations (except the DA-MCMC, which was run for 50 000 iterations) following a 5000 sample burn-in. (Note that the acceptance rates for some of the single repeats were so low we decided not to report the results here.) Summary results for the ABC-SMC routines are shown in Table 1b, using 2000 independent particles.

Effect of tolerance level

The mean posterior estimates obtained using the chi-squared metric for the full data are similar to those obtained from the DA-MCMC routine, even for reasonably large tolerance levels. This is encouraging, and suggests that the information from the chi-squared metric, combined across the three epidemic curves, manages to capture the average dynamics of the epidemic well. Nonetheless, the choice of tolerance has a marked effect on the precision of the approximate posterior. For example, the posterior standard deviations are approximately 2–2.5 times larger when using a tolerance of 1000 than when using a tolerance of 300.

In this paper we have deliberately chosen to use a model for which we can compare the approximate posterior with the exact posterior obtained from DA-MCMC. For ABC methods to be practical, they must be able to stand alone. In this situation, some other method for assessing the choice of tolerance is necessary. For the ABC-MCMC routines this may require running multiple chains using decreasing tolerances each time. This additional tuning is not necessary for DA-MCMC algorithms, so although ABC-MCMC routines can be made more efficient for a given tolerance level, there is a necessary decrease in computational efficiency due to extra time required to conduct repeated runs to evaluate alternative tolerances.

However, the ABC-SMC framework provides a natural way to deal with this issue, since we define each intermediate distribution in the sequence based on decreasing tolerance values. In each case in Table 1b we chose tolerances that were uniformly distributed between the maximum and minimum values shown in the second column of the table. Defining the intermediate distributions in this way allows the choice of tolerance to be assessed without having to re-run entire chains (as would be necessary if using MCMC-based methods). In addition it is straightforward to keep reducing the tolerance as required.

Overall, if a good metric is available, then accurate point estimates can be obtained even when using fairly large tolerances. However the precision of the approximate posterior requires some additional tuning of the tolerance. To highlight these patterns for the ABC-MCMC approaches, we can plot the

marginal Kullback–Liebler divergence against the efficiency of the chain for the different parameters of the model (Figure 2—excluding the outlying results for the single repeats). This measures the relative entropy of the approximate posterior compared to the exact posterior obtained through DA-MCMC. It can be seen that the general trend across each of the plots is that higher tolerances result in higher information-loss compared to the exact posterior obtained from DA-MCMC, but the sampling is more efficient.

Effect of repeated simulations

We can see that for a *given* tolerance level, increasing the number of repeats used to generate the Monte Carlo estimate of the approximate likelihood does not seem to produce markedly more precise posteriors. There is however an increase in the acceptance rate of the ABC-MCMC chains (Table 1a) when using multiple repeats. In most practical applications of ABC techniques (see e.g. Marjoram et al., 2003; Sisson et al., 2007a and Toni et al., 2009) it has been argued that for maximum computational efficiency only single repeats should be used. Since autocorrelated chains are produced for the MCMC models, we can derive a measure of chain efficiency by calculating the effective sample size (ESS), which gives a measure for the number of pseudo-independent samples in the posterior. We can then divide the time taken to run the chain (after burn-in) by the ESS to give an average time taken to produce one pseudo-independent sample. In this case, where the simulation algorithm is fast, there seems to be advantages in terms of chain efficiency if multiple repeats are used, even though more data-generation steps are required at each stage. For other large-scale epidemic systems (e.g. individual-level, or spatial models), the additional processing power required to evaluate multiple repeated simulations is likely to outweigh the computational advantage when simulating compared to calculating the likelihood.

The ABC-SMC algorithm performs much better when only single repeated simulations are used. The re-sampling step in Algorithm B (step B3) means that particles that do not produce epidemics that match the data well are eliminated in favour of those that do (Sisson et al., 2007a; Toni et al., 2009). In fact there is little noticeable difference between the accuracy and precision of the approximate posteriors obtained from using single and multiple repeats. If ABC frameworks are going to be useful for parameterising large-scale systems, then ideally we require as few data generation steps as possible. In this case the ABC-SMC algorithm shows some advantage over the ABC-MCMC approach.

Algorithm	Tolerance (ϵ)	Repeats R	β	$1/\delta$	$1/\gamma$	q	R_0	Acceptance Rate (%)
True values	-	-	0.20 -	5.0 -	7.0 -	0.20 -	1.4 -	-
DA-MCMC	-	-	0.19 (0.01)	4.7 (0.20)	6.6 (0.28)	0.17 (0.02)	1.2 (0.07)	2
ABC-MCMC	1000	1	0.20 (0.03)	4.3 (1.1)	6.1 (0.95)	0.18 (0.11)	1.2 (0.09)	0.12
		100	0.21 (0.05)	4.2 (1.2)	6.3 (1.5)	0.25 (0.15)	1.3 (0.11)	10
		250	0.21 (0.05)	4.1 (1.3)	6.4 (1.5)	0.27 (0.19)	1.3 (0.12)	18
		500	0.20 (0.04)	4.3 (1.2)	6.5 (1.4)	0.24 (0.13)	1.3 (0.11)	16
	500	1	0.21 (0.03)	4.2 (0.67)	6.2 (1.2)	0.34 (0.09)	1.3 (0.08)	0.02
		100	0.20 (0.03)	4.3 (0.74)	6.4 (0.95)	0.23 (0.11)	1.3 (0.09)	2.4
		250	0.20 (0.03)	4.3 (0.76)	6.4 (0.93)	0.24 (0.14)	1.3 (0.09)	5.3
		500	0.20 (0.03)	4.4 (0.72)	6.4 (0.90)	0.24 (0.10)	1.3 (0.09)	4.7
	300	1	- -	- -	- -	- -	- -	-
		100	0.19 (0.02)	4.4 (0.54)	6.5 (0.64)	0.19 (0.06)	1.2 (0.08)	0.18
		250	0.19 (0.02)	4.5 (0.49)	6.6 (0.65)	0.22 (0.09)	1.3 (0.09)	0.39
		500	0.19 (0.02)	4.5 (0.48)	6.5 (0.60)	0.21 (0.09)	1.3 (0.08)	0.80
CABC-MCMC	10	100	0.11 (0.05)	8.0 (3.0)	9.0 (2.3)	0.22 (0.16)	0.98 (0.39)	47
		250	0.10 (0.05)	8.5 (3.3)	9.3 (2.4)	0.21 (0.17)	0.93 (0.42)	53
		500	0.12 (0.04)	7.5 (3.4)	8.7 (2.0)	0.22 (0.13)	1.0 (0.36)	39
	5	100	0.17 (0.02)	5.1 (0.43)	6.8 (0.53)	0.21 (0.08)	1.2 (0.12)	1.1
		250	0.17 (0.02)	5.1 (0.56)	6.9 (0.64)	0.22 (0.10)	1.2 (0.15)	13
		500	0.17 (0.02)	5.1 (0.47)	6.8 (0.54)	0.21 (0.08)	1.2 (0.12)	5.1

(a) ABC-MCMC

Algorithm	Tolerance (ϵ)	Repeats R	β	$1/\delta$	$1/\gamma$	q	R_0	Acceptance Rate
ABC-SMC	1000-300	1	0.19 (0.02)	4.4 (0.42)	6.4 (0.59)	0.18 (0.04)	1.2 (0.08)	-
		100	0.20 (0.02)	4.4 (0.44)	6.4 (0.58)	0.18 (0.04)	1.2 (0.08)	-
		250	0.19 (0.02)	4.4 (0.42)	6.4 (0.60)	0.18 (0.04)	1.2 (0.08)	-
		500	0.20 (0.02)	4.4 (0.42)	6.4 (0.58)	0.18 (0.04)	1.2 (0.08)	-
CABC-SMC	10-3	100	0.20 (0.01)	4.7 (0.24)	6.6 (0.25)	0.18 (0.02)	1.3 (0.08)	-
		250	0.19 (0.01)	4.7 (0.26)	6.5 (0.36)	0.19 (0.03)	1.3 (0.09)	-
		500	0.19 (0.01)	4.7 (0.23)	6.5 (0.31)	0.19 (0.03)	1.2 (0.08)	-

(b) ABC-SMC

Table 1: (Approximate) posterior means (standard deviations in parentheses) from routines fitted to *complete* simulated data using competing exact and approximate routines. Note that the benchmarks for estimates obtained from ABC-MCMC are not the true parameter values but those estimated exactly using DA-MCMC. β is the per capita² infection rate, $1/\delta$ the mean latent period, $1/\gamma$ the mean infectious period, q governs the intervention effect and R_0 is the derived quantity, the pre-intervention basic reproduction number. For the ABC-SMC approaches, the tolerances defining the intermediate distributions are uniformly distributed in the range shown.

Figure 2: Marginal Kullback–Liebler divergence of approximate marginal posteriors generated from ABC-MCMC compared to the exact posterior obtained from DA-MCMC (shown by dashed lines). Darker colours represent higher tolerances, and for a given tolerance three points are shown—corresponding to generating the approximate likelihood ratio using 500, 250 and 100 independent simulations. Plots are of the divergence in a) transmission rate, b) incubation period, c) infectious period and d) intervention effect. Summaries of the posteriors for this data set are shown in Table 1.

Tolerance (ϵ)	β	$1/\delta$	$1/\gamma$	q	R_0
922	0.22 (0.04)	4.2 (1.1)	6.1 (1.2)	0.65 (0.52)	1.3 (0.10)
844	0.21 (0.04)	4.3 (0.96)	6.1 (1.1)	0.49 (0.37)	1.3 (0.10)
767	0.22 (0.03)	4.4 (0.85)	6.1 (1.0)	0.42 (0.28)	1.3 (0.10)
689	0.21 (0.03)	4.3 (0.76)	6.1 (0.97)	0.38 (0.23)	1.3 (0.09)
611	0.21 (0.03)	4.3 (0.70)	6.2 (0.92)	0.35 (0.19)	1.3 (0.09)
533	0.21 (0.03)	4.3 (0.66)	6.2 (0.86)	0.32 (0.16)	1.3 (0.09)
456	0.20 (0.02)	4.3 (0.57)	6.3 (0.79)	0.30 (0.14)	1.3 (0.09)
378	0.20 (0.02)	4.4 (0.50)	6.3 (0.67)	0.26 (0.11)	1.3 (0.09)
300	0.20 (0.02)	4.4 (0.44)	6.4 (0.59)	0.23 (0.08)	1.3 (0.08)

Table 2: ABC-SMC results for simulated data, using a chi-squared metric and uninformative $G(1, 1)$ priors, for a sequence of decreasing tolerance levels from 1000 to 300. Comparative posteriors generated using informative priors are shown in Table 1b.

Method	Tolerance (ϵ)	β	$1/\delta$	$1/\gamma$	q	R_0
DA-MCMC	-	0.18 (0.01)	5.7 (1.2)	6.6 (0.27)	0.17 (0.06)	1.3 (0.08)
ABC-SMC (Inf.)	800–200	0.19 (0.01)	5.4 (1.04)	6.5 (0.48)	0.20 (0.05)	1.2 (0.08)
ABC-SMC (Uninf.)	800–200	0.20 (0.02)	6.0 (1.32)	6.5 (0.45)	0.36 (0.14)	1.3 (0.08)

(a) Missing latent curve

Method	Tolerance (ϵ)	β	$1/\delta$	$1/\gamma$	q	R_0
DA-MCMC	-	0.18 (0.01)	5.1 (0.84)	6.6 (0.27)	0.19 (0.06)	1.2 (0.07)
ABC-SMC (Inf.)	400–267, 100–40, 100–40	0.22 (0.04)	4.6 (0.93)	5.9 (1.0)	0.23 (0.06)	1.3 (0.09)
ABC-SMC (Uninf.)	400–267, 100–40, 100–40	0.30 (0.09)	5.4 (1.4)	4.7 (1.4)	0.41 (0.18)	1.3 (0.09)

(b) Missing latent curve and infective/removal *incidence* only

Table 3: Comparative results for DA-MCMC and ABC-SMC for simulated data with a) missing latent curve, and b) missing latent curve and infective/removal incidence only. The approximation routines use the chi-squared metric with both informative and uninformative priors. For the ABC-SMC approaches, the tolerances defining the intermediate distributions are uniformly distributed in the range shown, and are in the order: a) ϵ_X , and b) $\epsilon_X, \epsilon_R, \epsilon_T$.

Figure 3: Sequential approximate posteriors obtained from an ABC-SMC algorithm for simulated data, using a chi-squared metric and uninformative $G(1, 1)$ priors, for a sequence of decreasing tolerance levels. Plots are for a) transmission rate, b) incubation period, c) infectious period and d) intervention effect. Summary results are given in Table 2.

Choice of metric

In the analysis (for which results are shown in Table 1b) we use informative priors. However, to explore the validity of the chi-squared metric for various tolerances we can repeat the model fitting using relatively (with respect to the likelihood) uninformative $G(1, 1)$ priors for each of the parameters. These results are shown in Table 2, and a plot showing the evolution of the approximate posteriors using ABC-SMC is shown in Figure 3. As the tolerance decreases over the sequence, the approximate posteriors begin to converge. Furthermore, we can see that the intervention effect (q) is more sensitive to the tolerance level than the other parameters, suggesting that the metric does not capture this aspect of the system dynamics as well when the tolerance is large. Nonetheless it does move rapidly away from the prior (which is similar to the red line—tolerance of 922—on Figure 3d) as the tolerance decreases and begins to stabilise at lower levels.

The results for the composite likelihood (Section 3.5.1) are also encouraging. In practice we were unable to obtain high enough acceptance rates for the CABC-MCMC when using tolerances of less than 5 for 100 repeats, and using single repeats proved highly inefficient for both CABC-MCMC and CABC-SMC approaches. Nonetheless the accuracy and precision of the approximate posteriors are arguably better than those produced by the chi-squared metric (though we need to exercise caution, since the ESSs are low compared to those produced using the chi-squared metric).

The results in Tables 1a and 1b suggest that the chi-squared metric is good at capturing the dynamics of the epidemic system when we have informative data. However, in real-life situations this level of detail is rarely available, and so instead we choose to look at what happens when we drop information from the data set. In the first instance we drop the latent curve only, but assume that we have fully observed infective and removal curves (Table 3a). In addition we also examine the effect of matching to infective and removal *incidence* only (Table 3b).

Due to its various advantages over ABC-MCMC, for the remainder of the paper we adopt ABC-SMC only, using single repeats ($R = 1$). Also, since we no longer know the exact state of the system at each time point, we can not use CABC, and so from now on we concentrate on the performance of the chi-squared metric only. We found that this worked reasonably well when we dropped the latent curve, though the approximate posterior standard deviation for the length of the infectious period ($1/\gamma$) is twice that from the DA-MCMC routine. When fitted to *incidence* data the chi-squared metric alone was not enough to fully capture the disease dynamics. However, an ad-

vantage with ABC techniques is that we can easily extend the system to use as much information as is available, and so we introduced further constraints by forcing the simulations additionally to match to the final epidemic size and the date of the final removal. This gives the metric for incidence data as:

$$\begin{aligned} \rho(\mathbf{D}'_r, \mathbf{D}) < \epsilon &= \rho_\chi(\mathbf{I}'_{\text{inc}}, \mathbf{I}_{\text{inc}}) < \epsilon_\chi \\ &\text{and } \rho_\chi(\mathbf{R}'_{\text{inc}}, \mathbf{R}_{\text{inc}}) < \epsilon_\chi \\ &\text{and } \rho_R(R'_{T'}, R_T) < \epsilon_R \\ &\text{and } \rho_T(T', T) < \epsilon_T, \end{aligned} \quad (4.2)$$

where for time-series counts $\mathbf{X} = (X_1, \dots, X_T)$ and $\mathbf{Y} = (Y_1, \dots, Y_T)$,

$$\rho_\chi(\mathbf{X}, \mathbf{Y}) = \sum_{j=1}^T \frac{(X_j - Y_j)^2}{Y_j}.$$

If T' and T are the simulated and observed dates of final removal respectively, and $R'_{T'}$ and R_T are the simulated and observed final epidemic sizes respectively, then,

$$\begin{aligned} \rho_R(R'_{T'}, R_T) &= |R'_{T'} - R_T|, \text{ and} \\ \rho_T(T', T) &= |T' - T|. \end{aligned}$$

Here $\mathbf{I}_{\text{inc}}$ and $\mathbf{R}_{\text{inc}}$ are the infective and removal *incidence* data from $t = 1, \dots, T$, and ϵ_χ , ϵ_R and ϵ_T correspond to the tolerance levels for each component of the metric (the incidence curves, final epidemic size and date of final removal respectively).

It can be seen that information is lost when using incidence data (Table 3b), particularly since the infectious period and intervention effect—which are highly correlated with each other—are less precise than before. This is perhaps not surprising because we now have no explicit information in the data regarding the latent curve, nor do we have information about actual numbers of infected or removed individuals at any time point. However, the estimates are still reasonable, and reflect the additional uncertainty inherent to the fact that we have a large degree of missing data.

It is worth noting that due to the missing data the DA-MCMC routines in Table 3b required additional tuning of the proposal distributions in order to achieve reasonable convergence and mixing. This was done by running one chain using independent normal proposals (with re-sampling to ensure positivity), and then running further chains using multivariate normal proposals, building in the correlation structure from the previous run (scaled by some factor < 1). Usually 2–3 re-estimations of the correlation structure were sufficient, and we used 1 000 000 burn-in iterations and a further 500 000 updates.

Method	Tolerance	β	$1/\delta$	$1/\gamma$	q	R_0
DA-MCMC	-	0.25 (0.03)	12.0 (1.2)	8.1 (0.92)	0.37 (0.17)	2.0 (0.21)
ABC-SMC	400–230, 40–5, 40–5	0.17 (0.03)	7.0 (1.8)	9.1 (1.8)	0.18 (0.05)	1.5 (0.17)
Lekone and Finkenstädt (2006)	-	0.24 (0.02)	9.4 (0.62)	5.7 (0.55)	0.16 (0.009)	1.4 (0.13)
Chowell et al. (2004)	-	0.33 (0.06)	5.3 (0.23)	5.6 (0.19)	-	1.8 (0.06)

Table 4: (Approximate) posterior means (standard deviations in parentheses) from routines fitted to the 1995 DRC Ebola outbreak data. For the ABC-SMC approach, the tolerances defining the intermediate distributions are uniformly distributed in the range shown, and are in the order: $\epsilon_\chi, \epsilon_R, \epsilon_T$.

4.3 1995 DRC Ebola outbreak

Overall the epidemic lasted 191 days from the first indexed case. Data are available from day 55 onwards and control interventions were introduced at around day 126 of the outbreak. We use metric (4.2), matching to final epidemic size and date, as well as a chi-squared metric on infective and removal incidence. The effect of the barren recording period at the outset introduces a further layer of missing information as we can only generate the chi-squared metric from day 55 onwards.

The results from fitting a full continuous-time DA-MCMC routine, as well as those from an ABC-SMC algorithm, applied to the real 1995 DRC Ebola outbreak data are given in Table 4, alongside estimates produced by Lekone and Finkenstädt (2006) and Chowell et al. (2004). For the DA-MCMC we used a burn-in of 1 000 000 iterations with a further 500 000 updates, and the proposal distributions had to be tuned in the same way as for the incomplete simulated data. The ABC-SMC routine used 2000 particles and single repeated simulations. We assume the first case became infective on the 1st January and that control interventions were introduced on 9th May (123 days afterwards) following Lekone and Finkenstädt (2006).

Interestingly the mean posterior estimate of R_0 from the continuous-time DA-MCMC routine is higher than both the discrete DA-MCMC and the ABC-SMC methods, and is closer to that of the deterministic model of Chowell et al. (2004). The three stochastic approaches produce arguably similar transmission rates, but the continuous-time DA-MCMC produces a longer incubation period, and there is even more uncertainty about the intervention effect than in either the ABC-SMC or discrete-time model. What is clear is that the sheer number of missing values in this data set drives up the inherent uncertainty in the system, with each of the four approaches in Table 4 settling around a slightly different configuration of interactions between the parameters.

One way to decide whether the posteriors are reasonable is to assess the

Figure 4: Posterior predictive distributions for removal incidence, for a) continuous-time DA-MCMC and b) ABC-SMC algorithm using metric (4.2). Summaries of posteriors for individual parameters are shown in Table 4. Credible intervals span 95% to 5%.

Figure 5: Posterior predictive distributions for a) final removal time, b) removal incidence for continuous-time DA-MCMC, and c) removal incidence for ABC-SMC using metric (4.2). The latter plots are *conditional* on an epidemic lasting > 120 days. Summaries of posteriors for individual parameters are shown in Table 4. Credible intervals span 95% to 5%.

goodness-of-fit of the models. Figure 4 shows the posterior *predictive* distributions for the removal incidence curves, for the continuous-time DA-MCMC and the ABC-SMC algorithms, generated from the full posterior with initial conditions corresponding to the first datum. The predictive posterior is highly skewed in both cases, with the mean generated from the DA-MCMC posteriors matching the data better than that from the ABC-SMC method. In both cases the medians and modes are almost flat compared to the observed curves.

The reason for this apparent disparity is due to the fact that there is a high probability of early extinction—caused by a combination of a low R_0 (≈ 1.5 – 1.6) and a low number of initial infectives (1 in this case). The predictive posterior distributions for the date of final removal—Figure 5a—are strongly bimodal, with ≈ 50 – 60% dying out within the first 60 days, and the remainder mostly in the period 120–250 days. Figure 4 suggests that based on this model and these data, sporadic importations of the Ebola virus from its wildlife reservoir rarely lead to sustained outbreaks, for the most part causing short-term, small-scale epidemics. The moderately-sized outbreak in the DRC in 1995 was one of those invasions that were successful.

The posterior predictive distributions presented in Figure 4 give useful information about the behaviour of the system, but the bimodality exhibited in Figure 5a makes them less useful as a model diagnostic tool, because in this case we know that the epidemic did take-off. To address the question of whether the data are consistent with predictions of successful invasions, we plot the posterior predictive distributions for the final epidemic curves, *conditional* on the epidemics taking-off. Our criterion for a successful outbreak is one lasting at least 120 days. The conditional trajectories are shown in Figures 5b and 5c, in which it can be seen that the data fall well within the range of plausible outbreaks for both the DA-MCMC and ABC-SMC algorithms.

5 Discussion

Central to the methodology discussed in this paper is the concept of estimation without likelihoods; instead a mathematical model of the underlying process is used to generate simulated data sets from which approximate likelihood ratios can be obtained. Incorporating the simulation steps into an MCMC or SMC algorithm, and using a metric to define relative degrees of ‘proximity’ between observed and simulated data sets, allows an *approximate* posterior distribution of the epidemic parameters to be obtained. Although there is an additional layer of uncertainty due to the approximation, these are arguably more informative than least-squares, an approximation in itself which is commonly

used for modelling these sorts of epidemic systems, and which conflates the observation process with population stochasticity. ABC methods also have the potential to be generalised to more complex systems where the concept of least-squares has little or no meaning (e.g. spatial/individual-level models).

We have demonstrated that the simulation-based approach works well in cases where the data are fully observed, at least in the sense of having regular (e.g. day-to-day) discrete time series counts for each of the stages of the lifecycle of the disease. We have also shown that they can provide reasonable parameter estimates in the much more common and challenging situation in which the observed data are not complete. Here likelihood-based approaches to parameter estimation would require data-augmentation procedures that can be difficult to implement efficiently. These simulation methods provide a more straightforward approach for tackling this problem. However, as with all MCMC/SMC routines, there are various computational and interpretation issues to be aware of, which may be exacerbated by the addition of an approximation to the process. For the MCMC routines there is the issue of convergence and mixing of the chains, which is affected not only by the proposal variance, but also the choice of metric and tolerance. For the SMC-type methods a choice of inappropriate metric and/or tolerance can result in particle degradation, although this can be mediated by altering the variance of the permutation kernel.

The first main issue to note is that of the ‘sufficiency’ of the metric used to match the simulated data sets with the observed data sets. The natural stochasticity in the simulation process, coupled with a large data set, means that requiring a precise match of the simulated and observed data sets leads to unacceptable acceptance rates. Furthermore, the censored nature of outbreak data prohibits the calculation of sufficient statistics. Nevertheless, it has been shown (Cook et al., 2008b) that, for non-spatial models at least, surprisingly low dimensional summaries of epidemic data are almost as informative as full epidemic data. This implies that *approximately* sufficient statistics, which capture the most salient aspects of the data in a small number of dimensions, could be used instead of the full data set. Nonetheless, as expected the uncertainty increases as information-loss in the data increases. Care must be taken to monitor convergence through the use of decreasing tolerance levels and post-hoc model checking procedures. Even metrics that capture the *data* well may not capture the *dynamics* well if the data are not informative. This can be seen most clearly in the analysis of the 1995 DRC Ebola outbreak, where even though the chi-squared metric captured the data well when the data were more informed, it was less successful when there were large amounts of missing data and only incidence curves were recorded. In the real Ebola

outbreak the data was even less complete due to the initial barren recording period. In this case we found that matching by the final epidemic size and date of the final removal in addition helped to mediate the problem, though it did result in more uncertainty about the approximate posteriors.

Another key issue with ABC methods is the choice of tolerance. Care needs to be taken to ensure that the tolerance is small enough that the metric is representative of the observed data, but large enough to ensure that there are a reasonable number of acceptances. Given a metric that accurately captures the dynamics of the system in question then ABC-SMC provides a useful way to “tune” the tolerances. In practice we found that there was a reasonably sharp cut-off between those tolerances that provided reasonable acceptance rates and those that produced very low acceptance rates, and the utility of ABC-SMC to allow a gradual decrease in tolerance levels provides a convenient way to track the impact of tolerance changes. In addition, it also allows the approximate posteriors to be monitored for different tolerance levels, giving information on the sensitivity and convergence of the approximate posteriors.

The choice of metric is also important, since we have shown that if a metric is available that captures the dynamics of the system well, then we obtain good point estimates even for “large” tolerances. On the other hand, if the metric does not capture the data well, or the amount of information in the data is limited (as is often the case for epidemic models, such as the 1995 DRC Ebola outbreak), then the accuracy of the parameter estimates will be compromised. However, if some aspects of the epidemic are simply not informed by limited data, exact (e.g. DA-MCMC) routines will also be affected—although possibly to a lesser degree—as can be seen in Table 4, and the problem is therefore more a modelling issue than a parameter estimation one.

Overall we found that ABC-SMC methods show more utility than ABC-MCMC. In particular the ABC-SMC routines allow a natural system for assessing the choice of tolerance level, and the convergence of the approximate posterior. In addition they were more efficient at sampling. For the MCMC routines we found that multiple simulations performed far better than single simulations in terms of both accuracy *and* efficiency, to the point where we got negligible acceptance rates when using single repeats for lower tolerance levels. The ABC-SMC routines on the other hand performed markedly better, and there was negligible difference between single and multiple repeats in terms of accuracy and precision, and the single repeats were much faster. This is because the ABC-SMC algorithm eliminates particles from areas of low probability mass through a re-sampling step at each stage.

The techniques described in this paper are intended as a possible alternative to likelihood-based inference when likelihoods are hard to generate or

computationally infeasible, as could be quite reasonably expected for very large scale disease outbreaks (Deardon et al., 2009). As far as possible we have tried to compare “vanilla” code. It would be possible to improve efficiency of the DA-MCMC algorithm by optimising the code so that parts of the likelihood are saved between iterations to prevent the computational cost of having to re-calculate. However, this process is complex, and gets much more complicated for larger-scale, more complex models (see, for example, Deardon et al., 2009). The simulation-based routines on the other hand are much simpler to code and optimise. They also incorporate natural model stochasticity as well as parameter uncertainty into the estimation process, produce full (albeit approximate) posterior distributions for the parameters and lend themselves naturally to parallel computation.

An advantage of the Bayesian framework is the ability to be able to ‘force’ identifiability onto various parameters, whilst still acknowledging parameter uncertainty, through the use of an informative prior distribution. This is particularly appealing for epidemic data, in which within-host dynamics are far better understood than between-host transmission rates. In cases where some aspects of an epidemic process are unobserved (such as in the case of the 1995 DRC Ebola outbreak), this can allow parameters for other aspects of the process to be estimated conditional on the choice of the identifiability-forcing prior. Epidemic parameters are often highly correlated, and using epidemiological and biological knowledge to fix reasonable priors is a sensible way to help with identifiability, and if interpreted correctly can help to make inference about infection parameters in an incompletely observed system. In these sorts of situations the use of ABC approaches are useful since they allow us to produce ‘approximate’ Bayesian posteriors in difficult circumstances.

Current on-going work involves the extension of these techniques to more complex systems, notably individual-level (spatial) models. An important issue is how to develop metrics that capture the dynamics of an epidemic effectively. For temporal count data we have shown that it is possible to capture a large degree of the variability by conditioning on the final epidemic size and date, however this may not be sufficient when dealing with a heterogeneous (and/or spatial) population, where information on individual (or spatial) risk would need to be captured. It remains to be seen how effective ABC methods would be in that situation. The use of kernel density based estimates of individual likelihood show promise, but challenges remain in combining approximate likelihood contributions across large numbers of individuals (particularly if data are missing), and/or developing spatio-temporal metrics to compare simulated and observed data. Where possible the use of data-augmentation and MCMC remains the preferred option.

Metric	Tolerance (ϵ)	β	$1/\delta$	$1/\gamma$	q	R_0	Initial time (s)	Total time (s)
Chi-squared	1000–300	0.20 (0.01)	4.5 (0.35)	6.5 (0.47)	0.25 (0.09)	1.3 (0.07)	553	1750
Absolute difference	6100–1300	0.19 (0.03)	4.6 (1.1)	6.8 (1.0)	0.32 (0.11)	1.3 (0.07)	574	2071
Sum-of-squares	220000–17000	0.20 (0.03)	4.7 (0.99)	6.6 (0.93)	0.36 (0.16)	1.3 (0.07)	510	1904
Envelope	55–25	0.20 (0.02)	4.6 (0.65)	6.3 (0.65)	0.31 (0.12)	1.3 (0.07)	548	1924

Table 5: (Approximate) posterior means (standard deviations in parentheses) from routines fitted to simulated data using ABC-SMC with various metrics. The tolerances defining the intermediate distributions are uniformly distributed in the range shown.

Appendix A: Comparison of metric choices

Table 5 shows the results for various metrics, based on: sums-of-squared and sums-of-absolute differences, as well as the chi-squared metric and one that uses an envelope around the data, all of which are discussed in Section 3.5. In order to make the comparison between metrics as fair as possible we used 10 intermediate distributions using tolerances uniformly-distributed across the ranges shown. We chose the ranges in order to standardise as best as possible the initial sampling of the 2000 particles (i.e. using the largest tolerances) and then the total run time (after all the intermediate stages). It is worth noting that we could reduce the tolerances further to produce better absolute fits, but wanted here to explore which metric seemed to better capture the data after similar run times.

We can see that they all produce similar point estimates, with the sum-of-absolute differences and sum-of-squares metrics producing arguably less precise approximate posteriors given similar run times. This is by no means a robust comparison, and we do not claim that the chi-squared metric is the optimal choice, rather that it is intuitive and worked well, and certainly none of the others seemed to work markedly better. Hence for the purposes of this paper we decided to stick with this metric and explore the other aspects of the system such as the effect of tolerances and repeated simulations.

References

- Anderson, R. and May, R. (1991), *Infectious Diseases of Humans*, Oxford University Press.
- Andersson, H. and Britton, T. (2000), *Stochastic Epidemic Models and their Statistical Analysis*, New York: Springer.

- Anon. (1978a), “Ebola haemorrhagic fever in Sudan, 1976. Report of a WHO/international study team,” *Bull World Health Organ*, 56, 247–270.
- Anon. (1978b), “Ebola haemorrhagic fever in Zaire, 1976. Report of a WHO/international study team,” *Bull World Health Organ*, 56, 271–293.
- Anon. (1995), “Ebola haemorrhagic fever in Zaire,” *Wkly Epidemiol Rec*, 70, 241–242.
- Bailey, N. T. (1975), *The Mathematical Theory of Infectious Diseases*, Charles Griffin and Company Ltd., London and High Wycombe, 2nd ed.
- Beaumont, M., Zhang, W., and Balding, D. (2002), “Approximate Bayesian Computation in population genetics,” *Genetics*, 162, 2025–2035.
- Beaumont, M.A. (2008), “A note on the ABC-PRC algorithm of Sisson *et al.* (2007),” <http://arxiv.org/abs/0805.3079v1>.
- Becker, N. (1976), “On a general stochastic epidemic model,” *Theoretical Population Biology*, 11, 23–26.
- Becker, N. (1989), *Analysis of Infectious Disease Data*, Chapman and Hall/CRC.
- Bortot, P., Coles, S., and Sisson, S. (2007), “Inference for stereological extremes,” *Journal of the American Statistical Association*, 102, 84–92.
- Buckland, S. T., Newman, K. B., Fernández, C., Thomas, L., and Harwood, J. (2007), “Embedding Population Dynamics Models in Inference,” *Stat Sci*, 22, 44–58.
- Cappé, O., Godsill, S., and Moulines, E. (2007), “An overview of existing methods and recent advances in Sequential Monte Carlo,” *Proceedings of the IEEE*, 95, 899–924.
- Cauchemez, S. and Ferguson, N. M. (2008), “Likelihood-based estimation of continuous-time epidemic models from time-series data: Application to measles transmission in London,” *Journal of the Royal Society Interface*, 5, 885–897.
- Chowell, G., Hengartner, N., Castillo-Chavez, C., Fenimore, P., and Hyman, J. (2004), “The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda,” *Journal of Theoretical Biology*, 229, 119–126.

- Chowell, G., Nishiura, H., Bettencourt, L.M.A. (2007), “Comparative estimation of the reproduction number for pandemic influenza from daily case notification data”, *J Roy Soc: Interface*, 4, 155–166.
- 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.
- Cook, A., Gibson, G., Gottwald, T., and Gilligan, C. (2008a), “Constructing the effect of alternative intervention strategies on historic epidemics,” *Journal of the Royal Society: Interface*, 5, 1203–1213.
- Cook, A. R., Gibson, G. J., and Gilligan, C. A. (2008b), “Optimal observation times in experimental epidemic processes,” *Biometrics*, 64, 860–868.
- Deardon, R., Brooks, S. P., Grenfell, B. T., Keeling, M. J., Tildesley, M. J., Savill, N. J., Shaw, D. J., and Woolhouse, M. E. (2009), “Inference for individual level models of infectious diseases in large populations,” *to appear in Statistica Sinica*.
- Diggle, P. (2006), “Spatio-temporal point processes, partial likelihood, foot-and-mouth disease,” *Statistical Methods in Medical Research*, 15, 325–336.
- Diggle, P. and Gratton, R. (1984), “MonteCarlo methods of inference for implicit statistical models (with discussion),” *Journal of the Royal Statistical Society. Series B*, 46, 193–227.
- Doucet, A., Freitas, N. D., and Gordon, N. (eds.) (2001), *Sequential Monte Carlo Methods in Practice*, Springer.
- 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.
- Filipe, J. and Gibson, G. (2001), “Comparing approximations to spatio-temporal models for epidemics with local spread,” *Bulletin of Mathematical Biology*, 63, 603–624.
- Formenty, P., Libama, F., Epelboin, A., Allaranger, Y., Leroy, E., Moudzeo, H., Tarangonia, P., Molamou, A., Lenzi, M., Ait-Ikhlef, K., Hewlett, B., Roth, C., Grein, T., and *l’equipe de lutte contre l’epidemie d’Ebola, au Congo*. (2003), “L’epidemie de fièvre hémorragique à virus Ebola en République du Congo, 2003: une nouvelle stratégie?”, *Med Trop*, 63, 291–295.

- Gamerman, D. and Lopes, H. F. (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, CRC Press, 2nd ed.
- 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.
- Gillespie, D. T. (1977), “Exact stochastic simulation of coupled chemical reactions,” *The Journal of Physical Chemistry*, 81, 2340–2361.
- Haydon, D., Chase-Topping, M., Shaw, D., Matthews, L., Friar, J., Wilesmith, J., and Woolhouse, M. (2003), “The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak,” *Proceedings Of The Royal Society Of London Series B - Biological Sciences*, 270, 121–127.
- Höhle, M., Jørgensen, E., and O’Neill, P. D. (2005), “Inference in disease transmission experiments by using stochastic epidemic models,” *Applied Statistics*, 54, 349–366.
- Ionides, E., Bretó, C., and King, A. (2006), “Inference for nonlinear dynamical systems,” *Proceedings of the National Academy of Sciences USA*, 103, 18438–18443.
- Jègat, C., Carrat, F., Lajaunie, C., and Wackernagel, H. (2008), “Early Detection and Assessment of Epidemics by Particle Filtering”, in Soares, A., Pereira, M. J. and Dimitrakopoulos, R. (eds) *geoENV VI – Geostatistics for Environmental Applications*, 16, 22–35, Springer Netherlands.
- Keeling, M. J. and Rohani, P. (2008), *Modeling Infectious Diseases in Humans and Animals*, Princeton University Press.
- Khan, A. S., Tshioko, F. K., Heymann, D. L., Guenno, B. L., Nabeth, P., Kerstiëns, B., Fleerackers, Y., Kilmarx, P. H., Rodier, G. R., Nkuku, O., Rollin, P. E., Sanchez, A., Zaki, S. R., Swanepoel, R., Tomori, O., Nichol, S. T., Peters, C., Muyembe-Tamfum, J., and Ksiazek, T. G. (1999), “The reemergence of Ebola Hemorrhagic Fever, Democratic Republic of Congo, 1995,” *The Journal of Infectious Diseases*, 179, 76–86.

- Krishnarajah, I., Cook, A., Marion, G., and Gibson, G. (2005), “Novel moment closure approximations in stochastic epidemics,” *Bulletin of Mathematical Biology*, 67, 855–873.
- Lekone, P. E. and Finkenstädt, B. F. (2006), “Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study,” *Biometrics*, 62, 1170–1177.
- Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag.
- 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.
- Moral, P. D., Doucet, A., and Jasra, A. (2006), “Sequential Monte Carlo samplers,” *Journal of the Royal Statistical Society, Series B*, 68, 411–436.
- Moser, M. R., Bender, T. R., Margolis, H. S., Noble, G. R., Kendal, A. P., and Ritter, D. G. (1979), “An outbreak of influenza aboard a commercial airliner,” *Am J Epidemiol*, 110, 1–6.
- Okware, S. I., Omaswa, F. G., Zaramba, S., Opio, A., Lutwama, J. J., Kamugisha, J., Rwaguma, E. B., Kagwa, P., and Lamunu, M. (2002), “An outbreak of Ebola in Uganda,” *Trop Med Int Health*, 7, 1068–1075.
- O’Neill, P. D. (2002), “A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods,” *Math Biosci*, 180, 103–114.
- 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 Roberts, G. O. (1999), “Bayesian inference for partially observed stochastic epidemics,” *Journal of the Royal Statistical Society. Series A (General)*, 162, 121–129.
- Plagnol, V. and Tavaré, S. (2002), “Approximate Bayesian computation and MCMC,” in *Monte Carlo and Quasi-Monte Carlo Methods 2002*, ed. Niederreiter, H., Springer-Verlag, 99–114.

- Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999), "Population growth of human Y chromosomes: a study of Y chromosome microsatellites," *Molecular Biology and Evolution*, 16, 1791–1798.
- R Development Core Team (2008), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
- Renshaw, E. (1993), *Modelling Biological Populations in Space and Time*, Cambridge University Press.
- Riley, S., Fraser, C., Donnelly, C. A., Ghani, A. C., Abu-Raddad, L. J., Hedley, A. J., Leung, G. M., Ho, L.-M., Lam, T.-H., Thach, T. Q., Chau, P., Chan, K.-P., Lo, S.-V., Leung, P.-Y., Tsang, T., Ho, W., Lee, K.-H., Lau, E. M., Ferguson, N. M., and Anderson, R. M. (2003), "Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions," *Science*, 300, 1961–1966.
- 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. (2007a), "Sequential Monte Carlo without likelihoods," *Proceedings of the National Academy of Sciences USA*, 104, 1760–1765.
- Sisson, S., Fan, Y., and Tanaka, M. M. (2007b), "A note on backward kernel choice for sequential Monte Carlo without likelihoods," http://web.maths.unsw.edu.au/~scott/papers/paper_smcabc_optimal.pdf.
- Streftaris, G. and Gibson, G. J. (2004), "Bayesian inference for stochastic epidemics in closed populations," *Statistical Modelling*, 4, 63–75.
- Tallmon, D. A., Luikart, G., and Beaumont, M. A. (2004), "Comparative evolution of a new effective population size estimator based on Approximate Bayesian Computation," *Genetics*, 167, 977–988.
- 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.

- Tavaré, S., Marshall, C. R., Will, O., Soligo, C., and Martin, R. D. (2002), “Using the fossil record to estimate the age of the last common ancestor of extant primates,” *Nature*, 416, 726–729.
- Tildesley, M. J., Savill, N. J., Shaw, D. J., Deardon, R., Brooks, S. P., Woolhouse, M. E., Grenfell, B. T., and Keeling, M. J. (2006), “Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK,” *Nature*, 440, 83–86.
- 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.
- Vynnycky, E., Trindall, A., and Mangtani, P. (2007), “Estimates of the reproduction number of Spanish influenza using morbidity data,” *Int J Epidemiol*, 38, 881–889.
- World Health Organisation (WHO) (2008), “Ebola haemorrhagic fever,” <http://www.who.int/mediacentre/factsheets/fs103/en/index.html>.