Pair-based likelihood approximations for stochastic epidemic models

JESSICA E. STOCKDALE

Department of Mathematics, Simon Fraser University, Burnaby, Canada

THEODORE KYPRAIOS, PHILIP D. O'NEILL[*]

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

philip.oneill@nottingham.ac.uk

Abstract

Fitting stochastic epidemic models to data is a non-standard problem because data on the infection processes defined in such models are rarely observed directly. This in turn means that the likelihood of the observed data is intractable in the sense that it is very computationally expensive to obtain. Although data-augmentated Markov chain Monte Carlo (MCMC) methods provide a solution to this problem, employing a tractable augmented likelihood, such methods typically deteriorate in large populations due to poor mixing and increased computation time. Here we describe a new approach that seeks to approximate the likelihood by exploiting the underlying structure of the epidemic model. Simulation study results show that this approach can be a serious competitor to data-augmented MCMC methods. Our approach can be applied to a wide variety of disease transmission models, and we provide examples with applications to the common cold, Ebola and foot-and-mouth disease.

Key words: Epidemic models; Likelihood approximation; Markov chain Monte Carlo methods; Stochastic epidemic models

1 Introduction

Mathematical models of infectious disease transmission are now routinely used as tools to assist with the analysis, prediction and control of real-life epidemics. Such models may be deterministic or stochastic, are usually mechanistic in the sense that they seek to describe the process of disease spread between individuals, and invariably contain parameters such as infection rates that must be assigned values. The natural statistical problem that arises is to estimate the parameters of an epidemic model given observed data on one or more outbreaks of disease.

In this paper we focus exclusively on stochastic epidemic models. Fitting such a model to data in a frequentist or Bayesian framework requires evaluation of the likelihood of the observed data. In many situations, this is a non-standard problem because the infection process defined in the model is not observed in reality. Consequently, evaluation of the likelihood involves integrating over the set

[*] To whom correspondence should be addressed

of all possible infection events that are compatible with the observed data, which is often a highly non-trivial exercise. Within the Bayesian framework, one solution is to use data augmentation, specifically including the unknown infection events as additional model parameters, which in turn leads to a computationally tractable likelihood. The posterior distribution of the model parameters given the data can then be explored via Markov chain Monte Carlo (MCMC) methods (Gibson and Renshaw, 1998; O'Neill and Roberts, 1999). However, such methods can struggle for large-scale problems, partly because the computational cost of evaluating the augmented-data likelihood increases, and partly because the missing data are strongly correlated to the model parameters which in turn creates mixing problems for the Markov chain. Although reparameterisation techniques such as non-centering can help (Kypraios, 2007), it is still desirable to find alternative approaches.

In this paper we describe a method for approximating the likelihood of a partially observed stochastic epidemic model without the need for data augmentation. The key idea is to approximate the likelihood by a product of tractable terms that relate to either single individuals or pairs of individuals. Our approach is somewhat inspired by, but distinct from, a likelihood expression derived in Eichner and Dietz (2003) for a stochastic smallpox transmission model. As explained later, this expression is actually not the true likelihood for the model, but as shown in Stockdale *and others* (2017) it yields very similar parameter estimates to those obtained from a full data-augmented MCMC approach using the correct likelihood. Note also that our methods are unrelated to pair-approximation techniques used in the analysis of deterministic epidemic models (Keeling *and others*, 1997).

Our main aim is to explore the potential utility of using approximate likelihood methods in fitting stochastic epidemic models to data. The work presented here is in some sense preliminary, since there are many possible future directions that could be taken, but also contains many promising results. The paper is structured as follows. Section 2 defines the epidemic model of interest and associated notation. The likelihood approximations are developed in Section 3 and illustrated via three applications in Section 4. Brief conclusions are given in Section 5. Details of proofs and results from an extensive simulation study can be found in the Supplementary Material.

2 Preliminary material

For ease of exposition, we shall describe likelihood approximations for a specific stochastic epidemic model defined below. However, similar approximations can be derived for more complex models, as illustrated in Section 4.

2.1 Stochastic epidemic model

The following epidemic model generalises the well-known general stochastic epidemic (see e.g. Bailey, 1975; Andersson and Britton, 2000) so that the infection rate between a pair of individuals is allowed to depend on who the individuals are, and infectious period distributions can vary between individuals.

Consider a population consisting of N individuals, labelled $1, \ldots, N$. At any time, each individual is either *susceptible*, meaning they are capable of contracting the disease in question, *infective*, meaning that they have the disease and can infect others, or *removed*, meaning that they are no longer able to infect others and cannot be re-infected. The precise interpretation of the removed state depends on the disease in question, examples including isolation, recovery, or death. Initially, the population is entirely susceptible apart from a few infective individuals. Each individual who becomes infective remains so for a period of time known as the infectious period. The infectious period of individual j is distributed according to some pre-specified random variable I_j. At the end of its infectious period, an individual becomes removed. The infectious periods of different individuals are assumed to be independent.

During its infectious period, a given infective individual j has contacts with susceptible individual k at times given by the points of a Poisson process of rate β_{jk}. All such Poisson processes are mutually independent and independent of the infectious periods. Any contact that occurs results in the susceptible individual k immediately becoming infective. We define the *infectious pressure* acting on susceptible k at time t as the hazard rate of infection at time t, in other words $\Sigma\beta_{jk}$ where the sum is over all individuals j who are infective at time t.

The epidemic continues until there are no infectives remaining. Thus at the end of the epidemic, each initially susceptible individual is either still susceptible, or removed. Finally, the population is assumed to be closed in the sense that no individuals may enter or leave during the epidemic.

The model defined above is rather general and contains $N(N-1)$ infection rate parameters corresponding to all possible choices of the ordered pair (j,k), $j \neq k$. For specific modelling situations we usually use a model with fewer parameters, which can be obtained by making suitable assignments for the β_{jk} and the parameters governing the infectious period distributions. Examples include the general stochastic epidemic, i.e. the standard homogeneously-mixing SIR (Susceptible-Infective-Removed) model, multi-type models, models with two or more levels of mixing, and spatial models.

2.2 Notation, data and likelihood

In real-life epidemics, the actual transmission process of infection between individuals is rarely observed. We therefore suppose henceforth that the observed data consist of the times of all removal events, and re-label members of the population such that individuals $1,\dots,n$ are those who are ultimately removed, and $n+1,\dots,N$ are those (if any) who remain susceptible, where $n \leq N$. We are thus implicitly assuming that each removal event in the model corresponds to a real-life observable event such as the appearance of symptoms in an individual, and furthermore that the individual is then unable to infect others, perhaps due to isolation. We are also assuming that the epidemic has come to an end, so that there are no unobserved removals.

For $j = 1,\dots,n$ let r_j denote the time of removal of individual j, with the convention that $r_j = \infty$ if j is never removed, i.e. if $j > n$. Similarly define i_j as the time at which j becomes infected, with $i_j = \infty$ if this never occurs. We assume that there is a single initial infective α, so that $\alpha \in \{1,\dots,n\}$, but we do not assume that α is known from the data. The assumption of a single initial infective is not necessary, but simplifies our exposition and is often realistic in practice. Let $\boldsymbol{r} = (r_1,\dots,r_n)$ and $\boldsymbol{i} = (i_1,\dots,i_{\alpha-1},i_{\alpha+1},\dots,i_n)$, so that $\boldsymbol{i}$ contains all infection times other than i_α. Denote by $\boldsymbol{\beta} = \{\beta_{jk} : 1 \leq j,k \leq N, j \neq k\}$ the set of infection rate parameters in the model.

Let θ_j denote the parameter vector of the infectious period distribution for individual j, $j = 1,\dots,n$. Our main focus will be upon the cases where the infectious periods are either exponential or Erlang with known shape parameter, where in both cases θ_j is one-dimensional. Finally let $\boldsymbol{\theta} = \{\theta_j : 1 \leq j \leq n\}$ denote the set of infectious period distribution parameters.

Our objective is to make inference for the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ given the data $\boldsymbol{r}$, assuming the population size N is known. Any likelihood-based approach therefore requires evaluation of the likelihood $\pi(\boldsymbol{r}|\boldsymbol{\beta},\boldsymbol{\theta})$ but calculating this quantity is highly computationally expensive. The reason for this is that any such calculation implicitly or explicitly involves integrating over all possible values of the unknown infection times, the set of which is non-trivial due to the constraints that (i) $i_j < r_j$ for $j = 1,\dots,n$ and (ii) at any time during the epidemic, there cannot be more removals than infections (see e.g. Clancy and O'Neill, 2008).

One solution to this problem is to instead work with the tractable augmented likelihood $\pi(\boldsymbol{i}, \boldsymbol{r}|\boldsymbol{\beta}, \boldsymbol{\theta}, \alpha, i_\alpha)$, given explicitly below. For instance, in a Bayesian framework the unknown infection times can be incorporated as extra parameters and an MCMC algorithm can be used to sample from the resulting posterior distribution, as described in O'Neill and Roberts (1999). Our objective here, however, is to find a way of approximating the likelihood $\pi(\boldsymbol{r}|\boldsymbol{\beta}, \boldsymbol{\theta})$ that avoids any data augmentation.

3 Pair-based likelihood approximation

3.1 Notation and augmented likelihood

Consider an individual j who becomes infected at time i_j and removed at time r_j. Define

$$\begin{aligned}
\psi_j &= P(j \text{ avoids infection until time } i_j),\\
\chi_j &= \text{infectious pressure acting on } j \text{ as they become infected},\\
\text{and } \phi_j &= P(j \text{ fails to infect any of the } N-n \text{ never-infected individuals}).
\end{aligned}$$

It follows from the definition of the epidemic model that

$$\psi_j = \exp\left(-\sum_{\substack{k=1\\k\neq j}}^{n} \beta_{kj}\tau_{kj}\right), \tag{1}$$

$$\chi_j = \sum_{\substack{k=1\\k\neq j}}^{n} \beta_{kj} 1_{\{i_k<i_j<r_k\}}, \tag{2}$$

$$\text{and } \phi_j = \exp\left(-\sum_{k=n+1}^{N} \beta_{jk}(r_j - i_j)\right) = \exp\left(-(r_j - i_j)B_j\right), \tag{3}$$

say, where, with $\wedge$ denoting minimum,

$$\tau_{kj} = r_k \wedge i_j - i_k \wedge i_j \tag{4}$$

is the length of time during which k is infective and j susceptible, and 1_A denotes the indicator function of the event A. Note that the summation terms in (1) and (2) could both be written without excluding $k \neq j$, since both τ_{jj} and $1_{\{i_j<i_j<r_j\}}$ are zero. However, when translating formulae into computer code it is helpful to know what can be excluded from sums or products, so our derivations will make this explicit.

For ease of exposition, we assume that the infectious period distributions are continuous and let f_j denote the probability density function of I_j; the arguments below also hold without this assumption. The augmented likelihood of all infection and removal events may be written as

$$\pi(\boldsymbol{i}, \boldsymbol{r}|\boldsymbol{\beta}, \boldsymbol{\theta}, \alpha, i_\alpha) = \left\{\prod_{\substack{j=1\\j\neq\alpha}}^{n} \chi_j\psi_j\phi_j f_j(r_j - i_j|\theta_j)\right\} \phi_\alpha f_\alpha(r_\alpha - i_\alpha|\theta_\alpha). \tag{5}$$

We now briefly explain Equation (5); a more detailed explanation for the special case of the general stochastic epidemic can be found in Andersson and Britton (2000). The product term accounts for each ever-infected individual j (other than α) avoiding infection until time i_j, becoming infected at time i_j, remaining infective until time r_j, and whilst infective failing to infect the $N-n$ individuals who avoid infection entirely. Note that the probability of j failing to infect another individual k, prior to k's own infection at time i_k, will be accounted for in ψ_k. Finally, the corresponding likelihood contribution for α is similar to that of j, but does not account for how α became infected since the epidemic model only describes events after the initial infection.

Equation (5) is not the only way to write the likelihood: for example, all of the ψ_j and ϕ_j terms could be combined together to give the total probability of individuals avoiding infection during the epidemic. However, (5) is in a form suitable for our purposes. Note also that the ψ_j and χ_j terms appearing in the product do not only depend on individual j, but also on the infection and removal times of other individuals.

3.2 Derivation of the pair-based approximation

We now derive an approximation to the desired likelihood $\pi(\boldsymbol{r}|\boldsymbol{\beta},\boldsymbol{\theta})$. First note that

$$\pi(\boldsymbol{r}|\boldsymbol{\beta},\boldsymbol{\theta}) = \int \pi(\boldsymbol{i},\boldsymbol{r}|\boldsymbol{\beta},\boldsymbol{\theta},\alpha,i_\alpha)\pi(i_\alpha,\alpha)d\boldsymbol{i}\; di_\alpha\; d\alpha,$$

where the integral represents integration over $\boldsymbol{i}$, i_α and summation over α, and where (i_α,α) is assumed to be independent of $(\boldsymbol{\beta},\boldsymbol{\theta})$ *a priori*. Thus,

$$\begin{aligned}
\pi(\boldsymbol{r}|\boldsymbol{\beta},\boldsymbol{\theta}) &= \sum_{\alpha=1}^{n}\pi(\alpha)\int \pi(\boldsymbol{i},\boldsymbol{r}|\boldsymbol{\beta},\boldsymbol{\theta},\alpha,i_\alpha)\pi(i_\alpha|\alpha)d\boldsymbol{i}\; di_\alpha \\
&= \sum_{\alpha=1}^{n}\pi(\alpha)\int \left\{\prod_{\substack{j=1\\ j\neq\alpha}}^{n}\chi_j\psi_j\phi_j\right\}\phi_\alpha\pi(i_\alpha|\alpha)\left\{\prod_{j=1}^{n}f_j(r_j-i_j|\theta_j)\right\}d\boldsymbol{i}\; di_\alpha \\
&= \sum_{\alpha=1}^{n}\pi(\alpha)\int \left\{\prod_{\substack{j=1\\ j\neq\alpha}}^{n}\chi_j\psi_j\right\}\pi(i_\alpha|\alpha)\left\{\prod_{j=1}^{n}\phi_j f_j(r_j-i_j|\theta_j)\right\}d\boldsymbol{i}\; di_\alpha, \qquad (6)
\end{aligned}$$

where $\pi(\alpha)$ denotes the prior probability that α is the initial infective. For $j=1,\ldots,n$,

$$\begin{aligned}
\phi_j f_j(r_j-i_j|\theta_j) &= \exp\left(-(r_j-i_j)B_j\right)f_j(r_j-i_j|\theta_j) \\
&= a(\theta,B_j)g_j(r_j-i_j|\theta_j),
\end{aligned}$$

say, where g_j is the probability density function defined for $x>0$ by

$$g_j(x|\theta_j) = \frac{\exp(-xB_j)f_j(x|\theta_j)}{\int\exp(-xB_j)f_j(x|\theta_j)\,dx} = \frac{\exp(-xB_j)f_j(x|\theta_j)}{a(\theta_j,B_j)}, \qquad (7)$$

and $g_j(x|\theta_j) = 0$ for $x \leq 0$. Thus $a(\theta_j, B_j)$ is the moment generating function of the infectious period I_j evaluated at B_j. Substituting into (6) yields

$$\begin{aligned}
\pi(\boldsymbol{r}|\boldsymbol{\beta},\boldsymbol{\theta}) &= \left\{\prod_{j=1}^{n} a(\theta_j, B_j)\right\} \sum_{\alpha=1}^{n} \pi(\alpha) \int \left\{\prod_{\substack{j=1\\j\neq\alpha}}^{n} \chi_j \psi_j\right\} \pi(i_\alpha|\alpha) \left\{\prod_{j=1}^{n} g_j(r_j - i_j|\theta_j)\right\} d\boldsymbol{i}\; di_\alpha \\
&= \left\{\prod_{j=1}^{n} a(\theta_j, B_j)\right\} \sum_{\alpha=1}^{n} \pi(\alpha) \mathbb{E}_{\boldsymbol{g}} \left[\pi(i_\alpha|\alpha) \prod_{\substack{j=1\\j\neq\alpha}}^{n} \chi_j \psi_j\right], \qquad (8)
\end{aligned}$$

where $\mathbb{E}_{\boldsymbol{g}}$ denotes expectation of $(r_1 - i_1, \ldots, r_n - i_n)$ with respect to the product density

$$\boldsymbol{g}(x_1, \ldots, x_n|\boldsymbol{\theta}) = \prod_{j=1}^{n} g_j(x_j|\theta_j). \qquad (9)$$

Note that here we regard $\boldsymbol{r}$ as fixed and the infection times as random variables, and thus sampling from $\boldsymbol{g}$ essentially generates a sample from $(\boldsymbol{i}, i_\alpha)$. Evaluating the required likelihood thus requires evaluation of the expectation term in (8).

Before turning to approximations we make some remarks about the exact equation (8). First, equation (6) can clearly be written as an expectation with respect to the product density obtained by multiplying the f_j terms together. The advantage of the approach leading to (8) is that the ϕ_j terms are absorbed into the expectation, and thus we do not need to evaluate or approximate them. Second, let $\mathcal{I}$ denote the set of values of $(\boldsymbol{i}, i_\alpha)$ such that the term inside the expectation in (8) is non-zero. Although the expectation is taken with respect to independent random variables, $\mathcal{I}$ is complicated in structure, which makes analytical progress difficult. Finally, a random sample from $\boldsymbol{g}$ is not guaranteed to lie inside $\mathcal{I}$, which makes standard Monte Carlo estimation inefficient. An importance sampling estimator could be constructed, although it is not obvious how to construct an efficient proposal distribution for the infection times. We therefore proceed via an approximation in which we assume independence of the product terms in the expectation term in (8), as follows.

First, we assume that

$$\mathbb{E}_{\boldsymbol{g}} \left[\pi(i_\alpha|\alpha) \prod_{\substack{j=1\\j\neq\alpha}}^{n} \chi_j \psi_j\right] \approx \mathbb{E}_{\boldsymbol{g}} \left[\pi(i_\alpha|\alpha)\right] \prod_{\substack{j=1\\j\neq\alpha}}^{n} \mathbb{E}_{\boldsymbol{g}} \left[\chi_j \psi_j\right]. \qquad (10)$$

Evaluation of the first expectation in (10) depends on the choice of prior density $\pi(i_\alpha|\alpha)$, but is often straightforward in practice. For the second expectation, first note that

$$\psi_j = \exp\left(-\sum_{\substack{l=1\\l\neq j}}^{n} \beta_{lj}\tau_{lj}\right) = \prod_{\substack{l=1\\l\neq j}}^{n} \exp(-\beta_{lj}\tau_{lj}) = \prod_{\substack{l=1\\l\neq j}}^{n} \psi_{jl},$$

say. Then

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{g}}\left[\chi_j\psi_j\right] &= \sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\mathbb{E}_{\boldsymbol{g}}\left[1_{\{i_k<i_j<r_k\}}\psi_j\right] \\
&= \sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\mathbb{E}_{\boldsymbol{g}}\left[1_{\{i_k<i_j<r_k\}}\prod_{\substack{l=1\\l\neq j}}^{n}\psi_{jl}\right] \\
&= \sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\mathbb{E}_{\boldsymbol{g}}\left[1_{\{i_k<i_j<r_k\}}\psi_{jk}\prod_{\substack{l=1\\l\neq j,k}}^{n}\psi_{jl}\right] \\
&\approx \sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\mathbb{E}_{\boldsymbol{g}}\left[1_{\{i_k<i_j<r_k\}}\psi_{jk}\right]\prod_{\substack{l=1\\l\neq j,k}}^{n}\mathbb{E}_{\boldsymbol{g}}\left[\psi_{jl}\right], \qquad (11)
\end{aligned}$$

where (11) only contains terms that concern pairs of individuals. For computational purposes, it is useful to re-write (11) as

$$\mathbb{E}_{\boldsymbol{g}}\left[\chi_j\psi_j\right] \approx \left\{\prod_{\substack{l=1\\l\neq j}}^{n}\mathbb{E}_{\boldsymbol{g}}\left[\psi_{jl}\right]\right\}\sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\mathbb{E}_{\boldsymbol{g}}\left[1_{\{i_k<i_j<r_k\}}\psi_{jk}\right]\left(\mathbb{E}_{\boldsymbol{g}}\left[\psi_{jk}\right]\right)^{-1}, \qquad (12)$$

to avoid computing the product terms in (11) separately for each term in the sum.

Definition 1 *We refer to the approximation arising from equations and (8), (10) and (12) as the* standard *pair-based likelihood approximation (PBLA).*

To evaluate (12) we need to compute, for $j \neq k$,

$$\mathbb{E}_{\boldsymbol{g}}\left[\psi_{jk}\right] = \mathbb{E}_{g_j,g_k}\left[\exp(-\beta_{kj}\tau_{kj})\right], \qquad (13)$$

which is the probability that j avoids infection from k while k is infective and j susceptible, and the related quantity

$$\mathbb{E}_{\boldsymbol{g}}\left[1_{\{i_k<i_j<r_k\}}\psi_{jk}\right] = \mathbb{E}_{g_j,g_k}\left[1_{\{i_k<i_j<r_k\}}\exp(-\beta_{kj}\tau_{kj})\right]. \qquad (14)$$

Explicit expressions for (13) and (14) for the case of exponential or Erlang infectious period distributions are given below in Sections 3.4 and 3.5, respectively.

Finally, it is important to note that the standard approximation, and those described below, become exact in the special case where infectious periods are non-random, since the expectations are redundant. This in turn suggests that the less variability an infectious period distribution has, the more accurate the approximation will be.

3.3 Alternative approximations

The derivation of (12) is clearly not the only way to derive an approximate likelihood, and here we mention some alternatives.

3.3.1 Use f_j for expectations

As mentioned previously one could proceed without introducing the change of density from f_j to g_j, so that (8) becomes

$$\pi(\boldsymbol{r}|\boldsymbol{\beta},\theta) = \sum_{\alpha=1}^{n} \pi(\alpha)\mathbb{E}_{\boldsymbol{f}} \left[\phi_\alpha \pi(i_\alpha|\alpha) \prod_{\substack{j=1 \\ j\neq\alpha}}^{n} \chi_j \psi_j \phi_j \right] \tag{15}$$

where the expectation is with respect to the product density $\prod_{j=1}^{n} f_j(x_j|\theta_j)$ (cf. (9)). Following the arguments above leads naturally to terms of the form

$$\begin{aligned} \mathbb{E}_{\boldsymbol{f}}[\chi_j\psi_j\phi_j] &\approx \mathbb{E}_{\boldsymbol{f}}[\chi_j\phi_j]\mathbb{E}_{\boldsymbol{f}}[\psi_j] \\ &\approx \mathbb{E}_{\boldsymbol{f}}[\chi_j\phi_j] \prod_{\substack{k=1 \\ k\neq j}}^{n} \mathbb{E}_{\boldsymbol{f}}[\psi_{jk}], \end{aligned}$$

evaluation of which requires the equivalent of (13) with $\boldsymbol{f}$ replacing $\boldsymbol{g}$, and

$$\mathbb{E}_{\boldsymbol{f}}[\chi_j\phi_j] = \mathbb{E}_{\boldsymbol{f}} \left[1_{\{i_k<i_j<r_k\}} \exp\left(-(r_j - i_j)B_j\right) \right].$$

3.3.2 Separate all χ and ψ terms

We could write

$$\begin{aligned} \mathbb{E}_{\boldsymbol{g}}\left[\chi_j\psi_j\right] &\approx \mathbb{E}_{\boldsymbol{g}}\left[\psi_j\right]\mathbb{E}_{\boldsymbol{g}}\left[\chi_j\right] \\ &\approx \left\{ \prod_{\substack{l=1 \\ l\neq j}}^{n} \mathbb{E}_{\boldsymbol{g}}\left[\psi_{jl}\right] \right\} \sum_{\substack{k=1 \\ k\neq j}}^{n} \beta_{kj} \mathbb{E}_{\boldsymbol{g}} \left[1_{\{i_k<i_j<r_k\}} \right], \end{aligned}$$

following the arguments leading to (12). We found that this formulation is numerically fairly similar to (12), although it involves marginally more approximation because the indicator function and ψ_{jk} terms are separated.

3.3.3 Approximate product of ψ terms

We could attempt to approximate the entire product of ψ_j terms. The expectation in equation (8) may be approximated by

$$\mathbb{E}_{\boldsymbol{g}} \left[\pi(i_\alpha|\alpha) \prod_{\substack{j=1 \\ j\neq\alpha}}^{n} \chi_j\psi_j \right] \approx \mathbb{E}_{\boldsymbol{g}}\left[\pi(i_\alpha|\alpha)\right] \left\{ \prod_{\substack{j=1 \\ j\neq\alpha}}^{n} \mathbb{E}_{\boldsymbol{g}}\left[\chi_j\right] \right\} \left\{ \mathbb{E}_{\boldsymbol{g}} \left[\prod_{\substack{j=1 \\ j\neq\alpha}}^{n} \psi_j \right] \right\}, \tag{16}$$

where as above we have

$$\mathbb{E}_{\boldsymbol{g}}\left[\chi_j\right] = \sum_{\substack{k=1 \\ k\neq j}}^{n} \beta_{kj} \mathbb{E}_{\boldsymbol{g}} \left[1_{\{i_k<i_j<r_k\}} \right], \tag{17}$$

while

$$\mathbb{E}_{\boldsymbol{g}}\left[\prod_{\substack{j=1\\j\neq\alpha}}^{n}\psi_j\right]=\mathbb{E}_{\boldsymbol{g}}\left[\exp\left(-\sum_{\substack{j=1\\j\neq\alpha}}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\tau_{kj}\right)\right]. \tag{18}$$

In Section 3.4.2 below we describe methods that exploit (18) for the case where $\beta_{kj}=\beta/N$ and infectious periods are exponentially distributed.

3.3.4 Eichner and Dietz approximation

Eichner and Dietz (2003) define a stochastic model for smallpox transmission, and give a likelihood expression which is used for maximum likelihood estimation of the model parameters. Although not presented as such, their expression is actually an approximation to the true likelihood. For our model, their method is as follows. For $j=1,\dots,n$ define

$$\lambda_j(u)=\sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}1_{\{i_k<u<r_k\}}$$

as the infectious pressure acting on individual j at time u. Note that $\chi_j=\lambda_j(i_j)$. The Eichner-Dietz (ED) likelihood approximation is

$$\begin{aligned}\pi(\boldsymbol{r}|\boldsymbol{\beta},\theta)\approx\pi_{ED}(\boldsymbol{r}|\boldsymbol{\beta},\theta) \quad = \quad & \left\{\prod_{j=1}^{n}\left(\int_{-\infty}^{r_j}\mathbb{E}[\lambda_j(i_j)]\exp\left(-\int_{-\infty}^{i_j}\mathbb{E}[\lambda_j(s)]\,ds\right)f_j(r_j-i_j|\theta_j)\,di_j\right)\right\}\\ & \times\prod_{j=n+1}^{N}\exp\left(-\int_{-\infty}^{r_n}\mathbb{E}[\lambda_j(s)]\,ds\right),\end{aligned} \tag{19}$$

where

$$\mathbb{E}[\lambda_j(u)]=\sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}P(i_k<u<r_k)=\sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}1_{\{u<r_k\}}\int_{r_k-u}^{\infty}f_k(s|\theta_k)\,ds, \tag{20}$$

since r_k-i_k has density f_k.

The ED approximation could be derived as follows. The starting point is to assume that the likelihood takes the form

$$L_{ED}=\mathbb{E}_{\boldsymbol{f}}\left[\prod_{j=1}^{n}\chi_j\psi_j\phi_j\right], \tag{21}$$

which bears some resemblance to the exact expression at (15) (e.g. by setting $\pi(\alpha)=1_{\{\alpha=1\}}$), but differs because the initial infective is not explicitly included. Now,

$$\prod_{j=1}^{n}\phi_j=\exp\left(-\sum_{j=1}^{n}\sum_{k=n+1}^{N}\beta_{jk}(r_j-i_j)\right)=\prod_{k=n+1}^{N}\exp\left(-\sum_{j=1}^{n}\beta_{jk}\tau_{jk}\right)=\prod_{k=n+1}^{N}\tilde{\phi}_k,$$

say, since $\tau_{jk}=(r_k-i_k)$ for $k>n$. Note that $\tilde{\phi}_k$ is the probability that individual k avoids infection throughout the epidemic. Thus (21) may be written as

$$\begin{aligned} L_{ED} &= \mathbb{E}_{\boldsymbol{f}}\left[\left\{\prod_{j=1}^{n}\chi_j\psi_j\right\}\left\{\prod_{k=n+1}^{N}\tilde{\phi}_k\right\}\right] \\ &\approx \left\{\prod_{j=1}^{n}\mathbb{E}_{\boldsymbol{f}}[\chi_j\psi_j]\right\}\left\{\prod_{k=n+1}^{N}\mathbb{E}_{\boldsymbol{f}}[\tilde{\phi}_k]\right\} \\ &\approx \left\{\prod_{j=1}^{n}\mathbb{E}_{f_j}\left[\mathbb{E}_{\boldsymbol{f}}[\chi_j|i_j]\mathbb{E}_{\boldsymbol{f}}[\psi_j|i_j]\right]\right\}\left\{\prod_{k=n+1}^{N}\mathbb{E}_{\boldsymbol{f}}[\tilde{\phi}_k]\right\}, \end{aligned} \qquad (22)$$

where $\mathbb{E}_{f_j}$ denotes expectation of i_j with respect to $f_j(r_j - i_j|\theta_j)$. The expectations in (22) are then evaluated using further approximations as shown in (19).

The main difficulty with (19) is that, in practice, it involves numerical integration. As shown below, even for the general stochastic epidemic the integral appears to be analytically intractable. However, conditioning on i_j is an attractive feature of the ED approximation, since it removes one of the sources of approximation in the pair-based methods.

3.4 Exponential infectious periods

In this section we assume that for $j = 1, \ldots, n$ the infectious period random variable I_j is exponential with rate parameter $\theta_j = \gamma_j$, denoted by $I_j \sim Exp(\gamma_j)$. Such models, particularly the case where $\gamma_j = \gamma$ for all j, appear frequently in the epidemic modelling literature, largely because of their relative mathematical tractability.

Let $x > 0$. Since $f_j(x|\theta_j) = \gamma_j \exp(-\gamma_j x)$ then for $j = 1, \ldots, n$, (7) gives

$$g_j(x|\theta_j) = \frac{\gamma_j \exp(-(\gamma_j + B_j)x)}{\gamma_j/(\gamma_j + B_j)} = \delta_j \exp(-\delta_j x),$$

say, so that g_j is the probability density function of an exponential random variable with rate $\delta_j = \gamma_j + B_j$, and

$$a(\theta_j, B_j) = a(\gamma_j, B_j) = \gamma_j/(\gamma_j + B_j) = \gamma_j/\delta_j.$$

3.4.1 The standard pair-based approximation with exponential infectious periods

We require expressions for (13) and (14). Recall from (4) that $\tau_{kj} = r_k \wedge i_j - i_k \wedge i_j$.

Lemma 1 *Let $1 \leq j, k \leq n$ with $j \neq k$, and $\beta > 0$. Then*

$$\begin{aligned} &\mathbb{E}_{g_j,g_k}\left[\exp(-\beta\tau_{kj})\right] \\ &\quad = \begin{cases} 1 - \beta\delta_j\left\{(\delta_j+\delta_k)(\beta+\delta_k)\right\}^{-1}\exp(-\delta_k(r_k - r_j)) & \textit{if } r_j < r_k, \\ \delta_k(\beta+\delta_k)^{-1} + \beta\delta_k\left\{(\delta_j+\delta_k)(\beta+\delta_k)\right\}^{-1}\exp(-\delta_j(r_j - r_k)) & \textit{if } r_j > r_k, \end{cases} \end{aligned}$$

$$\begin{aligned} &\mathbb{E}_{g_j,g_k}\left[1_{\{i_k<i_j<r_k\}}\exp(-\beta\tau_{kj})\right] \\ &\quad = \begin{cases} \delta_j\delta_k\left\{(\delta_j+\delta_k)(\beta+\delta_k)\right\}^{-1}\exp(-\delta_k(r_k - r_j)) & \textit{if } r_j < r_k, \\ \delta_j\delta_k\left\{(\delta_j+\delta_k)(\beta+\delta_k)\right\}^{-1}\exp(-\delta_j(r_j - r_k)) & \textit{if } r_j > r_k. \end{cases} \end{aligned}$$

Lemma 1 can be proved either by direction calculation or by probability arguments; see the Supplementary Material for details.

3.4.2 Approximations for the general stochastic epidemic

Suppose now that for $1 \leq j, k \leq N$, $\beta_{kj} = \beta/N$ and $\gamma_j = \gamma$, so that the epidemic model is the general stochastic epidemic. From (3) we have $B_j = (N-n)\beta/N$, and thus $\delta_j = \gamma + (N-n)\beta/N = \delta$, say. This in turn leads to some simplifications in the expressions in Lemma 1 for the standard approximation.

We now focus on approximations that involve the product of ψ terms. Note that (18) becomes

$$\mathbb{E}_{\boldsymbol{g}}\left[\prod_{\substack{j=1\\j\neq\alpha}}^{n}\psi_j\right] = \mathbb{E}_{\boldsymbol{g}}\left[\exp\left(-(\beta/N)\sum_{\substack{j=1\\j\neq\alpha}}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\tau_{kj}\right)\right]. \tag{23}$$

Recall from (4) that τ_{kj} is the length of time during which k is able to infect j. Thus for a given set of infection times, one of τ_{kj} and τ_{jk} is zero. To exploit this dependency we rewrite the double sum in (23), also using the facts that $\tau_{k\alpha} = \tau_{jj} = 0$, to give

$$\sum_{\substack{j=1\\j\neq\alpha}}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\tau_{kj} = \sum_{j=1}^{n-1}\sum_{k=j+1}^{n}(\tau_{kj}+\tau_{jk}) = \sum_{j=1}^{n-1}\sum_{k=j+1}^{n}\omega_{jk} = X, \tag{24}$$

say, where, since $r_j < r_k$ for $j < k$, we have

$$\omega_{jk} = \begin{cases} i_j - i_k & \text{if } i_k < i_j, \\ i_k - i_j & \text{if } i_j < i_k < r_j, \\ r_j - i_j & \text{if } i_k > r_j. \end{cases} \tag{25}$$

Note that ω_{jk} is the length of time that j exerts infectious pressure on k, or vice versa. We are thus concerned with the behaviour of X given that $r_1 - i_1, \ldots, r_n - i_n$ are independent exponential distributions with parameter δ. The following result, proved in the Supplementary Material, provides an explicit distribution for the total infectious pressure time among any subset of individuals in $\{1, \ldots, n\}$.

Lemma 2 *Let $\mathcal{K}$ be any subset of $\{1, \ldots, n\}$ with $K \geq 2$ elements. If $\{r_j - i_j : j \in \mathcal{K}\}$ is a set of independent Exp(δ) random variables, then*

$$\sum_{\substack{j,k\in\mathcal{K}\\j<k}}\omega_{jk} \sim \sum_{j=1}^{K-1} Y_j$$

where $Y_j \sim$ Exp(δ/j) and $Y_1, \ldots, Y_{K-1}$ are independent.

Setting $K = 2$ in Lemma 2 yields that $\omega_{jk} \sim$ *Exp*(δ) for any $j \neq k$, and setting $K = n$ yields an explicit distribution for X in (24). Furthermore, (23) and Lemma 2 yield the result

$$\mathbb{E}_{\boldsymbol{g}}\left[\prod_{\substack{j=1\\j\neq\alpha}}^{n}\psi_j\right] = \prod_{j=1}^{n-1}\left(\frac{\delta}{(\beta j/N)+\delta}\right).$$

The behaviour of X as $n \to \infty$ is given in the following result, proved in the Supplementary Material.

Lemma 3 *If $r_1 - i_1, \ldots, r_n - i_n \sim Exp(\delta)$ are independent then*

$$\frac{\sum_{j=1}^{n-1}\sum_{k=j+1}^{n}(\omega_{jk} - \mathbb{E}_{\boldsymbol{g}}[\omega_{jk}])}{s_n} = \frac{\sum_{j=1}^{n-1}\sum_{k=j+1}^{n}(\omega_{jk} - \delta^{-1})}{s_n}$$

converges in distribution to a standard Gaussian random variable as $n \to \infty$, where

$$\begin{aligned} s_n^2 &= \sum_{j=1}^{n-1}\sum_{k=j+1}^{n}\sum_{l=1}^{n-1}\sum_{m=j+1}^{n}(\mathbb{E}_{\boldsymbol{g}}[\omega_{jk}\omega_{lm}] - \mathbb{E}_{\boldsymbol{g}}[\omega_{jk}]\mathbb{E}_{\boldsymbol{g}}[\omega_{lm}]) \\ &= \frac{n(n-1)(2n-1)}{6\delta^2}. \end{aligned}$$

Lemma 3 implies that, for large n, X is approximately Gaussian with mean $\binom{n}{2}\delta^{-1}$ and variance s_n^2, and thus the right-hand side of (23) is approximately equal to the moment generating function of this Gaussian distribution evaluated at the point $-\beta/N$. This yields the approximation

$$\mathbb{E}_{\boldsymbol{g}}\left[\prod_{\substack{j=1\\ j\neq\alpha}}^{n}\psi_j\right] \approx \exp\left\{-\frac{\beta}{N\delta}\binom{n}{2} + \frac{\beta^2}{12\delta^2N^2}n(n-1)(2n-1)\right\},$$

which along with equations (8), (16) and (17) yields a likelihood approximation for large n.

3.4.3 Eichner and Dietz approximation with exponential infectious periods

We now consider the ED approximation given in (19) under the assumption that $I_j \sim Exp(\gamma_j)$. First, (20) becomes

$$\mathbb{E}[\lambda_j(u)] = \sum_{\substack{k=1\\ k\neq j}}^{n}\beta_{kj}\exp(-\gamma_k(r_k - u))1_{\{u<r_k\}},$$

and direct calculation yields that

$$\int_{-\infty}^{t}\mathbb{E}[\lambda_j(s)]\,ds = \sum_{\substack{k=1\\ k\neq j}}^{n}\beta_{kj}\gamma_k^{-1}\exp\{-\gamma_k(r_k - (t\wedge r_k))\} = A_j(t),$$

say. It follows that (19) becomes

$$\begin{aligned} \pi_{ED}(\boldsymbol{r}|\boldsymbol{\beta},\theta) &= \left\{\prod_{j=1}^{n}\gamma_j\left(\int_{-\infty}^{r_j}\sum_{\substack{k=1\\ k\neq j}}^{n}\beta_{kj}\exp\{-\gamma_j(r_j-t)-\gamma_k(r_k-t)-A_j(t)\}\,1_{\{t<r_k\}}\,dt\right)\right\} \\ &\quad\times\exp\left(-\sum_{j=n+1}^{N}\sum_{k=1}^{n}\beta_{kj}\gamma_k^{-1}\right), \end{aligned}$$

where the integral term does not appear to be available in closed form and thus must be evaluated numerically.

3.5 Erlang infectious periods

In this section we assume that infectious period random variable I_j has an Erlang distribution, i.e. a Gamma distribution with positive integer shape parameter m_j and rate parameter ν_j. Thus $\theta_j = (m_j, \nu_j)$, and we write $I_j \sim \Gamma(m_j, \nu_j)$. Such a model is usually more appropriate for real-life diseases than the assumption of exponential infectious period distributions.

Let $x > 0$. We have $f_j(x|\theta_j) = \nu_j^{m_j} x^{m_j-1} \exp(-\nu_j x)/(m_j-1)!$ and for $j = 1, \dots, n$, (7) becomes

$$g_j(x|\theta_j) = \frac{\nu_j^{m_j} x^{m_j-1} \exp(-(\nu_j + B_j)x)/(m_j-1)!}{[\nu_j/(\nu_j + B_j)]^{m_j}} = \delta_j^{m_j} x^{m_j-1} \exp(-\delta_j x)/(m_j-1)!,$$

where $\delta_j = \nu_j + B_j$, so that g_j is the probability density function of a $\Gamma(m_j, \delta_j)$ random variable, and

$$a(\theta_j, B_j) = a((m_j, \nu_j), B_j) = [\nu_j/(\nu_j + B_j)]^{m_j} = (\nu_j/\delta_j)^{m_j}.$$

3.5.1 The standard pair-based approximation with Erlang infectious periods

Lemma 4 *Let* $1 \le j, k \le n$ *with* $j \ne k$, *and* $\beta > 0$. *Then*

$$\begin{aligned}
&\mathbb{E}_{g_j,g_k}\left[\exp(-\beta\tau_{kj})\right] \\
&= \begin{cases}
1 + \exp(-\delta_k(r_k - r_j))\delta_j^{m_j} \sum_{l=0}^{m_k-1} \delta_k^l \left[\left(\frac{\delta_k}{\delta_k+\beta}\right)^{m_k-l} - 1\right] & \\
\times \sum_{p=0}^{l} \frac{1}{(l-p)!} \binom{m_j+p-1}{p} \frac{(r_k-r_j)^{l-p}}{(\delta_j+\delta_k)^{m_j+p}} & \textit{if } r_j < r_k, \\
1 - F_{m_j,\delta_j}(r_j - r_k)\left[1 - \left(\frac{\delta_k}{\delta_k+\beta}\right)^{m_k}\right] & \\
+ \exp(-\delta_k(r_j - r_k))\delta_j^{m_j} \sum_{l=0}^{m_k-1} \delta_k^l \left[\left(\frac{\delta_k}{\delta_k+\beta}\right)^{m_k-l} - 1\right] & \\
\times \sum_{p=0}^{m_j-1} \frac{1}{(m_j-p-1)!} \binom{l+p}{p} \frac{(r_j-r_k)^{m_j-p-1}}{(\delta_j+\delta_k)^{l+p+1}} & \textit{if } r_j > r_k,
\end{cases}
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{g_j,g_k}\left[1_{\{i_k<i_j<r_k\}} \exp(-\beta\tau_{kj})\right] \\
&= \begin{cases}
\exp(-\delta_k(r_k - r_j))\delta_j^{m_j} \delta_k^{m_k} \sum_{l=0}^{m_k-1} \left(\frac{1}{\delta_k+\beta}\right)^{m_k-l} & \\
\times \sum_{p=0}^{l} \frac{1}{(l-p)!} \binom{m_j+p-1}{p} \frac{(r_k-r_j)^{l-p}}{(\delta_j+\delta_k)^{m_j+p}} & \textit{if } r_j < r_k, \\
\exp(-\delta_k(r_j - r_k))\delta_j^{m_j} \delta_k^{m_k} \sum_{l=0}^{m_k-1} \left(\frac{1}{\delta_k+\beta}\right)^{m_k-l} & \\
\times \sum_{p=0}^{m_j-1} \frac{1}{(m_j-p-1)!} \binom{l+p}{p} \frac{(r_j-r_k)^{m_j-p-1}}{(\delta_j+\delta_k)^{l+p+1}} & \textit{if } r_j > r_k,
\end{cases}
\end{aligned}$$

where $F_{m,\nu}$ denotes the distribution function of a $\Gamma(m, \nu)$ random variable.

3.5.2 The Eichner and Dietz approximation with Erlang infectious periods

First note that

$$\mathbb{E}[\lambda_j(u)] = \sum_{\substack{k=1 \\ k \ne j}}^{n} \beta_{kj} 1_{\{u<r_k\}} \exp(-\nu_k(r_k - u)) \sum_{l=0}^{m_k-1} \frac{(\nu_k(r_k - u))^l}{l!},$$

from which we obtain

$$\begin{aligned}\int_{-\infty}^{t} \mathbb{E}[\lambda_j(s)]\, ds &= \sum_{\substack{k=1\\k\neq j}}^{n} \beta_{kj}\nu_k^{-1} \exp\{-\nu_k(r_k-(t\wedge r_k))\} \sum_{l=0}^{m_k-1} \frac{[\nu_k(r_k-(t\wedge r_k))]^l}{l!}(m_k-l)\\ &= C_j(t),\end{aligned}$$

say. It follows that

$$\begin{aligned}&\pi_{ED}(\boldsymbol{r}|\boldsymbol{\beta},\boldsymbol{\theta}) =\\ &\quad\left\{\prod_{j=1}^{n}\left(\int_{-\infty}^{r_j} f_j(r_j-t|m_j,\nu_j)\sum_{\substack{k=1\\k\neq j}}^{n}\beta_{kj}\exp\{-\nu_k(r_k-t)-C_j(t)\}\, 1_{\{t<r_k\}}\sum_{l=0}^{m_k-1}\frac{[\nu_k(r_k-t)]^l}{l!}\,dt\right)\right\}\\ &\quad\times\exp\left(-\sum_{j=n+1}^{N}\sum_{k=1}^{n}\beta_{kj}m_k\nu_k^{-1}\right),\end{aligned}$$

where $f_j(r_j-t|m_j,\nu_j) = \nu_j^{m_j}(r_j-t)^{m_j-1}\exp(-\nu_j(r_j-t))/(m_j-1)!$.

4 Applications to data

Having derived PBLA methods, it is natural to assess their performance for both simulated and real data. Here we briefly describe the findings of a simulation study, and then illustrate the PBLA methods via three examples involving real-life data. In each case of the latter, the setting goes beyond that of a simple SIR model in a homogeneously-mixing population, thus illustrating the potential flexibility of the PBLA methods. For comparison, in each case we also provide results from an alternative analysis such as standard MCMC with data-augmentation.

4.1 Simulation study

Details of an extensive simulation study can be found in the Supplementary Material, in which the performance of the PBLA methods is explored for the SIR model across a range of data sets and parameter values. Comparisons with other methods are also described. The focus is on the homogenously-mixing case, since it seems natural to assess the methods in this setting. Broadly speaking the methods (i) are found to work well in situations where the proportion of individuals infected is not larger than around 70%, (ii) are competitive with data-augmented MCMC methods for large population sizes, and (iii) improve in accuracy as the shape parameter of the Erlang distribution increases. As an example, Figure 1 shows maximum likelihood estimates taken from 1000 simulated data sets for both the PBLA and the Eichner-Dietz methods. Full details are given in the Supplementary Material.

4.2 Respiratory disease in Tristan da Cunha

We now apply our methods to a data set described in full and analysed in Becker and Hopper (1983) and Hayakawa *and others* (2003). The data set consists of case diagnosis times of individuals who contracted the common cold during an outbreak which occurred between October and November of 1967 on the island of Tristan da Cunha in the South Atlantic. The population of 255 islanders comprised three age groups, namely infants, children and adults, which we label 1, 2 and

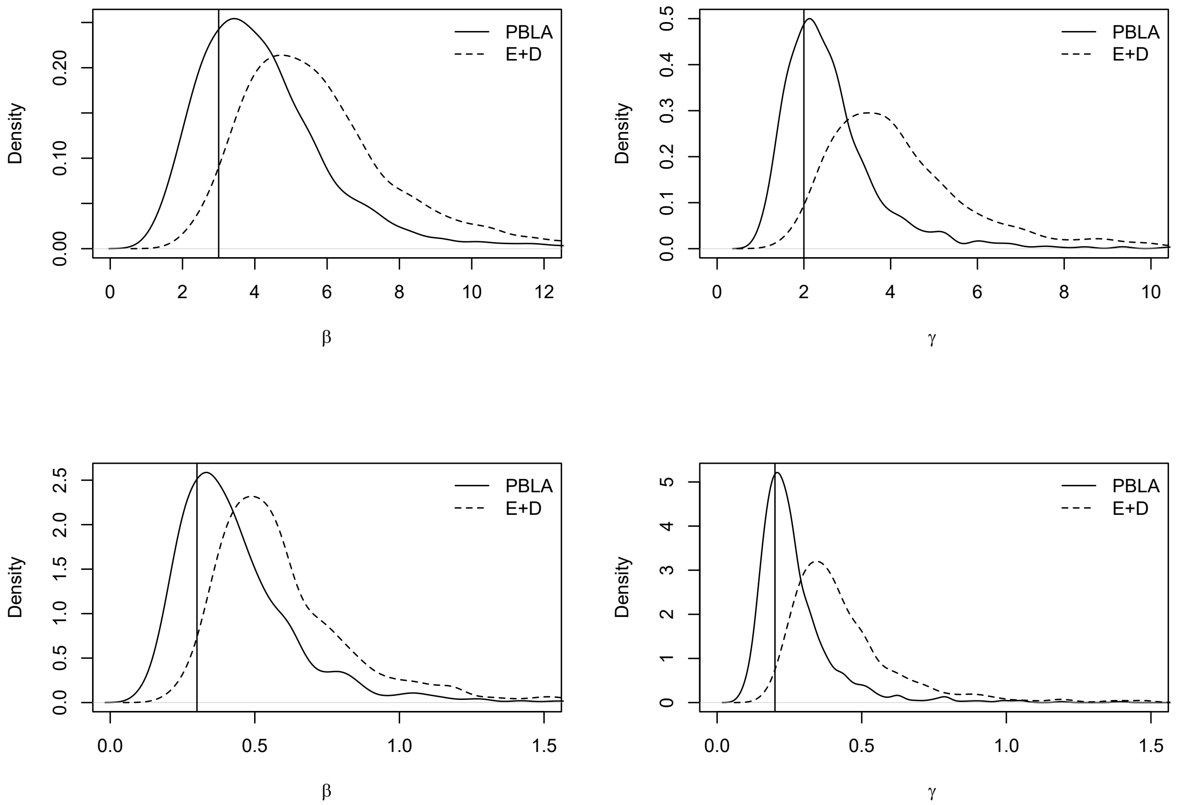

Figure 1: Comparison of maximum likelihood estimates as β and γ vary, using simulated data sets with exponential infectious periods and $N = 100$, $R_0 = 1.5$. Top panels: $\beta = 3$, $\gamma = 2$. Bottom panels: $\beta = 0.3$, $\gamma = 0.2$. Vertical lines show the true values.

3 respectively. As there was one unidentified case, we suppose that $N = 254$. The initial number of susceptibles in each group are $N_1 = 25$, $N_2 = 36$ and $N_3 = 192$. We assume that the initial infective is the individual who was diagnosed first, which is not unreasonable since one week elapsed between the first and second diagnosed cases. The number of cases in each group was $n_1 = 9$, $n_2 = 6$ and $n_2 = 25$.

4.2.1 Transmission model

Following Hayakawa *and others* (2003) we consider a multi-type stochastic SIR model in which the population is divided into three groups. Infectious periods are exponentially distributed with mean γ^{-1}, and the infection rate from individual i to j is $\beta_{ij} = \beta_{G(j)}$, where $G(j)$ denotes the group (1, 2 or 3) of individual j. This model assumes that the population mixes homogeneously and that all infectives are equally infectious, but that the susceptibility of individuals depends on their age group. We relate this model to the data by assuming that case diagnosis times correspond to removal times. We carried out a Bayesian analysis using MCMC in which the target density is the posterior distribution of the four model parameters given the observed removal times under the assumption of the PBLA likelihood. We used the independent prior distributions in Hayakawa *and others* (2003), namely that $\gamma \sim \Gamma(10^{-4}, 10^{-3})$, $\beta_j \sim \Gamma(10^{-8}, 10^{-5})$ for $j = 1, 2, 3$.

	PBLA mean	MCMC mean	E-D MAP
β_1	0.00641	0.00451	0.0104
β_2	0.00239	0.00181	0.00408
β_3	0.00171	0.00121	0.00289
γ	0.499	0.371	0.879
R_0	1.2	1.2	1.1

Table 1: Posterior means from PBLA method and from data-augmented MCMC methods (Hayakawa *and others*, 2003), and maximum *a posteriori* (MAP) estimates using the Eichner-Dietz method, for the Tristan da Cunha data set.

4.2.2 Results

Figure 2 shows marginal posterior distributions of the four model parameters from the PBLA analysis and for R_0, and also the corresponding posterior means from the analysis in Hayakawa *and others* (2003) which was carried out using MCMC methods featuring data augmentation for the unknown infection times. The latter can be regarded as the 'gold-standard' results in the Bayesian setting. Table 1 contains numerical values for the posterior means from the PBLA and data-augmented MCMC approaches. There is good agreement which shows that the PBLA methods provide a good approximation in this case. For comparison, Figure 2 also shows maximum *a posteriori* estimates using the Eichner-Dietz approach, which appears to perform rather less well than PBLA here.

4.3 Ebola in West Africa

Our second example uses publicly available data from the Centers for Disease Control and Prevention (CDC) on the outbreaks of Ebola virus in Guinea, Sierra Leone and Liberia during 2014. Each data set consists of the numbers of deaths each day due to Ebola. For comparison, we also fit these data to a deterministic epidemic model similar to one proposed by Althaus (2014). This model features latent periods, i.e. it is a Susceptible-Exposed-Infective-Removed (SEIR) model, and also a time-dependent infection rate. We thus have to adapt the PBLA approach to incorporate both these features.

4.3.1 Latent periods in the PBLA framework

The model described in Section 2.1 can be extended to an SEIR model by stipulating that when a susceptible individual j is contacted by an infective, the susceptible immediately enters a latent (or exposed) period at time e_j, say, before becoming infective at time i_j. During the latent period, the individual is unable to infect others, and cannot themselves be re-infected. In the following, we shall assume that latent periods are of a known fixed duration c. Although this is partly for analytical convenience, in reality it is pragmatic to make strong assumptions about either infectious or latent periods if the only available data are removal times. This is essentially because a single data point r_j is insufficient to estimate both e_j and i_j separately without additional assumptions.

Without latent periods, infection times such as i_j play two roles in the PBLA approximation, namely (i) the start of j's infectious period and (ii) the time at which j becomes infected. With latent periods, these times are i_j and e_j, respectively. For example, the quantity τ_{kj} defined at (4) now becomes

$$\tau_{kj} = r_k \wedge e_j - i_k \wedge e_j.$$

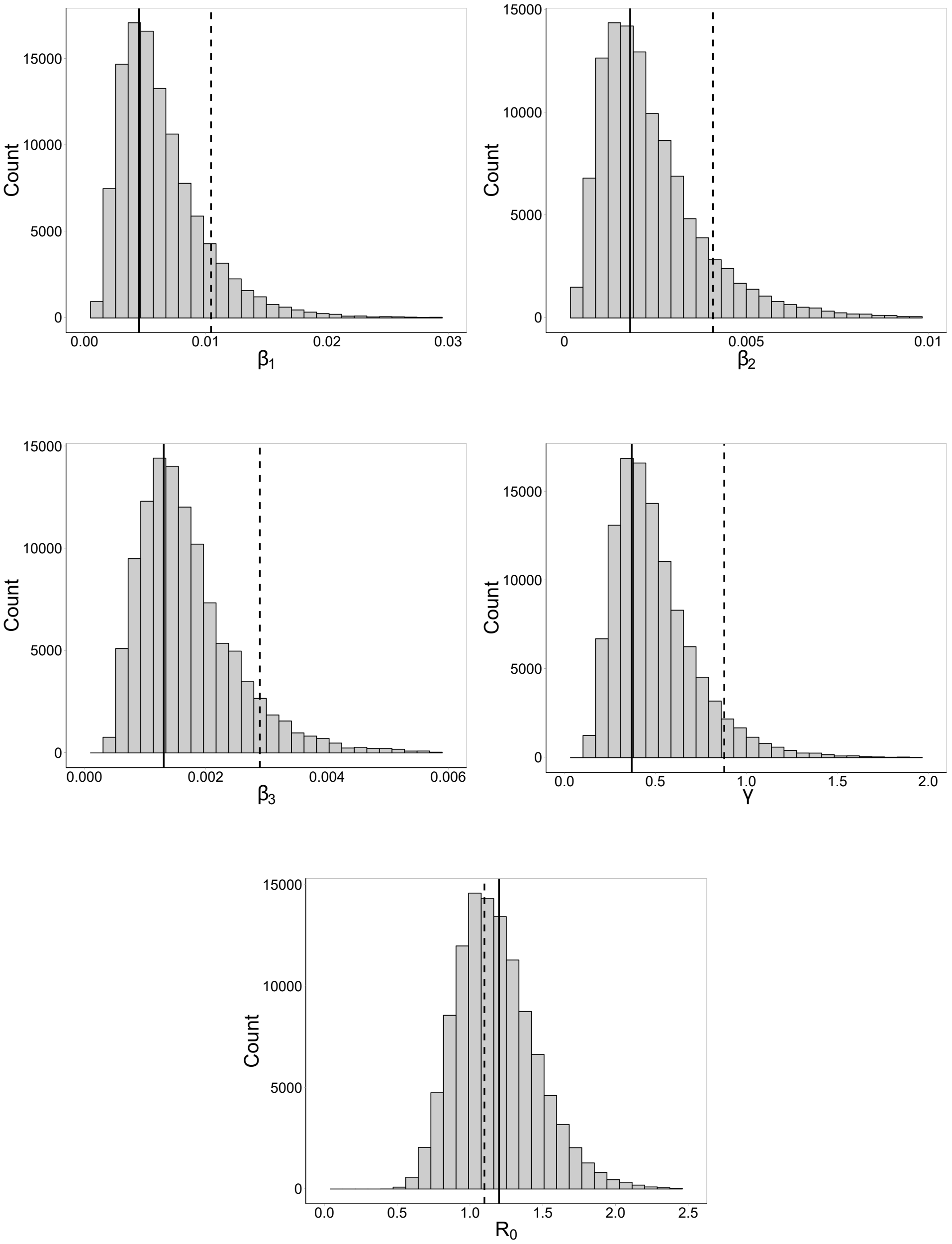

Figure 2: Marginal posterior distributions for the Tristan Da Cunha data using the PBLA method. For comparison, solid lines show the posterior means from Hayakawa *and others* (2003) and dashed lines show maximum *a posteriori* estimates from the Eichner-Dietz method.

However, $e_j = i_j - c$, and so the probability distribution of e_j given that r_j is identical to the distribution of i_j given $r_j - c$. By following such arguments, it can be shown the results in lemmas 1 and 4 can be used for the SEIR model by simply replacing r_j with $r_j - c$ throughout.

4.3.2 Deterministic model for Ebola

Let $s(t)$, $e(t)$, $i(t)$ and $r(t)$ denote respectively the numbers of susceptible, exposed, infective and removed individuals in a population of size N at time t and assume that $s(t)+e(t)+i(t)+r(t) = N$ for all $t \geq \tau_0$, where τ_0 denotes the initial time point of the epidemic. We define a deterministic model by the set of differential equations

$$\begin{aligned} \frac{ds}{dt} &= -\beta(t)\frac{si}{N}, \\ \frac{de}{dt} &= \beta(t)\frac{si}{N} - \sigma e, \\ \frac{di}{dt} &= \sigma e - \gamma i, \\ \frac{dr}{dt} &= \gamma i, \end{aligned}$$

with initial conditions $s(\tau_0) = N - 1$, $i(\tau_0) = 1$, where

$$\beta(t) = b_0 \exp(-k(t + \tau_0)).$$

This model is a simplification of that proposed by Althaus (2014), the difference being that the latter also accounts for non-fatal cases. This means we can compare our methods (which are designed for data on one kind of observation, namely removal times), directly with the deterministic modelling approach without having to make extra assumptions for how to deal a second kind of observation, namely non-fatal cases.

The time-dependent infection rate $\beta(t)$ is motivated by the impact of control measures; τ_0 is the time at which the initial infective appears, and is relevant because when fitting the model to data, it is necessary to decide when the epidemic begins. The parameters σ and γ are the rates at which individuals move from the exposed to infective and infective to removed classes, respectively.

Following Althaus (2014), we assume that (i) the average lengths of the latent and infectious periods are $\sigma^{-1} = 5.3$ and $\gamma^{-1} = 5.61$ days, respectively, and that $N = 10^6$ for each country, while (ii) the parameter τ_0 is known for the Guinea outbreak but unknown for the other outbreaks. The remaining parameters b_0 and k, and τ_0 if required, are estimated from the data as follows, again using the approach of Althaus (2014). First note that the data take the form $\{r(t_i) : i = 1, \ldots, M\}$, i.e. M observations of the total number of removals, which are assumed to correspond to deaths. A likelihood can be constructed by assuming the observed number of removals at time t is drawn from a Poisson distribution with mean $r(t)$, where $r(t)$ can be computed by numerical solution of the differential equation system, and with independence between different observations. It is then straightforward to obtain numerical maximum likelihood estimates of the model parameters.

4.3.3 PBLA method

To implement the PBLA method we assume that infectious periods are exponential with mean $\gamma^{-1} = 5.61$ days and latent periods are all $c = \sigma^{-1} = 5.3$ days. As usual we assume that individuals $1, \ldots, n$ are those who become infected and set $r_1 < \ldots < r_n$. For simplicity we set τ_0 equal to the

		b_0	k
PBLA	Guinea	0.243	0.00105
	Sierra Leone	0.335	0.00289
	Liberia	0.266	0.00214
Althaus	Guinea	0.231	0.000712
	Sierra Leone	0.277	0.00180
	Liberia	0.303	0.00251

Table 2: Maximium likelihood estimates from the PBLA method and from the Althaus model for the Ebola deaths data from CDC.

estimated values from Althaus (2014). Since the PBLA method assumes that the infection rate between any two individuals is fixed, and not time-dependent, we set

$$\beta_{jk} = b_0 \exp(-k(T_{jk} + \tau_0))$$

where T_{jk} is the expected mid-point of the time during which j can infect k. Thus for $1 \leq j, k \leq n$,

$$\begin{aligned} T_{jk} &= (\mathbb{E}[r_j \wedge e_k] + \mathbb{E}[i_j \wedge e_k])/2 \\ &= \begin{cases} r_k - \gamma^{-1} - \sigma^{-1} - (4\gamma)^{-1} \exp(-\gamma(r_j - r_k + \sigma^{-1})) & \text{if } r_j > r_k - \sigma^{-1}, \\ r_j - (2\gamma)^{-1} + 3(4\gamma)^{-1} \exp(-\gamma(r_k - r_j - \sigma^{-1})) & \text{if } r_j < r_k - \sigma^{-1}, \end{cases} \end{aligned}$$

while for $1 \leq j \leq n$ and $k > n$, $T_{jk} = r_j - (2\gamma)^{-1}$. We found in practice that other reasonable definitions of T_{jk}, e.g. taking into account the exponentially decaying nature of $\beta(t)$, gave similar results.

4.3.4 Results

Since the PBLA method is not designed to approximate a Poisson likelihood for an ordinary differential equation model, it is interesting to see how the two approaches compare. Figure 3 shows profile log-likelihood plots for the PBLA method, with maximum likelihood estimates from the Althaus model for comparison. Table 2 contains the numerical values of the maximum likelihood estimates for both approaches. It can be seen that the PBLA method gives reasonably similar results. Point estimates of the basic reproduction number, $R_0 = \beta_0/\gamma$, for PBLA (Althaus) are 1.4 (1.3), 1.9 (1.6) and 1.7 (1.5) for Guinea, Sierra Leone and Liberia respectively, which again show reasonable agreement.

4.4 Foot-and-Mouth disease in Cumbria, UK

Our final application concerns a large data set taken from the 2001 Foot and Mouth disease outbreak in the UK county of Cumbria. In this outbreak, the disease spread between farms, and if detected on a farm the animals there were culled in order to prevent further transmission. The particular data that we consider are described in Kypraios (2007), Jewell *and others* (2009) and Xiang and Neal (2014). In summary, for each farm in Cumbria the data comprise (i) its geographic location, (ii) the numbers of cattle and sheep, and (iii) the culling date if the farm was deemed to have been infected. In total, $n = 1021$ of $N = 5378$ farms were infected.

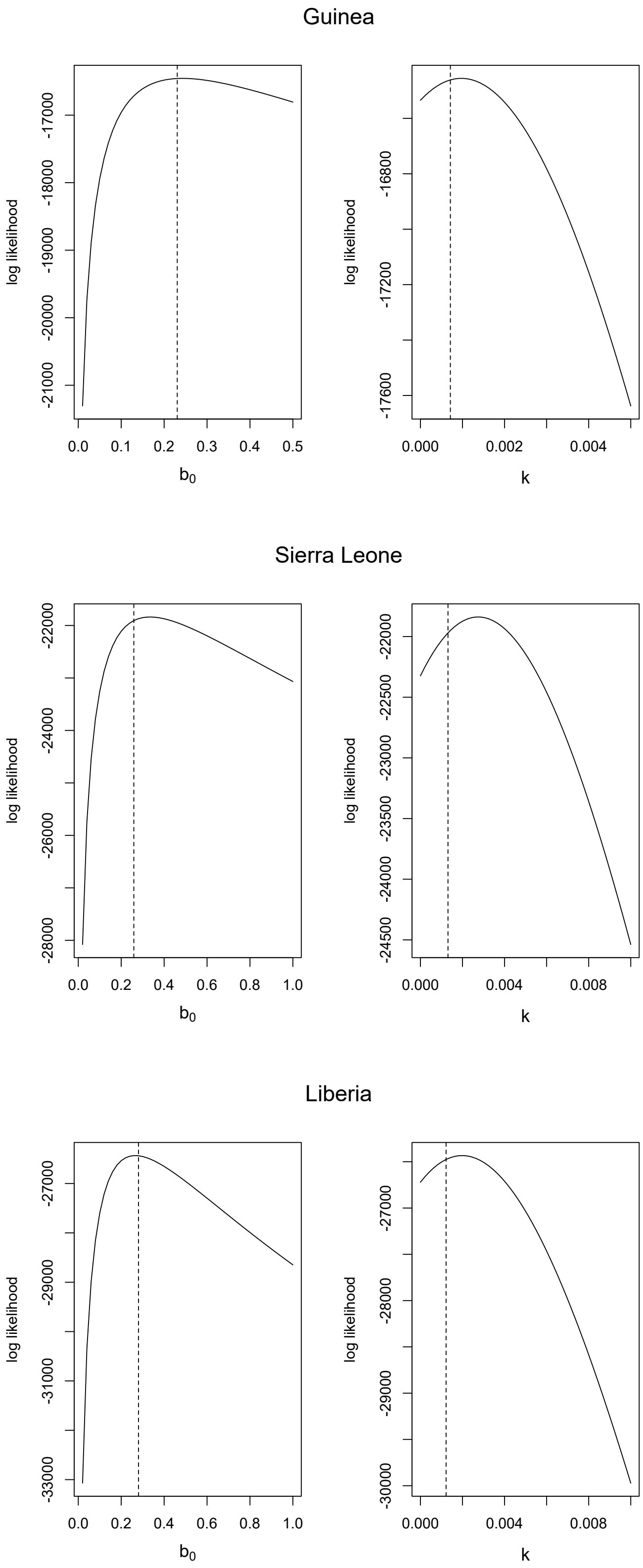

Figure 3: Profile log-likelihoods for b_0 and k for the PBLA log-likelihood, using Ebola deaths data from CDC. For comparison, the dashed lines show the maximum likelihood estimates using the Althaus model.

θ	$\hat{\theta}_{MAP}$ (PBLA)	$E[\theta\|\boldsymbol{r}]$	Rel. Diff.	$\hat{\theta}_{MAP}$ (E-D)
β_0	7.05×10^{-7}	6.07×10^{-7}	0.16	7.10×10^{-7}
γ	0.45	0.52	0.13	0.54
v	0.0048	0.0065	0.26	0.0049
ε	1.57	1.45	0.08	2.32
ξ	2.39	2.32	0.03	2.20
ζ	0.32	0.32	0	0.31

Table 3: MAP estimates (PBLA method) and posterior means from Kypraios (2007) for the Foot and Mouth data set, along with the relative difference $|\hat{\theta}_{MAP} - E[\theta|\boldsymbol{r}]|/E[\theta|\boldsymbol{r}]$. For comparison, the final column shows the Eichner-Dietz MAP estimates.

4.4.1 Transmission model

Kypraios (2007) describes a stochastic SIR transmission model in which individuals are farms, infectious periods are assumed to follow independent $\Gamma(4, \gamma)$ distributions, and the infection rate from farm i to farm j is given by

$$\beta_{ij} = \beta_0 \frac{v}{\rho_{ij}^2 + v^2} (\varepsilon (n_i^c)^\zeta + (n_i^s)^\zeta)(\xi (n_j^c)^\zeta + (n_j^s)^\zeta),$$

where ρ_{ij} denotes the Euclidean distance between farms i and j, while n_i^c and n_i^s denote respectively the number of cattle and sheep on farm i. The model is thus explicitly spatial and multi-type, and has six parameters. This model is related to the data by assuming that culling dates correspond to removal events. Kypraios (2007) carries out parameter estimation in a Bayesian framework by augmenting the parameter space with the unknown infection times. This is a very computationally demanding approach, due to the combination of a large number of infected farms, a six-dimensional model parameter space, and the inherent posterior correlations between the infection times and the model parameters.

We adopted the same independent prior distributions as those of Kypraios (2007), namely that $\beta_0, \gamma \sim \Gamma(0.001, 0.001)$, $v \sim Exp(0.1)$ and $\varepsilon, \xi, \zeta \sim Exp(0.001)$. We used the PBLA method to obtain maximum *a posteriori* (MAP) point estimates for all six parameters.

4.4.2 Results

Figure 4 shows profile log-likelihood plots along with the posterior mean estimates from Kypraios (2007), and table 3 compares the latter with the MAP estimates from the PBLA analysis. MAP estimates from the Kypraios analysis are not available, but since the marginal posterior density plots reported are reasonably symmetric then the posterior means would presumably be fairly close. It can be seen that there is reasonable agreement between the PBLA approach and the Kypraios analysis. For comparison we also present the corresponding MAP estimates from the Eichner-Dietz approximation, which are slightly less accurate than the PBLA estimates.

5 Conclusions

We have developed likelihood approximation methods for partially observed stochastic epidemic models. We regard such methods as an addition to the toolkit for analysing infectious disease data, with potential to provide parameter estimates in situations where other methods (such as the 'gold-standard' of data-augmented MCMC, or likelihood-free methods such as Approximate

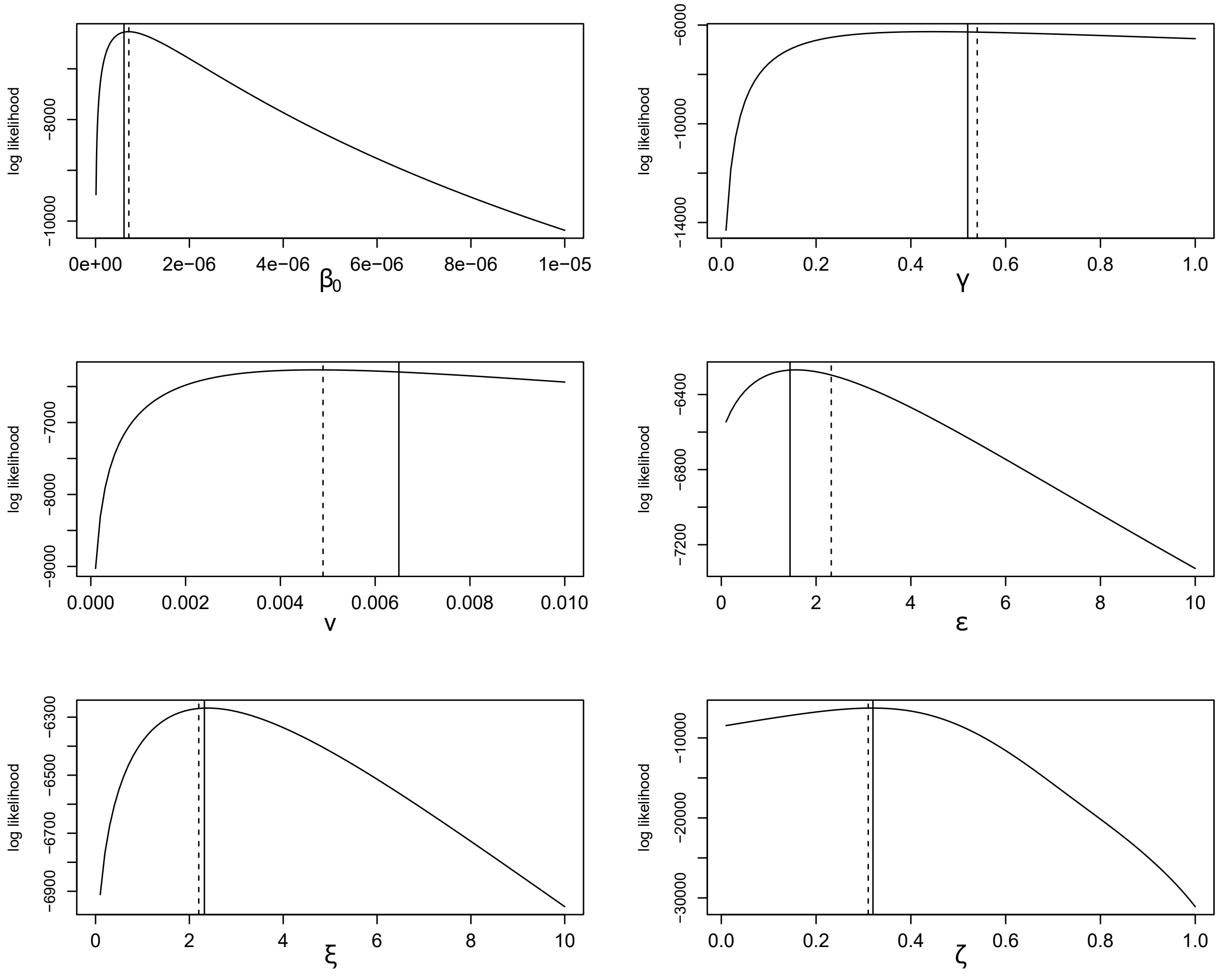

Figure 4: Profile log-likelihoods for the Foot and Mouth disease model parameters. For comparison, the solid lines show posterior mean estimates from Kypraios (2007) and the dashed lines show corresponding Eichner-Dietz MAP estimates.

Bayesian Computation) may struggle due to their computational burden. Our work is somewhat preliminary but demonstrates that likelihood-approximation approaches have useful potential.

Our approach can be summarised as follows. Broadly speaking, the true likelihood can be described by considering all individuals who ever become infected. It then takes account of these individuals avoiding infection, becoming infected, remaining infected for a period of time, and either infecting or failing to infect others whilst infective. Due to the fact that these events are dependent for different individuals, the likelihood is then typically intractable. The key to our approximation method is to consider pairs of ever-infected individuals in the population, since the likelihood contribution from such pairs are often tractable, depending on the choice of infectious period distribution. We then assume independence between such pairs in order to form an approximate likelihood.

The PBLA approach appears to work reasonably well in settings encountered in practice, specifically where the total proportion of the population infected is not too large. This makes intuitive sense, since the extent to which the likelihood can be approximated by independent components will clearly become less plausible as the proportion infected increases. The general idea of basing approximations on the interactions of pairs of individuals is widely applicable, as we have demonstrated via several examples. It seems likely that such methods could also be applied to models where individuals enter and leave the population, for instance models for nosocomial infections in hospital wards (Wei *and others*, 2018). Another possible extension is to epidemics in progress, relaxing our assumption that the observed outbreak has terminated. The challenge in that situation is that both the infection times and total number of ever-infected individuals in the population are both unknown.

Unlike data-augmented MCMC, the PBLA approach gives a relatively fast way to obtain approximate maximum likelihood values for model parameters. For small data sets the PBLA method can produce maximum likelihood estimates in seconds or less, while even the Ebola data set with $N = 10^6$ individuals takes around one minute on a standard laptop. Further work is needed to develop approaches that can improve the speed and accuracy of the approximation across more scenarios, for instance by taking greater account of the fact that not all configurations of the unknown infection times are possible if the epidemic is not to die out prematurely.

6 Supplementary Material

Supplementary material is available online at `http://biostatistics.oxfordjournals.org` and code and data sets from the paper are available at `https://github.com/jessicastockdale/PBLA`

7 Funding

Jessica Stockdale was supported by Engineering and Physical Sciences Research Council grant EP/M506588/1.

8 Acknowledgements

We thank an Associate Editor and two reviewers for helpful comments that have improved the paper.

References

Althaus, Christian L. (2014). Estimating the reproduction number of ebola virus (ebov) during the 2014 outbreak in west africa. *PLOS Current Outbreaks* **6**.

Andersson, Håkan and Britton, Tom. (2000). *Stochastic Epidemic Models and their Statistical Analysis*, Volume 4. New York: Springer.

Bailey, Norman T. J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications*. New York: Charles Griffin & Company Ltd.

Becker, Niels G. and Hopper, John L. (1983). Assessing the heterogeneity of disease spread through a community. *American Journal of Epidemiology* **117**, 362–374.

Clancy, Damian and O'Neill, Philip D. (2008). Bayesian estimation of the basic reproduction number in stochastic epidemic models. *Bayesian Analysis* **3**(4), 737–757.

Eichner, Martin and Dietz, Klaus. (2003). Transmission potential of smallpox: estimates based on detailed data from an outbreak. *American Journal of Epidemiology* **158**(2), 110–117.

Gibson, Gavin J and Renshaw, Eric. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology* **15**(1), 19–40.

Hayakawa, Yu, O'Neill, Philip D., Upton, Darren and Yip, Paul S.F. (2003). Bayesian inference for a stochastic epidemic model with uncertain numbers of susceptibles of several types. *Australian & New Zealand Journal of Statistics* **45**(4), 491–502.

Jewell, Chris P, Kypraios, Theodore, Neal, Peter and Roberts, Gareth O. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis* **4**(3), 465–496.

Keeling, Matt J, Rand, David A and Morris, Andrew J. (1997). Correlation models for childhood epidemics. *Proceedings of the Royal Society, Series B (Biological Sciences)* **264**, 1149–1156.

Kypraios, Theodore. (2007). Efficient bayesian inference for partially observed stochastic epidemics and a new class of semi-parametric time series models [Ph.D. Thesis]. Lancaster University.

O'Neill, Philip D. and Roberts, Gareth O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **162**(1), 121–129.

Stockdale, Jessica E, O'Neill, Philip D and Kypraios, Theodore. (2017). Modelling and Bayesian analysis of the Abakaliki smallpox data. *Epidemics* **19**, 13–23.

Wei, Yinghui, Kypraios, Theodore, O'Neill, Philip D, Huang, Susan S, Rifas-Shiman, Sheryl L and Cooper, Ben S. (2018). Evaluating hospital infection control measures for antimicrobial-resistant pathogens using stochastic transmission models: Application to vancomycin-resistant enterococci in intensive care units. *Statistical Methods in Medical Research* **27**(1), 269–285.

Xiang, Fei and Neal, Peter. (2014). Efficient MCMC for temporal epidemics via parameter reduction. *Computational Statistics & Data Analysis* **80**, 240–250.

Pair-based likelihood approximations for stochastic epidemic models Supplementary Material

J. E. Stockdale[*], T. Kypraios[†], P. D. O'Neill[†]

1 Proofs

1.1 Proofs of Lemmas 1 and 4

We give two methods of proof for the first result in lemma 1; the other parts of both lemmas can be proved in a similar fashion. First note that

$$\tau_{kj} = r_k \wedge i_j - i_k \wedge i_j = \begin{cases} 0 & \text{if } i_j < i_k, \\ i_j - i_k & \text{if } i_k < i_j < r_k, \\ r_k - i_k & \text{if } i_j > r_k. \end{cases}$$

Thus

$$\begin{aligned} &\mathbb{E}_{g_j,g_k}\left[\exp(-\beta\tau_{kj})\right] \\ &\quad = \mathbb{E}_{g_j,g_k}\left[\exp(-\beta\tau_{kj})1_{\{i_j<i_k\}}\right] + \mathbb{E}_{g_j,g_k}\left[\exp(-\beta\tau_{kj})1_{\{i_k<i_j<r_k\}}\right] + \mathbb{E}_{g_j,g_k}\left[\exp(-\beta\tau_{kj})1_{\{i_j>r_k\}}\right] \\ &\quad = \mathbb{E}_{g_j,g_k}\left[1_{\{i_j<i_k\}}\right] + \mathbb{E}_{g_j,g_k}\left[\exp(-\beta(i_j-i_k))1_{\{i_k<i_j<r_k\}}\right] + \mathbb{E}_{g_j,g_k}\left[\exp(-\beta(r_k-i_k))1_{\{i_j>r_k\}}\right] \end{aligned} \tag{1}$$

and all three terms in (1) can be evaluated directly; for instance, if $r_j < r_k$ then

$$\begin{aligned} \mathbb{E}_{g_j,g_k}\left[1_{\{i_j<i_k\}}\right] &= \int_{-\infty}^{r_j}\int_{i_j}^{r_k} \delta_k \exp(-\delta_k(r_k-i_k))\delta_j \exp(-\delta_j(r_j-i_j))\, di_k di_j \\ &= 1 - \delta_j(\delta_j+\delta_k)^{-1}\exp(-\delta_k(r_k-r_j)), \end{aligned} \tag{2}$$

$$\begin{aligned} &\mathbb{E}_{g_j,g_k}\left[\exp(-\beta(i_j-i_k))1_{\{i_k<i_j<r_k\}}\right] \\ &\quad = \int_{-\infty}^{r_j}\int_{-\infty}^{i_j} \exp(-\beta(i_j-i_k))\delta_k \exp(-\delta_k(r_k-i_k))\delta_j \exp(-\delta_j(r_j-i_j))\, di_k di_j \\ &\quad = \delta_j\delta_k\left\{(\delta_j+\delta_k)(\delta_k+\beta)\right\}^{-1}\exp(-\delta_k(r_k-r_j)), \end{aligned} \tag{3}$$

[*] Department of Mathematics, Simon Fraser University

[†] School of Mathematical Sciences, University of Nottingham

and the third term is zero since $i_j < r_j < r_k$, whence $1_{\{i_j > r_k\}} = 0$. Combining (2) and (3) yields the first result in lemma 1.

Alternatively, a direct probabilistic proof is as follows. From (1) we have

$$\begin{aligned} &\mathbb{E}_{g_j,g_k}\left[\exp(-\beta\tau_{kj})\right] \\ &\quad = \mathbb{P}_{g_j,g_k}(i_j < i_k) + \mathbb{E}_{g_j,g_k}\left[\exp(-\beta(i_j - i_k))|i_k < i_j < r_k\right]\mathbb{P}_{g_j,g_k}(i_k < i_j < r_k). \end{aligned} \tag{4}$$

Both terms in (4) can be evaluated by considering a reverse-time construction of i_j and i_k, as follows (see also the proof of lemma 2 below for a similar approach). Recall that $r_j < r_k$. Start at time r_k and run a Poisson process of rate δ_k backwards in time. If a point occurs at time $t \in (r_j, r_k)$ then set $i_k = t$, and run a new Poisson process of rate δ_j backwards in time from r_j and set i_j equal to the time of its first point. If not, at time r_j increase the Poisson process rate to $\delta_j + \delta_k$; when the first point appears at time $t < r_j$, set $i_j = t$ with probability $\delta_j(\delta_j + \delta_k)^{-1}$, otherwise set $i_k = t$. In the former case reduce the Poisson process rate to δ_k, and in the latter reduce it to δ_j. Continue back in time; the next point to appear will be the remaining infection time. It is straightforward to see that this construction yields i_j and i_k such that $r_j - i_j$ and $r_k - i_k$ are independent $Exp(\delta_j)$ and $Exp(\delta_k)$ random variables, respectively.

Under this construction, $i_k < i_j$ if and only if (i) the Poisson process has no point in (r_j, r_k), which has probability $\exp(-\delta_k(r_k - r_j))$, and (ii) independently of this, i_j is chosen when the first point of the rate-$(\delta_j + \delta_k)$ Poisson process appears, which has probability $\delta_j(\delta_j + \delta_k)^{-1}$. Thus the first term in (4) is

$$\mathbb{P}_{g_j,g_k}(i_j < i_k) = 1 - \mathbb{P}_{g_j,g_k}(i_k < i_j) = 1 - \delta_j(\delta_j + \delta_k)^{-1}\exp(-\delta_k(r_k - r_j)),$$

in agreement with (2). Next, since $i_j < r_j < r_k$ we have

$$\mathbb{P}_{g_j,g_k}(i_k < i_j < r_k) = \mathbb{P}_{g_j,g_k}(i_k < i_j) = \delta_j(\delta_j + \delta_k)^{-1}\exp(-\delta_k(r_k - r_j)). \tag{5}$$

Finally, to evaluate the expectation term in (4), consider a Poisson process of rate β that runs backwards in time from i_j, independently of the rate-δ_k Poisson process that is used if $i_k < i_j$. The expectation is simply the probability that the former process has no points before the first point of the latter process, yielding

$$\mathbb{E}_{g_j,g_k}\left[\exp(-\beta(i_j - i_k))|i_k < i_j < r_k\right] = \delta_k(\delta_k + \beta)^{-1}$$

which along with (5) yields the same expression as (3).

The remaining cases in lemma 1, and 4, can be proved using either direct calculation or probabilistic arguments. For the case of Erlang distributed infectious periods, the probabilistic arguments require splitting infectious periods into exponentially-distributed stages. Full details can be found in Stockdale (2018).

1.2 Proof of Lemma 2

Note that in the following, we focus exclusively on the individuals in $\mathcal{K}$ and ignore all other individuals in the population. For ease of exposition, re-label the members of $\mathcal{K}$ as $1, \ldots, K$. Given these individuals' removal times $r_1 < \ldots < r_K$, their infection times $i_1, \ldots, i_K$ can be constructed as follows. Start at time r_K and consider the Markov process $\left\{(\tilde{S}(t), \tilde{I}(t)) : t < r_K\right\}$ which runs backwards in time such that (i) $(\tilde{S}(t), \tilde{I}(t)) \to (\tilde{S}(t) + 1, \tilde{I}(t) - 1)$ at rate $\delta\tilde{I}(t)$; (ii) at each removal time, $\tilde{I}(t)$ increases by 1. Here $\tilde{S}(t)$ and $\tilde{I}(t)$ denote respectively the numbers of susceptibles and infectives in $\mathcal{K}$ at time t and initially (i.e. just before time r_K) there are no susceptibles and one infective. Transitions of type 1 generate infection times, and at each such event one of the currently infected individuals is selected uniformly at random to be the individual who is infected at that time. The process terminates as soon as $S(t) = K$, which occurs when the last of the required infection times is generated; call this time i_a.

Consider the total time during which infectious pressure is exerted, i.e.

$$\sum_{j<k} \omega_{jk} = \int_{i_a}^{r_K} \tilde{S}(t)\tilde{I}(t)\ dt.$$

This quantity can be similarly constructed in reverse time as follows. For $u < r_K$ define

$$T(u) = \int_{u}^{r_K} \tilde{S}(t)\tilde{I}(t)\ dt$$

and observe that, starting at r_K and moving backwards in time, T is a piecewise linear function that increases at rate $\tilde{S}(t)\tilde{I}(t)$ at time t.

Next, consider a random time-transformation in which (i) the clock runs at rate $\tilde{I}^{-1}$ when there are $\tilde{I} \geq 1$ infectives, and (ii) the clock stops running whenever $\tilde{I} = 0$. This has no impact on the ultimate value of T, because in each time period during which $\tilde{S}(t)\tilde{I}(t)$ does not change, T increases by the same amount in both the original and transformed time scales. In the transformed time scale, still in reverse time, (i) T increases at rate $\tilde{S}$ when there are $\tilde{S}$ susceptibles, and removal times do not affect this rate because they do not change the number of susceptibles, while (ii) the number of susceptibles increases by one at rate δ, corresponding to an increase at rate $\delta\tilde{I}(t)$ in the original time-scale.

Combining these observations means that, in the transformed time-scale, T is as a function which starts at zero, and as soon as the first susceptible appears increases at rate 1; then increases at rate 2 once the second susceptible appears, and so on until all K susceptibles appear. Since the times at which susceptibles appear are the points of a Poisson process of rate δ, it follows that

$$\sum_{j<k} \omega_{jk} = T(i_a) = \sum_{j=1}^{K-1} jY_j$$

where the Y_j are independent $Exp(\delta)$ random variables, and the result follows since $jExp(\delta) \sim Exp(\delta/j)$.

1.3 Proof of Lemma 3

1.3.1 Preliminaries

For $n \geq 1$ let $D_n = \{(i,j) : 1 \leq i < j \leq n\}$ denote the set of 2-element subsets of $\mathcal{N}_n = \{1, \dots, n\}$. For $(i,j), (k,l) \in D_n$, $|(i,j) \cap (k,l)|$ is the number of elements of $\mathcal{N}_n$ that (i,j) and (k,l) have in common, which can equal either 0, 1 or 2. We require the following result, which is essentially Theorem 2.1 from Barbour and Eagleson (1985) restricted to our setting.

Lemma 1 *Let $\{X_{ij} : (i,j) \in D_n\}$ be a collection of zero-mean random variables such that $E[|X_{ij}|^3] < \infty$ for all $(i,j) \in D_n$, and X_{ij} and X_{kl} are independent if $|(i,j) \cap (k,l)| = 0$. Let*

$$\sigma_n^2 = \sum_{(i,j),(k,l) \in D_n} E[X_{ij}X_{kl}].$$

Then $\sigma_n^{-1}\sum_{(i,j)\in D_n} X_{ij}$ *converges in distribution as* $n\to\infty$ *to a standard normal random variable if*

$$\sigma_n^{-3}\sum_{(i,j)\in D_n}(E[|X_{ij}|^3])^{1/3}\left(\sum_{(k,l):|(i,j)\cap(k,l)|=0}(E[|X_{kl}|^3])^{1/3}\right)^2\to 0. \qquad (6)$$

1.3.2 Proof of Lemma 3

For $(i,j)\in D_n$ set $X_{ij}=\omega_{ij}-E[\omega_{ij}]=\omega_{ij}-\delta^{-1}$, since $\omega_{ij}\sim Exp(\delta)$ from lemma 2. Then $E[X_{ij}]=0$, $E[X_{ij}^2]=\delta^{-2}$, $(E[|X_{ij}|^3])^{1/3}=c>0$, say, $E[X_{ij}X_{kl}]=E[\omega_{ij}\omega_{kl}]$ and $\sigma_n^2=s_n^2$. Thus the left hand side of (6) equals $\sigma_n^{-3}c^3\binom{n}{2}\binom{n-2}{2}^2$ and the result follows if $\sigma_n^3=O(n^{4+\eta})$ for some $\eta>0$. Now

$$\begin{aligned}
\sigma_n^2 &= \sum_{(i,j),(k,l)\in D_n} E[X_{ij}X_{kl}]\\
&= \sum_{\substack{(i,j),(k,l)\in D_n\\ |(i,j)\cap(k,l)|=0}} E[X_{ij}X_{kl}]+\sum_{\substack{(i,j),(k,l)\in D_n\\ |(i,j)\cap(k,l)|=1}} E[X_{ij}X_{kl}]+\sum_{\substack{(i,j),(k,l)\in D_n\\ |(i,j)\cap(k,l)|=2}} E[X_{ij}X_{kl}]\\
&= \sum_{\substack{(i,j),(k,l)\in D_n\\ |(i,j)\cap(k,l)|=0}} E[X_{ij}]E[X_{kl}]+\sum_{\substack{(i,j),(k,l)\in D_n\\ |(i,j)\cap(k,l)|=1}} E[X_{ij}X_{kl}]+\sum_{(i,j)\in D_n} E[X_{ij}^2]\\
&= \sum_{\substack{(i,j),(k,l)\in D_n\\ |(i,j)\cap(k,l)|=1}} E[X_{ij}X_{kl}]+\binom{n}{2}\delta^{-2}.
\end{aligned}$$

For $1\le j<k<l\le n$ define

$$\Omega_{jkl}=\omega_{jk}\omega_{jl}+\omega_{jk}\omega_{kl}+\omega_{jl}\omega_{kl},$$

whence

$$\sum_{\substack{(i,j),(k,l)\in D_n\\ |(i,j)\cap(k,l)|=1}} E[X_{ij}X_{kl}]=\sum_{1\le j<k<l\le n}E[\Omega_{jkl}].$$

Now

$$\begin{aligned}
E[\Omega_{jkl}] &= \operatorname{cov}(\omega_{jk},\omega_{jl})+\operatorname{cov}(\omega_{jk},\omega_{kl})+\operatorname{cov}(\omega_{jl},\omega_{kl})+E[\omega_{jk}]E[\omega_{jl}]+E[\omega_{jk}]E[\omega_{kl}]+E[\omega_{jl}]E[\omega_{kl}]\\
&= \operatorname{cov}(\omega_{jk},\omega_{jl})+\operatorname{cov}(\omega_{jk},\omega_{kl})+\operatorname{cov}(\omega_{jl},\omega_{kl})+3\delta^{-2}\\
&= \left(\operatorname{var}(\omega_{jk}+\omega_{jl}+\omega_{kl})-\operatorname{var}(\omega_{jk})-\operatorname{var}(\omega_{jl})-\operatorname{var}(\omega_{kl})\right)/2+3\delta^{-2}\\
&= \big(5\delta^{-2}-3\delta^{2})\big)/2+3\delta^{-2}\\
&= 4\delta^{-2},
\end{aligned}$$

since $\operatorname{var}(\omega_{jk}+\omega_{jl}+\omega_{kl})=\operatorname{var}(Exp(\delta))+\operatorname{var}(Exp(\delta/2))=5\delta^{-2}$ from lemma 2 with $K=3$. Thus

$$\sigma_n^2=4\binom{n}{3}\delta^{-2}+\binom{n}{2}\delta^{-2}=\frac{n(n-1)(2n-1)}{6\delta^2}=O(n^3),$$

and so $\sigma_n^3=O(n^{9/2})$ as required.

2 Simulation study

In this section we assess the performance of the PBLA and ED approximations using simulated data. Throughout we assume that $\beta_{ij} = \beta N^{-1}$ for all i, j, and that infectious periods are identically distributed, since it seems most natural to assess the methods in the most basic setting. More complex settings are considered in the real-life data examples in the main text. We will also be concerned with the basic reproduction number R_0, defined as the number of infections caused by a single infective in an infinite population of susceptibles (see e.g. Andersson and Britton, 2000).

Our main focus is on the standard PBLA method, since we found that the alternative approximations described in section 3.3 in the main text were either numerically similar, or performed less well (see 2.4 below for details).

2.1 Exponential infectious periods

Suppose that $I_j \sim Exp(\gamma)$ for all $j = 1, \ldots, n$. Then $R_0 = \beta\gamma^{-1}$.

We first consider the performance of the PBLA and ED methods as various combinations of β, γ, R_0 and N vary. Specifically, in each setting we simulate a large number of data sets, typically 500 or 1000, each consisting of removal times $r_1, \ldots, r_n$. For each data set we then find maximum likelihood estimates of (β, γ) for both the PBLA and ED approximation using standard numerical optimisation methods. To provide a visual representation of the estimates we use them to create kernel density estimates.

2.1.1 Varying β and γ with N and R_0 fixed

Figure 1 illustrates that both the PBLA and ED methods are largely unaffected by variations in the actual values of β and γ, at least while N and R_0 are fixed. The PBLA method appears to perform slightly better in terms of being able to estimate the true parameter values.

2.1.2 Varying N with β and γ fixed

Figure 2 shows the impact of varying N while β and γ are fixed. Again we see that there is little variation, and that PBLA performs better than the ED method. Note that since R_0 is fixed, increasing N leads to larger values of the outbreak size n, and so here we see evidence that both approximation methods are able to cope with large n values.

2.1.3 Varying R_0

Figure 3, in conjunction with the top panel in Figure 2, shows how variation in R_0 impacts the performance of the PBLA and ED methods, specifically for $R_0 = 0.5, 1.5$ and 2. For both approximations, the performance deteriorates as R_0 increases. To understand this, first note that the value of $N - n$ contributes to the true likelihood (5), in this case via the product term

$$\prod_{j=1}^{n} \phi_j = \exp\left\{ -\beta N^{-1}(N-n) \sum_{j=1}^{n} (r_j - i_j) \right\}.$$

As the proportion of the population infected, n/N, increases, so the relative importance of the ϕ terms compared to the other terms in (5) decreases, and thus to obtain a good overall approximation it becomes increasingly important to have a good approximation for the ψ terms. However, the latter are themselves the subject of the least accurate approximation, essentially because each of

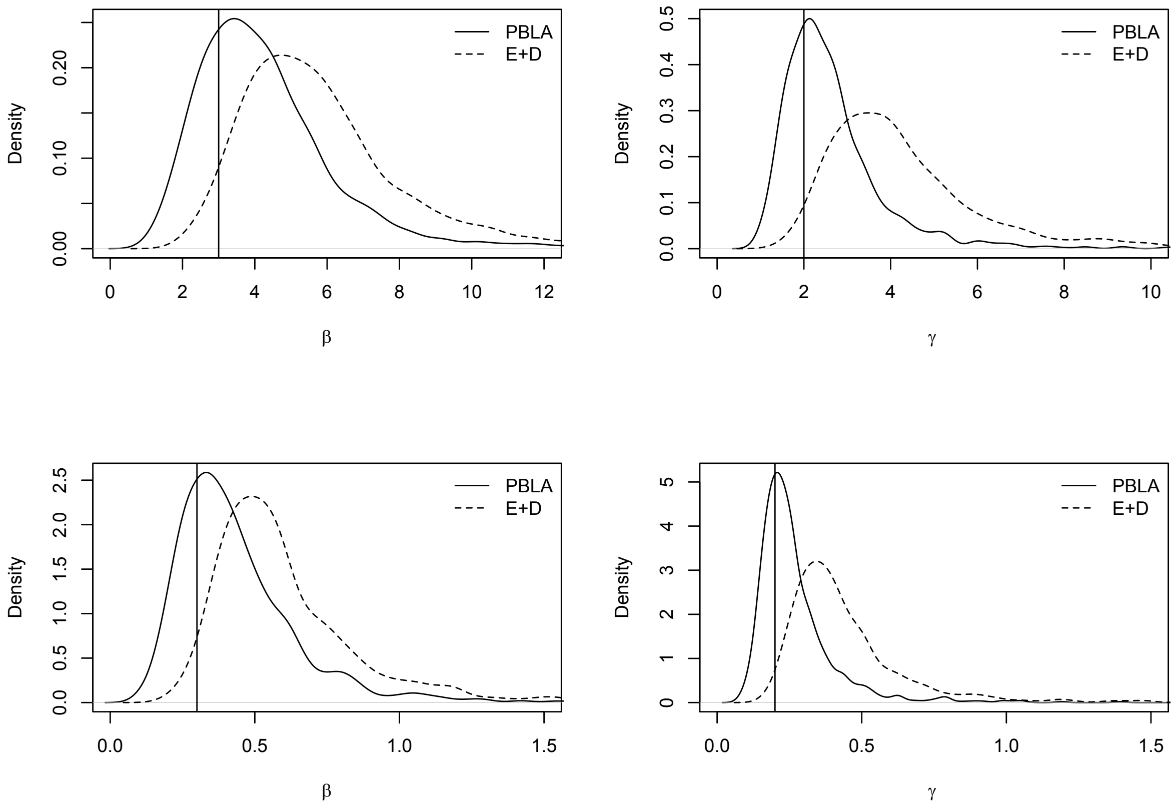

Figure 1: Comparison of maximum likelihood estimates as β and γ vary, using simulated data sets with exponential infectious periods and $N = 100$, $R_0 = 1.5$. Top panels: $\beta = 3$, $\gamma = 2$. Bottom panels: $\beta = 0.3$, $\gamma = 0.2$. Vertical lines show the true values.

the ψ_j terms is itself approximated by a product. Roughly speaking, we found that provided n/N is less than around 0.7, then the relative importance of the ϕ terms is sufficient to provide good approximations for β and γ. However, for larger values of n/N this is no longer the case, and the overall approximation suffers. From a practical point of view this does not seem to be particularly restrictive, since real-life outbreaks rarely infect such high proportions of the susceptible population.

2.2 Erlang infectious periods

Suppose now that $I_j \sim \Gamma(m, \gamma)$ for all $j = 1, \dots, n$, where m is a positive integer. Then $R_0 = \beta m \gamma^{-1}$. We now briefly show that the broad conclusions of the simulation study for the case of exponential infectious periods also hold true in this setting. We also show that the approximations improve as the variance of the infectious period decreases while its mean is kept fixed.

2.2.1 Varying β and γ with N and R_0 fixed

Figure 4 illustrates that both the PBLA and ED methods are largely unaffected by variations in the actual values of β and γ, at least while N and R_0 are fixed. The PBLA method appears to perform slightly better in terms of being able to estimate the true parameter values.

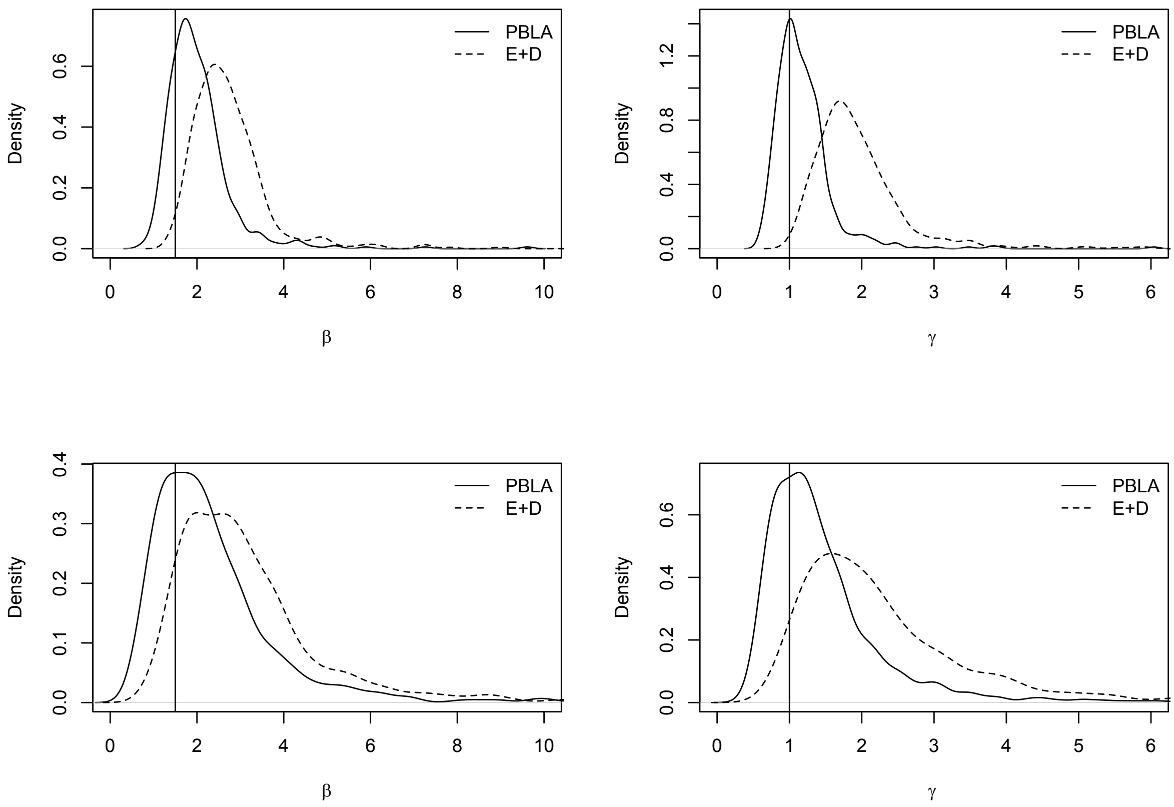

Figure 2: Comparison of maximum likelihood estimates as N varies, using simulated data sets with exponential infectious periods and $\beta = 1.5$, $\gamma = 1$, $R_0 = 1.5$. Top panels: $N = 500$. Bottom panels: $N = 40$. Vertical lines show the true values.

2.2.2 Varying N with β and γ fixed

Figure 5 shows the impact of varying N while β and γ are fixed. The results are broadly similar to the exponential infectious period case, with relatively little variation in the mode of the estimates as N increases, but that the variance of the estimates decreases with N. The latter feature suggests that having more data enables better estimation in the sense that the average error is reduced. We illustrate this graphically in figure 6 which shows mean square error values as N varies. This error reduction is clearly a desirable property of both the PBLA and ED approximation methods, but one that is not obviously true at first sight since larger data sets lead to more approximations being used in both methods.

2.2.3 Varying R_0 with m and γ fixed.

Figure 7 shows the impact of varying R_0 whilst keeping the infectious period distribution parameters m and γ fixed. As for the exponential infectious periods case, and for the same underlying reasons, estimation for both β and γ becomes less accurate as R_0 increases.

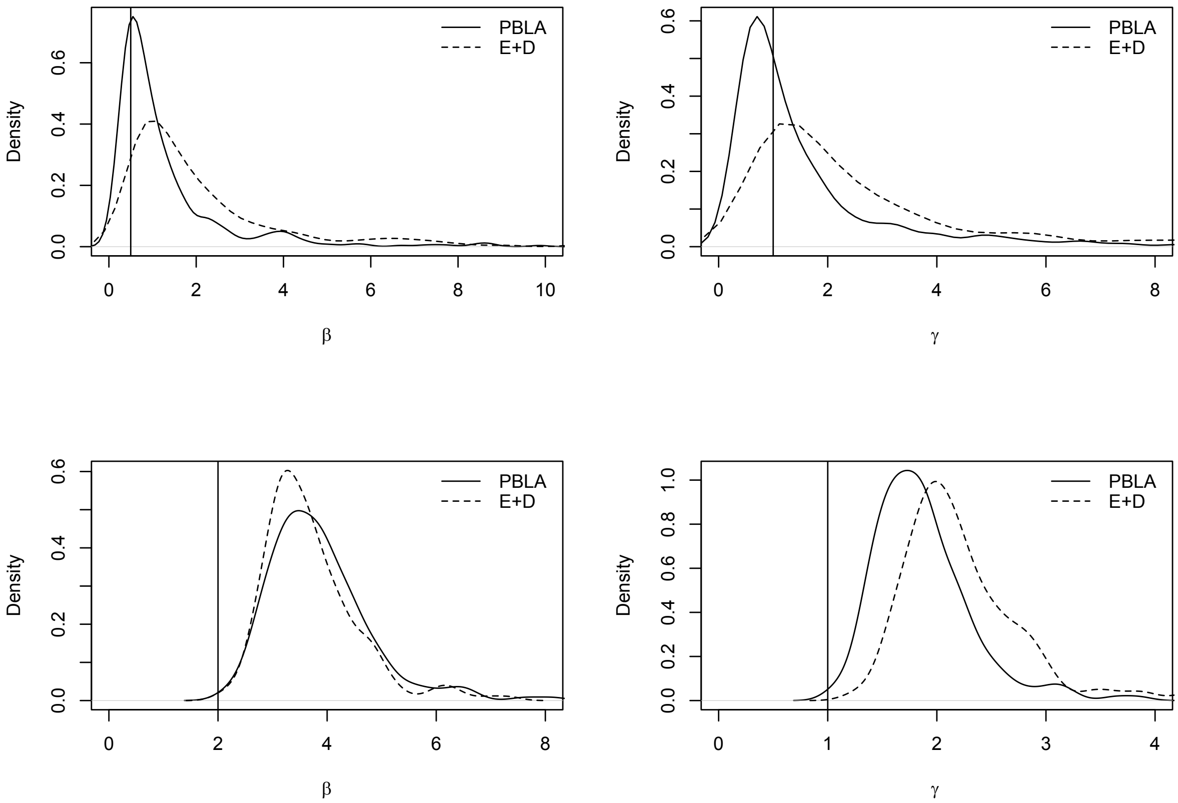

Figure 3: Comparison of maximum likelihood estimates as R_0 varies, using simulated data sets with exponential infectious periods and $N = 500$. Top panels: $\beta = 0.5$, $\gamma = 1$, $R_0 = 1.5$. Bottom panels: $\beta = 2$, $\gamma = 1$, $R_0 = 2$. Vertical lines show the true values.

2.2.4 Varying m with N and R_0 fixed.

If the mean of the infectious period distribution, m/γ, is kept fixed, then increasing m will reduce the variability of the infectious period. In this situation we expect the accuracy of both the ED and PBLA methods to improve, since the underlying assumptions of independence between different components of the likelihood become less unrealistic. Figures 8 and 9 illustrate that this is indeed the case. What is most striking is that marked improvement in estimation from $m = 1$ and $m = 2$, while increasing m further only provides modest further gains.

2.3 Computational performance

One motivation for the PBLA approach is to dispense with the need for data augmentation. However, since the PBLA likelihood can be costly to compute, it is natural to ask how it compares to the standard data-augmented MCMC (DA-MCMC) method described in O'Neill and Roberts (1999) which generates samples from the posterior density $\pi(\beta, \gamma|\boldsymbol{r})$. To address this question we first generated several simulated data sets, consisting of removal times, from the SIR model. On each data set we then ran both the standard DA-MCMC algorithm, with β, γ and all unknown infection times updated at each iteration, and an MCMC algorithm with the PBLA likelihood in which β and

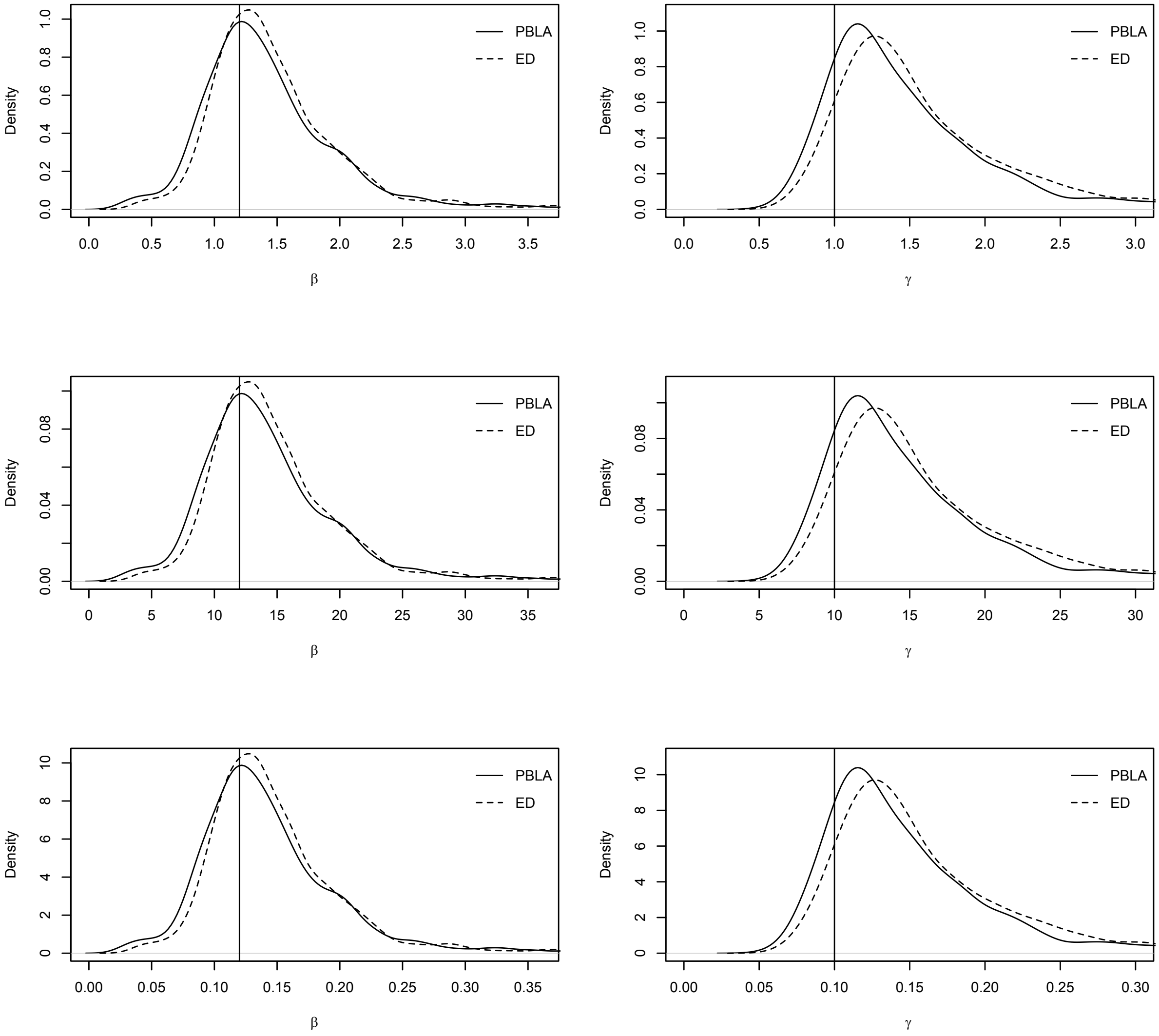

Figure 4: Comparison of maximum likelihood estimates as β and γ vary, using simulated data sets with Erlang infectious periods and $N = 100$, $m = 2$, $R_0 = 2.4$. Top panels: $\beta = 12$, $\gamma = 10$. Middle panels: $\beta = 1.2$, $\gamma = 1$. Bottom panels: $\beta = 0.12$, $\gamma = 0.1$. Vertical lines show the true values.

γ were updated separately at each iteration using a Gaussian random walk proposal mechanism. Both algorithms were initialised in stationarity, following burn-in, and then run for the same number of iterations, n_s. We then calculated the effective sample size $ESS = n_s(1 + 2\sum_{k=1}^{\infty} \rho(k))^{-1}$, where $\rho(k)$ denotes the sample correlation at lag k, and where the sum was truncated at lag M if $\rho(M + 1) < 0.05$. We assigned independent $Exp(10^{-4})$ prior distributions to both β and γ.

Figure 10 shows the effective sample size per second for DA-MCMC and PBLA for exponential infectious periods and Erlang infectious periods with shape parameters $m = 2, 5$. In each case PBLA eventually outperforms DA-MCMC as N increases, although the improvement is less marked as m increases, due to the increasing cost of computing the PBLA likelihood.

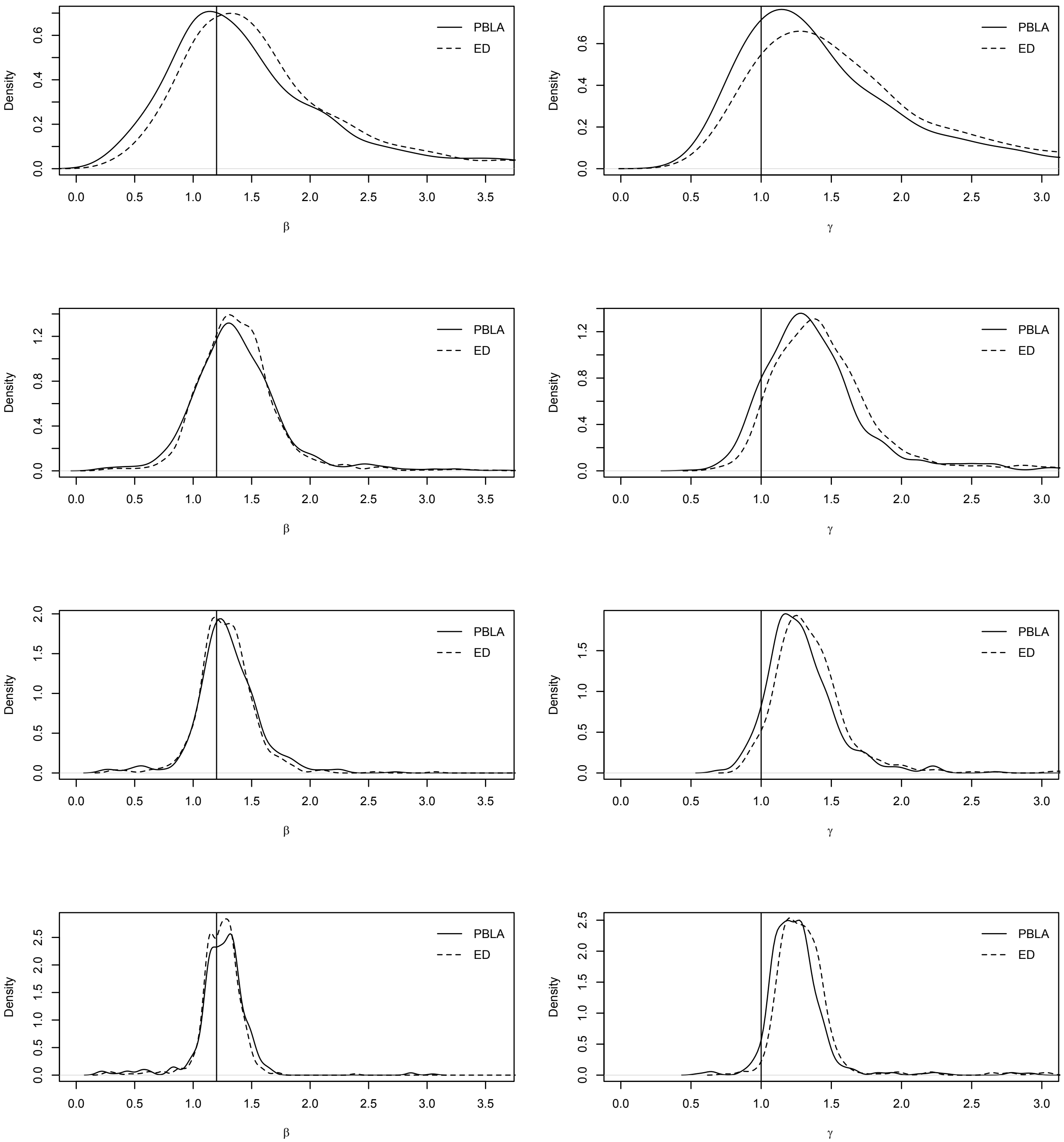

Figure 5: Comparison of maximum likelihood estimates as N varies, using simulated data sets with Erlang infectious periods and $\beta = 1.2$, $\gamma = 1$, $m = 2$, $R_0 = 2.4$. From top to bottom panels: $N = 15$, $N = 100$, $N = 250$, $N = 500$. Vertical lines show the true values.

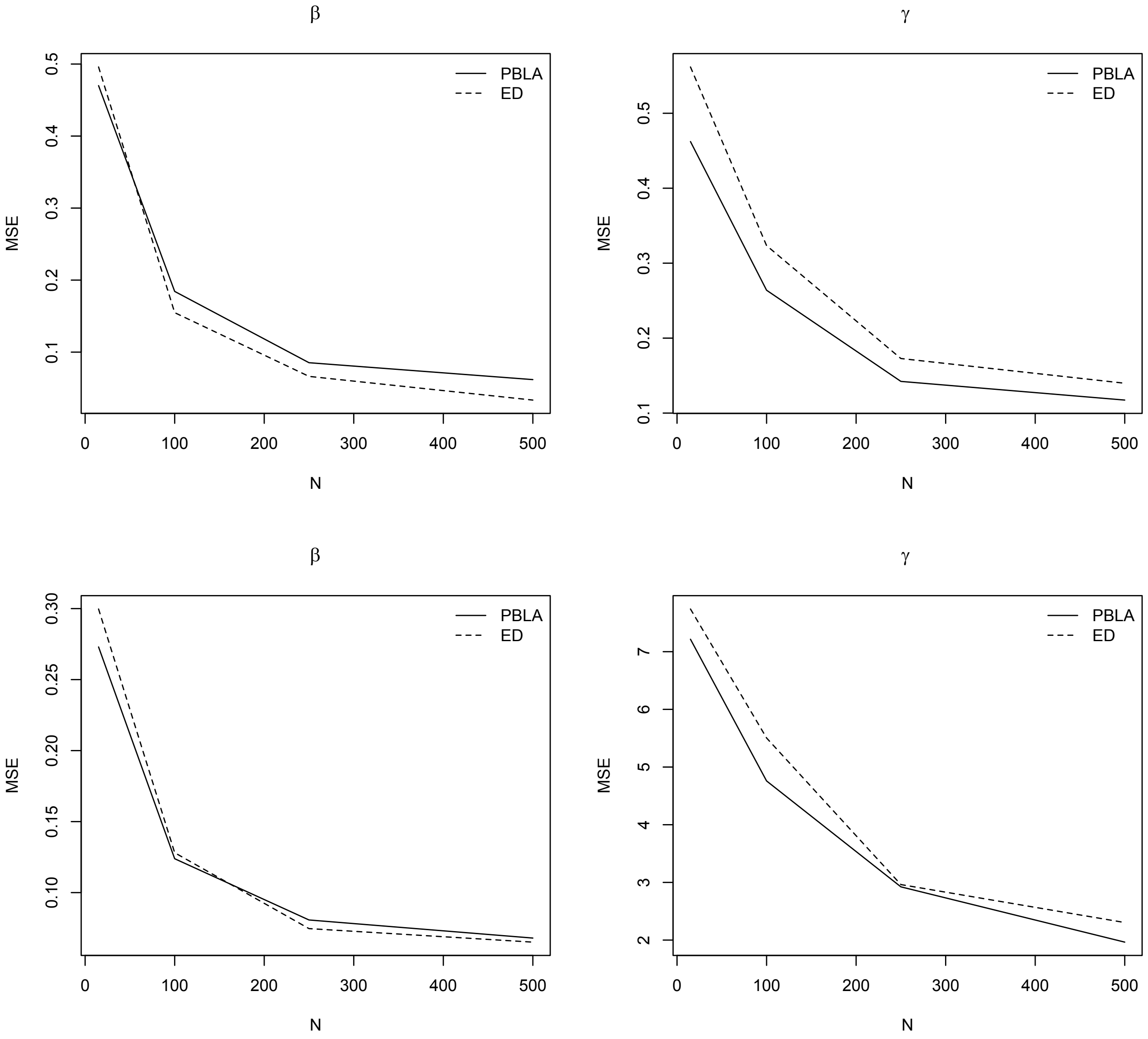

Figure 6: Mean square error (MSE) values as N varies, using simulated data sets with Erlang infectious periods and $\beta = 1.2$, $\gamma = 1$, $m = 2$, $R_0 = 2.4$.

2.4 Comparison of alternative approximations

The main text describes five possible likelihood approximations, namely (i) the standard approximation (section 3.2), (ii) using f_i for expectations (3.3.1), (iii) separating all χ and ψ terms (3.3.2),(iv) approximating the product of ψ terms (3.4.2) and (v) use of the central limit theorem (3.4.2). Note that the last two approximations were developed for the special case of the general stochastic epidemic whereas the first three are more general.

Our numerical experiments showed that the standard approximation generally performed best, although approximations (iv) and (v) are less numerically intensive. Figure 11 shows typical results based on 1000 simulations for the general stochastic epidemic model. Two key observations are that

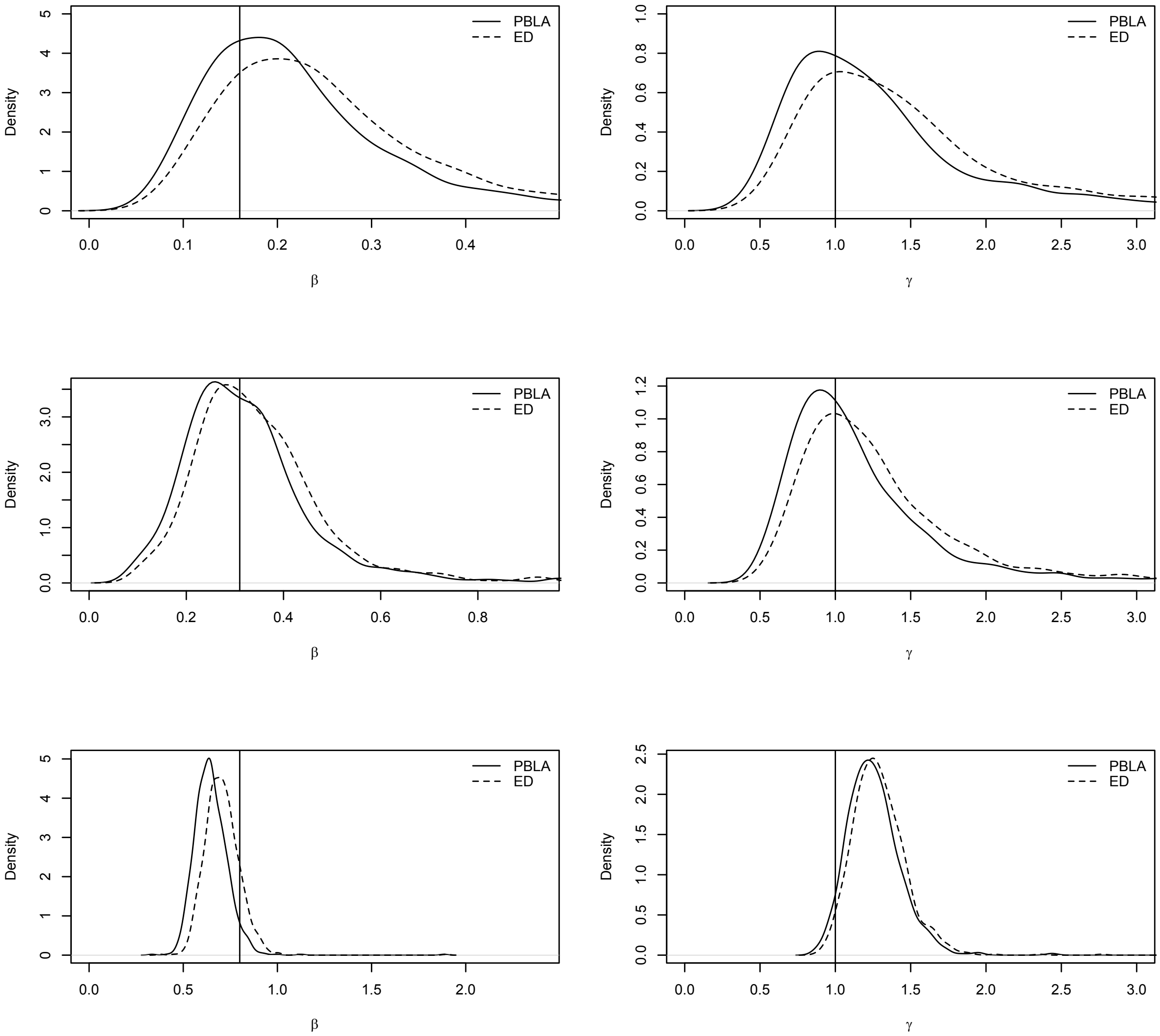

Figure 7: Comparison of maximum likelihood estimates as R_0 varies, using simulated data sets with Erlang infectious periods and $N = 80$, $\gamma = 1$, $m = 5$. Top panel: $\beta = 0.16$, $R_0 = 0.8$. Middle panel: $\beta = 0.31$, $R_0 = 1.55$. Bottom panel: $\beta = 0.8$, $R_0 = 2$. Vertical lines show the true values.

the f_i expectation method is generally inferior to the others, and that the central limit theorem approximation improves as N increases, as expected.

References

Andersson, H. and T. Britton (2000). *Stochastic Epidemic Models and their Statistical Analysis*, Volume 4. New York: Springer.

Barbour, A. D. and G. K. Eagleson (1985). Multiple comparisons and sums of dissociated random variables. *Advances in Applied Probability 17*(1), 147–162.

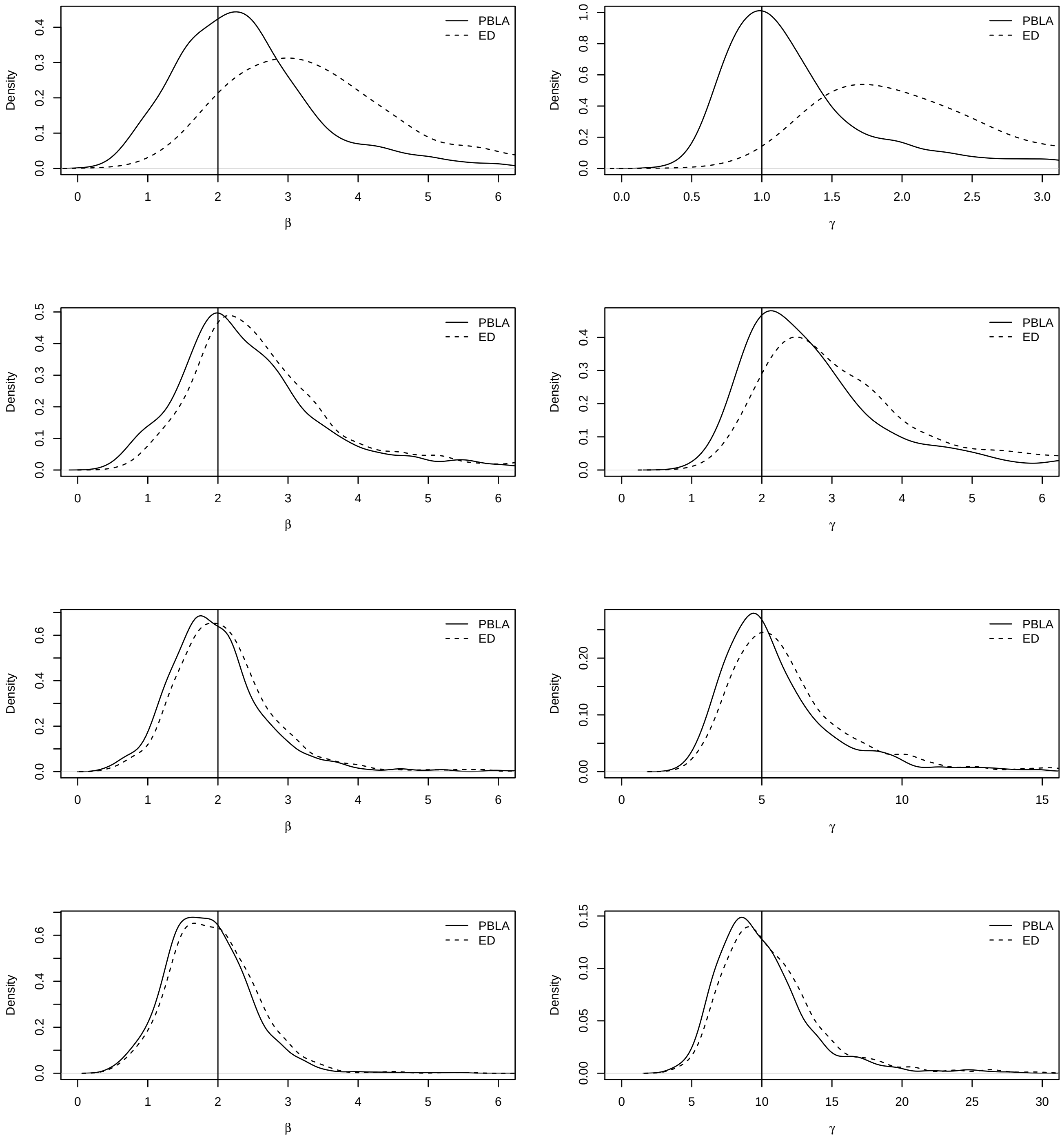

Figure 8: Comparison of maximum likelihood estimates as m varies, using simulated data sets with Erlang infectious periods and $N = 50$, $\beta = 1$, $R_0 = 2$. From top to bottom: $\gamma = m = 1, 2, 5, 10$. Vertical lines show the true values.

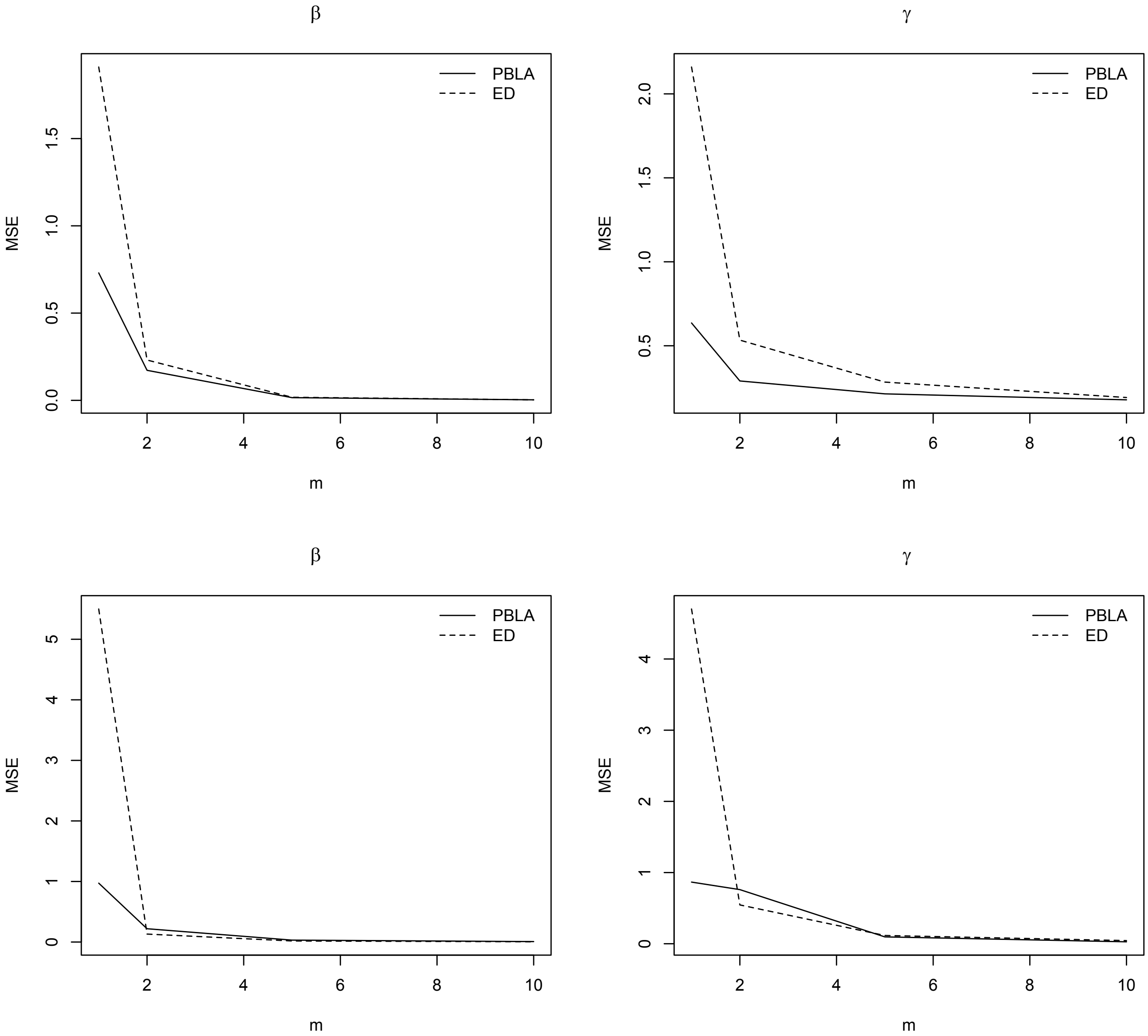

Figure 9: MSE values as m varies, using simulated data sets with Erlang infectious periods and $N = 50$, $\beta = 1$, $R_0 = 2$.

O'Neill, P. D. and G. O. Roberts (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 162*(1), 121–129.

Stockdale, J. E. (2018). *Bayesian Computational Methods for Stochastic Epidemics*. Ph. D. thesis, University of Nottingham.

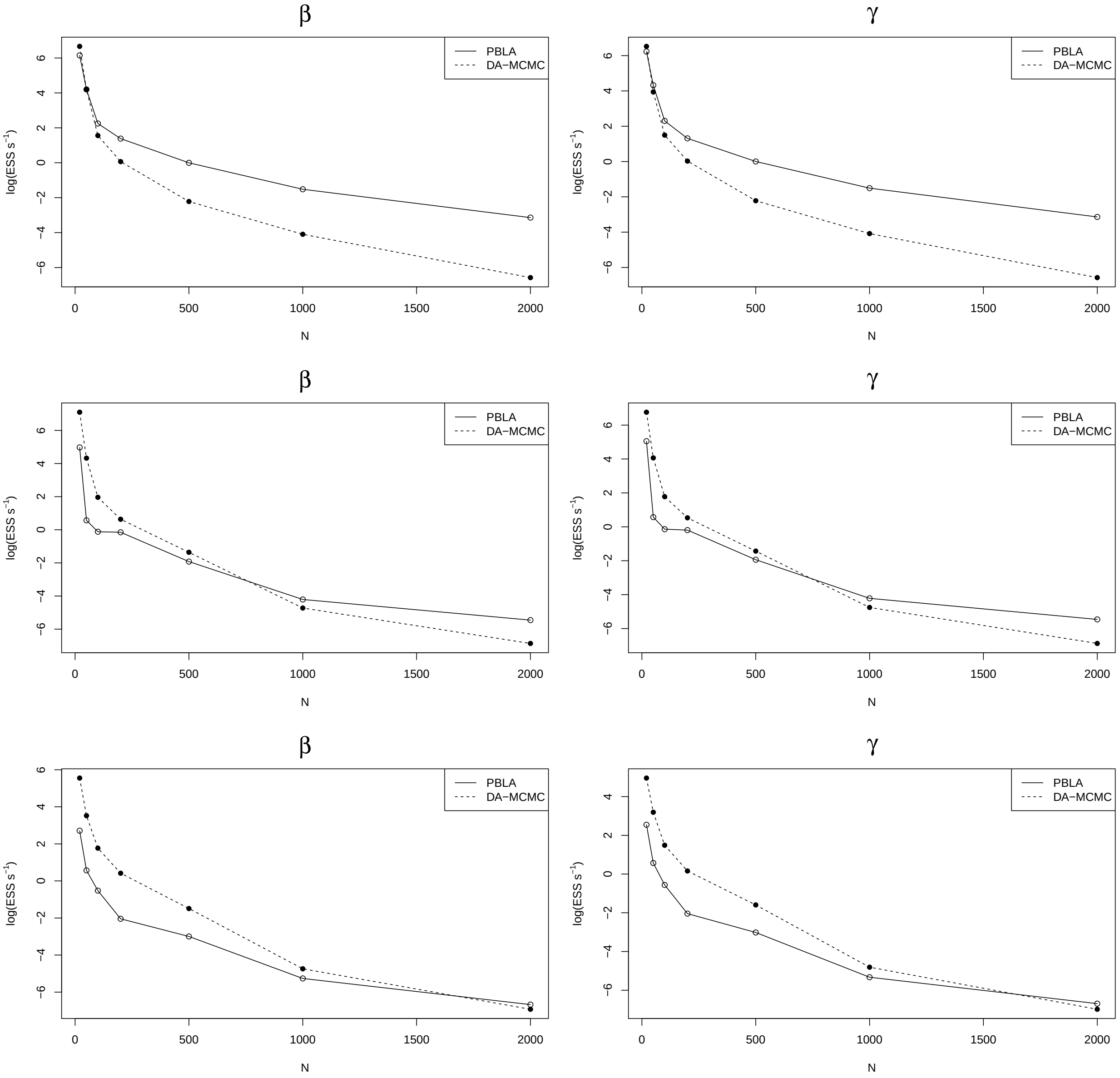

Figure 10: Log effective sample size per second of PBLA and DA-MCMC algorithms as N varies for exponential (top panels), Erlang shape $m = 2$ (middle panels) and Erlang shape $m = 5$ (bottom panels) infectious period distributions.

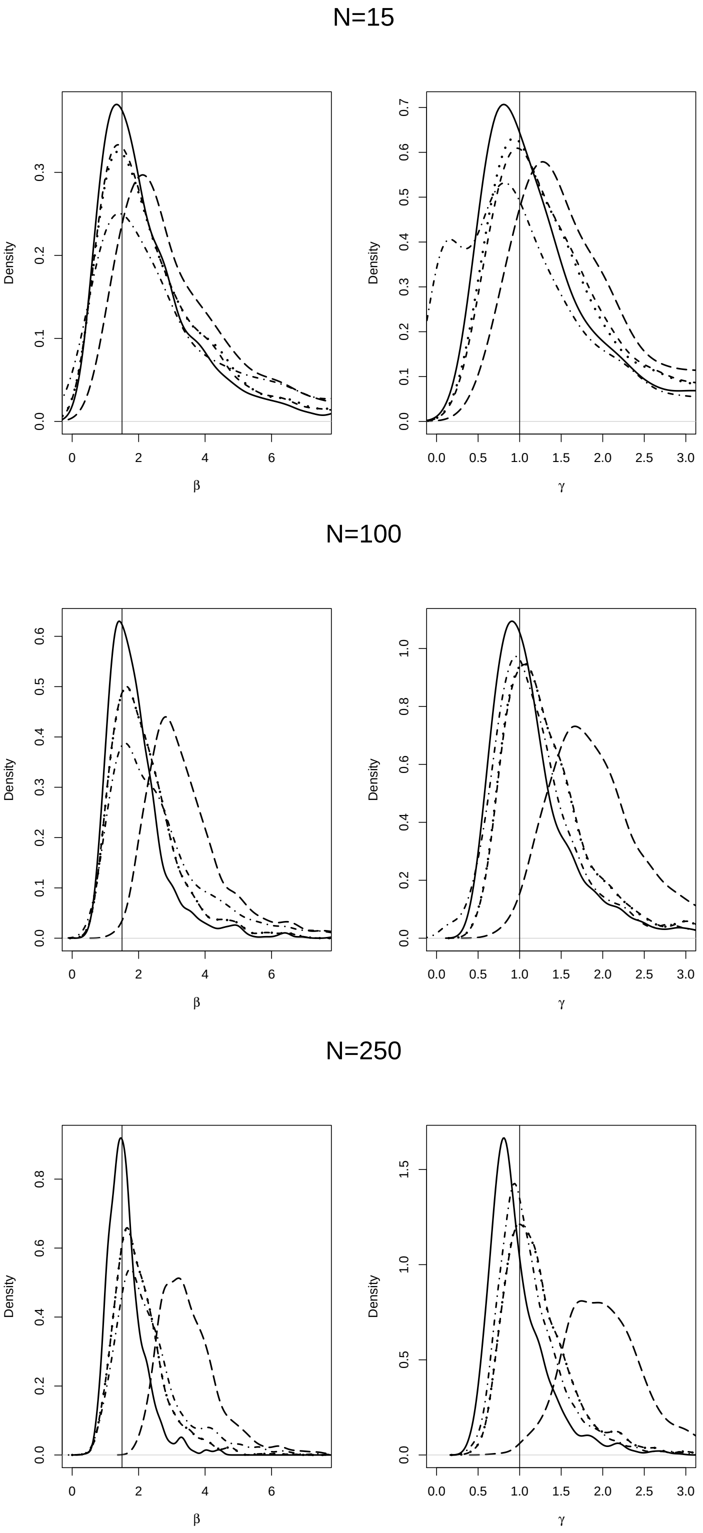

Figure 11: Comparison of different approximations. Plots show maximum likelihood estimates with $\beta = 1.5$ and $\gamma = 1$ for 1000 simulations. Solid line = standard, long dashes = f_i expectations, dots = $\chi - \psi$ separate, short dashes = ψ product and dot-dashes = central limit theorem. Vertical lines show the true values.