Statistical Inference for Stochastic Processes (2025) 28:5
https://doi.org/10.1007/s11203-024-09323-4

Fast likelihood calculations for emerging epidemics

Frank Ball[1] · **Peter Neal**[1]

Received: 5 September 2024 / Accepted: 31 December 2024

Abstract
Statistical inference for epidemic outbreaks is often complicated by only partial observation of the epidemic process. Recently in Ball and Neal (Adv Appl Probab 55:895-926, 2023) the distribution of the number of infectives (individuals alive) given only the times of removals (death) in a Markovian SIR epidemic (time-inhomogeneous birth–death process) was derived. We show that this allows us to derive an explicit expression for the likelihood of the observed inter-removal times of the epidemic without recourse to data augmentation techniques. Moreover, the time-inhomogeneous birth–death process provides a good approximation for the SIR epidemic model for which we are able to obtain both, the exact likelihood of the inter-arrival death times, and a fast to compute Gaussian-based approximation of the likelihood. The explicit expressions for the likelihood enable us to reveal bi-modality in the likelihood of the ongoing Markovian SIR epidemic model and to devise scaleable MCMC algorithms which are applied to the emergence of the Covid-19 epidemic in Europe (March–May 2020).

Keywords Markovian SIR epidemic · Approximate likelihood · Scaleable MCMC · Bi-modaliy · Covid-19

1 Introduction

The Covid-19 pandemic has brought into sharp focus the strengths and weaknesses of current approaches for modelling infectious disease spread, especially the emergence of a new disease. In the early stages of a pandemic data are often sparse, Shadbolt et al. (2022), and only available at a coarse, national or regional, level of granularity. Therefore, models are often fairly simple to capture the key features of the emerging disease such as the growth rate and reproduction number and to avoid spurious results caused by over-fitting.

One approach to modelling the time-course of an epidemic is to use ODEs (ordinary differential equations), see, for example, Reed et al. (2021) for the emergence of Covid-19. The use of ODE models for infectious diseases date back to the pioneering work of Kermack and McKendrick, see Kermack and McKendrick (1927), and are a useful tool for

✉ Frank Ball
frank.ball@nottingham.ac.uk

✉ Peter Neal
peter.neal@nottingham.ac.uk

[1] School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK

Published online: 29 January 2025

capturing the dynamics of an epidemic, especially if there are a large number of individuals in each compartment of the model. However, the initial stages of an epidemic are inherently stochastic. This has led to the development of stochastic models for epidemics from the 1940's onwards (Bartlett 1949), see Bailey (1975) for a summary of the early developments of stochastic epidemic modelling. This paper focusses on parameter estimation for stochastic models for emerging diseases.

Three key components to modelling the spread of an infectious disease are; the transmission of the disease (from infectious individuals to susceptibles), the life history of infected individuals (latent and infectious periods) and the detection of infected individuals (symptomatic and asymptomatic individuals). The simplest assumptions are that these three components do not change over time and that all infected individuals are ultimately detected, see, for example, O'Neill and Roberts (1999), Fearnhead and Meligkotsidou (2004) and McKinley et al. (2014), for statistical inference methodology designed for such scenarios. However, for an emerging disease the transmission dynamics will usually vary over time owing to a variety of factors such as introduction of interventions, and the probability of detection of cases will vary due to changes in awareness and surveillance of the disease, see, for example (Schneble et al. 2021). Thus key summaries of the epidemic such as the reproduction number will be varying over time and hence the development of methodologies, for example, Wallinga and Teunis (2004) and Cori et al. (2013), to estimate the reproduction number.

Consider an SEIR epidemic model, where individuals start off susceptible (S) before potentially becoming infected, and transitioning sequentially through the exposed (E) state, where the individual is infected but not yet infectious, the infectious (I) state, where the individual is able to infect other individuals, and finishing in the removed (R) state, where the individual is no longer infectious and plays no further part in the epidemic. The SIR epidemic model is the special case where there is no exposed state and individuals transition directly from being susceptible to infectious. A major complication with statistical inference for SIR and SEIR epidemic models is that typically only partial information about the epidemic process is available. A common assumption is that the data consists of (partial) observation of the removal ($I \to R$) process, see, for example, O'Neill and Roberts (1999), Fearnhead and Meligkotsidou (2004) and McKinley et al. (2014). Partial observation of other transitions are considered in the literature such as $E \to I$ in Nguyen-Van-Yen et al. (2021) and incidence data ($S \to E$) in Fintzi et al. (2021). The presence of asymptomatic individuals implies that there are individuals who play a role in the spread of the epidemic that we never observe. Furthermore, the data are usually aggregated count data of the number of observations over a given time period (day or week) and the length of reporting time periods may vary over time. However, for continuous time models we ideally require the exact transition times. All of the above has meant that data augmentation Markov chain Monte Carlo (MCMC) algorithms, for example, O'Neill and Roberts (1999), have frequently been applied to infer parameters for epidemic models.

Data augmentation MCMC generally works well for small epidemics but struggles with larger epidemic outbreaks due to the computational cost of large scale data augmentation, Ho et al. (2018). The posterior dependence between the augmented data and parameters of interest leads to poor mixing of MCMC algorithms. This is particularly pertinent to ongoing epidemics where the number of infectious individuals is unknown, necessitating the use of reversible jump MCMC, see, for example (Jewell et al. 2009). These problems are further exacerbated when there are asymptomatic cases so the number of removed cases is also unknown. There have been efforts to circumvent the computational burden through approximations of the likelihood, such as the pair-based likelihood approximation (Stockdale et al.

2021) and diffusion-based approximations (Cauchemez and Ferguson 2008; Fintzi et al. 2021). Diffusion-based approximations use a continuous approximation of the discrete state-space and are less appropriate when there are a small number of infectives, Ho et al. (2018). This provides motivation for the current work, a likelihood-based method for inference for epidemics which is applicable when there are few, and possibly asymptomatic, infectives in the population but is also scalable to major epidemic outbreaks.

In Ball and Neal (2023), Theorem 3.3, it was shown that for a time-inhomogeneous Markovian SIR epidemic model (with piecewise-constant, time varying infection and removal rates) it is possible to derive the distribution of the number of infectives at time t given only the times of removals ($I \to R$ transitions) up to and including time t. A by-product of Ball and Neal (2023), Theorem 3.3, is an explicit expression for the likelihood of the observed inter-arrival of the removal times given in (1) in Sect. 3.1. The exact likelihood is computationally intensive to compute as it involves multiplication of matrix exponentials with the size of the matrices depending on the size of the population. In Ball and Neal (2023), Theorem 3.1, the distribution of the number of individuals alive at time t in a time-inhomogeneous birth–death process, given only the times of deaths up to and including time t, is derived. This gives rise in Ball and Neal (2023), Corollary 3.1, to an explicit expression for the likelihood of the observed inter-arrival times of deaths up to time t. This likelihood involves matrix multiplication with the largest matrix being of size $(K-1) \times K$, where K is the number of deaths up to time t. Hence, the computation of the likelihood for the birth–death process is usually much faster than for the likelihood for the time-inhomogeneous Markovian SIR epidemic model, supporting the use of a time-inhomogeneous birth–death process approximation of the epidemic model. Birth–death process approximations, and more generally branching process approximations, of epidemics have a long history dating back to the 1950 s, see, for example, Whittle (1955). However, by allowing time varying birth and death rate parameters in the birth–death process we can construct approximations of the epidemic process which are useful beyond the initial stages of the epidemic, see Ball and Neal (2023), Section 7, and Sect. 2 of this paper. By using (Ball and Neal 2023), Theorem 3.2, we are able in Sect. 3.2 below to extend computation of the likelihood of the time-inhomogeneous birth–death process to the case where not all deaths are observed and to allow for the probability that a death is observed to vary over time. In an epidemic context individuals whose deaths are not observed correspond to asympotomatic individuals who remain undetected throughout the course of the epidemic. By allowing the probability of detection of removals to vary over time we allow for changes in surveillance due to changes in monitoring and testing of the disease.

The derivation of the likelihood for the time-inhomogeneous Markovian SIR epidemic model without requiring data augmentation of the infection times enables us to give fresh insight into the likelihood for SIR epidemic models. In particular, in Sect. 4 we show that the likelihood of an ongoing Markovian SIR epidemic outbreak can be bi-modal, a feature which has not previously been noted in the literature. This supports using a Bayesian approach with an informative prior to estimate the parameters of an ongoing Markovian SIR epidemic model as the maximum likelihood estimate of the parameters often corresponds to implausible parameter values with excessively high infection rates and low removal rates.

The likelihood for the birth–death process derived in Sect. 3.2 has two important limitations for its direct application to large epidemic outbreaks, which we address as follows. Firstly, whilst computation of the likelihood for the time-inhomogeneous birth–death process is much faster than for the SIR epidemic model it becomes more computationally demanding to compute with every observed death and the time taken to compute the likelihood grows approximately quadratically in the number of observed deaths. Therefore, in Sect. 3.3, we

derive a Gaussian-based approximation to the likelihood given in Sect. 3.2 whose computation time grows linearly in the observed number of deaths. Secondly, the likelihood is based on knowing the exact time of the the observed removals, whereas typically epidemic data on removals will be in the form of the number of observed removals in a given time period. Often such data will be collected regularly, say daily or weekly, and for the examples considered in this paper these data consist of daily aggregated counts of observed removals. Given aggregated daily data we use the approximate birth–death likelihood within data augmentation MCMC algorithms where the exact removals times are imputed from the daily counts, see Sects. 4.2 and 5.3. The data augmentation algorithms perform well with efficient mixing over the parameter space. This is because the aggregated daily counts of removals are informative about when the removal times occur unlike imputing the unknown infection times in, for example, O'Neill and Roberts (1999) or Jewell et al. (2009). Moreover, we do not require reversible jump MCMC as we require only the exact removal times of the observed individuals.

The remainder of the paper is structured as follows. In Sect. 2 we outline the epidemic and birth–death process models. In particular, we discuss how a time-inhomogeneous birth–death process can be used to approximate an epidemic throughout its entire course. In Sect. 3, we start with explicit derivation of the likelihood of the inter-arrival times of removals for the Markovian SIR epidemic. We then turn to the time-inhomogeneous birth–death process for which we derive the likelihood of the inter-arrival times of deaths, both the exact likelihood and the Gaussian based approximation. In Sect. 4, we apply the likelihoods derived in Sect. 3 to the much-studied Abakiliki smallpox data set, see Bailey (1975) and references given in Sect. 4. The small size of the Abakiliki outbreak, 30 individuals infected out of a population of 120 individuals, enables us to compare the approximations to the SIR epidemic likelihood given by the time-inhomogeneous birth–death process likelihood and approximate likelihood, and demonstrate their usefulness. The Abakiliki data also allow us to illustrate the possible bi-modality of the likelihood of an ongoing Markovian SIR epidemic outbreak. In Sect. 5, we analyse the emergenece of Covid-19 in Europe in March–May 2020 using an MCMC algorithm with the Gaussian-based approximate likelihood for the time-inhomogeneous birth–death process. Finally in Sect. 6 we present some concluding remarks and discuss extensions of the current work.

2 Epidemic and birth–death process models

In this section we present the time-inhomogeneous Markovian SIR (general stochastic) epidemic and birth–death process models which are considered in this paper. We focus mainly on the case where the infection (birth) and removal (death) parameters are piecewise-constant between removals (deaths).

We start with the time-inhomogeneous Markovian SIR epidemic model in a closed population of size N. We assume that the epidemic is initiated by a single infective in an otherwise susceptible population at time s_0. We set time $t = 0$ to be the time of the first removal in the epidemic process, so $s_0 < 0$. Throughout we assume that s_0 is unknown. For $t \geq s_0$, let $S(t)$, $I(t)$ and $R(t) = N - I(t) - S(t)$ denote the numbers of susceptibles, infectives and removed individuals, respectively, in the population at time t. For $t \geq s_0$, let β_t and γ_t be the rates at which an individual, infectious at time t, makes infectious contacts and becomes removed, respectively. (Note that β_t is the overall rate that an infective makes infectious contacts, so the individual-to-individual infection rate is $N^{-1}\beta_t$ including possible self contacts.) Thus

for all $t \geq s_0$ and $h \geq 0$, we have that, for $s, i = 0, 1, \ldots, N$ with $s + i \leq N$,

$$\mathbb{P}((S(t+h), I(t+h)) = (x, y) \mid (S(t), I(t)) = (s, i)) = \begin{cases} \frac{s}{N}\beta_t ih + o(h) & (x, y) = (s-1, i+1) \\ \gamma_t ih + o(h) & (x, y) = (s, i-1) \\ 1 - \left\{\frac{s}{N}\beta_t + \gamma_t\right\} ih + o(h) & (x, y) = (s, i) \\ o(h) & \text{otherwise.} \end{cases}$$

For $t \geq s_0$, let δ_t be the probability that a removal which occurs at time t is observed, with removals being observed independently. Let $V(t)$ denote the number of observed removed individuals up to, and including, time t with $V(t) \leq R(t)$. Throughout this paper we assume that there exists $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)$ and $\mathbf{d} = (d_1, d_2, \ldots, d_N)$ such that $\beta_t = \alpha_{V(t)+1}$, $\gamma_t = \mu_{V(t)+1}$ and $\delta_t = d_{V(t)+1}$. That is, the infection and removal rates, along with the probability that a removal is observed, are piecewise-constant between observed removal times. This is for ease of exposition in developing the likelihood; the derivation of the likelihood is easily adapted to the parameters changing at alternative changepoints. Note that if $R(t) = N$ then the epidemic is over with everybody having been infected, so $\boldsymbol{\alpha}$, $\boldsymbol{\mu}$ and $\mathbf{d}$ are sufficient to define the epidemic process and its observation process. For brevity, we drop the subscript and use α, μ and δ to denote the infection and removal rates and detection probability in the time-homogeneous case.

We utilise time-inhomogeneous birth–death processes to approximate the epidemic process. As with the epidemic process, we assume the birth–death process starts from a single individual at time $s_0 < 0$ (unknown) with the first observed death occuring at time $t = 0$. We allow for the possibility that not all deaths are observed. Then for $t \geq s_0$, let $B(t)$ and $K(t)$ denote the number of individuals alive at time t and the number of observed deaths up to and including time t, respectively. For $t \geq s_0$, let $\tilde{\beta}_t$ and $\tilde{\gamma}_t$ be the rates at which an individual, alive at time t, gives birth and dies, respectively. Thus, for all $t \geq s_0$ and $h \geq 0$, we have that, for $i = 0, 1, \ldots,$

$$\mathbb{P}(B(t+h) = x \mid B(t) = i) = \begin{cases} \tilde{\beta}_t ih + o(h) & x = i+1 \\ \tilde{\gamma}_t ih + o(h) & x = i-1 \\ 1 - \left\{\tilde{\beta}_t + \tilde{\gamma}_t\right\} ih + o(h) & x = i \\ o(h) & \text{otherwise.} \end{cases}$$

For $t \geq s_0$, let $\tilde{\delta}_t$ be the probability that a death which occurs at time t is observed, with deaths being observed independently. We assume that there exists $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots)$, $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \tilde{\mu}_2, \ldots)$ and $\tilde{\mathbf{d}} = (\tilde{d}_1, \tilde{d}_2, \ldots)$ such that $\tilde{\beta}_t = \tilde{\alpha}_{K(t)+1}$, $\tilde{\gamma}_t = \tilde{\mu}_{K(t)+1}$ and $\tilde{\delta}_t = \tilde{d}_{K(t)+1}$. That is, the birth and death rates, along with the probability that a death is observed, are piecewise-constant between observed death times mirroring the epidemic process.

We assume that the only data which are observed are the observed removal times in the epidemic model and the observed death times in the birth–death process. For notational convenience, we use the inter-arrival times of observed removals (deaths). For $k = 2, 3, \ldots,$ let T_k ($\tilde{T}_k$) denote the distribution of the inter-arrival time from the $(k-1)^{st}$ observed removal (death) until the kth observed removal (death), with t_k ($\tilde{t}_k$) denoting a realisation of T_k ($\tilde{T}_k$). For $T \geq 0$, let $\mathbf{t}_T = (t_2, t_3, \ldots t_{V(T)})$ ($\tilde{\mathbf{t}}_T = (\tilde{t}_2, \tilde{t}_3, \ldots \tilde{t}_{K(T)})$) denote the set of inter-arrival times of observed removals (deaths) up to and including time T, with $\tau_T = T - t_{V(T)}$ ($\tilde{\tau}_T = T - \tilde{t}_{K(T)}$) denoting the time since the last observed removal (death) prior to time T.

We construct the time-inhomogeneous birth–death process as an approximation for the epidemic process with for $T > 0$, $\tilde{\mathbf{t}}_T = \mathbf{t}_T$, (observed deaths in the birth–death process corresponding to observed removals in the epidemic process). It is well-known, see for

example (Whittle 1955), that the time-homogeneous SIR epidemic model with $\beta_t = \alpha$ and $\gamma_t = \mu$ can be approximated in its initial stages by a time-homogeneous birth–death process with $\tilde{\beta}_t = \alpha$ and $\tilde{\gamma}_t = \mu$. More explicitly, infections and removals in the epidemic process can be coupled to births and deaths in the birth–death process such that $I(t) = B(t)$ up until there is an attempt in the epidemic process to infect a non-susceptible individual. For large N, with high probability, this will not occur until $O(\sqrt{N})$ of the population have been infected, see Ball and Donnelly (1995). If we know $S(t)$, then setting $\tilde{\beta}_t = \beta_t \times S(t)/N$ in the birth–death process would result in a process that has a birth rate which exactly matches the infection rate in the epidemic process. However, given only removal (death) times we do not know $S(t)$ but if we have a good approximation, $\tilde{S}(t)$ say, for the number of susceptibles at time t, then we can approximate the epidemic process with a birth–death process with birth rate $\tilde{\beta}_t = \beta_t \times \tilde{S}(t)/N$. Note that removals (deaths) only depend upon the number of infectives (individuals alive), so we set $\tilde{\gamma}_t = \mu$.

We estimate $\tilde{S}(t)$ as follows. We set $\tilde{S}(t)$ to be constant between removal times, which results in the approximating birth–death process having piecewise constant parameters between death times. For $t > s_0$, let $D(t)$ denote the total number of deaths up to, and including time t, and note that if all deaths are detected $D(t) = K(t)$. At time $s_k = \sum_{i=2}^{k} t_i$ $(= \tilde{s}_k)$, the time of the kth observed removal we calculate $\mathbb{E}[B(s_k) \mid \tilde{\mathbf{T}}_{2:k} = \tilde{\mathbf{t}}_{2:k}]$ and $\mathbb{E}[D(s_k) \mid \tilde{\mathbf{T}}_{2:k} = \tilde{\mathbf{t}}_{2:k}]$, the expected number of individuals alive at time s_k and the expected number of deaths up to time s_k, respectively, in the birth–death process. Since whether or not a death (removal) is detected is independent of all other deaths (removals) and the probability of a death (removal) being detected is constant between observed deaths, it follows, by properties of the geometric distribution, that

$$\mathbb{E}[D(s_k) \mid \tilde{\mathbf{T}}_{2:k} = \tilde{\mathbf{t}}_{2:k}] = \sum_{i=1}^{k} \frac{1}{d_i}.$$

Note that if all deaths (removals) are detected then $\mathbb{E}[D(s_k) \mid \tilde{\mathbf{T}}_{2:k} = \tilde{\mathbf{t}}_{2:k}] = k$. We approximate $I(s_k)$ and $R(s_k)$ by $\mathbb{E}[B(s_k) \mid \tilde{\mathbf{T}}_{2:k} = \tilde{\mathbf{t}}_{2:k}]$ and $\mathbb{E}[D(s_k) \mid \tilde{\mathbf{T}}_{2:k} = \tilde{\mathbf{t}}_{2:k}]$, respectively. Then, for $t \in [s_k, s_{k+1})$ we set

$$\tilde{S}(t) = \max\{0, N - \mathbb{E}[D(s_k) \mid \tilde{\mathbf{T}}_{2:k} = \tilde{\mathbf{t}}_{2:k}] - \mathbb{E}[B(s_k) \mid \tilde{\mathbf{T}}_{2:k} = \tilde{\mathbf{t}}_{2:k}]\}.$$

That is, we estimate the number of susceptibles in the population by the total population size minus the (expected) number of births in the approximating birth–death process up to time s_k, with the restriction that if the estimated number of individuals born (alive or dead) in the birth-death process up to time t exceeds N, we set the estimated number of susceptibles equal to 0. This approximation is reasonable if the number of susceptible individuals does not change dramatically between removal times.

3 Likelihood

In this section, we discuss calculation of the likelihood for the epidemic and birth–death process models introduced in Sect. 2. The main reason for introducing the birth–death process approximation of the epidemic process is that it is much faster to calculate the birth–death process likelihood of observing $\tilde{\mathbf{t}}_T$ rather than the epidemic likelihood of observing $\mathbf{t}_T$. We calculate the likelihood for the epidemic and birth–death process models in Sects. 3.1 and 3.2, respectively. However, calculation of the likelihood of the birth–death process grows at least quadratically in the number of observed deaths. Therefore, in Sect. 3.3 we present a Gaussian

approximation of the likelihood of the birth–death process, the calculation of which grows linearly in the number of observed deaths.

3.1 Epidemic likelihood

We derive the epidemic likelihood in the case that all removals are detected, $\delta_t = 1$, $(t > s_0)$. This is because the calculation of the likelihood become a lot more computationally burdensome with partial detection of the removal process, as we discuss further at the end of this section.

For $k = 1, 2, \ldots, N$, let $\boldsymbol{\Theta}_k = \{\theta_1, \theta_2, \ldots, \theta_k\}$, where $\theta_j = (\alpha_j, \mu_j)$, the parameters of the epidemic process up to and including the kth removal. For $T \geq 0$, $\boldsymbol{\Theta}_{R(T)+1}$ represents the parameters of the epidemic up to and including time T.

The likelihood for the epidemic process is constructed by considering the progress of the epidemic between and at removal times by following (Ball and Neal 2023), Section 3.5. For $k = 0, 1, \ldots, N-1$, let

$$\Omega_k = \{(N-k-i, i) : i = 1, 2, \ldots, N-k\}$$

be the set of states of $\{(S(t), I(t))\}$ in which the epidemic is still going (i.e. there is at least one infective) and precisely k removals have occurred. Give the states in Ω_k the labels $k_1, k_2, \ldots, k_{N-k}$, where the state $(N-k-i, i)$ has label k_i $(i = 1, 2, \ldots, N-k)$. Then for $k = 0, 1, \ldots, N-1$, let $\mathbf{Q}_{k,k} = [q_{k_i,k_j}]$ be the $(N-k) \times (N-k)$ sub-stochastic transition-rate matrix with entries,

$$q_{k_i,k_j} = \begin{cases} -\left(\frac{\alpha_{k+1}}{N}(N-k-i)i + \mu_{k+1}i\right) & \text{if } j = i, \\ \frac{\alpha_{k+1}}{N}(N-k-i)i & \text{if } j = i+1, \\ 0 & \text{otherwise,} \end{cases}$$

see Ball and Neal (2023), (3.30), governing the transition in the number of infectives in the population between the kth and $(k+1)^{st}$ removal. The sub-stochasticity is due to the possibility of removal. For $k = 0, 1, \ldots, N-2$, let $\mathbf{Q}_{k,k+1} = [q_{k_i,(k+1)_j}]$ be the $(N-k) \times (N-k-1)$ transition-rate matrix with entries,

$$q_{k_i,(k+1)_j} = \begin{cases} i\mu_{k+1} & \text{if } i \in \{2, 3, \ldots, N-k\} \text{ and } j = i-1, \\ 0 & \text{otherwise,} \end{cases}$$

governing the transition from Ω_k to Ω_{k+1} (a removal). Note that if a removal occurs in state $k_1 = (N-k-1, 1)$ then the epidemic terminates and there is no transition to Ω_{k+1}. Let $\tilde{\mathbf{Q}}_{k,k}$ be the $(N-k+1) \times (N-k+1)$ transition-rate matrix constructed from $\mathbf{Q}_{k,k}$ by the addition of an initial row and column of 0 s. Finally, let $\tilde{\mathbf{Q}}_{k,k+1}$ be the diagonal $(N-k) \times (N-k)$ matrix with successive diagonal elements $\mu_{k+1}, 2\mu_{k+1}, \ldots, (N-k)\mu_{k+1}$.

Let $L(\mathbf{t}_T; \boldsymbol{\Theta}_{R(T)+1})$ denote the likelihood of observing the inter-arrival removal times $\mathbf{t}_T$ given the parameters $\boldsymbol{\Theta}_{R(T)+1}$. Then $L(\mathbf{t}_T; \boldsymbol{\Theta}_{R(T)+1})$ is given by

$$L(\mathbf{t}_T; \boldsymbol{\Theta}_{R(T)+1}) = -\mathbf{u}_1 \mathbf{Q}_{0,0}^{-1} \left(\prod_{i=1}^{R(T)-1} \mathbf{Q}_{i-1,i} \exp(\mathbf{Q}_{i,i} t_{i+1}) \right) \times \tilde{\mathbf{Q}}_{R(T)-1,R(T)} \exp(\tilde{\mathbf{Q}}_{R(T),R(T)} \tau_T) \cdot \mathbf{1}_{N+1-R(T)}^{\top}, \quad (1)$$

where $\mathbf{u}_1$ is the row vector of length N whose first element is 1 and all other elements are 0 (the initial state) and $\mathbf{1}_{N+1-R(T)}^{\top}$ is the column vector of 1 s of length $N+1-R(T)$ (the

likelihood is summed over all possible number of infectives in the population at time T). (The product is set equal to 1 if $R(T) = 1$.) The matrix $-\mathbf{Q}_{0,0}^{-1}$ describes the transitions before the first removal (unknown time from initial infection to first removal), whilst the product term captures the transitions in states between the successive removals. Finally, the matrices after $\times$ in (1) describe the transitions after the final removal before time T. The right-hand side of (1) is defined as $c_k(T)$ in Ball and Neal (2023), (3.32) with k in place of $R(T)$.

In the event of a completed epidemic, i.e. no more infectives in the epidemic process, we can set $T = \infty$, and hence, $\tau_T = \infty$. The final terms in (1) after $\times$ can be replaced by

$$\times \mu_{R(\infty)} \exp\left(-t_{R(\infty)}\left\{\alpha_{R(\infty)}\frac{N - R(\infty)}{N} + \mu_{R(\infty)}\right\}\right)\bar{\mathbf{u}}_1^\top,$$

where $\bar{\mathbf{u}}_1^\top$ is the column vector of length $N + 1 - R(\infty)$ whose first element is 1 and all other elements are 0. The computation of the likelihood of the completed epidemic can be speeded up by noting that the minimum number of susceptible individuals at any point in time is $N - R(\infty)$. Hence the dimensionality of the matrices in (1) can be reduced to cover only the states with at least $N - R(\infty)$ susceptibles since the contribution to the likelihood of other states is 0.

In the case where all removals are detected we know $R(t)$ for all $t \in \mathbb{R}$. Hence, if we additionally know $I(t)$, then, since the population is closed with $N = S(t) + I(t) + R(t)$, we know the state of the epidemic process at time t. However, if the removal process is only partially observed we only know that $R(t) \geq V(t)$. Hence, we require both $R(t)$ and $I(t)$ to know the state of the epidemic. This means keeping track of possible $(I(t), R(t))$ states and we need to extend the matrices $\mathbf{Q}_{k,k}$ to be $(N-k)(N+1-k)/2 \times (N-k)(N+1-k)/2$, where $(N-k)(N+1-k)/2$ is the total number of possible states for $(I(t), R(t))$ given that $I(t) \geq 1$ and $R(t) \geq k$. Similar changes are required for $\tilde{\mathbf{Q}}_{k,k}$ and $\tilde{\mathbf{Q}}_{k,k+1}$.

3.2 Birth–death process likelihood

We turn to calculation of the likelihood for the time-inhomogeneous birth–death process. The likelihood for the time-inhomogeneous birth–death process in the case where all deaths are observed is given in Ball and Neal (2023), Corollary 3.1. In Lemma 1 we present the likelihood allowing for partial observation of the death process, and present the likelihood in a form which is amenable for the derivation of the Gaussian approximation presented in Sect. 3.3.

For $k = 1, 2, \ldots$, let $\tilde{\boldsymbol{\Theta}}_k = \{\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_k\}$, where $\tilde{\theta}_j = (\tilde{\alpha}_j, \tilde{\mu}_j, \tilde{d}_j)$, the parameters of the birth–death process up to and including the kth observed death. For $T \geq 0$, $\tilde{\boldsymbol{\Theta}}_{K(T)+1}$ represents the parameters of the birth–death process up to and including time T. We factorise the likelihood to express it as

$$L(\tilde{\mathbf{t}}_T; \tilde{\boldsymbol{\Theta}}_{K(T)+1}) = f_{\tilde{T}_2}(\tilde{t}_2 \mid \tilde{\boldsymbol{\Theta}}_2)\left\{\prod_{k=3}^{K(T)} f_{\tilde{T}_k}(\tilde{t}_k \mid \tilde{\mathbf{T}}_{2:(k-1)} = \tilde{\mathbf{t}}_{2:(k-1)}, \tilde{\boldsymbol{\Theta}}_k)\right\}$$
$$\times \mathbb{P}\left(\tilde{T}_{K(T)+1} > \tau_T \mid \tilde{\mathbf{T}}_{2:K(T)} = \tilde{\mathbf{t}}_{2:K(T)}, \tilde{\boldsymbol{\Theta}}_{K(T)+1}\right). \quad (2)$$

That is, we consider sequentially the distribution of the time between successive detected deaths conditional upon the previous times between detected deaths.

The key tool to deriving an explicit expression for (2), given in Lemma 1, is the distribution of the number of individuals alive in the birth–death process immediately following the kth

observed death, $\tilde{X}_k$. For $k = 1, 2, \ldots$, let $\tilde{G}_k$ denote a geometric random variable with probability mass function $\mathbb{P}(\tilde{G}_k = x) = (1-\tilde{\pi}_k)^x \tilde{\pi}_k$ $(x = 0, 1, \ldots)$, where $0 < \tilde{\pi}_k < 1$ is defined in (5) below. In Ball and Neal (2023), Theorem 3.2, it is shown that, $\tilde{X}_1 \sim \tilde{G}_1$, and for $k = 2, 3, \ldots$,

$$\left\{ \tilde{X}_k \mid \tilde{\mathbf{T}}_{2:k} = \tilde{\mathbf{t}}_{2:k} \right\} = \sum_{j=1}^{\tilde{R}_k} G_{k,j} \tag{3}$$

where $\tilde{R}_k$ is a random variable having support $\{2, 3, \ldots, k\}$, and the $G_{k,j}$'s are independent and identically distributed according to $\tilde{G}_k$.

It follows from (3) that, for $k \geq 2$, $\tilde{X}_k$ is a mixture of negative binomial distributions of the form $\{\text{NegBin}(l, \tilde{\pi}_k); l = 2, 3, \ldots, k\}$, where the mixing weights are determined by $\tilde{R}_k$. Therefore in deriving $L(\tilde{\mathbf{t}}_T; \tilde{\boldsymbol{\Theta}}_{K(T)+1})$ it suffices to focus on $\tilde{R}_k$. Specifically, we follow (Ball and Neal 2023) in stating that after the $(k-1)$th detected death, the population consists of $\tilde{R}_{k-1}$ family groups, where each family group has an independent and identically distributed number of members according to $\tilde{G}_{k-1}$. Note that a family group can consist of no individuals. Conditional upon $\tilde{R}_{k-1} = l$, for some $l = 2, 3, \ldots, k-1$, the interarrival time until the kth detected death will be distributed according to the minimum of l independent and identically distributed random variables each distributed according to Y_k, say. Then Y_k is the distribution of the time until the next detected death in a time-homogeneous birth–death process which at time 0 consists of $\tilde{G}_{k-1}$ individuals and has parameters $\tilde{\theta}_k$. Note that $\mathbb{P}(Y_k = \infty) \geq \mathbb{P}(\tilde{G}_{k-1} = 0) = \tilde{\pi}_{k-1}$. We continue by giving the notation required to derive the distribution of Y_k, and hence, $f_{\tilde{T}_k}(\tilde{t}_k \mid \tilde{\mathbf{T}}_{2:(k-1)} = \tilde{\mathbf{t}}_{2:(k-1)}, \tilde{\boldsymbol{\Theta}}_k)$.

For $k = 1, 2, \ldots$, let $\tilde{q}_k = \tilde{\mu}_k/(\tilde{\alpha}_k + \tilde{\mu}_k)$, the probability that an event is a death with parameters $\tilde{\theta}_k$, and let $\tilde{p}_k = 1 - \tilde{q}_k$. Then, recapping (Ball and Neal 2023), (3.24), let

$$\begin{aligned} u_k &= \sqrt{1 - 4\tilde{p}_k\tilde{q}_k(1-\tilde{d}_k)}, & \lambda_k &= \tfrac{1+u_k-2\tilde{p}_k}{1+u_k}, \\ \nu_k &= \tfrac{1-u_k}{2\tilde{p}_k}, & \tilde{\phi}_k(\tau) &= \exp(-[\tilde{\alpha}_k + \tilde{\mu}_k]u_k\tau), \\ \tilde{\psi}_k(\tau) &= \tfrac{(1-\lambda_k)(1-\tilde{\phi}_k(\tau))}{1-\nu_k(1-\lambda_k)\tilde{\phi}_k(\tau)}. \end{aligned} \tag{4}$$

We define the sequence of probabilities $\{\tilde{\pi}_k\}$ iteratively using (4). Let $\tilde{\pi}_1 = \lambda_1$ and for $k = 2, 3, \ldots$, let

$$\tilde{\pi}_k = \frac{\lambda_k[1 - \nu_k(1-\tilde{\pi}_{k-1})] - (1-\nu_k)[\lambda_k - \tilde{\pi}_{k-1}]\tilde{\phi}_k(\tilde{t}_k)}{1 - \nu_k(1-\tilde{\pi}_{k-1}) + \nu_k[\lambda_k - \tilde{\pi}_{k-1}]\tilde{\phi}_k(\tilde{t}_k)}. \tag{5}$$

In general, the terms in (4) do not have an intuitive explanation for their importance. However, we make the following observations regarding the quantities defined in (4) and $\{\tilde{\pi}_k\}$. Firstly, if the parameters are time-homogeneous then for all $k = 1, 2, \ldots, \tilde{\pi}_k = \lambda_1$. Secondly if $d_k = 1$, i.e. all deaths are detected, $u_k = 1$, $\nu_k = 0$ and $\lambda_k = \tilde{q}_k$ with $\tilde{\phi}_k(\tau)$ being the probability an individual neither gives birth or dies in an interval of length τ and (5) simplifying to

$$\tilde{\pi}_k = \lambda_k[1 - \tilde{\phi}_k(\tilde{t}_k)] + \tilde{\pi}_{k-1}\tilde{\phi}_k(\tilde{t}_k),$$

a weighted average between $\lambda_k (= \tilde{q}_k)$ and $\tilde{\pi}_{k-1}$.

For $t \geq 0$, we have that

$$\mathbb{P}(Y_k > t) = \frac{\tilde{\pi}_{k-1}\left[\lambda_k + (1-\lambda_k)(1-\nu_k)\tilde{\phi}_k(t)\right]}{\lambda_k[1 - \nu_k(1-\tilde{\pi}_{k-1})] - (1-\nu_k)(\lambda_k - \tilde{\pi}_{k-1})\tilde{\phi}_k(t)} = \tilde{r}_k(t), \text{ say,}$$

see, for example, Ball and Neal (2023), Theorem 3.2 or (5.10). Therefore if $\tilde{R}_{k-1} = l$, we have, for $\tau > 0$, that

$$\mathbb{P}(\tilde{T}_k > \tau \mid \tilde{R}_{k-1} = l, \tilde{\mathbf{T}}_{2:k-1} = \tilde{\mathbf{t}}_{2:k-1}, \tilde{\boldsymbol{\Theta}}_k) = \tilde{r}_k(\tau)^l.$$

For $k = 2, 3, \ldots$ and $t \geq 0$, let

$$\begin{aligned}\tilde{g}_k(t) &= -\frac{d}{dt}\tilde{r}_k(t) \\ &= \frac{\tilde{\pi}_{k-1}\lambda_k(1-\nu_k)\left\{(\lambda_k - \tilde{\pi}_{k-1}) + (1-\lambda_k)[1-\nu_k(1-\tilde{\pi}_{k-1})]\right\}(\tilde{\alpha}_k + \tilde{\mu}_k)u_k\tilde{\phi}_k(t)}{\left\{\lambda_k[1-\nu_k(1-\tilde{\pi}_{k-1})] - (1-\nu_k)[\lambda_k - \tilde{\pi}_{k-1}]\tilde{\phi}_k(t)\right\}^2}.\end{aligned}$$

Then it follows that, if $\tilde{R}_{k-1} = l$,

$$f_{\tilde{T}_k}(\tau \mid \tilde{R}_{k-1} = l, \tilde{\mathbf{T}}_{2:k-1} = \tilde{\mathbf{t}}_{2:(k-1)}, \tilde{\boldsymbol{\Theta}}_k) = \tilde{g}_k(\tau) l \tilde{r}_k(\tau)^{l-1}. \tag{6}$$

We summarise the calculation of $L(\tilde{\mathbf{t}}_T; \tilde{\boldsymbol{\Theta}}_{K(T)+1})$ in Lemma 1.

Lemma 1 *For $T > 0$, given inter-death times $\tilde{t}_T$ and parameters $\tilde{\boldsymbol{\Theta}}_{K(T)+1}$, the likelihood satisfies*

$$L(\tilde{\mathbf{t}}_T; \tilde{\boldsymbol{\Theta}}_{K(T)+1}) = \tilde{g}_2(\tilde{t}_2)\left\{\prod_{k=3}^{K(T)} \tilde{g}_k(\tilde{t}_k)\mathbb{E}\left[\tilde{R}_{k-1}\tilde{r}_k(\tilde{t}_k)^{\tilde{R}_{k-1}-1}\right]\right\}\mathbb{E}\left[\tilde{r}_{K(T)+1}(\tilde{\tau}_T)^{\tilde{R}_{K(T)}}\right]. \tag{7}$$

The probability mass functions of $\tilde{R}_k$ $(k = 2, 3, \ldots, K(T))$ are defined iteratively as follows. For $k = 2, 3, \ldots$ and $j = 2, 3, \ldots, k$, let $\tilde{B}_{k,j} = \mathbb{P}(\tilde{R}_k = j \mid \tilde{\mathbf{T}}_{2:k} = \tilde{\mathbf{t}}_{2:k}, \tilde{\boldsymbol{\Theta}}_{2:k})$. For $k = 2, 3, \ldots$, let $\tilde{\mathbf{B}}_k = (\tilde{B}_{k,2}, \tilde{B}_{k,3}, \ldots, \tilde{B}_{k,k})$, with $\tilde{\mathbf{B}}_2 = (1)$ and, for $k = 3, 4, \ldots$, $\tilde{\mathbf{B}}_k$ satisfying

$$\tilde{\mathbf{B}}_k = \left\{\tilde{\mathbf{B}}_{k-1}\tilde{\mathbf{M}}_{k-1}\left(\tilde{t}_k\right)\cdot\mathbf{1}_{k-1}^\top\right\}^{-1}\tilde{\mathbf{B}}_{k-1}\tilde{\mathbf{M}}_{k-1}\left(\tilde{t}_k\right), \tag{8}$$

where $\mathbf{1}_{k-1}^\top$ denotes a column vector of 1s of length $k-1$ and for $\tau \geq 0$, $\tilde{\mathbf{M}}_{k-1}(\tau)$ is the $(k-2)\times(k-1)$ matrix with (i, j)th element

$$\left[\tilde{\mathbf{M}}_{k-1}(\tau)\right]_{i,j} = \begin{cases}(i+1)\binom{i}{j-1}h_k(\tau)^{j-1}[1-h_k(\tau)]^{i+1-j}\tilde{r}_k(\tau)^i & \text{for } j \leq i+1,\\ 0 & \text{otherwise},\end{cases}$$

where

$$h_k(\tau) = \frac{1-\tilde{\pi}_k - \tilde{\psi}_k(\tau)}{[1-\tilde{\pi}_k][1-\tilde{\psi}_k(\tau)]}. \tag{9}$$

Proof An immediate consequence of (6) is that

$$f_{\tilde{T}_k}(\tilde{t}_k \mid \tilde{\mathbf{T}}_{2:k-1} = \tilde{\mathbf{t}}_{2:k-1}, \tilde{\boldsymbol{\Theta}}_k) = \tilde{g}_k(\tilde{t}_k)\mathbb{E}\left[\tilde{R}_{k-1}\tilde{r}_k(\tilde{t}_k)^{\tilde{R}_{k-1}-1}\right].$$

Since $\tilde{R}_1 \equiv 1$, it follows that $f_{\tilde{T}_2}(\tilde{t}_2 \mid \tilde{\boldsymbol{\Theta}}_2) = \tilde{g}_2(\tilde{t}_2)$. Since the probability of no death between times $t_{K(T)}$ and T is $\mathbb{E}\left[\tilde{r}_{K(T)+1}(\tilde{\tau}_T)^{\tilde{R}_{K(T)}}\right]$, (7) follows. The lemma is completed by noting that the probability mass function of $\tilde{R}_k$ $(k = 2, 3, \ldots, K(T))$ is given by Ball and Neal (2023), Theorem 3.2. □

It is straightforward using (8) to iterate back from k to 2 to obtain the expression for $\tilde{\mathbf{B}}_k$ given in Ball and Neal (2023), (3.26). There are three advantages though of using (8) and computing $\tilde{\mathbf{B}}_2, \tilde{\mathbf{B}}_3, \ldots, \tilde{\mathbf{B}}_{K(T)}$ iteratively. Firstly, $\tilde{\mathbf{B}}_2, \tilde{\mathbf{B}}_3, \ldots, \tilde{\mathbf{B}}_{K(T)}$ are required to calculate $L(\tilde{\mathbf{t}}_T; \tilde{\mathbf{\Theta}}_{K(T)+1})$. Secondly, using (8) repeatedly is more numerically stable than using (Ball and Neal 2023), (3.26) as the probability mass function is normalised at each step. Finally, and most importantly, (8) forms the basis for a fast approximation of the likelihood presented in Sect. 3.3.

3.3 Approximation of birth–death process likelihood

In Lemma 1 the calculation of $\tilde{\mathbf{B}}_k$ from $\tilde{\mathbf{B}}_{k-1}$ involves the multiplication of a vector of length $k-2$ by a $(k-2)\times(k-1)$ matrix, see (8). Hence, the calculation of the likelihood given in Lemma 1 at time T is at least $O(K(T)^2)$, and simulation studies suggest that the computation time of the likelihood grows at a rate between quadratic and cubic in the observed number of deaths. The exact likelihood becomes prohibitively slow to use in algorithms which require repeated calculation of the likelihood, such as MCMC, as the number of deaths observed enters the thousands. Therefore, we consider a fast-to-compute approximation of the likelihood, based upon a Gaussian approximation of $\tilde{R}_k$, and, in particular, a Gaussian approximation of the moment generating function of $\tilde{R}_k$ and its first three derivatives.

We define the necessary notation to present the approximate likelihood in Lemma 2. Let $(\eta_2, \sigma_2^2) = (2, 0)$. For $k = 3, 4, \ldots$, define (η_k, σ_k^2) in terms of $(\eta_{k-1}, \sigma_{k-1}^2)$, $h(\tilde{t}_k)$ and $\varphi_k = \log(\tilde{r}_k(\tilde{t}_k))$ with

$$\eta_k = 2 + h_k(\tilde{t}_k)\left[\eta_{k-1} + \varphi_k\sigma_{k-1}^2 + \frac{\sigma_{k-1}^2}{\eta_{k-1}+\varphi_k\sigma_{k-1}^2} - 1\right] \tag{10}$$

and

$$\begin{aligned}\sigma_k^2 = {} & h_k(\tilde{t}_k)\{1-h_k(\tilde{t}_k)\}\left[\eta_{k-1} + \varphi_k\sigma_{k-1}^2 + \frac{\sigma_{k-1}^2}{\eta_{k-1}+\varphi_k\sigma_{k-1}^2} - 1\right] \\ & + h_k(\tilde{t}_k)^2\sigma_{k-1}^2\left[1 - \frac{\sigma_{k-1}^2}{[\eta_{k-1}+\varphi_k\sigma_{k-1}^2]^2}\right].\end{aligned} \tag{11}$$

For $k = 2, 3, \ldots$ and $\rho \in \mathbb{R}$, let

$$\xi_k(\rho) = \exp\left(\rho\eta_{k-1} + \frac{\rho^2}{2}\sigma_{k-1}^2\right). \tag{12}$$

Lemma 2 *For $T > 0$, given inter-death times $\tilde{\mathbf{t}}_T$ and parameters $\tilde{\mathbf{\Theta}}_{K(T)+1}$, the approximate likelihood satisfies*

$$\hat{L}(\tilde{\mathbf{t}}_T; \tilde{\mathbf{\Theta}}_{K(T)+1}) = \tilde{g}_2(\tilde{t}_2)\left\{\prod_{k=3}^{K(T)} \tilde{g}_k(\tilde{t}_k)\frac{[\eta_{k-1}+\varphi_k\sigma_{k-1}^2]\xi_k(\varphi_k)}{\tilde{r}_k(\tilde{t}_k)}\right\}\xi_{K(T)+1}(\log\{\tilde{r}_{K(T)+1}(\tilde{\tau}_T)\}). \tag{13}$$

Proof The mean and variance of $\tilde{R}_k$ can be expressed in terms of the moment generating function of $\tilde{R}_{k-1}$, its first three derivatives and $h_k(\tilde{t}_k)$. For $n = 0, 1, 2, 3$ and $k = 2, 3, \ldots$, let

$$\chi_k^n = \mathbb{E}\left[\tilde{R}_{k-1}^n \tilde{r}_k(\tilde{t}_k)^{\tilde{R}_{k-1}}\right] = \mathbb{E}\left[\tilde{R}_{k-1}^n \exp\left(\varphi_k\tilde{R}_{k-1}\right)\right].$$

Then

$$\mathbb{E}\left[\tilde{R}_k\right] = 2 + h_k(\tilde{t}_k)\left[\frac{\chi_k^2}{\chi_k^1} - 1\right] = \tilde{\eta}_k, \quad \text{say}, \tag{14}$$

and

$$\text{var}\left(\tilde{R}_k\right) = h_k(\tilde{t}_k)\{1 - h_k(\tilde{t}_k)\}\left[\frac{\chi_k^2}{\chi_k^1} - 1\right] + h_k(\tilde{t}_k)^2\left[\frac{\chi_k^3}{\chi_k^1} - \left(\frac{\chi_k^2}{\chi_k^1}\right)^2\right] = \tilde{\sigma}_k^2, \quad \text{say}. \tag{15}$$

The derivations of (14) and (15) are presented in the Supplementary Material.

The likelihood given in Lemma 1, (7), can be expressed in terms of χ_k^0 and χ_k^1 with

$$L(\tilde{\mathbf{t}}_T; \tilde{\boldsymbol{\Theta}}_{K(T)+1}) = \tilde{g}_2(\tilde{t}_2)\left\{\prod_{k=3}^{K(T)} \tilde{g}_k(\tilde{t}_k)\frac{\chi_k^1}{r_k(\tilde{t}_k)}\right\}\chi_{K(T)+1}^0. \tag{16}$$

Let W_k denote the Gaussian approximation of $\tilde{R}_k$ based upon matching the mean and variance of W_k to those of $\tilde{R}_k$. For $k = 2, 3, \ldots$, we set $W_k \sim N(\eta_k, \sigma_k^2)$, where η_k and σ_k satisfy (10) and (11). Thus $\xi_k(\rho) = \mathbb{E}[\exp(\rho W_{k-1})]$, given in (12), is the moment generating function of W_{k-1}.

For $n = 1, 2, 3$, let $\xi_k^{(n)}(\rho) = \mathbb{E}[W_{k-1}^n \exp(\rho W_{k-1})]$, the nth derivative of $\xi_k(\rho)$ with respect to ρ. Therefore, if W_{k-1} is an approximation for $\tilde{R}_{k-1}$, we have that $\xi_k^{(n)}(\varphi_k) \approx \chi_k^n$ $(n = 0, 1, 2, 3)$. Given that

$$\frac{\xi_k^{(2)}(\varphi_k)}{\xi_k^{(1)}(\varphi_k)} = \eta_{k-1} + \varphi_k\sigma_{k-1}^2 + \frac{\sigma_{k-1}^2}{\eta_{k-1} + \varphi_k\sigma_{k-1}^2}$$

$$\frac{\xi_k^{(3)}(\varphi_k)}{\xi_k^{(1)}(\varphi_k)} = 3\sigma_{k-1}^2 + \left[\eta_{k-1} + \varphi_k\sigma_{k-1}^2\right]^2,$$

it follows that if $W_{k-1} \approx \tilde{R}_{k-1}$, then $\eta_k \approx \tilde{\eta}_k$ and $\sigma_k^2 \approx \tilde{\sigma}_k^2$. Therefore replacing χ_k^1 and $\chi_{K(T)+1}$ in (16) by $\xi_k^{(1)}(\varphi_k) = [\eta_{k-1} + \varphi_k\sigma_{k-1}^2]\xi_k(\varphi_k)$ and $\xi_{K(T)+1}(\log\{\tilde{r}_{K(T)+1}(\tilde{\tau}_T)\})$, we obtain (13). □

We discuss the approximate likelihood $\hat{L}(\tilde{\mathbf{t}}_T; \tilde{\boldsymbol{\Theta}}_{K(T)+1})$ given by Lemma 2. Since the likelihood is constructed iteratively we can define a hybrid likelihood where, for the first $H(\geq 3)$ detected deaths, the exact likelihood is used before turning to the Gaussian approximation. For $H \geq 3$ and $k = 1, 2, \ldots$, let $\hat{\eta}_{H,k}$ and $\hat{\sigma}_{H,k}^2$ denote the (estimated) mean and variance for the hybrid likelihood with for $k \leq H$, $\hat{\eta}_{H,k} = \mathbb{E}[\tilde{R}_k]$ and $\hat{\sigma}_{H,k}^2 = \text{var}(\tilde{R}_k)$, and for $k > H$, $\hat{\eta}_{H,k}$ and $\hat{\sigma}_{H,k}^2$ computed from $(\hat{\eta}_{H,k-1}, \hat{\sigma}_{H,k-1}^2)$ using (10) and (11). We then have

$$\hat{L}_H(\tilde{\mathbf{t}}_T; \tilde{\boldsymbol{\Theta}}_{K(T)+1}) = \tilde{g}_2(\tilde{t}_2)\left\{\prod_{k=3}^{H} \tilde{g}_k(\tilde{t}_k)\frac{\chi_k^1}{\tilde{r}_k(\tilde{t}_k)}\prod_{k=H+1}^{K(T)} \tilde{g}_k(\tilde{t}_k)\frac{\hat{\xi}_{H,k}^{(1)}(\varphi_k)}{\tilde{r}_k(\tilde{t}_k)}\right\}$$
$$\times\, \hat{\xi}_{H,K(T+1)}(\log\{\tilde{r}_{K(T)+1}(\tilde{\tau}_T)\}, \tag{17}$$

with $\hat{\xi}_{H,k}(\varphi_k)$ and $\hat{\xi}_{H,k}^{(1)}(\varphi_k)$ defined in the obvious fashion. The computational time required to compute $\hat{L}_H(\tilde{\mathbf{t}}_T; \tilde{\boldsymbol{\Theta}}_{K(T)+1})$ is increasing in H, but practical to use, for example, within an

MCMC algorithm provided that H is not too large. This will give an improved approximation of $L(\tilde{\mathbf{t}}_T; \tilde{\boldsymbol{\Theta}}_{K(T)+1})$, which is useful if for small k, $\xi_k^{(1)}(\varphi_k)$ is a poor approximation for χ_k^1.

We observe in practice that, even for small k, $\xi_k^{(1)}(\varphi_k)$ provides a good approximation for χ_k^1, and in the cases $k = 4, 5$, we provide explicit bounds in the Supplementary Material for the time-homogeneous model. (Note that for $k = 3$, $\tilde{R}_{k-1} \equiv 2$ with $\eta_{k-1} = 2$ and $\sigma_{k-1}^2 = 0$ giving $\chi_k^1 = \xi_k^{(1)}(\varphi_k)$.) Also in the Supplementary Material we present a simulation study which shows that the gains in accuracy from using a hybrid likelihood are small and hence in the analysis in Sects. 4 and 5, we focus solely on using (13) as the gains in accuracy from using a hybrid likelihood are outweighed by the additional computational time.

We will often be interested in supercritical birth–death processes, or birth–death processes which are at least initially supercritical. Therefore we will have that $\tilde{\eta}_{k-1}$ is typically increasing with k and we explore the impact of $\tilde{\eta}_{k-1} \gg 1$ on the approximating likelihood. For large j and $\tau \approx 0$, we have using a Maclaurin expansion of $\tilde{r}(\cdot)$, that

$$\begin{aligned}\mathbb{P}(\tilde{T}_k > \tau | \tilde{R}_{k-1} = j, \tilde{\mathbf{T}}_{2:k-1} = \tilde{\mathbf{t}}_{2:k-1}, \tilde{\boldsymbol{\Theta}}_k) &= \tilde{r}_k(\tau)^j \\ &\approx [1 + \tilde{r}_k'(0)\tau]^j \approx \exp(j\tau\tilde{r}_k'(0))\end{aligned}$$

For $\tilde{\eta}_{k-1} \gg 1$ with $\tilde{\sigma}_{k-1}^2 = O(\tilde{\eta}_{k-1})$, we have, for $\tau = O(\tilde{\eta}_{k-1}^{-1})$, using a Taylor series expansion and $\mathbb{E}[\tilde{R}_{k-1}] = \tilde{\eta}_{k-1}$, that

$$\begin{aligned}&\mathbb{P}(\tilde{T}_k > \tau | \tilde{\mathbf{T}}_{2:k-1} = \tilde{\mathbf{t}}_{2:k-1}, \tilde{\boldsymbol{\Theta}}_k) \\ &\quad\approx \mathbb{E}\left[\exp\left(\tau\tilde{r}_k'(0)\tilde{R}_{k-1}\right)\right] \\ &\quad= \exp\left(\tau\tilde{r}_k'(0)\tilde{\eta}_{k-1}\right)\left\{1 + \frac{\tilde{r}_k'(0)^2\tau^2}{2}\tilde{\sigma}_{k-1}^2 + O(\tau^3)\mathbb{E}\left[\left(\tilde{R}_{k-1} - \tilde{\eta}_{k-1}\right)^3\right]\right\}. \qquad (18)\end{aligned}$$

For $\tau = O(\tilde{\eta}_{k-1}^{-1})$, the right-hand side of (18) is equal to $\exp\left(\tau\tilde{r}_k'(0)\tilde{\eta}_{k-1}\right)\{1 + O(\tilde{\eta}_{k-1}^{-1})\}$. Therefore it follows that $\tilde{\eta}_{k-1}\tilde{T}_k$ is approximately exponentially distributed with rate $-\tilde{r}_k'(0)$, and hence, $\tilde{t}_k$ is $O(\tilde{\eta}_{k-1}^{-1})$. It is then straightforward to show that $\varphi_k = \log r_k(\tilde{t}_k) = O(\tilde{\eta}_{k-1}^{-1})$.

Lemma 3 *Suppose that* $\tilde{\eta}_{k-1} \gg 1$ *with* $\tilde{t}_k = O(\tilde{\eta}_{k-1}^{-1})$, $\tilde{\sigma}_{k-1}^2 = O(\tilde{\eta}_{k-1})$, $\mathbb{E}[(\tilde{R}_{k-1} - \tilde{\eta}_{k-1})^3] = O(\tilde{\eta}_{k-1})$ *and* $\mathbb{E}[(\tilde{R}_{k-1} - \tilde{\eta}_{k-1})^4] = O(\tilde{\eta}_{k-1}^2)$. *Then for* $m = 1, 2, 3$,

$$\log(\chi_k^m / \tilde{\xi}_k^{(m)}(\varphi_k)) = O(\tilde{\eta}_{k-1}^{-2}),$$

where $\tilde{\xi}_k(\rho) = \exp(\rho\tilde{\eta}_{k-1} + \rho^2\tilde{\sigma}_{k-1}^2/2)$, *the moment generating function of* $\tilde{W}_{k-1} \sim N(\tilde{\eta}_{k-1}, \tilde{\sigma}_{k-1}^2)$.

The proof of Lemma 3 is provided in the Supplementary material. Then provided (η_k, σ_k^2) given by (10) and (11) are close to $(\tilde{\eta}_k, \tilde{\sigma}_k^2)$, we can control the difference between $\log(\chi_k^m / \xi_k^{(m)}(\varphi_k))$.

4 Abakiliki data

In this section, we study properties of the SIR epidemic likelihood and the approximations of the likelihood given by the birth–death process. We use the Abakiliki smallpox data set, see Bailey (1975), page 125, and O'Neill and Roberts (1999), Section 3.2, to illustrate the

findings. This data set assumes all removals are detected and has been used by many authors to illustrate new statistical methodology for epidemic models, for example, a forward-backward filtering algorithm, Fearnhead and Meligkotsidou (2004), a non-centered MCMC algorithm, Neal and Roberts (2005) and an importance sampling algorithm, McKinley et al. (2014), have all been applied to the Abakiliki data set to fit a general stochastic epidemic model. For comparison purposes we note that we use α to denote the overall rate of infectious contacts made by an infective whereas the cited references typically report α/N the individual-to-individual infection rate.

The Abakiliki data comprises 30 cases of smallpox in a population of size 120 with removal times taking place over a 76 day period. Therefore there are 29 inter-removal times which are reported in O'Neill and Roberts (1999), Section 3.2, and repeated below:

$$\mathbf{t}_{2:30} = (13, 7, 2, 3, 0, 0, 1, 4, 5, 3, 2, 0, 2, 0, 5, 3, 1, 4, 0, 1, 1, 1, 2, 0, 1, 5, 0, 5, 5). \quad (19)$$

The Abakiliki data presented in (19) are given on a discrete, daily timescale and can alternatively be represented as aggregated daily counts. We begin by following (O'Neill and Roberts 1999) and Neal and Roberts (2005) in analysing the Abakiliki data as if they were continuous allowing time 0 between successive removals on the same day. This enables us to explore the behaviour of the likelihood without recourse to data augmentation and to provide fresh insight in the form of bimodality of the likelihood. For comparison purposes we also follow (McKinley et al. 2014) in treating the Abakiliki data as daily aggregated counts of removals with imputation of the exact removal times. As we observe in Sect. 4.2 there is very little difference estimation in the parameters between the two approaches.

We primarily analyse the data set as an ongoing epidemic focussing attention on $T = 46.99$, just before the 16th removal, and $T = 90$, 2 weeks after the last removal. We also considered $T = 47$ (just after the 16th removal) and $T = 80, 100$ (4 and 24 days, respectively, after the last removal), along with the completed epidemic, $T = \infty$. The absence of data augmentation in computing the likelihood gives substantial flexibility in the analysis which can be undertaken. We can easily compute the likelihood over a grid (*c.f.* Fearnhead and Meligkotsidou (2004)) to estimate the maximum likelihood estimator, or by computing the normalising constant, the posterior distribution. The likelihood can be utilised within an MCMC algorithm or a rejection sampling algorithm *c.f.* Clancy and O'Neill (2007).

4.1 Abakiliki data: exact likelihood

In Fig. 1 we present contour plots of the epidemic likelihood on day $T = 46.99$ and $T = 90$ on the range $0.025 \leq \alpha \leq 0.25$ and $0.0005 \leq \mu \leq 0.2$ using (1). The likelihood has been calculated at 10,000 points corresponding to each combination of $\alpha = 0.0025i$ $(i = 1, 2, \ldots, 100)$ and $\mu = 0.0005i$ $(i = 1, 2, \ldots, 25)$; $\mu = 0.0025(i - 20)$ $(i = 26, 27, \ldots, 100)$. For $T = 90$, the contour plot reveals two modes in the likelihood at approximately $(\alpha, \mu) = (0.09, 0.0775)$ and $(\alpha, \mu) = (0.16, 0.004)$ with the latter being the maximum likelihood estimate (MLE). For $T = 46.99$, the likelihood has a single mode (MLE) at $(\alpha, \mu) = (0.155, 0.004)$ very close to the MLE for $T = 90$, but the likelihood drops off slowly along a ridge which approximately goes from $(\alpha, \mu) = (0.07, 0.02)$ to $(\alpha, \mu) = (0.11, 0.085)$.

We study the case $T = 90$ in detail with more refined searches of the two modes. The mode at $(\alpha, \mu) = (0.0889, 0.0761)$ with log-likelihood -60.019 yields $R_0 = 1.168$ and has parameter values close to the posterior means (appropriately rescaled for α) reported in Fearnhead and Meligkotsidou (2004), Neal and Roberts (2005) and McKinley et al. (2014)

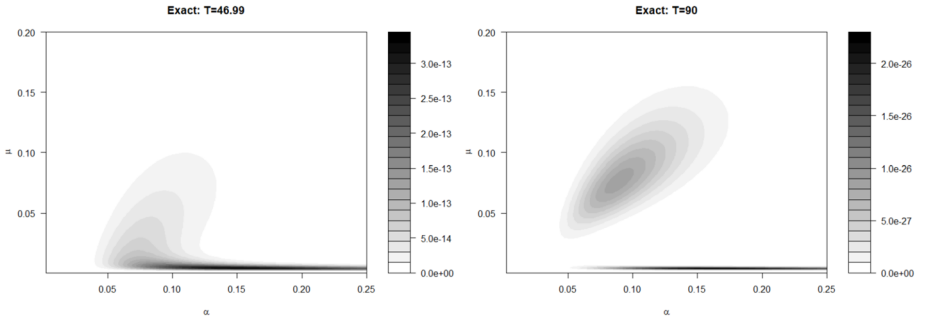

Fig. 1 Contour plots of the likelihood for Abakiliki data analsyed at day $T = 46.99$ (left) and $T = 90$ (right) on the range $0 \leq \alpha \leq 0.25$ and $0 \leq \mu \leq 0.2$

for the completed Abakiliki epidemic. The absence of a removal in the previous 2 weeks is consistent with the epidemic being over, or possibly a small number of infectives remaining. The second mode at $(\alpha, \mu) = (0.1628, 0.00382)$ with log-likelihood -59.014 yields $R_0 = 42.618$. This is consistent with a severe epidemic which infects the whole population some considerable time before $T = 90$ and individuals having long (mean) infectious periods so it is not surprising to see no removals in a 2 week period. Once everybody in the population becomes infected, the likelihood is driven solely by the inter-removal times of the infectives. The mode at $(\alpha, \mu) = (0.1628, 0.00382)$ is the maximum likelihood estimate of the parameters but the mode at $(\alpha, \mu) = (0.0889, 0.0761)$ represents more plausible epidemic parameters even if we do not know for certain that the epidemic is over.

We consider how the likelihood behaves as the infection rate $\alpha \to \infty$. Unless $\mu \to \infty$, we have, with probability tending to 1, that all individuals are infected before the first removal. Suppose that all $N = 120$ individuals become infected before the first removal. Then the distribution of the waiting time between the $(k-1)^{st}$ and kth removals will be exponentially distributed with rate $[N + 1 - k]\mu$. Therefore, if there have been K removals, up to and including time T, with inter-removal times $\mathbf{e}_{2:K} = (e_2, e_3, \ldots, e_K)$, then

$$f(\mathbf{e}_{2:K}, T \mid \mu) = \exp(-[N - K]\mu[T - e_K]) \prod_{k=2}^{K} [(N + 1 - k)\mu] \exp(-[N + 1 - k]\mu e_k).$$

For the Abakiliki data on day 90, we have for $\alpha \approx \infty$, $\hat{\mu} = 0.00308$ with $\log(f(\mathbf{e}_{2:K}, T \mid \hat{\mu})) = -61.82$. Thus the likelihood at $(\alpha, \mu) = (\infty, 0.00308)$ is approximately 0.060 times the maximum value of the likelihood and approximately 0.165 times the value of the likelihood at the second mode.

There are a few points to draw out from the above analysis. For an ongoing general stochastic epidemic, the maximum likelihood estimate (MLE) of the parameters often corresponds to an unrealistically high R_0 scenario with a high infection rate and a low removal rate. We only see the mode about $(\alpha, \mu) = (0.0889, 0.0761)$ becoming the MLE at day $T = 94$. This provides support for using a Bayesian approach for analysing ongoing epidemics with an informative prior. For a Bayesian analysis, owing to the behaviour of the likelihood as $\alpha \to \infty$, we require a proper prior on α for an ongoing general stochastic epidemic, as otherwise we will obtain an improper posterior distribution. Previous analyses of the Abakiliki data, or more generally the general stochastic epidemic model, have not identified/reported the bimodality of the likelihood. There are several reasons for this. Firstly, the bimodality does not exist for completed epidemics. Secondly, an informative prior as used in O'Neill

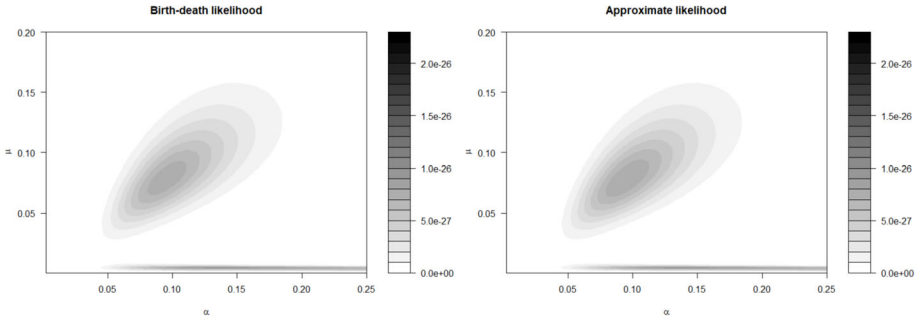

Fig. 2 Contour plot of likelihood for Abakiliki data analsyed at day $T = 90$ on $0 \leq \alpha \leq 0.25$ and $0 \leq \mu \leq 0.2$ using the birth–death likelihood (left) and the approximate birth–death likelihood (right)

and Roberts (1999), Fearnhead and Meligkotsidou (2004) and McKinley et al. (2014) leads to a unimodal posterior distribution. Thirdly, data-augmentation MCMC algorithms initiated close to $(\alpha, \mu) = (0.09, 0.0775)$ are extremely unlikely to move to the mode close to $(\alpha, \mu) = (0.155, 0.004)$ owing to the strong dependence between the augmented data and parameter values.

4.2 Abakiliki data: approximate likelihood

It took approximately 45 min to produce the 100×100 likelihood grid for $T = 90$ using the exact likelihood. (All computations throughout the paper were performed on a desktop PC with Intel(R) Core(TM) i5-12400 Six core 2.50 GHz processor.) Reproducing the likelihood grid using the birth–death process likelihood (7), and the approximate birth–death process likelihood (13), took 35 s and 3 s, respectively, with the birth rate changing at removal times as outlined at the end of Sect. 2. In Fig. 2 we present contour plots of the estimated epidemic likelihood on day $T = 90$ on the range $0.025 \leq \alpha \leq 0.25$ and $0.0005 \leq \mu \leq 0.2$ using (7) and (13). The scale on the contour plots is the same as in Fig. 1 for $T = 90$, and we note that the contour plots are qualitatively similar with correct identification of the two modal regions and modes at approximately $(0.14, 0.0045)$ and $(0.095, 0.0775)$. A more refined search found the modes at $(0.1406, 0.00454)$ and $(0.0947, 0.0779)$ with corresponding R_0 estimates of 30.969 and 1.216, respectively. There is underestimation of the likelihood about the mode at $(0.1406, 0.00454)$ but this is to be expected as the birth–death process approximation is not designed for the highly infectious scenario. There is much better estimation of the likelihood about the mode at $(0.0947, 0.0779)$.

Given that the birth–death and the approximate birth–death likelihoods are approximately 75 and 900 times faster than the exact general stochastic epidemic likelihood in the case $T = 90$, we considered how using the approximate likelihoods within an MCMC algorithm affected estimates of the posterior distribution of the parameters. We use the same informative, independent gamma prior distributions for $\alpha \sim \text{Gamma}(10, 500/6)$ and $\mu \sim \text{Gamma}(10, 100)$ as used in O'Neill and Roberts (1999), Fearnhead and Meligkotsidou (2004) and McKinley et al. (2014). Note that the prior on α corresponds to a $\text{Gamma}(10, 10^4)$ prior on α/N, the individual-to-individual infection rate.

We used a random walk Metropolis algorithm with a Gaussian proposal distribution to obtain samples from the joint posterior distribution of (α, μ) and ran the algorithm for $m_B R_B + R$ iterations. The first $m_B R_B$ iterations were used as burn-in and the final R iter-

Table 1 Estimates of the posterior means and standard deviations of α, μ and R_0 at $T = 46.99$ using the MCMC algorithms with the approximate birth–death, the birth–death and the exact general stochastic epidemic likelihoods treating the data both as exact removal times and aggregated daily removal counts. Numerical integration of the posterior quantities using the grid of likelihood values computed for Fig. 1 is also included

Likelihood	$\bar{\alpha}$	$\bar{\mu}$	$\bar{R}_0$	sd α	sd μ	sd R_0	$cor(\alpha, \mu)$
Approximate	0.1150	0.0846	1.4667	0.0276	0.0252	0.5370	0.2511
Birth–death	0.1143	0.0837	1.4682	0.0265	0.0243	0.5229	0.2425
Exact	0.1107	0.0841	1.4166	0.0258	0.0247	0.4997	0.2640
Numerical integration	0.1103	0.0839	1.4114	0.0257	0.0245	0.4874	0.2701
Approximate (Aggregate)	0.1122	0.0840	1.4403	0.0266	0.0249	0.5261	0.2374
Birth–death (Aggregate)	0.1126	0.0840	1.4468	0.0267	0.0246	0.5296	0.2300
Exact (Aggregate)	0.1108	0.0836	1.4230	0.0261	0.0248	0.4918	0.2836

ations were kept as samples from the posterior distribution. Throughout the burn-in period, Σ, the variance-covariance matrix for the proposal in the random walk algorithm is automatically tuned to produce an efficient MCMC algorithm. Ideally, we would choose $\Sigma = l^2\Sigma^*$, where Σ^* is the variance-covariance matrix of the posterior of the parameters and l is an appropriate scalar, see Roberts and Rosenthal (2001), Section 7.3. Optimal scaling of the random walk Metropolis algorithm would suggest $l = 2.4/\sqrt{2}$ (see Roberts et al. 2009). During the burn-in, after each R_B iterations, we calculate $\mathbf{S}$ and $\mathbf{V}$, the variance-covariance matrix of the parameters and a diagonal matrix with the variance of the parameters on the diagonal, respectively, using the last R_B iterations of the MCMC algorithm. Then we set

$$\Sigma = \frac{2.4^2}{2}\left[0.95\mathbf{S} + 0.05\mathbf{V}\right]. \tag{20}$$

Thus we approximate Σ^* by $0.95\mathbf{S} + 0.05\mathbf{V}$ rather than by $\mathbf{S}$, *c.f.* Sherlock et al. (2010), Algorithm 6, to avoid issues with over-estimation of the correlation between the parameters. Note that Σ is updated m_B times during the burn-in. Throughout we take $m_B = 3$, $R_B = 5,000$ and $R = 50,000$, so that 65,000 iterations (likelihood calculations) are used per MCMC run.

The estimated posterior means and standard deviations for α and μ, along with the correlation between α and μ, using each of the three likelihoods are reported in Tables 1 and 2 for $T = 46.99$ and $T = 90$, respectively. (Similar results were obtained for $T = 47, 80, 100$ and $T = \infty$.) In Tables 1 and 2, we also give calculation of the posterior quantities using numerical integration over the grid of likelihood values computed for Fig. 1 and posterior estimates of the parameters from running the MCMC algorithm using the three likelihoods and taking the Abakiliki data as aggregated daily removal counts (*c.f.* McKinley et al. (2014)). All MCMC algorithms resulted in acceptance rates for the parameters of between 30% and 40% and effective sample sizes for the parameters of approximately 6,000.

For the aggregated daily removal counts we include an additional step in the MCMC algorithm which involves updating 5 randomly selected removal times. The removal times to be updated are chosen uniformly at random from the set of all removal times and for each selected removal time, a new time for the removal is proposed on the day on which the removal occurs. This step has an acceptance rate of approximately 97% across all MCMC runs demonstrating that the parameter estimates are largely insensitive to the exact removal times within a given day.

Table 2 Estimates of the posterior means and standard deviations of α, μ and R_0 at $T = 90$ using the MCMC algorithms with the approximate birth–death, the birth–death and the exact general stochastic epidemic likelihoods treating the data both as exact removal times and aggregated daily removal counts. Numerical integration of the posterior quantities using the grid of likelihood values computed for Fig. 1 is also included

Likelihood	$\bar{\alpha}$	$\bar{\mu}$	$\bar{R}_0$	sd α	sd μ	sd R_0	$cor(\alpha, \mu)$
Approximate	0.1105	0.0918	1.2453	0.0243	0.0214	0.3114	0.4198
Birth–death	0.1106	0.0927	1.2320	0.0239	0.0211	0.3025	0.4148
Exact	0.1072	0.0924	1.1955	0.0234	0.0210	0.2827	0.4559
Numerical integration	0.1072	0.0924	1.1943	0.0231	0.0208	0.2784	0.4542
Approximate (Aggregate)	0.1108	0.0920	1.2449	0.0241	0.0210	0.3158	0.3924
Birth–death (Aggregate)	0.1103	0.0920	1.2383	0.0240	0.0208	0.3075	0.4093
Exact (Aggregate)	0.1073	0.0924	1.1948	0.0230	0.0206	0.2755	0.4554

The estimation of μ is very similar across the three likelihoods, which is unsurprising, since the infectious period/lifetime distribution, governed by μ, does not change between the epidemic model and the birth–death process model. The birth–death process approximation uses the estimated mean number of individuals alive immediately after a removal (death) to estimate the number of susceptibles until the next removal (death). The distribution of the number of individuals alive in the birth–death process approximation has continuously decreasing mean between death times with an upward jump at each death time, see Ball and Neal (2023), Fig. 2 for an illustration. This suggests that in the birth–death approximation the piecewise-constant approximation of the number of susceptibles (infectives) under-estimates (over-estimates) the true distribution of the number of susceptibles, and hence, a slightly higher value of α is found to compensate.

5 Covid-19 data

5.1 Introduction

The main reason for deriving the birth–death process likelihood and its approximation in Sects. 3.2 and 3.3 was to introduce approximations of the epidemic likelihood which are scalable to large epidemic outbreaks, such as the emergence of the Covid-19 pandemic. We use a time-inhomogeneous birth–death process to model the early spread of Covid-19, up to and including the 4th May 2020, in 11 European countries (Austria, Belgium, Denmark, France, Germany, Italy, Norway, Spain, Sweden, Switzerland and United Kingdom) and to evaluate the effect of non-pharmaceutical interventions (NPIs) on the spread of Covid-19. A similar approach was taken in Flaxman et al. (2020). There were a total of 128,141 Covid-19 deaths across the 11 countries up to and including 4th May 2020, with four countries (France, Italy, Spain and the United Kingdom) each experiencing over 24,000 deaths. The large number of Covid-19 deaths renders using the exact likelihood derived in Sect. 3.2 impractical, so we focus on the approximate likelihood given in Sect. 3.3. We use the reported numbers of deaths rather than case counts as the former are likely to be far more reliable than case counts, Flaxman et al. (2020), especially early in the pandemic. Therefore, the observed data, Covid-19 deaths, is a subset of the removal data which includes symptomatic cases which did not lead to death and asymptomatic cases.

Control measures in the form of NPIs varied between the 11 European countries but can be classified into 5 categories. The NPIs were mandatory case-based self-isolation, social distancing measures introduced, public event bans, school closures ordered and lockdown implemented. With the exception of no lockdown being implemented in Sweden, all interventions were implemented across the countries during March 2020 and remained in place at least until the 4th May 2020 when lockdown restrictions were eased in Italy and Spain.

The remainder of the section is structured as follows. In Sect. 5.2 we introduce the Covid-19 model. Details of the MCMC algorithm are given in Sect. 5.3. In Sect. 5.4, we analyse the initial spread of Covid-19 in Europe. This is supported by a simulation study presented in the Supplementary Material, which investigates, and confirms, the identifiability of the model.

5.2 Covid-19 model

The model for Covid-19 has similarities to the model used in Flaxman et al. (2020) in that the time-varying transmission rates depend on a country's baseline R_0 and the NPIs in place. A key difference is that we use a Markovian SIR epidemic model rather than a discrete-time renewal process model, see Cori et al. (2013), for the progression of Covid-19. Given we are modelling the early stages of the pandemic, we make the assumption that there is no significant depletion of susceptible individuals in the countries up to and including the 4th May 2020, and hence we use a time-inhomogeneous birth–death process to model the spread of Covid-19. As previously, we utilise the birth–death process to model the infection (birth) and removal (death) of individuals. To avoid confusion with death of individuals from Covid-19 we use infection and removal rates in place of birth and death rates when describing the time-inhomogeneous birth–death process model for Covid-19.

We assume that after the initial seeding of Covid-19 in a country, which we take to be a single introductory case, the evolution of Covid-19 within each country is independent. Given the restrictions placed on movement between countries during the early stages of the Covid-19 pandemic after the initial seeding of the disease, transmission was driven by internal spread. We briefly discuss the possibility of including trans-country transmission in Sect. 6.

For $i = 1, 2, \ldots, 11$, let α_i denote the baseline infection (birth) rate in country i with the countries numbered according to alphabetic order, 1 = Austria through to 11 = United Kingdom. The removal rate, denoted μ, is assumed to be constant across the 11 countries and constant over time, so throughout we have that the infectious period distribution is $\text{Exp}(\mu)$. In addition to the five NPIs stated, for each country we consider two functions of the NPIs, the dates on which the first and last NPI were introduced in that country. For $j = 1, 2, \ldots, 6$, let $0 \leq \zeta_j \leq 1$ denote the effect of the jth NPI (1 =self isolation, 2 =social distancing, 3 =public event ban, 4 =school closure, 5 =lockdown, 6 =first NPI) on transmission of Covid-19 with $\zeta_j = 0$ if the NPI completely eliminates transmission and $\zeta_j = 1$ if the NPI has no effect on transmission. For $i = 1, 2, \ldots, 11$, let $\xi_i \in \mathbb{R}$ denote the relative (log) effect on the infection rate in country i once all the NPIs implemented during March 2020 have been introduced. For $i = 1, 2, \ldots, 11$, $j = 1, 2, \ldots, 7$ and $t \in \mathbb{Z}$, let $\omega_{i,j}(t) = 1$ if the jth NPI is in force on day t in country i and $\omega_{i,j}(t) = 0$ otherwise, with $j = 7$ corresponding to the last NPI being implemented. Then the infection rate $\beta_i(t)$ ($i = 1, 2, \ldots, 11$; $t \in \mathbb{Z}$) in country i on day t is given by

$$\beta_i(t) = \alpha_i \times \prod_{j=1}^{6} \zeta_j^{\omega_{i,j}(t)} \times \exp(-\xi_i \omega_{i,7}(t)).$$

We equate removal of individuals with the end of their effective infectious period which can be taken to be the time during which they are active in the population and able to infect other individuals. This will typically correspond to the emergence of Covid-19 symptoms. We assume that death owing to Covid-19 occurs some time after the removal of the individual and let D denote the distribution for the time from onset of symptoms to death. We set $\mathbb{E}[D] = 18$ which is in line with Flaxman et al. (2020) and focus on the case where $\mathbb{P}(D \equiv 18) = 1$, death is assumed to occur 18 days after removal.

Let $\delta_i(t)$ denote the probability that a Covid-19 case in country i which is removed on day t results in death. Since not all cases of Covid-19 lead to death, $\delta_i(t) < 1$. It is reasonable to assume that, during the early stages of the pandemic, within a country the probability of death is approximately constant over time, $\delta_i(t) = d$. As noted in the Supplementary Material the model is largely insensitive to the choice of d. Therefore for a parsimonious model we fix $\delta_i(t) = 0.1$ throughout. This is consistent with approximately one death for every eight observed cases in the European data and approximately 30% of cases being asymptomatic, see, for example, Alene et al. (2021).

5.3 Covid-19 MCMC

The MCMC algorithm for analysing the European Covid-19 data (Sect. 5.4) is similar to the MCMC algorithm described in Sect. 4.2. We include a data augmentation step into the MCMC algorithm to impute the detection (arrival) times of cases, and hence, derive the inter-arrival times of cases. We initiate the augmented data by assigning to each detected case, a detection time which is drawn uniformly at random from the day on which they were detected. We updated the detection times of cases in one country at a time given the current parameter values. Since the spread of Covid-19 is assumed to be independent in different countries, this meant that the only likelihood which changed was in the country where the detection times of cases were updated. For each country we proposed to update the detection times of a uniform random sample of 10% of those infected in a single update with the new detection times for the selected individuals drawn uniformly at random from the day on which they were detected. This typically resulted in acceptance rates per country of between 30% and 85% for the independence sampler, which are sufficiently in line with optimal scaling results for the independence sampler given in Lee and Neal (2018) to preclude the need for further tuning. We employed a random walk Metropolis step to update the parameters of the model given the inter-arrival times of the detected cases as in Sect. 4.2. Pilot runs of the MCMC algorithm, typically 20,000 iterations, were used to identify the approximate posterior mode of the parameters with a proposal variance being a scaled multiple of the identity matrix. The random walk Metropolis was then tuned through a further burn-in period of 30,000 iterations, divided into 3 batches of 10,000 iterations, using (20) to derive the proposal covariance matrix, before fixing the covariance matrix to draw a sample of size 50,000 iterations from the posterior distribution of the parameters.

5.4 Covid-19 data Europe

In this section we apply the time-inhomogeneous birth–death process model to the early stages of the Covid-19 model with a delay of 18 days from removal to death of individuals. The time inhomogeneous birth–death process model is not fully appropriate for modelling Covid-19, with the absence of a latent period and the assumption of an exponential infectious period. Unlike the simulation study, presented in the Supplementary Material, attempts to

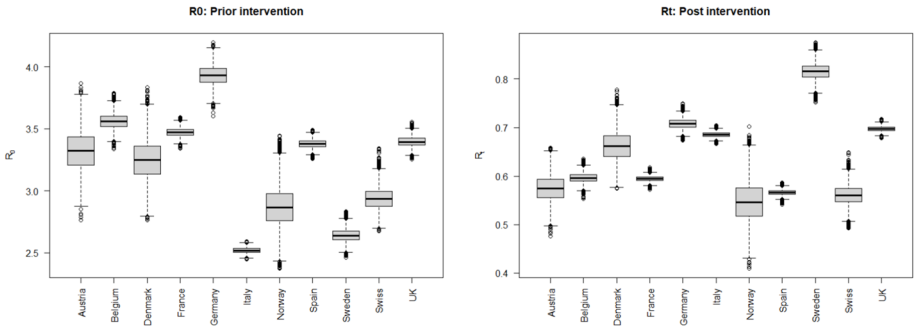

Fig. 3 Boxplots of the MCMC estimates of R_0 (left), prior to, and R_t (right), post, the implementation of NPIs in 11 European countries with $\mu = 0.1$

estimate the removal rate μ of the infectious period along with the other parameters were problematic with the MCMC algorithm returning infeasibly high values of μ. Therefore, we fixed the removal rate μ in the MCMC algorithm and explored the sensitivity of the results by considering a range of plausible values for μ. This approach is in line with Flaxman et al. (2020), who also focus their attention on estimation of the transmission rates and the effects of NPIs.

Each run of the MCMC algorithm was initiated with $\alpha_i = 3\mu$, $\xi_i = 0$ $(i = 1, 2, \ldots, 11)$ and $\zeta_j = 0.8$ $(j = 1, 2, \ldots, 6)$, corresponding to an initial $R_0 = 3$ in each country, no country level variation in the effect of the implementation of the final NPI and each NPI reducing infectivity by 20%. The priors on α_i $(i = 1, 2, \ldots, 11)$ were Gamma$(10, 25)$, corresponding to a prior mean on the basic reproduction number, $R_0^i = \alpha_i/\mu$, of 4. The priors on ξ_i $(i = 1, 2, \ldots, 11)$ were $N(0, 0.1^2)$. Finally, Beta$(1, 1)$ (uniform) priors were chosen for ζ_j $(j = 1, 2, \ldots, 6)$.

In Fig. 3, boxplots of the estimates of the reproduction number R_0 (prior to the implementation of NPIs) and R_t (after the implementation of NPIs) for the 11 European countries from the final 50,000 iterations of the MCMC algorithm with $\mu = 0.1$ and a fixed 18 day delay from removal to death. We observe that in all countries the outbreak of Covid-19 goes from being super-critical, before the implementation of NPIs, to sub-critical, after the implementation of NPIs. Unsurprisingly we observe less variability in the estimation of R_0 and R_t in countries with larger outbreaks than those with countries with fewer deaths. The posterior means $\boldsymbol{\zeta} = (0.9859, 0.7286, 0.9984, 0.9950, 0.3481, 0.8059)$ show that self-isolation, public event bans and school closure had very little effect on reducing transmission of Covid-19 beyond the cases where these were the first NPI. The implementation of the first NPI reduced transmission by approximately 20% with social distancing reducing transmission by approximately 27%. The most effective NPI was lockdown which reduced transmission by almost two-thirds.

In Fig. 4, we give comparisons of the estimates of the growth rate of the epidemic pre-and post-intervention in the United Kingdom for $\mu = 1/8,\ 1/10$ and $1/12$ corresponding to mean infectious periods of 8, 10 and 12 days, respectively. We note that the estimates of the growth rates are consistent across the choice of μ in line with the comment at the end of Sect. 5.2 and Parag et al. (2022). Similar plots are obtained for the other European countries. The estimations of the country parameters α_i and ξ_i do not experience significant changes with the change in μ. The relative effects of the NPIs do not change but for smaller μ we observe a greater change in transmission because for a smaller value of μ a bigger change in

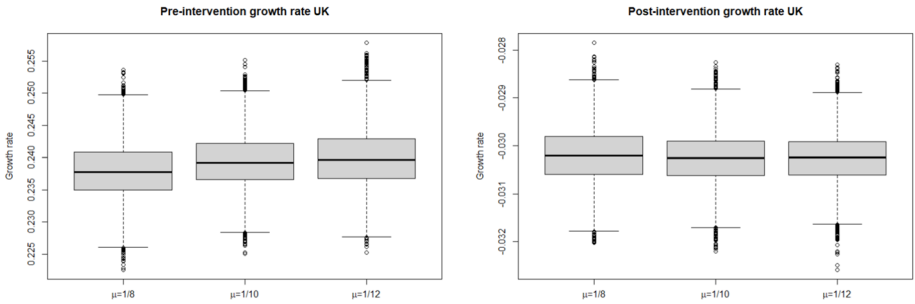

Fig. 4 Boxplots of the MCMC estimates of the growth rate of Covid-19 in the United Kingdom (UK), pre-and post-intervention with $\mu = 1/8$, $\mu = 1/10$ and $\mu = 1/12$

the reproduction rate is required to obtain the same change in the growth rate. For example, the posterior mean estimates of ζ_5, the lockdown effect, are 0.4360 ($\mu = 1/8$) and 0.2735 ($\mu = 1/12$).

The conclusions of the analysis are in broad agreement with the findings of Flaxman et al. (2020), in that lockdown is the dominant control measure with similar overall reductions in R_t owing to NPIs. There are also differences, primarily owing to modelling differences, with the estimates of R_0 generally slightly higher in Flaxman et al. (2020) than those given above. Also in Flaxman et al. (2020) the estimated mean effects of all non-lockdown NPIs are similar.

6 Concluding remarks

We have presented an approximate likelihood for the Markovian SIR epidemic which does not require data augmentation for infection times and is scalable to large epidemic outbreaks. The only data augmentation used within the MCMC algorithms constructed is the removal times of detected cases through the day of detection and this approach can easily be extended to cases where the aggregation of data is not on a regular, daily timescale. In cases where the aggregation of counts is over a longer period of time more sophisticated updating schemes are likely to be needed to take account of whether the epidemic outbreak is growing or shrinking.

Both the exact and approximate likelihoods could be used to update parameters using a sequential Monte Carlo algorithm with daily aggregated data. The Markovian nature of the model means that the likelihood of a specific set of removal times on a given day only depends upon the time of the last removal prior to that day and the model parameters. Therefore within a particle filter it is straightforward to update the weights of particles on a daily basis. The updating of the posterior of the parameters can then be done either using an approximate *online* method which reweight the parameters such as Liu and West (2001) or using an exact method such as SMC^2, see Chopin et al. (2013), which targets the posterior distribution using MCMC rejuvenation steps.

There are limitations to the epidemic model presented above. The birth–death process approximation implies in the homogeneous case an exponential infectious period, and more generally that the recovery rate does not depend on how long an individual has been infectious. In Ball and Neal (2025) we extend Theorems 3.1 and 3.2 of Ball and Neal (2023) to phase-type lifetime distributions. However, the distribution of the number of individuals alive (infectious) immediately following the kth detected death (removal) is a mixture of

$(k-1)!$ distributions, so rapidly becomes impractical to use. An approximation of the distribution of the number of individuals alive (infectious) immediately following the kth detected death (removal) consisting of a mixture of k negative binomial distributions is presented in Ball and Neal (2025) and through numerical examples shown to perform well for Erlang lifetime (infectious period) distributions. This provides a promising approach for deriving an approximate likelihood in the spirit of Sect. 3.3 for Erlang lifetime (infectious period) distributions. An alternative approach to dealing with a general infectious period (lifetime) distribution is to use a time varying removal rate in the birth–death process to capture the changing infectious age profile of individuals over time. Specifically, suppose that individuals have independent and identically distributed lifetimes according to a random variable L with probability density function $f_L(\cdot)$ and cumulative distribution function $F_L(\cdot)$. We can approximate the number of individuals infectious at time t by

$$\int_{-\infty}^{t} \tilde{\beta}_s \mathbb{E}[B(s)|\tilde{\mathbf{T}}_{2:K(s)} = \tilde{\mathbf{t}}_{2:K(s)}]\{1 - F_L(t-s)\}\, ds,$$

and hence, set

$$\tilde{\gamma}_t = \frac{\int_{-\infty}^{t} \tilde{\beta}_s \mathbb{E}[B(s)|\tilde{\mathbf{T}}_{2:K(s)} = \tilde{\mathbf{t}}_{2:K(s)}] f_L(t-s)\, ds}{\int_{-\infty}^{t} \tilde{\beta}_s \mathbb{E}[B(s)|\tilde{\mathbf{T}}_{2:K(s)} = \tilde{\mathbf{t}}_{2:K(s)}]\{1 - F_L(t-s)\}\, ds}.$$

Finally, we have assumed independence of the spread of Covid-19 between countries. However, if we have m communities we can approximate the interactions between communities using an inhomogeneous birth–death process as follows. Suppose that α_{ij} $(1 \le i, j \le m)$ is the rate at which an individual in community i makes contact with individuals in community j and, for simplicity, assume that there is a common removal rate γ. Let $S_i(t)$ and $I_i(t)$ denote the number of susceptibles and infectives in community i at time t, respectively, with N_i denoting the number of individuals in community i. Then the *birth* rate of new infectives in community i at time t will be approximately

$$\beta_i(t) = \frac{\mathbb{E}[S_i(t)]}{N_i} \sum_{j=1}^{N} \alpha_{ji} \frac{\mathbb{E}[I_j(t)]}{\mathbb{E}[I_i(t)]},$$

where for $j = 1, 2, \ldots, N$, $\mathbb{E}[S_j(t)]$ and $\mathbb{E}[I_j(t)]$ are the estimated mean numbers of susceptibles and infectives in community j at time t.

Supplementary information

Supplementary text material covering derivations of results and simulation studies is available.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11203-024-09323-4.

Author Contributions F.B and P.N contributed to the study conception and design. P.N wrote the R code and performed the analysis. F.B and P.N jointly wrote the manuscript and approved the final submission.

Data availability The epidemic data for Abikiliki and Covid-19 analysed in the paper along with the R code used in the analysis of the data are available at: https://github.com/peteneal77/BDLikelihood

Declarations

Conflict of interest The authors declare no competing interests.

References

Alene M, Yismaw L, Assemie MA, Ketema DB, Mengist B, Kassie B, Birhan TY (2021) Magnitude of asymptomatic Covid-19 cases throughout the course of infection: a systematic review and meta-analysis. PLoS One 16:e0249090

Bailey N (1975) The mathematical theory of infectious diseases Griffin. London

Ball F, Donnelly P (1995) Strong approximations for epidemic models. Stoch Proc Appl 55:1–21

Ball F, Neal P (2023) The size of a Markovian SIR epidemic given only removal data. Adv Appl Probab 55:895–926

Ball F, Neal P (2025) The number of individuals alive in a branching process given only times of deaths. To appear in Adv. Appl, Probab

Bartlett MS (1949) Some evolutionary stochastic processes. J R Stat Soc Ser B 11:211–229

Cauchemez S, Ferguson NM (2008) Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission. J R Soc Interface 5:885–897

Chopin N, Jacob PE, Papaspiliopoulos O (2013) SMC^2: an efficient algorithm for sequential analysis of state space models. J R Stat Soc B 75:397–425

Clancy D, O'Neill PD (2007) Exact Bayesian inference and model selection for stochastic models of epidemics among a community of households. Scan J Stat 34:259–274

Cori A, Ferguson NM, Fraser C, Cauchemez S (2013) A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol 178:1505–1512

Fearnhead P, Meligkotsidou L (2004) Exact filtering for partially-observed continuous time models. J R Stat Soc Ser B 66:771–789

Fintzi J, Wakefield J, Minin VN (2021) A linear noise approximation for stochastic epidemic models fit to partially observed incidence counts. Biometrics 78:1530–1541

Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, Whittaker C, Zhu H, Berah T, Eaton JW, Monod M, Imperical College COVID-19 Response Team, Ghani AC, Donnelly CA, Riley SM, Vollmer MAC, Ferguson NM, Okell LC, Bhatt S (2020) Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature 584, 257–261

Ho LST, Crawford FW, Suchard MA (2018) Direct likelihood-based inference for discretely observed stochastic compartmental models of infectious disease. Ann Appl Stat 12:1993–2021

Jewell CP, Kypraios T, Neal P, Roberts GO (2009) Bayesian analysis for emerging infectious diseases. Bayesian Anal 4:465–496

Kermack WO, McKendrick AG (1927) A contribution to mathematical theory of epidemics. Proc R Soc Lond Ser A 115:700–721

Lee C, Neal P (2018) Optimal scaling of the independence sampler: theory and practice. Bernoulli 24:1636–1652

Liu J, West M (2001) Combined parameter and state estimation in simulation-based filtering. In: Doucet A, de Freitas N, Gordon NJ (eds) Sequential Monte Carlo methods in practice. Springer, New York, pp 197–223

McKinley TJ, Ross JV, Deardon R, Cook AR (2014) Simulation-based Bayesian inference for epidemic models. Comput Stat Data Anal 71:434–447

Neal P, Roberts GO (2005) A case study in non-centering for data augmentation: stochastic epidemics. Stat Comput 15:315–327

Nguyen-Van-Yen B, Del Moral P, Cazelles B (2021) Stochastic epidemic models inference and diagnosis with Poisson random measure data augmentation. Math Biosci 335:108583

O'Neill PD, Roberts GO (1999) Bayesian inference for partially observed stochastic epidemics. J R Stat Soc Ser A 162:121–129
Parag KV, Thompson RN, Donnelly CA (2022) Are epidemic growth rates more informative than reproduction numbers? J R Stat Soc Ser A 185:1–11
Reed JM, Brigden JRE, Cummings DAT, Ho A, Jewell CP (2021) Novel coronavirus 2019-nCoV (COVID-19): early estimation of epidemiological parameters and epidemic size estimates. Phil Trans R Soc B 376:20200265
Roberts GO, Rosenthal JS (2001) Optimal scaling for various metropolis-hastings algorithms. Stat Sci 16:351–367
Roberts GO, Gelman A, Gilks WR (2009) Weak convergence and optimal scaling of random walk metropolis algorithms. Ann Appl Probab 7:110–120
Schneble M, De Nicola G, Kauermann G, Berger U (2021) A statistical model for the dynamics of COVID-19 infections and their case detection ratio in 2020. Biom J 63:1623–1632
Shadbolt N, Brett A, Chen M, Marion G, McKendrick IJ, Panovska-Griffiths J, Pellis L, Reeve R, Swallow B (2022) The challenges of data in future pandemics. Epidemics 40:100612
Sherlock C, Fearnhead P, Roberts GO (2010) The random walk metropolis: linking theory and practice through a case study. Stat Sci 25:172–190
Stockdale JE, Kypraios T, O'Neill PD (2021) Pair-based likelihood approximations for stochastic epidemic models. Biostatistics 22:575–597
Wallinga J, Teunis P (2004) Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am J Epidemiol 160:509–516
Whittle P (1955) The outcome of a stochastic epidemic—a note on Bailey's paper. Biometrika 42:116–122