INTERFACE FOCUS

royalsocietypublishing.org/journal/rsfs

Research

Cite this article: KhudaBukhsh WR, Choi B, Kenah E, Rempała GA. 2019 Survival dynamical systems: individual-level survival analysis from population-level epidemic models. *Interface Focus* **10**: 20190048.
http://dx.doi.org/10.1098/rsfs.2019.0048

Accepted: 7 October 2019

One contribution of 9 to a theme issue 'Multi-scale dynamics of infectious diseases'.

Subject Areas:
computational biology, biomathematics, biometrics

Keywords:
epidemic models, survival analysis, stochastic processes, dynamical systems, multiscale models

Author for correspondence:
Eben Kenah
e-mail: kenah.1@osu.edu

Survival dynamical systems: individual-level survival analysis from population-level epidemic models

Wasiur R. KhudaBukhsh[1], Boseung Choi[2], Eben Kenah[3] and Grzegorz A. Rempała[4]

[1]Mathematical Biosciences Institute, The Ohio State University, Columbus, OH, USA
[2]Division of Economics and Statistics, Department of National Statistics, Korea University Sejong campus, Sejong Special Autonomous City, Republic of Korea
[3]Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, OH, USA
[4]Division of Biostatistics, College of Public Health and Mathematical Biosciences Institute, The Ohio State University, Columbus, OH, USA

WRK, 0000-0003-1803-0470; EK, 0000-0002-7117-7773; GAR, 0000-0002-6307-4555

In this paper, we show that solutions to ordinary differential equations describing the large-population limits of Markovian stochastic epidemic models can be interpreted as survival or cumulative hazard functions when analysing data on individuals sampled from the population. We refer to the individual-level survival and hazard functions derived from population-level equations as a survival dynamical system (SDS). To illustrate how population-level dynamics imply probability laws for individual-level infection and recovery times that can be used for statistical inference, we show numerical examples based on synthetic data. In these examples, we show that an SDS analysis compares favourably with a complete-data maximum-likelihood analysis. Finally, we use the SDS approach to analyse data from a 2009 influenza A(H1N1) outbreak at Washington State University.

1. Introduction

Despite their ubiquity in modern epidemiology, mathematical models of epidemics suffer many theoretical and practical drawbacks. Due to the need for mathematical tractability, such models often ignore important characteristics of disease transmission patterns and the underlying populations. This often leads to poor predictions. During the SARS epidemic of 2002–2003, the number of cases in China was predicted to reach 30 000 during the first four months of the epidemic. In fact, there were fewer than 800 cases reported during that time [1]. A more recent example is the Centers for Disease Control and Prevention (CDC) prediction of the 1 400 000 cases of Ebola in West Africa during 2013–2016 outbreak [2,3]. Although the CDC team did indicate that their prediction was the 'worst-case scenario', the inaccuracy of this upper bound prediction has highlighted the need for better mathematical models of epidemics and their control.

A typical challenge in the problem of epidemic control is how to relate the global, population-level dynamics of infection transmission to local, individual-level intervention (e.g. vaccination). This dichotomy is reflected in two distinct approaches to modelling epidemiological processes. Agent-based models capture individual-level histories of infection and removal. By contrast, ecological models look at the population at an aggregate level, keeping track of summary statistics such as the counts of susceptible, infected and recovered/removed individuals. Although both agent-based and ecological models are routinely used in practice and in the literature, the two scales of analysis are almost always considered separately [4].

THE ROYAL SOCIETY PUBLISHING

Table 1. List of symbols.

symbol	meaning
β	infection rate
γ	recovery rate
ρ	fraction of initially infected population
τ	final size of the epidemic
T	end of observation period
$\mathcal{R}_0$	basic reproduction number
$S_i(t)$, $I_i(t)$, $T_i(t)$	indicator functions taking value 1 if at time t, i is, respectively, susceptible, infected or removed and 0 otherwise
$S(t)$, $I(t)$, $R(t)$	numbers of susceptible, infected and recovered individuals at time t
$T_{i,I}$, $T_{i,R}$	the times of infection and recovery of i
T_I, T_R	the times of infection and recovery of a randomly chosen individual
W	the infectious period, i.e. $W := T_R - T_I$
f_τ	the density of T_I conditional on $T_I < \infty$
g_τ	the density of T_R

The Kermack–McKendrick model [5] is the most fundamental example of an ecological model. It assumes the population is segregated into susceptible (S), infected (I) and recovered/removed (R) compartments. The time evolutions of the population proportions in compartments (denoted by S_t, I_t and R_t) are described by the following well-known system of ordinary differential equations (ODEs):

$$\left.\begin{aligned} \dot{S}_t &= -\beta S_t I_t, \\ \dot{I}_t &= \beta S_t I_t - \gamma I_t \\ \text{and} \quad \dot{R}_t &= \gamma I_t. \end{aligned}\right\} \tag{1.1}$$

Here, β and γ are the infection and recovery rates, respectively. Solutions to equation (1.1) are often called the susceptible–infected–recovered (SIR) curves (figure 1). The law of mass action has been implicitly assumed, so any infectious individual can infect any susceptible individual. The ODEs model in equation (1.1) averages out individual dynamics, so it does not capture the stochastic fluctuation of epidemic processes in real life. In particular, the practical problems of applying equation (1.1) to data are:

1. **Population size.** Since the quantities in the SIR equations are proportions, it is not immediately clear how to apply them to real epidemics, which occur in *finite* susceptible populations. Moreover, the size of the population is often unknown.
2. **Likelihood.** Since the SIR equations are deterministic, we cannot write a likelihood for epidemic data without further, often *ad hoc*, statistical assumptions about the form of the likelihood function.
3. **Aggregation over individuals.** The SIR model represents the mean-field equations for (scaled) population counts, aggregating out individual characteristics.

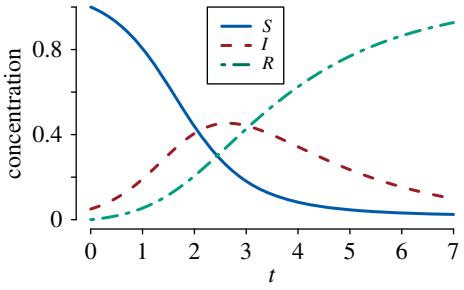

Figure 1. SDS interpretation of the SIR curves. The S_t curve is the survival function for time to infection: $S_t = P(T_I > t)$ where T_I is the time at which an individual moves from the susceptible to the infected compartment. The R_t curve, upon multiplication with $\mathcal{R}_0$, gives the corresponding cumulative hazard. Finally, the convolution of the infection time T_I and the infectious time T_R (time spent in the infected compartment) is given by the I_t curve, after adjustment for the initial infecteds. Parameter values: $\beta = 2$, $\gamma = 0.5$ with initial condition $S_0 = 1$, $I_0 = 0.05$ and $R_0 = 0$. (Online version in colour.)

4. **Aggregation over time.** The real data are typically aggregated not just over the population but also over observed time periods, leading to interval censoring[1] that cannot be easily incorporated into the SIR equations.

In this paper, we show that simple algebraic manipulation of the SIR equation (1.1) uncovers a precise probability law for the individual transitions between compartments. We refer to this interpretation of the solutions of equation (1.1) as a survival dynamical system (SDS). This new interpretation allows us to apply tools from survival analysis to population-level epidemic data. It directly addresses the first two problems listed above, and it lays a theoretical foundation for addressing the latter two problems. We focus on Markovian mass-action SIR models in this paper, but the SDS approach generalizes to non-Markov and network-based epidemic models.

The rest of the paper is structured as follows. First, we briefly review the relevant background on mathematical modelling in epidemiological literature. In §2, we make the SDS interpretation of the SIR equation (1.1) precise. In §3, we show how this approach can be used for statistical inference and compare the performance of estimators based on SDS likelihoods to those based on standard complete-data likelihoods. In §4, we use an SDS likelihood to analyse 2009 influenza A(H1N1) outbreak data from Washington State University. Finally, we conclude the paper with a brief discussion in §5. Additional mathematical preliminaries, statistical inference results and other material are provided in the appendices. A list of symbols used in the paper is provided in table 1.

1.1. Individual level: agent-based susceptible–infected–recovered model

Suppose we have n susceptible and m infectious individuals initially. Infectious individuals infect susceptible individuals, who change state from susceptible to infected. Infected individuals recover after an exponential infectious period. All infectious contacts and recoveries are assumed independent of each other. For the i-th individual, define the process S_i

such that $S_i(t) = 1$ if he or she is in the susceptible compartment at time t and $S_i(t) = 0$ otherwise. Similarly, define the processes I_i for the infected compartment and R_i for the recovered compartment. Naturally, $S_i(t) + I_i(t) + R_i(t) = 1$. For time $T \in (0, \infty)$, we assume that the process $\{(S_i(t), I_i(t), R_i(t))\}_{i=1,\ldots,n+m;\ t\in[0,T]}$ is a continuous-time Markov chain (CTMC). For notational convenience, we have labelled the initial susceptible individuals 1, 2, …, n and the initial infectious individuals $n+1, n+2, \ldots, n+m$. Then the random time change representation of a CTMC (see [6, ch. 6, pp. 326–328], [7, eqn 5.2, ch. 5, p. 41] and [8, eqn 1.8, ch. 1, p. 11]) allows us to write, for each $i \in \{1, \ldots, n+m\}$,

$$\left.\begin{aligned} S_i(t) &= S_i(0) - Y_i\left(\int_0^t \frac{\beta}{n} S_i(s) \sum_{j=1}^{n+m} I_j(s)\,\mathrm{d}s\right), \\ I_i(t) &= I_i(0) + Y_i\left(\int_0^t \frac{\beta}{n} S_i(s) \sum_{j=1}^{n+m} I_j(s)\,\mathrm{d}s\right) - Z_i\left(\int_0^t \gamma I_i(s)\,\mathrm{d}s\right) \\ \text{and}\quad R_i(t) &= Z_i\left(\int_0^t \gamma I_i(s)\,\mathrm{d}s\right), \end{aligned}\right\} \quad (1.2)$$

where $Y_1, Y_2, \ldots, Y_{n+m}$ and $Z_1, Z_2, \ldots, Z_{n+m}$ are independent unit-rate Poisson processes. Models of this form are often called agent-based models in the literature [9,10].

An intuitive explanation behind the random time change represetation in equation (1.2) is as follows: consider individual i who is initially susceptible. He or she will change status from susceptible to infected as soon as one of the infected individuals make an infectious contact. Because infected individuals make infectious contacts independently, the amount of time the i-th individual will remain susceptible has an exponential distribution with rate $n^{-1}\beta\sum_{j=1}^{n+m} I_j$. Once infected, he/she cannot be infected again. Therefore, the jump of the local process S_i from 1 to 0 can be equivalently described by the jump of the process $Y_i(\int_0^t n^{-1}\beta S_i(s)\sum_{j=1}^{n+m} I_j(s)\,\mathrm{d}s)$, where Y_i is a unit-rate Poisson process. Note that when the local process S_i jumps from 1 to 0, the process I_i also jumps from 0 to 1. When i is in infected status, he/she will recover after an exponentially distributed amount of time with rate γ. Therefore, the jump of the local process I_i from 1 to 0 can be equivalently described by the jump of $Z_i(\int_0^t \gamma I_i(s)\,\mathrm{d}s)$, where Z_i is a unit-rate Poisson process. Similar arguments give the equation for the local process R_i. The random time change representation in equation (1.2) for the entire ensemble $\{(S_i(t), I_i(t), R_i(t))\}_{i=1,\ldots,n+m;\ t\in[0,T]}$ follows from these considerations.

An equivalent construction of the agent-based model in equation (1.2) was proposed by Sellke [11]. Let $T_{i,I}$ denote the amount of time i remains susceptible, provided he or she was susceptible initially. Given the history of the infection process $I(s) = \sum_{j=1}^{m+n} I_j(s)$ up to time t, the conditional probability that individual i remains susceptible until time t is given by

$$P(T_{i,I} > t \mid (I(s))_{s\in[0,t]}) = \exp\left(-\frac{\beta}{n}\int_0^t I(s)\,\mathrm{d}s\right). \quad (1.3)$$

Therefore, to each susceptible individual i, we can assign an independent Exponential(1) random variable Q_i and change his/her status from susceptible to infected when

$$Q_i > \Lambda(t) := \frac{\beta}{n}\int_0^t I(s)\,\mathrm{d}s.$$

Once a susceptible individual gets infected, he or she recovers after an infectious period that follows an exponential distribution with rate γ. If we denote the recovery time of the i-th individual by $T_{i,R}$, it follows immediately from equation (1.2) that $T_{i,R} - T_{i,I}$ and $T_{i,I}$ are independent and $T_{i,R} - T_{i,I}$ has an exponential distribution with rate γ. Symbolically,

$$T_{i,R} - T_{i,I} \perp T_{i,I} \quad \text{and} \quad T_{i,R} - T_{i,I} \sim \text{Exponential}(\gamma). \quad (1.4)$$

The fate of an individual is entirely described by the statistical distributions given in equations (1.3) and (1.4). The Sellke construction can also be derived using a statistical representation of agent-based models under the law of mass action based on contact intervals [12,13]. In this case, the contact interval distribution is Exponential(β).

These considerations lead to algorithm 1.1 for simulating the process in equation (1.2), which is known as the *Sellke construction* [7,14,15]. It can be easily verified that algorithm 1.1 is equivalent to simulating the system in equation (1.2).

Algorithm 1.1. Pseudocode for the Sellke construction.

1: Assume you have initially m infectives and n susceptibles. Arrange all n susceptibles according to the order statistics $Q_{(1)} < \cdots < Q_{(n)}$ of an iid random sample from Exponential(1)
2: Simulate $m+n$ infectious periods y_i as iid sample from Exponential(γ)
3: Calculate $\Lambda(t) = \frac{\beta}{n}\int_0^t I(u)du$ with removal times from Step 2 for initial infectives
4: **for** $i = 1, 2, \ldots, n$ **do**
5: Calculate $t_i = \inf\{t: Q\_(i) > \Lambda(t)\}$.
6: **If** $t_i < \infty$ **then**
7: Change i-th susceptible to infective
8: Set removal time $r_i = t_i + y_i$ for i
9: Update $\Lambda(t)$ with new infection and removal times
10: **else**
11: Stop and break loop
12: **end if**
13: Set $i = i + 1$.
14: **end for**

1.2. Population level: ecological susceptible–infected–recovered model

The simplest way to derive an ecological model from the agent-based model in equation (1.2) is via lumping or aggregation of states. When the aggregation of states is *strongly lumpable* [16,17] (also see appendix A), the resulting aggregated process remains Markovian for any choice of the initial distribution. For the SIR process, let $\mathcal{X} := \{S, I, R\}$ denote the state space of each individual. Then, $\mathcal{X}^{n+m}$ is the state space of the ensemble of individual-based S_i, I_i, R_i processes. Define the macro-level processes

$$S(t) = \sum_{i=1}^{n+m} S_i(t),\ I(t) = \sum_{i=1}^{n+m} I_i(t) \quad \text{and} \quad R(t) = \sum_{i=1}^{n+m} R_i(t), \quad (1.5)$$

which keep track of the total counts of susceptible, infected and recovered individuals. Let $L := \binom{n+m+2}{2}$. Partition $\mathcal{X}^{n+m}$ into $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_L$ such that any two states in each $\mathcal{X}_l$ produce the same counts for $S(t)$, $I(t)$, $R(t)$, for $l = 1, 2, \ldots, L$. It is easy to see that the Markov chain described in equation (1.2) is (strongly) lumpable with respect to the partition $\{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_L\}$ (see [10,17,18]). That is, the lumped process (S, I, R) is also Markovian for any choice of the initial distribution. Therefore, we can write

$$\left.\begin{aligned} S(t) &= S(0) - Y\left(\int_0^t \frac{\beta}{n} S(s)I(s)\,\mathrm{d}s\right), \\ I(t) &= I(0) + Y\left(\int_0^t \frac{\beta}{n} S(s)I(s)\,\mathrm{d}s\right) - Z\left(\int_0^t \gamma I(s)\,\mathrm{d}s\right) \\ \text{and}\quad R(t) &= Z\left(\int_0^t \gamma I(s)\,\mathrm{d}s\right), \end{aligned}\right\} \quad (1.6)$$

where Y and Z are independent unit-rate Poisson processes. This system can be simulated using the Doob–Gillespie algorithm (see algorithm B.1 in appendix B).

This ecological model is convenient in that it is amenable to asymptotic analysis. Indeed, for very large populations, we can approximate the scaled stochastic SIR dynamics by a system of ODEs [19,20]. This is sometimes called *mean-field* or *fluid limit* of the Markov jump process. For our SIR system in equation (1.6), the scaled process $(S_n, I_n, R_n) := (S/n, I/n, R/n)$ satisfies

$$\left.\begin{aligned} S_n(t) &= S_n(0) - \frac{1}{n} Y\left(n\int_0^t \beta S_n(s)I_n(s)\,\mathrm{d}s\right), \\ I_n(t) &= I_n(0) + \frac{1}{n} Y\left(n\int_0^t \beta S_n(s)I_n(s)\,\mathrm{d}s\right) - \frac{1}{n} Z\left(n\int_0^t \gamma I_n(s)\,\mathrm{d}s\right) \\ \text{and}\quad R_n(t) &= \frac{1}{n} Z\left(n\int_0^t \gamma I_n(s)\,\mathrm{d}s\right). \end{aligned}\right\} \quad (1.7)$$

By virtue of the Poisson law of large numbers (LLN) [6], which asserts that $n^{-1}V(nt) \approx t$ for a unit-rate Poisson process V when n is large, the processes in equation (1.7) converge to the solution of the following system of ODEs as $n \to \infty$ and $m/n \to \rho \in (0, 1)$:

$$\dot{s}_t = -\beta s_t \iota_t, \quad \dot{\iota}_t = \beta s_t \iota_t - \gamma \iota_t \quad \text{and} \quad \dot{r}_t = \gamma \iota_t. \quad (1.8)$$

These are identical to the Kermack–McKendrick ODEs in equation (1.1). The introduction of ρ is convenient because it sets $s_0 = 1$, $\iota_0 = \rho$ and $r_0 = 0$. The rate of convergence to this LLN ODEs limit can be computed using sample path large deviations principle (LDP) of the Markov process in equation (1.7). Standard tools from [21–23] as well as related results from [24–26] can be borrowed for this purpose.

2. Survival dynamical systems

The ODEs in equation (1.8) that describe the large-population limit of the ecological SIR model can be given an agent-based probabilistic interpretation. It is convenient to rewrite equation (1.8) as follows:

$$\left.\begin{aligned} s_t &= \exp\left(-\beta\int_0^t \iota_u\,\mathrm{d}u\right) = \exp(-\mathcal{R}_0 r_t), \\ \iota_t &= \rho \mathrm{e}^{-\gamma t} - \int_0^t \dot{s}_u \mathrm{e}^{-\gamma(t-u)}\,\mathrm{d}u \\ \text{and}\quad r_t &= \gamma\int_0^t \iota_u\,\mathrm{d}u, \end{aligned}\right\} \quad (2.1)$$

where $\mathcal{R}_0 = \beta/\gamma$ is the basic reproduction number. Here, the first two equations are obtained by partially solving the ODEs system using the integrating factor (first equation) and variation of parameter (second equation) methods.

In the limit of a large population, the time of infection T_I of a randomly chosen susceptible individual has the survival function

$$P(T_I > t) = s_t = \exp(-\mathcal{R}_0 r_t). \quad (2.2)$$

This is a direct analogue of equation (1.3) where the stochastic quantity $n^{-1}\int_0^t I(u)\,\mathrm{d}u$ is replaced by its deterministic limit $\int_0^t \iota_u\,\mathrm{d}u$ from equation (2.1). Similarly, $\mathcal{R}_0 r_t = \beta\int_0^t \iota_s \mathrm{d}s$ may be thought of as the cumulative hazard and $\beta\iota_t$ as the hazard function of the random variable T_I. This hazard is sometimes called the *force of infection*. In the limit of large n, the units become independent due to the phenomenon known as *mean-field independence* or *propagation of chaos* [27–29].

Because T_I is an improper random variable, its survival, cumulative hazard and hazard functions are also improper. The probability that $T_I = \infty$ equals s_∞, which is the limiting proportion of individuals who remain susceptible. Setting $s_\infty = 1 - \tau$ and $\tau = r_\infty - \rho$ where r_∞ is the limiting proportion of recovered individuals, we see that τ must satisfy the deterministic final size equation

$$1 - \tau = \exp(-\mathcal{R}_0(\tau + \rho)). \quad (2.3)$$

The final size equation is a contraction map, so it is amenable to numerically efficient fixed-point iteration schemes. Because $0 \le \tau < 1$, we may interpret τ as the probability that $T_I < \infty$. Given that $T_I < \infty$, its conditional survival function is

$$\tilde{s}_t = \frac{s_t - (1-\tau)}{\tau} \quad (2.4)$$

and its probability density is

$$f_\tau(t) = -\frac{\dot{s}_t}{\tau}. \quad (2.5)$$

Let T_R be the time of removal of an infected individual who is infected at time T_I (with $T_I < T_R$), and let

$$\tilde{\iota}_t = \iota_t - \rho\exp(-\gamma t) \quad (2.6)$$

be the infected proportion of the population excluding the remaining initial infecteds. From equations (2.1) and (2.5), we obtain

$$\frac{\gamma\tilde{\iota}_t}{\tau} = \int_0^t f_\tau(u)\,\gamma \mathrm{e}^{-\gamma(t-u)}\,\mathrm{d}u. \quad (2.7)$$

Because $f_\tau(u)$ is a density function, the right-hand side above is a convolution of the conditional density f_τ of T_I

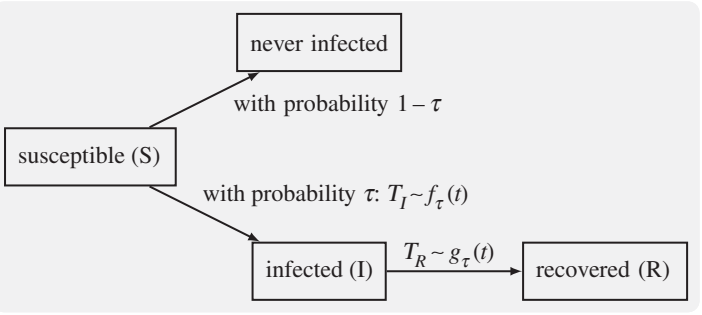

Figure 2. SDS derived from SIR equation (1.1). To each individual, we assign random variables T_I and T_R specifying his/her infection and recovery times, respectively. The laws of T_I and T_R are given by equations (2.5) and (2.8).

and the (exponential) density of $T_R - T_I$, the infectious period. It follows that the right-hand side quantity

$$g_\tau(t) = \frac{\gamma \tilde{\iota}_t}{\tau} \tag{2.8}$$

is itself a density of the variable T_R, which is the sum of the independent random variables T_I and $T_R - T_I$. Note the analogy of this result with equation (1.4). Let us denote the infectious period by the random variable $W := T_R - T_I$. These considerations give us algorithm 2.1 for simulating individual histories in the SIR model. See figure 2 for a pictorial representation of the idea.

Algorithm 2.1. Pseudocode for simulating a single SDS trajectory.

1: Calculate (s_t, ι_t, r_t) as given by equation (1.8)
2: With probability $1 - \tau$, where τ is given by equation (2.5), leave the unit in S state forever. With probability τ move to Step 3
3: Simulate infection time $T_I \sim f_\tau(t)$ where the density $f_\tau(t)$ is given by equation (2.5)
4: Independent of T_I, simulate infectious period $T_R - T_I \sim \text{EXPONENTIAL}(\gamma)$
5: Record the pair (T_I, T_R).

Analysing timepoints (T_I, T_R) according to algorithm 2.1 addresses all four issues of macro SIR model in equation (1.1) described in §1. Algorithm 2.1 no longer requires the population size (problem 1). Generation of individual trajectories according to algorithm 2.1 allows us to specify a likelihood function (problem 2), account for differences in individual characteristics (problem 3), and overcome issues with censoring or interval-based data (problem 4). Algorithm 2.1 brings us back from ecological to agent-based models and completes a conceptual 'micro-macro-micro' loop. The SDS interpretation has similarities with *symbolic dynamical systems* [30–32].

3. Parameter estimation

Under the stochastic agent-based SIR model equation (1.2) or its aggregated ecological version in equation (1.6), the vector of parameters of interest is $\theta = (\beta, \gamma, \rho)$ with $m = I(0) = \rho n$. The parameter τ is expressible in terms of θ via equation (2.3). The size of the initial susceptible population (n) is usually unknown and may be considered a nuisance parameter. The estimation of this nuisance parameter is often problematic, and popular methods such as profile likelihoods do not always yield good estimates. In order to address this problem, we propose a likelihood based on the SDS interpretation of the SIR model in equation (1.1) that does not require n (although n still may be estimated, see algorithm 4.1 in §4). Before going into the details of SDS likelihood, we describe the exact likelihood based on the Doob–Gillespie algorithm (see algorithm B.1 in appendix B). To emphasize the utility of the SDS likelihood, we compare its performance to an exact likelihood that is given the correct value of n.

3.1. Exact (Doob–Gillespie) likelihood

Assume we observe a total of $z = z_I + z_R$ events $(k_i, t_i)_{i=0}^z$ at times $0 < t_1 < \cdots < t_z = T$ where $k_i \in \{\text{I}, \text{R}\}$ denotes the type of event. Of these events, z_I are infections and z_R are removals. Put $X(t) = (S(t), I(t), R(t))$. Then, following algorithm B.1, the exact log-likelihood for θ is

$$\begin{aligned}\ell_1(\theta \mid X(t)_{t\in[0,T]}) &= \sum_{i=1}^{z} \log(\lambda_{k_i}(X(t_i))) - \int_0^T [\lambda_I(X(t)) + \lambda_R(X(t))]\,\mathrm{d}t \\ &= z_I \log(\beta) + z_R \log(\gamma) + \sum_{i:k_i=\text{I}} \log(S(t_i)/n) \\ &\quad + \sum_{i=1}^{z} \log(I(t_i)) - \int_0^T \frac{\beta}{n} S(t) I(t)\,\mathrm{d}t - \int_0^T \gamma I(t)\,\mathrm{d}t,\end{aligned} \tag{3.1}$$

where the last two integrals may be also written as finite sums. It is important to note that the above likelihood is conditional on the initial value $X(0) = (n, \rho n, 0)$, which we assume to be known. From equation (3.1), the maximum-likelihood estimate (MLE) for β and γ can be derived as

$$\hat{\beta} = \frac{n z_I}{\int_0^T S(t) I(t)\,\mathrm{d}t} \quad \text{and} \quad \hat{\gamma} = \frac{z_R}{\int_0^T I(t)\,\mathrm{d}t}. \tag{3.2}$$

Because we know the population size n and the trajectory $X(t)_{t\in[0,T]}$ when using the exact likelihood, the parameter $\rho = n^{-1} I(0)$ is also known exactly.

3.2. Survival dynamical system likelihood

Following the discussion in §2, an approximation of the exact likelihood function $\ell_1(\theta)$ in equation (3.1) can be obtained from equation (1.3) by replacing the process $n^{-1} I(u)$ with its asymptotic limit ι (as $n \to \infty$) and considering the individual trajectories as independent. Since we let $n \to \infty$, the exact value of the initial size of the susceptible population is no longer needed.

Assume we randomly sample $N+M$ individuals of whom N are initially susceptible and M initially infected. We observe these $N+M$ individuals up to the cut-off time T and record their infection or recovery times. Suppose K out of the N initially susceptible individuals get infected at times $t_1, t_2, \ldots, t_K$ and L of them recover by time T. Pair each infection time t_i with the corresponding duration of infectious period w_i if the individual recovers by time T. If the individual does not recover by time T, pair t_i with the censored recovery period $w_i = T - t_i$. Among the M initially infected individuals, suppose $\tilde{L}$ individuals recover by the cut-off T at times $\epsilon_1, \epsilon_2, \ldots, \epsilon_{\tilde{L}}$. Then, following algorithm 2.1, we have the following SDS likelihood:

$$\ell_2(\theta \mid \{t_i, w_i\}_{i=1}^{K}, \{\epsilon_j\}_{j=1}^{\tilde{L}}) = (N-K)\log(s_T) + \sum_{i=1}^{K}\log(\tau f_\tau(t_i)) + (L+\tilde{L})\log(\gamma) - \gamma\left(\sum_{i=1}^{K} w_i + \sum_{j=1}^{\tilde{L}}\epsilon_j + (M-\tilde{L})T\right), \tag{3.3}$$

where, as described in §2,

$$f_\tau(t) = \beta\tau^{-1}\iota_t \exp(-\mathcal{R}_0 t_t), \quad s_t = \exp(-\mathcal{R}_0 r_t),$$

and $\tau = r_\infty - \rho$ satisfies equation (2.3). In the next section, we evaluate the performance of the SDS likelihood from equation (3.3) in MLE and Markov chain Monte Carlo (MCMC) implementations.

3.3. Bayesian estimation using Markov chain Monte Carlo

In order to construct a posterior distribution for θ, we assign gamma priors to the parameters β, γ and ρ:

$$\left.\begin{aligned} \beta &\sim \text{GAMMA}(a_\beta, b_\beta), \\ \gamma &\sim \text{GAMMA}(a_\gamma, b_\gamma) \\ \text{and} \quad \rho &\sim \text{GAMMA}(a_\rho, b_\rho). \end{aligned}\right\} \tag{3.4}$$

The positive quantities a_β, b_β, a_γ, b_γ, a_ρ and b_ρ are appropriately chosen hyper-parameters. The posterior distribution of θ is obtained by Bayes' rule: it is proportional to the product of the likelihood function given in equation (3.3) and the three priors above.

Unfortunately, the posterior distribution of the SDS likelihood cannot be written in closed form. Even if a conditional posterior distribution is obtained, any closed-form expression for the probability density function would require solutions s_t, ι_t, r_t to equation (2.1), which are themselves functions of θ. Thus, we cannot employ a generic Gibbs sampler method [33,34], and we need a more efficient updating algorithm than the standard Metropolis–Hastings algorithm. Here, we adopt the robust adaptive metropolis (RAM) algorithm [35,36], which adapts the tuning constant and the variance–covariance matrix of the proposal distribution to maintain a consistent acceptance ratio in the Metropolis steps, which helps achieve good mixing of the chain. The variance–covariance matrix is updated during the MCMC iterations. In algorithm B.2, in appendix B, we provide pseudocode for implementing an MCMC procedure for drawing posterior samples using RAM.

3.4. Simulation study

The SDS likelihood presented in the previous section has several theoretical advantages. Two of the main advantages are: (a) it does not require knowledge of the number of initially susceptible individuals n and (b) it works with partial data in that it requires trajectories of *only* a randomly chosen sample of individuals. Nevertheless, the SDS likelihood is based on an LLN approximation of a large population, so it is important to evaluate the accuracy of this approximation. In this section, we compare the accuracy of the inference based on the SDS likelihood (without n) to that of the exact likelihood (with n). Though the comparison is deliberately unfair in that exact value of n and full data trajectories are supplied only to the exact likelihood, our objective is to see how much worse the inferences from the SDS likelihood are due to the approximation error as well as lack of n and full data trajectories. The data used for parameter inference are generated according to algorithm 1.1.

We compare three different inference methods:

1. **Method 1** uses the Doob–Gillespie likelihood given in equation (3.1) and calculates MLE according to equation (3.2).
2. **Method 2** also uses the Doob–Gillespie likelihood given in equation (3.1), but implements an MCMC scheme with the priors listed in equation (3.4) to infer θ. Because of conjugacy of the gamma priors, the posteriors are also gamma distributions [33]. In particular, they are given by

$$\beta \mid (X(t)_{t\in[0,T]}) \sim \text{GAMMA}\left(\text{nz}_\text{I} + \text{a}_\beta, \int_0^\text{T} \text{S(t)I(t)}\,\text{dt} + \text{b}_\beta\right)$$

$$\text{and} \quad \gamma \mid (X(t)_{t\in[0,T]}) \sim \text{GAMMA}\left(\text{z}_\text{R} + \text{a}_\gamma, \int_0^\text{T} \text{I(t)}\,\text{dt} + \text{b}_\gamma\right).$$

3. **Method 3** uses the SDS likelihood given in equation (3.3) and follows the MCMC procedure described in algorithm B.2.

For all MCMC-based methods, we constrain the proposed values of ρ in the MCMC iteration steps so that ρ remains within (0, 1) and satisfies equation (2.3). We have a total of 18 simulation scenarios based on combinations of the following:

— Three values of $\theta = (\beta, \gamma, \rho)$: $\theta_1 = (2.0, 0.5, 0.05)$, $\theta_2 = (2.0, 1.0, 0.05)$ and $\theta_3 = (1.5, 1.0, 0.05)$ yielding $\mathcal{R}_0$ equal to 4, 2 and 1.5, respectively.
— Two cut-off times T. Since the epidemic curve sees an exponential growth phase near the beginning, one often runs into problems such as overestimation of the size of the outbreak if inference is done using data collected when the epidemic is at or just before its peak. In order to see the impact of the censoring time T, we choose two cut-off times. One cut-off time is chosen around the half-time of the epidemic duration (near the peak of the infection process) and another one towards the end. The chosen values of T in our simulation set-up are 3 and 9 for θ_1, 3 and 7 for θ_2, and 3 and 6 for θ_3. See figure 8 for the SIR curves for different parameter values and cut-off times. The vertical line in each plot represents the cut-off time.
— Three values of the size of the susceptible population n: 10^2, 10^3 and 10^4.

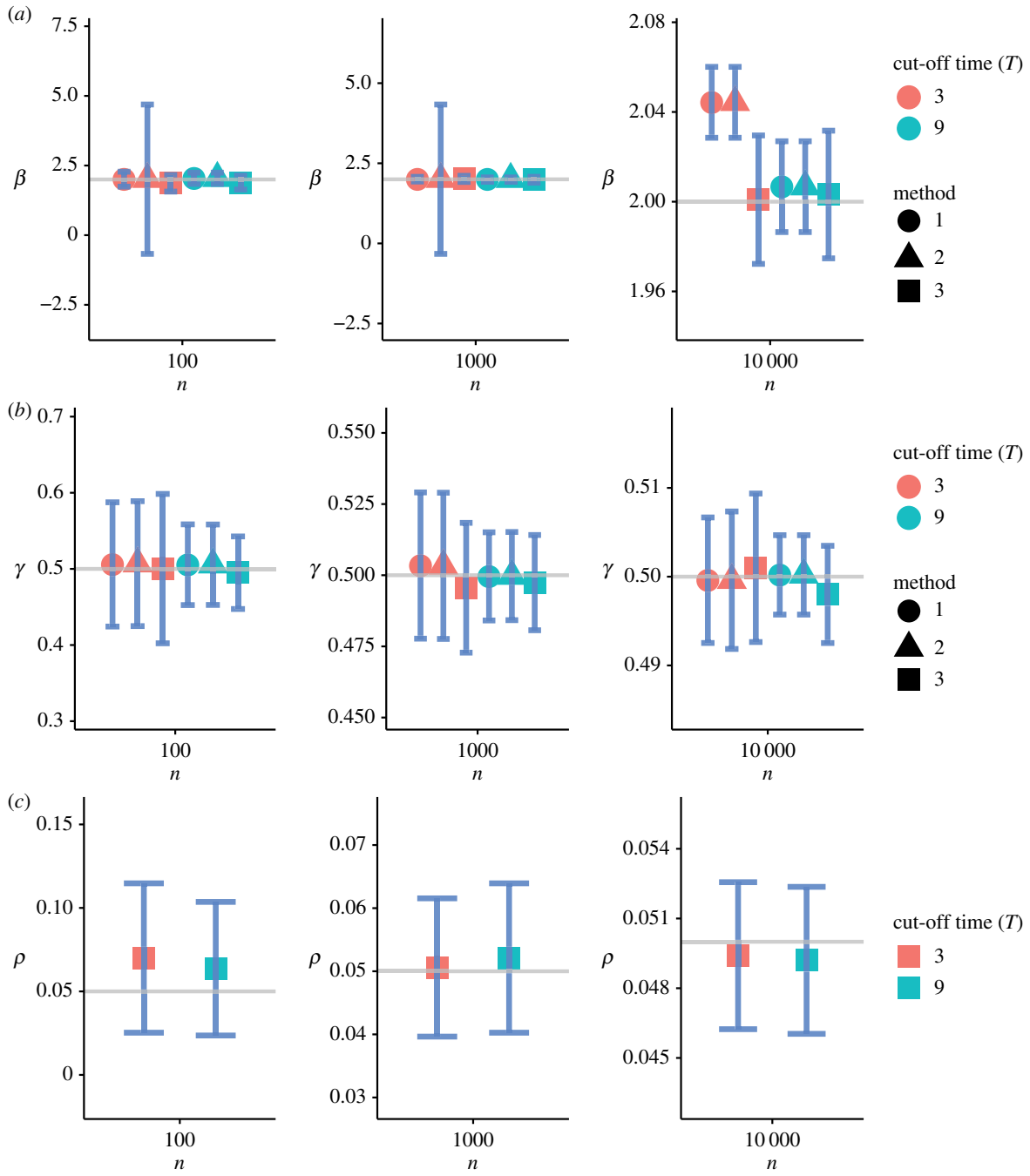

Figure 3. Inference under the parameter setting θ_1. (*a–c*) The parameters β, γ and ρ, respectively. The solid grey lines correspond to the true parameter values. The error bars correspond to ±1.96 s.d. of each estimate. The parameter ρ is estimated by Method 3 only. (Online version in colour.)

For each of the 18 scenarios, we generate 100 sets of synthetic epidemic data using algorithm 1.1. Each generated dataset has $n + n \times \rho$ rows (one for each individual in the epidemic) and two columns (one for T_I and one for T_R). To ensure the prior distributions in our Bayesian inference are uninformative, we set $a_i = i \times 0.01$ and $b_i = 0.01$ for $i = \beta, \gamma$ and ρ. For Method 2, we generate 1000 samples without any burn-in phase or thinning because Monte Carlo simulations are sufficient. For Method 3, we iterated the MCMC procedures 11 000 times. The first 1000 iterations are removed as burn-in. After burn-in, every 10th iteration is stored as a posterior sample. In total, 1000 posterior samples are used for estimation. For the Bayesian methods (i.e. Method 2 and Method 3), we estimate the parameters β, γ and ρ by taking the means of 1000 posterior samples.

Figure 3 summarizes the numerical results of the parameter setting θ_1. Figure 4 shows the results of the parameter setting θ_2, and figure 5 shows the results of the parameter setting θ_3. In addition to the parameter estimates (posterior means), error bars (1.96 s.d.) are also provided. These figures show that Method 3 based on the SDS likelihood fares well against Methods 1 and 2 based on the exact likelihood. Barring minor exceptions, Method 3 yielded accurate estimates for all three parameters β, γ and ρ even for relatively small values of n. The results for $n = 10^2$ are particularly encouraging. Tables 2 and 3 show that the mean squared error (MSE) decreases with increasing n across all three methods. As expected, the quality of inferences for the large cut-off time settings is better than that for the small cut-off time settings.

Since ρ is assumed known for Methods 1 and 2, it is estimated only in Method 3. Figures 3*c*, 4*c* and 5*c* show that the quality of estimation is sometimes poor when n is small. Note the $n = 10^2$ case in particular. Nevertheless, it is estimated accurately when n is moderately large.

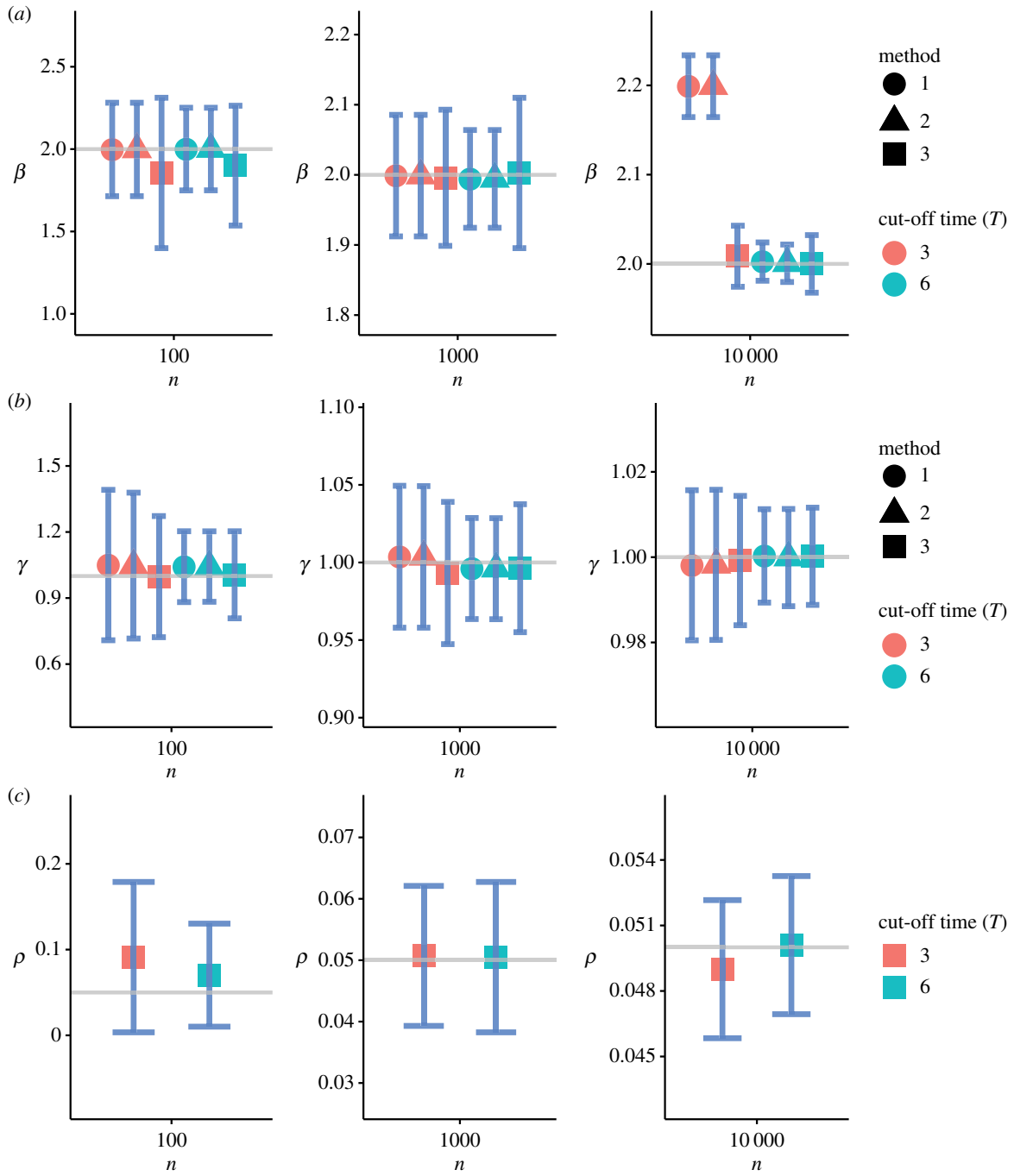

Figure 4. Inference under the parameter setting θ_2. (*a–c*) The parameters β, γ and ρ, respectively. The solid grey lines correspond to the true parameter values. The error bars correspond to ±1.96 s.d. of each estimate. The parameter ρ is estimated by Method 3 only. (Online version in colour.)

Further numerical results and explanations are provided in appendix C. Method 3 seems to have a slightly larger variance than the other two methods. Even though visual inspection suggests that Method 3 achieves comparable performance against Method 2 and Method 3, a more objective criterion would be useful. Such a criterion should take into account both the biases and the MSE of the methods. For instance, information criteria such as the focused information criterion [37] can be used for this purpose. However, our intention here is not to find which method performs the best, but rather to find how the approximate SDS likelihood performs against the exact likelihoods. Since figures 3–5 and the additional results in appendix C provide satisfactory evidence in favour of Method 3 and give adequate insight into its performance, we do not perform any further comparative analysis. Instead, we apply the SDS likelihood to a real dataset in the next section.

4. Data analysis

In the autumn of 2009, a new strain of influenza spread around the world after its initial outbreak in the state of Veracruz, Mexico in April 2009. The influenza A(H1N1)pdm09 virus was a triple reassortment of bird, swine and human flu viruses further combined with a Eurasian pig influenza virus [38]. Unlike most strains of influenza, this influenza A(H1N1) virus did not disproportionately infect adults older than 60 years, and it spread easily among young, healthy adults. This feature of the virus resulted in multiple outbreaks of the disease on college campuses across the continental USA. An outbreak on the campus of Washington State University (WSU) in

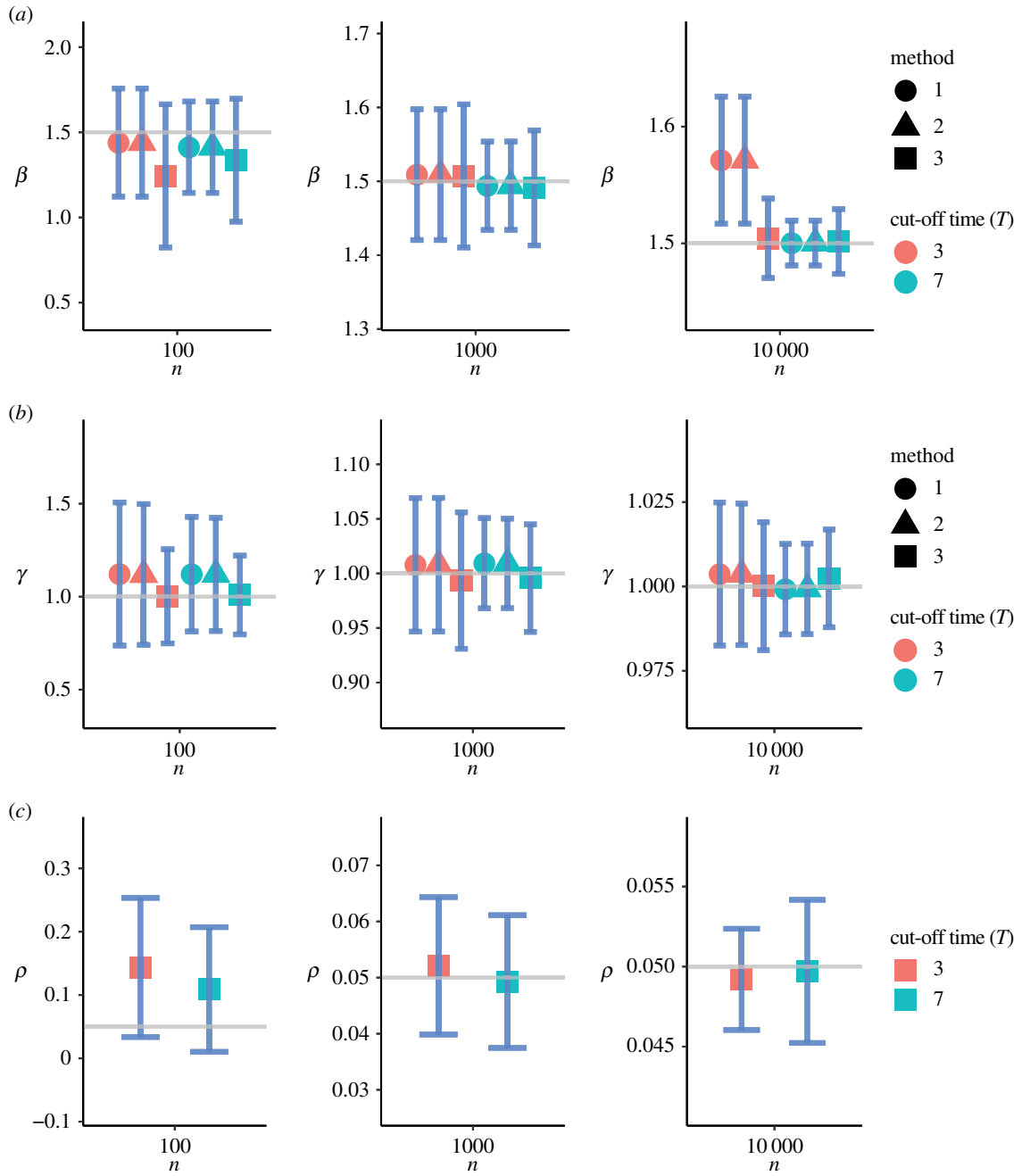

Figure 5. Inference under the parameter setting θ_3. (*a–c*) The parameters β, γ and ρ, respectively. The solid grey lines correspond to the true parameter values. The error bars correspond to ±1.96 standard deviations of each estimate. The parameter ρ is estimated by Method 3 only. (Online version in colour.)

Pullman, Washington began in late August 2009, upon the return of students for the autumn semester. Over a period of slightly more than three months, almost 2300 students were seen at the campus health centre with influenza-like illnesses that were treated as influenza A(H1N1) infections.[2]

Figure 6 shows daily counts of new infections for 105 days beginning on 22 August 2009. The counts were obtained directly from the cases of 'influenza-like illness' among students who visited or called the WSU Student Health Services seeking care. In our statistical analysis, the collected daily counts were considered as records of 'new infectives'. This particular dataset is interesting because it was obtained from an approximately closed population. The WSU campus is located in a town with a large student population (around 18 000 students) and a relatively small resident population (around 9000 residents). The location is relatively remote, with an average population density of only eight households per square mile in the surrounding rural areas.

As discussed in an earlier analysis of this dataset [38], these data may have been subject to both over-reporting and under-reporting: Some students may have assumed they had H1N1 when they had other influenza-like illnesses, while some students infected with H1N1 may not have sought medical care. However, such misreporting was considered to be relatively minor compared to the overall counts in the dataset [39]. This dataset was analysed earlier using a stochastic SIR model with parameters estimated using both likelihood-based and least-squares methods. Here, we re-analyse it using the SDS likelihood, emphasizing its multilevel nature by showing how the shape of the epidemic curve reflects changes in risk of infection in students who were susceptible.

Table 2. Summary of the numerical results for the longer cut-off times. Here, the values of T are 9 for θ_1, 6 for θ_2 and 7 for θ_3 such that T is near the end of the epidemic process (also see figure 8). Method 3 yields accurate estimates without requiring knowledge of the size of the susceptible population n. Values in italics indicate the results corresponding to the best performing method.

			β			γ			ρ
	n	statistics	Method 1	Method 2	Method 3	Method 1	Method 2	Method 3	Method 3
$\beta = 2$	10^4	Avg.	2.0067	2.0067	*2.0032*	*0.5002*	0.5002	0.4980	*0.0492*
		(MSE)	(0.00046)	(0.00046)	(0.00082)	(0.00002)	(0.00002)	(0.00004)	(0.00001)
$\gamma = 0.5$	10^3	Avg.	*2.0033*	2.0033	1.9868	0.4996	*0.4997*	0.4974	*0.0521*
		(MSE)	(0.00334)	(0.00334)	(0.00883)	(0.00024)	(0.00024)	(0.00028)	(0.00014)
$\rho = 0.05$	10^2	Avg.	2.0433	*2.0432*	1.8890	0.5055	0.5055	*0.4950*	*0.0636*
		(MSE)	(0.04238)	(0.04236)	(0.07655)	(0.07502)	(0.00284)	(0.00230)	(0.00178)
$\beta = 2$	10^4	Avg.	2.0026	2.0007	*2.0000*	1.0003	*0.9999*	1.0002	*0.0501*
		(MSE)	(0.00046)	(0.00044)	(0.00101)	(0.00012)	(0.00013)	(0.00013)	(0.00001)
$\gamma = 1$	10^3	Avg.	1.9942	1.9942	*2.0027*	0.9961	0.9960	*0.9963*	*0.0505*
		(MSE)	(0.00489)	(0.00489)	(0.01151)	(0.00107)	(0.00108)	(0.00172)	(0.00015)
$\rho = 0.05$	10^2	Avg.	*2.0002*	2.0005	1.8997	1.0425	1.0431	*1.0056*	*0.0702*
		(MSE)	(0.06295)	(0.0628)	(0.14231)	(0.02772)	(0.02748)	(0.03915)	(0.00402)
$\beta = 1.5$	10^4	Avg	*1.5003*	1.5003	1.5016	0.9992	*0.9993*	1.0024	*0.0497*
		(MSE)	(0.00037)	(0.00037)	(0.00078)	(0.00018)	(0.00018)	(0.00022)	(0.00002)
$\gamma = 1$	10^3	Avg.	1.4940	*1.4941*	1.4911	1.0094	*1.0092*	0.9957	*0.0493*
		(MSE)	(0.00362)	(0.00362)	(0.00615)	(0.00180)	(0.00177)	(0.00245)	(0.00014)
$\rho = 0.05$	10^2	Avg.	1.4126	*1.4127*	1.3362	1.1211	1.1199	*1.0090*	*0.1087*
		(MSE)	(0.0796)	(0.07962)	(0.15705)	(0.10955)	(0.10715)	(0.04502)	(0.01313)

The density of the infection time (conditional on $T_I < \infty$) is given by $f_\tau(t) = -\dot{s}_t/\tau$ (see equation (2.5)). Consequently, for the collection of n individuals at risk out of which k are seen to be infected at times $t_1 < \cdots < t_k < T$ where $T < \infty$ in the observation time horizon (i.e. censoring time), we have the log-likelihood function for infection times

$$\ell_I(t_1, \ldots, t_k|\theta, n) = (n-k)\log s_T + \sum_{i=1}^{k} \log f_\tau(t_i),$$

where $\theta = (\beta, \gamma, \rho)$ is the vector of free parameters, with τ being an implicit function of θ according to equation (2.3). Note that the above likelihood is conditional on the number of individuals at risk n, which is also typically unknown, and that the value $0 \le k \le n$ is a random variable. In particular, if T is sufficiently large, we have approximately $k \sim \text{BINOMIAL}(n, \tau)$. Note that this implies in particular that if we do not know the value of n but have observed k, a reasonable estimate of the former is k/τ. In general, to impute a value of n, we could take $n \sim \text{NEGBINOM}(k, \tau)$, the negative binomial distribution. Conditionally on the value of k the (unobserved) recovery likelihood is then the usual log likelihood for the exponential survival model. Assuming r individuals have recovered after infectious periods $w_1 < \cdots < w_r < T$, we have

$$\ell_R(w_1, \ldots, w_r|\theta, k) = (k-r)\log H_\gamma(T) + \sum_{i=1}^{r} \log h_\gamma(w_i),$$

where $H_\gamma(\cdot)$ and $h_\gamma(\cdot)$ are, respectively, the survival function and the probability density function of the exponential distribution with rate γ. Averaging the infectious periods used in the previous analysis [38,39], we assume here that the recovery times have an exponential distribution with mean $\gamma^{-1} = 5.5$ days (see also [40,41]), so γ was not estimated. The complete log-likelihood conditional on the population size n, the parameters and observables is then

$$\ell_0(t_1 \ldots, t_k, w_1, \ldots, w_r|\theta, n) = \ell_I(t_1, \ldots, t_k|\theta, n) + \ell_R(w_1, \ldots, w_r|\theta, k).$$

Based on this SDS likelihood, algorithm 4.1 may be used for obtaining the posterior distributions of the parameters θ and n given the WSU dataset.

Algorithm 4.1. Estimation of joint posterior distribution of $\theta = (\beta, \gamma, \rho)$ and n.

1: Initiate $\theta = (\beta, \gamma, \rho)$ from the prior distribution and set $n = k$.
2: **repeat**
3: Generate new r and ℓ_R value based on independent sample of recovery times $w_i \sim g_\gamma(s)$ $s = 1, \ldots, k$.
4: Perform Metropolis-Hastings step for the target conditional distribution of $(\theta|n)$ using the complete log-likelihood $\ell_0 = \ell_I + \ell_R$.
5: Calculate τ based on the current value of θ using formula (2.3)
6: Sample the conditional distribution of $(n|\theta)$ by drawing $n \sim \text{NEGBINOM}(k, \tau)$.
7: **until** convergence

Table 3. Summary of the numerical results for the shorter cut-off times. Here, we fix $T = 3$ so that the epidemic process is near its peak at T (also see figure 8). Method 3 yields accurate estimates without requiring knowledge of the size of the susceptible population n. Values in italics indicate the results corresponding to the best performing method.

			β			γ			ρ
	n	statistics	Method 1	Method 2	Method 3	Method 1	Method 2	Method 3	Method 3
$\beta = 2$	10^4	Avg.	2.0443	2.0443	*2.0009*	0.4996	0.4996	*0.5010*	*0.0494*
		(MSE)	(0.00221)	(0.00221)	(0.00082)	(0.00006)	(0.00006)	(0.00007)	(0.00001)
$\gamma = 0.5$	10^3	Avg.	2.0041	*2.0040*	2.0134	0.5034	*0.5033*	0.4956	*0.0506*
		(MSE)	(0.00545)	(5.44670)	(0.00940)	(0.00067)	(0.00067)	(0.00053)	(0.00012)
$\rho = 0.05$	10^2	Avg.	2.0101	*2.0100*	1.8631	0.5059	0.5069	*0.5004*	*0.0700*
		(MSE)	(0.07191)	(7.19654)	(0.10753)	(0.00669)	(0.00677)	(0.00962)	(0.00240)
$\beta = 2$	10^4	Avg.	2.1991	2.1991	*2.0086*	0.9981	0.9982	*0.9992*	*0.0490*
		(MSE)	(0.04083)	(0.04083)	(0.00124)	(0.00031)	(0.00031)	(0.00023)	(0.00002)
$\gamma = 1$	10^3	Avg.	*1.9989*	1.9989	1.9958	1.0037	*1.0036*	0.9932	*0.0507*
		(MSE)	(0.00751)	(0.00751)	(0.00945)	(0.00210)	(0.00210)	(0.00214)	(0.00013)
$\rho = 0.05$	10^2	Avg.	1.9979	*1.9980*	1.8553	1.0499	1.0474	*0.9973*	*0.0912*
		(MSE)	(0.08047)	(0.08043)	(0.22925)	(0.11915)	(0.11203)	(0.07570)	(0.00939)
$\beta = 1.5$	10^4	Avg	1.5713	1.5713	*1.5044*	1.0037	1.0036	*1.0001*	*0.0492*
		(MSE)	(0.00804)	(0.00804)	(0.00118)	(0.00046)	(0.00046)	(0.00036)	(0.00001)
$\gamma = 1$	10^3	Avg.	1.5091	1.5091	*1.5073*	1.0079	1.0080	*0.9935*	*0.0521*
		(MSE)	(0.00794)	(0.00794)	(0.00945)	(0.00381)	(0.00381)	(0.00396)	(0.00016)
$\rho = 0.05$	10^2	Avg.	*1.4398*	1.4398	1.2439	1.1220	1.1192	*1.0020*	*0.1435*
		(MSE)	(0.10451)	(0.10461)	(0.24216)	(0.16303)	(0.15773)	0.06412)	(0.02082)

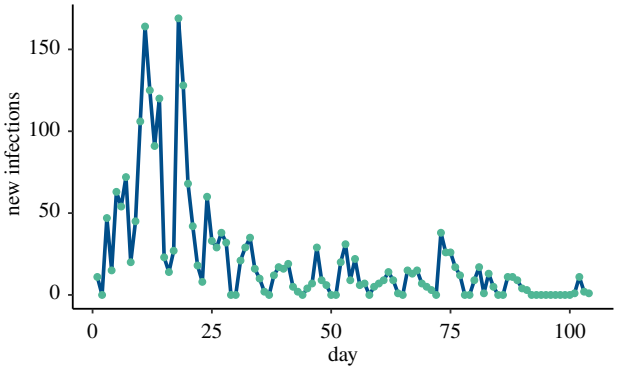

Figure 6. Daily new infection counts from WSU H1N1 outbreak. (Online version in colour.)

Table 4. The values of posterior parameter estimates and their credibility bounds based on the hybrid Gibbs sampler given the WSU data in figure 6.

parameter	MAP	90% credibility
n	7051	(6602, 7581)
β	0.1887	(0.185, 0.196)
ρ	0.0423	(0.04, 0.045)
$\mathcal{R}_0$	1.06	(1.04, 1.09)

The results of applying algorithm 4.1 to the WSU dataset are summarized in table 4 and in figure 7. As in previous sections, independent, non-informative gamma priors were used for θ. The uniform (improper) prior was used for n. The maximum *a posteriori* estimate (MAP) of the effective population size (population at risk) was found to be $n = 7051$. This is much smaller than the value of approximately 18 000 (total WSU student body) assumed in the previous analyses [38,39]. Consequently, the MAP value of $\mathcal{R}_0 \approx 1.06$ is slightly smaller than that obtained in the previous analysis, and the SDS-based MAP for ρ is substantially larger than other estimates of the initially infected. Contrary to previous analysis [39], these values suggest that the high peak of an epidemic in early days of the academic year was not caused by high infectivity among newly infected students but rather by a high number of already infected individuals (high value of ρ). This point was already made in [38].

5. Discussion

In this paper, we present a new way of using classical SIR-type epidemic models for statistical inference. Our method addresses all four problems identified in §1. Indeed, parameter estimation based on the SDS likelihood (described in §3) does not require the effective population size n, addressing problem 1. The SDS likelihood, being a direct consequence of the SDS interpretation of the SIR equation (1.1), provides a principled way of specifying a likelihood function from epidemiological field data where the effective population size is unknown but large, addressing problem 2. Although we do not explicitly illustrate this here, the independence of individuals' contributions to the SDS likelihood also addresses the problem of aggregation over individuals

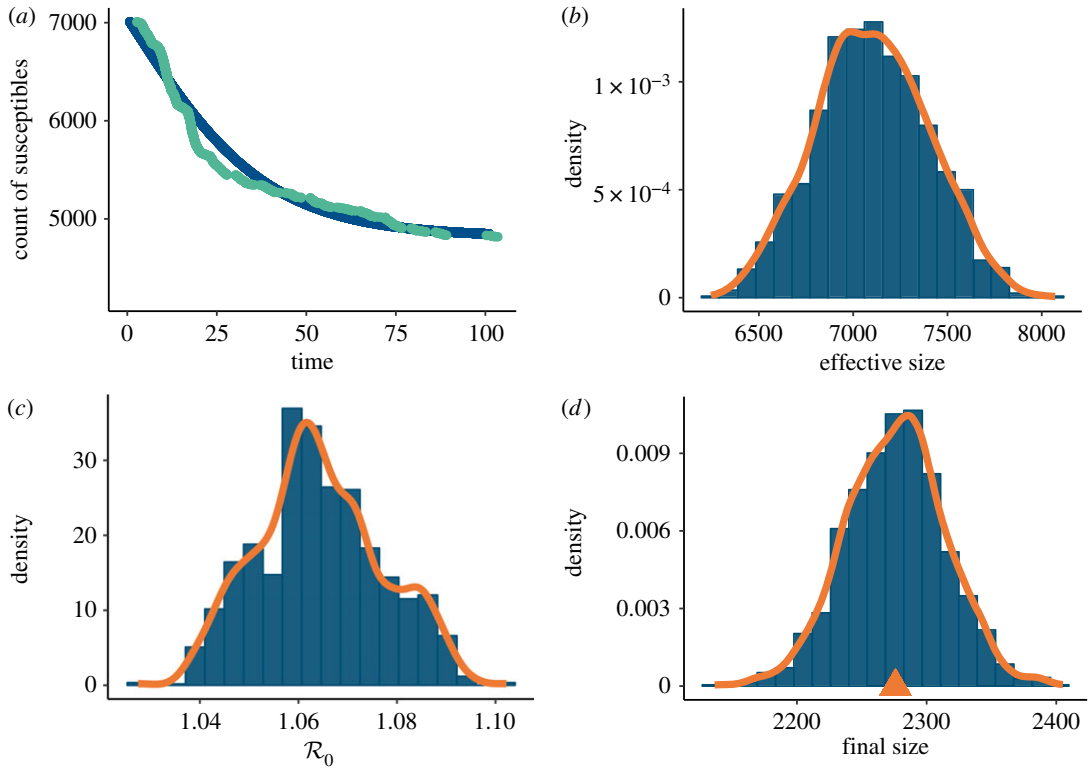

Figure 7. (*a*) Fitted (blue) versus observed (green) s_t curves and (*b*) the posterior distribution of the effective population size (n). Both curves are conditional on the effective population size. Posterior distributions of (*c*) the basic reproduction number $\mathcal{R}_0$ and (*d*) the final epidemic size. The distribution of the latter may be used to validate the model against actually observed data. The mark is placed at the actually observed epidemic size of 2276.

(problem 3) and over time (problem 4). Moreover, due to its product form, the SDS likelihood method is easier to implement and analyse than methods based on partially observed CTMC (e.g. the Doob–Gillespie likelihood).

The SDS method allows a novel approach to the monitoring of epidemics. Instead of longitudinally counting the number of infections, a random sample of individuals can be monitored continuously for changes in their health status. This is akin to a sentinel sensor network. Similar ideas have been routinely explored in communication networks literature in computer science (e.g. network probing and monitoring) [42]. The use of individual-level longitudinal data rather than counts allows much greater flexibility in estimating the effects of covariates (e.g. vaccination status) on infectiousness and susceptibility, and it extends easily to non-Markov transmission models.

Using the SDS likelihood, it typically suffices to have much smaller sample of transition data than other inference methods such as the exact likelihood method. Due to the asymptotic independence of infection and recovery times of individuals (see §2), the SDS likelihood takes a particularly simple form, facilitating a convenient implementation of a suitable MCMC scheme. We have made our code implementation of the SDS likelihood and MCMC scheme publicly available [43].

The SDS framework proposed here can be readily extended to accommodate a wide class of compartmental models with some partial ordering among compartments. The classical SIR model has been chosen here as an important example to illustrate the ideas underpinning SDS likelihoods. Indeed, the machinery developed in the present paper goes beyond compartmental SIR models, and it can be applied to more general epidemic processes as well as to many compartmental models arising in physics and chemistry. In particular, we believe SDS likelihoods can be applied to certain subclasses of chemical reaction network models in which the individual species molecules can be tracked as they undergo chemical reactions.

In many studies of epidemiological field data, the effective population size is assumed to be very large. For instance, a total population size of 10^6 was assumed in [44,45]. Our method is particularly appropriate for such settings. For smaller populations, knowledge of the rate of convergence of the scaled processes to the LLN limit is crucial for assessing the quality of inference based on the SDS likelihood. Therefore, to fully evaluate the appropriateness of the SDS approximation, one should first establish an LDP for the scaled process of interest. This is particularly important for small-scale epidemics. Even though our numerical results are encouraging for values of n as small as 100, quantifying the rate of convergence will be useful. Although we did not consider an LDP in this paper, we believe that the standard techniques [21–23,25,26,46] can be applied for this purpose in our context.

Another direction of future investigation will be to consider network-based systems and non-Markovian systems. For many epidemiological scenarios, the mass-action assumption is untenable. Several network-based models have been proposed in recent times [47–49]. Asymptotic study of those models in the form of various large-graph limits has also been done [50–52]. Therefore, extending our method to network-based models appears to be a natural next step.

Data accessibility. This article has no additional data.

Authors' contributions. G.A.R., E.K. and W.K.B. conceived and designed the research. B.C. provided numerical examples, contributed analysis tools and helped write the paper. W.K.B., E.K. and G.A.R. wrote the paper. All authors helped in editing and proofreading the final manuscript.

Competing interests. The authors declare no competing interests.

Funding. B.C. was supported by a Korea University Grant. G.A.R. was supported by National Science Foundation (NSF) grant nos. NSF-DMS 1440386 and NSF-DMS 1513489. E.K. and G.A.R. were supported by NSF grant no. NSF-DMS 1853587. E.K. was supported by National Institute of General Medical Sciences (NIGMS) grant no. U54 GM111274. E.K. and W.K.B. were supported by National Institute of Allergy and Infectious Diseases (NIAID) grant no. R01 AI116770. The content is solely the responsibility of the authors and does not represent the official views of NRF, NSF, NIGMS or NIAID.

Acknowledgements. The authors acknowledge the anonymous reviewers whose constructive feedback significantly improved the expository quality of the paper. In particular, the reference [37] was pointed out to the authors by one of the reviewers. A large part of this research was conducted during the Mathematical Biosciences Institute (MBI) semester-long programme on modelling infectious diseases in spring 2018. The authors thank MBI and its staff for their hospitality.

Endnotes

[1]A random variable is said to be interval-censored when it cannot be observed exactly and is only known to lie within an interval.
[2]In fact, as described in [39], for the first 10 days of the outbreak, all suspected cases were tested and laboratory-confirmed to be H1N1, after which all cases were considered H1N1.
[3]There is also a notion of weak lumpability in the theory of Markov processes.

Appendix A. Mathematical background

A.1. Lumpability of a Markov chain

Consider a CTMC C_t on a state space $\mathcal{Y} := \{1, 2, \ldots, K\}$ for a finite positive integer K. Given a partition $\{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_M\}$ of $\mathcal{Y}$, define the process $\tilde{C}_t$ such that $\tilde{C}_t := i$ whenever $C_t \in \mathcal{Y}_i$, for $i = 1, 2, \ldots, M$. The original CTMC C_t is said to be strongly lumpable with respect to the partition $\{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_M\}$ of $\mathcal{Y}$ if the process $\tilde{C}_t$ is also a CTMC for every choice of initial distribution of C_t. The process $\tilde{C}_t$ is often called the aggregated or the lumped process. Intuitively, lumpability is the property that disjoint sets of states can be identified by representative states such that the induced stochastic process on the representative states (which we call the aggregated or lumped process) is also Markovian for every choice of initial distribution of the original CTMC. In our individual-level model described in §1.1, the representative states are given by the partition $\{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_L\}$ of the state space $\mathcal{X}^{n+m}$. The representative states then correspond to the population counts $S(t)$, $I(t)$, $R(t)$.

The (strong) lumpability[3] of a CTMC can also be described in terms of lumpability of a linear system of ODEs. Consider the linear system $\dot{y} = yA$, where $A = ((a_{i,j}))$ is a $K \times K$ matrix (representing the transition rate or the infinitesimal generator matrix of the corresponding CTMC on state space $\mathcal{Y}$).

Definition A.1 (lumpability of a linear system [10,18]). The linear system $\dot{y} = yA$ is said to be lumpable with respect to a partition $\{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_M\}$ of $\mathcal{Y}$, if there exists an $M \times M$ matrix $B = ((b_{i,j}))$ satisfying Dynkin's criterion (i.e. if $b_{i,j} = \sum_{l \in \mathcal{Y}_j} a_{u,l} = \sum_{l \in \mathcal{Y}_j} a_{v,l}$ for all $u, v \in \mathcal{Y}_i$). The matrix B is often called a lumping of A. The following is immediate: if B is a lumping of A, then there exists an $K \times M$ matrix V such that $AV = VB$.

Refer to [16,17,53] for further reading and numerous characterizations of Markov chain lumpability.

Appendix B. Additional pseudocode

For the sake of completeness, we provide some additional pseudocode for implementing popular statistical procedures. The first pseudocode is for simulating trajectories of a CTMC following the well-known Doob–Gillespie algorithm.

Algorithm B.1. Pseudocode for Doob–Gillespie algorithm.

1: Initiate $(S(0), I(0), R(0))$
2: Assume you have the process value $(S(t), I(t), R(t))$ at $t \geq 0$
3: Calculate rates $\lambda_I(t) = \beta S(t) I(t)/n$ and $\lambda_R(t) = \gamma I(t)$
4: Set next transition time Δt as $\textsc{Exponential}(\lambda_I(t) + \lambda_R(t))$
5: Select transition type (infection or recovery) as

$$\textsc{Bernoulli}\left(\frac{\lambda_I(t)}{\lambda_I(t) + \lambda_R(t)}\right)$$

6: Update $(S(t'), I(t'), R(t'))$ at $t' = t + \Delta t$ and go to Step 2.

The MCMC procedure for drawing posterior samples using the RAM algorithm can be implemented by the following pseudocode.

Algorithm B.2. MCMC for drawing posterior sample using RAM method.

1: Initialize (β, γ, ρ) and the variance-covariance matrix of proposal distribution
2: **repeat** ▷ adjust for burn-in etc.
3: Draw candidate samples of (β, γ, ρ) from the proposal distribution
4: Solve equation (1.8) and store the solutions at the observed infection times $t_1, t_2, \ldots, t_K$
5: Perform one step of the Metropolis algorithm and determine whether the candidate samples are accepted
6: Perform one step of the RAM method to update the variance-covariance matrix of the proposal distribution
7: **until** convergence.

Appendix C. Additional numerical results

Here, we provide additional numerical results. In particular, we show the posterior plots and crucial diagnostic statistics for the MCMC methods.

The cut-off times are chosen based on figure 8. The idea is to study the impact of censoring on the quality of inference.

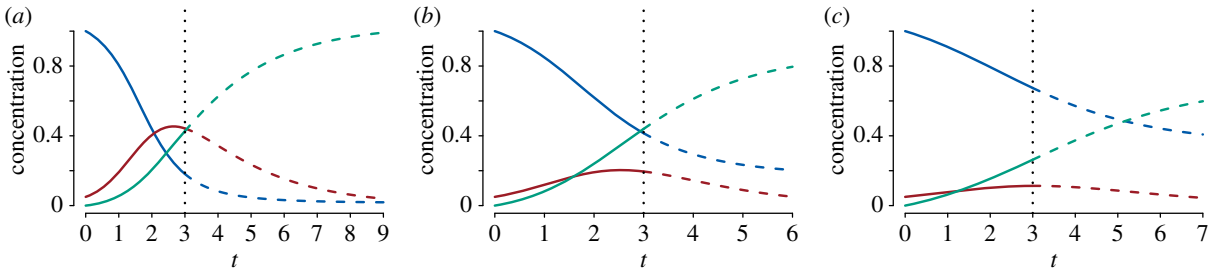

Figure 8. The SIR curves for the three different parameter values considered in §3.4. The initial values are $S_0 = 1$, $R_0 = \rho$ and $R_0 = 0$. The vertical dotted lines represent the cut-off times. (*a*) $\beta = 2$, $\gamma = 0.5$, $\rho = 0.05$, (*b*) $\beta = 2$, $\gamma = 1.0$, $\rho = 0.05$, (*c*) $\beta = 1.5$, $\gamma = 1.0$, $\rho = 0.05$. (Online version in colour.)

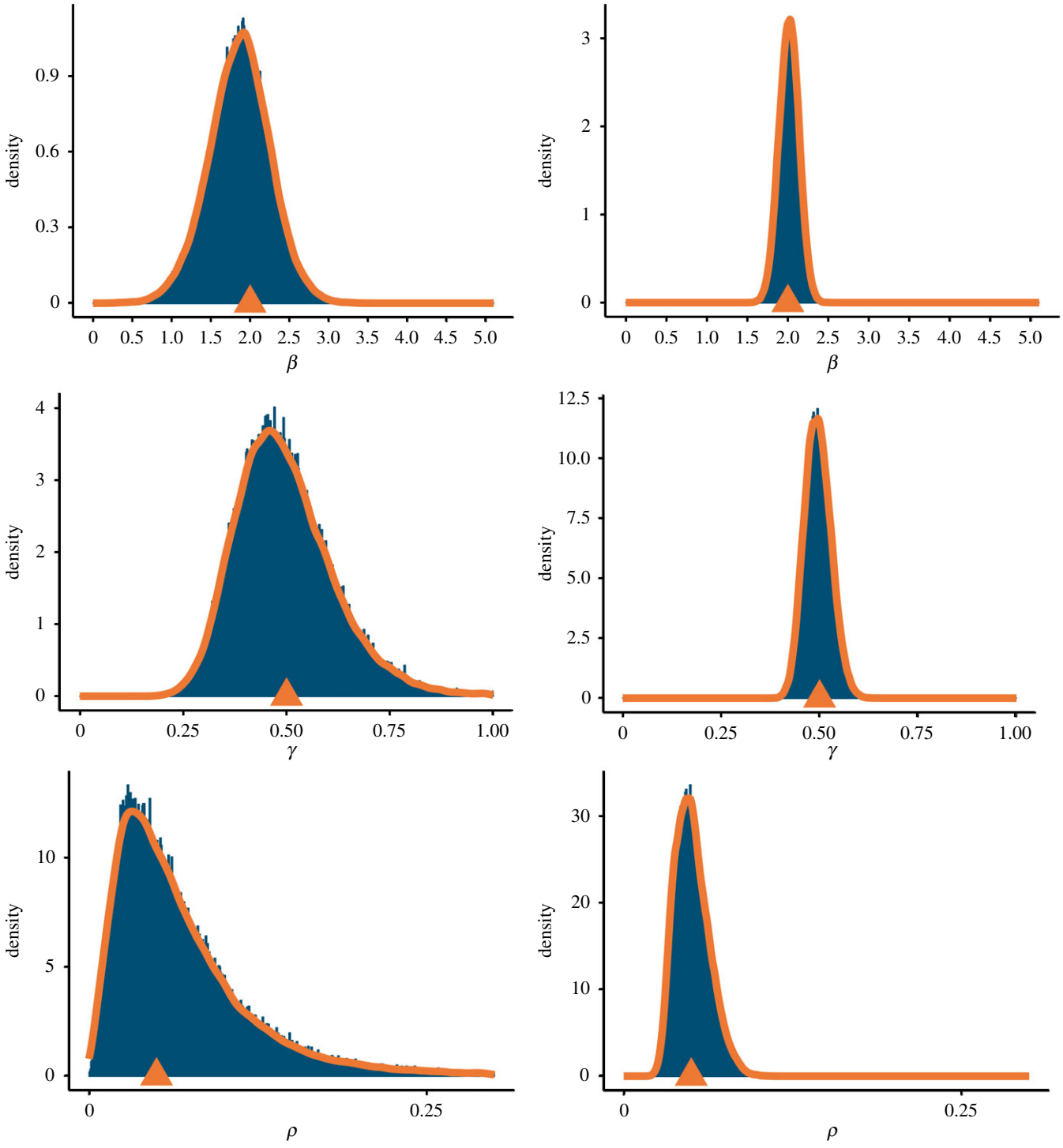

Figure 9. The posterior distributions of the Method 3 estimators of β, γ and ρ based on the SDS likelihood for the smaller cut-off time ($T = 3$). The left-hand panels correspond to $n = 10^2$, and the right-hand panels correspond to $n = 10^3$. The true parameter values are $\beta = 2$, $\gamma = 0.5$ and $\rho = 0.05$ (parameter setting θ_1). (Online version in colour.)

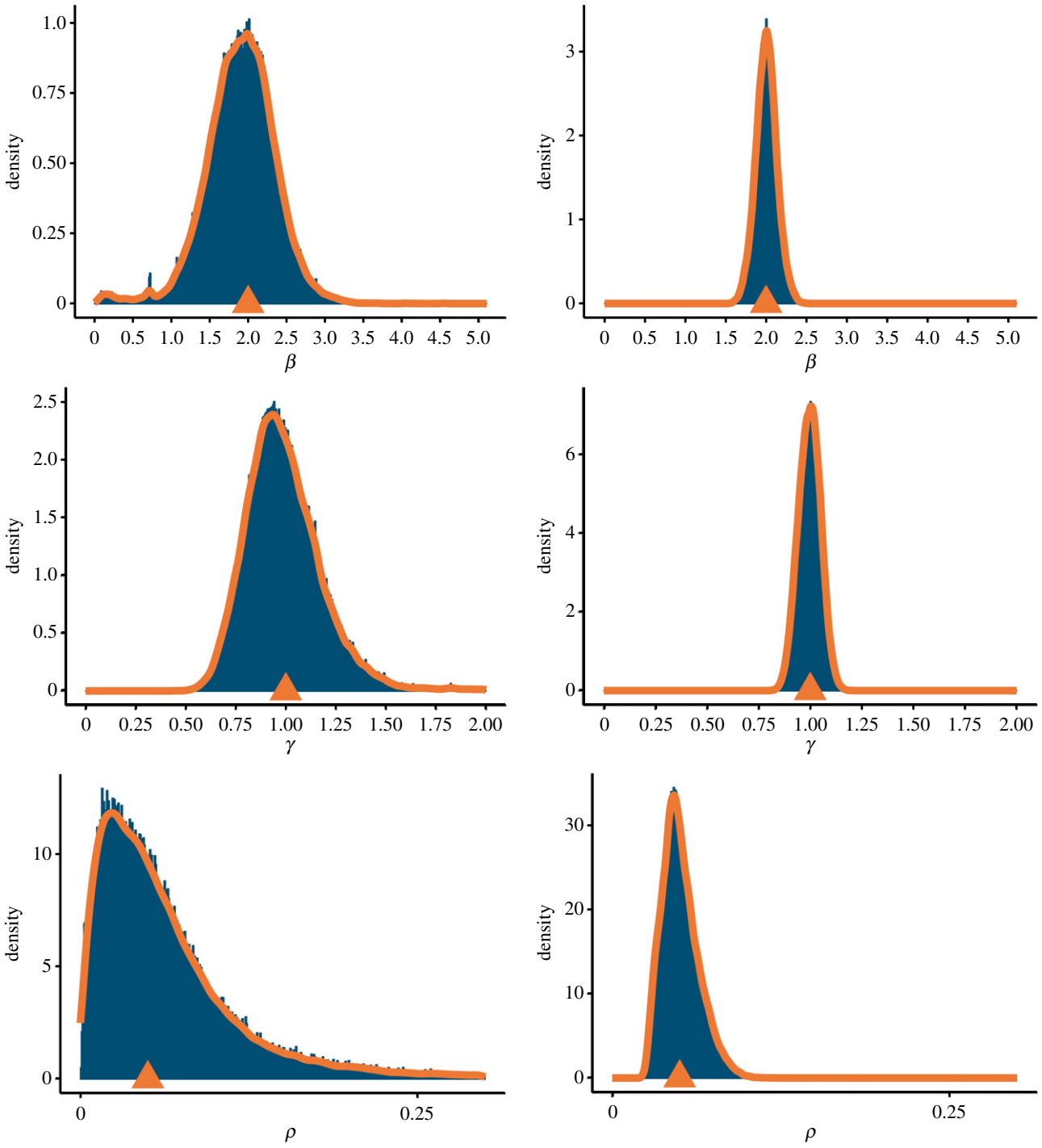

Figure 10. The posterior distributions of the Method 3 estimators of β, γ and ρ based on the SDS likelihood for the larger cut-off time ($T = 6$). The left-hand panels correspond to $n = 10^2$, and the right-hand panels correspond to $n = 10^3$. The true parameter values are $\beta = 2$, $\gamma = 1$ and $\rho = 0.05$ (parameter setting θ_2). (Online version in colour.)

For each parameter setting, we chose two cut-off times: one near the peak of the epidemic and one near the end of the epidemic. The vertical lines in figure 8 indicate the smaller cut-off time for each of the three settings of the parameter values.

Our numerical results are summarized in tables 2 and 3. For the longer cut-off times, table 2 provides a summary of the simulation results for the three parameter settings and different initial numbers of susceptibles n. Here, the values of T are 9 for θ_1, 6 for θ_2 and 7 for θ_3. In each case, the epidemic is almost at its end by time T (figure 8). The first three columns show estimates of β from Methods 1, 2 and 3. Similarly, the next three columns show estimates of γ. The last column shows Method 3 estimates of ρ (recall that ρ is known exactly for Method 1 and Method 2). The rows of the table are divided into three parts corresponding the parameter settings θ_1, θ_2 and θ_3. Each of the three parts is further subdivided into the three different susceptible population sizes $n = 10^2$, 10^3 and 10^3. Finally, in each cell, we show the average of 100 posterior means and the MSE of parameter estimators. As we can see, Method 3 based on the SDS likelihood yields accurate estimates for all three parameters β, γ and ρ even for relatively small values of n (see the results for $n = 10^2$).

Whereas table 2 considers data collected until a T near the end of an epidemic, table 3 considers data with cut-off $T = 3$ that is close to the peak of an epidemic. (See figure 8 for a visualization of the SIR curves corresponding to these three parameter settings truncated at $T = 3$ by a vertical line.) The table formats are identical. Since the inference is based on

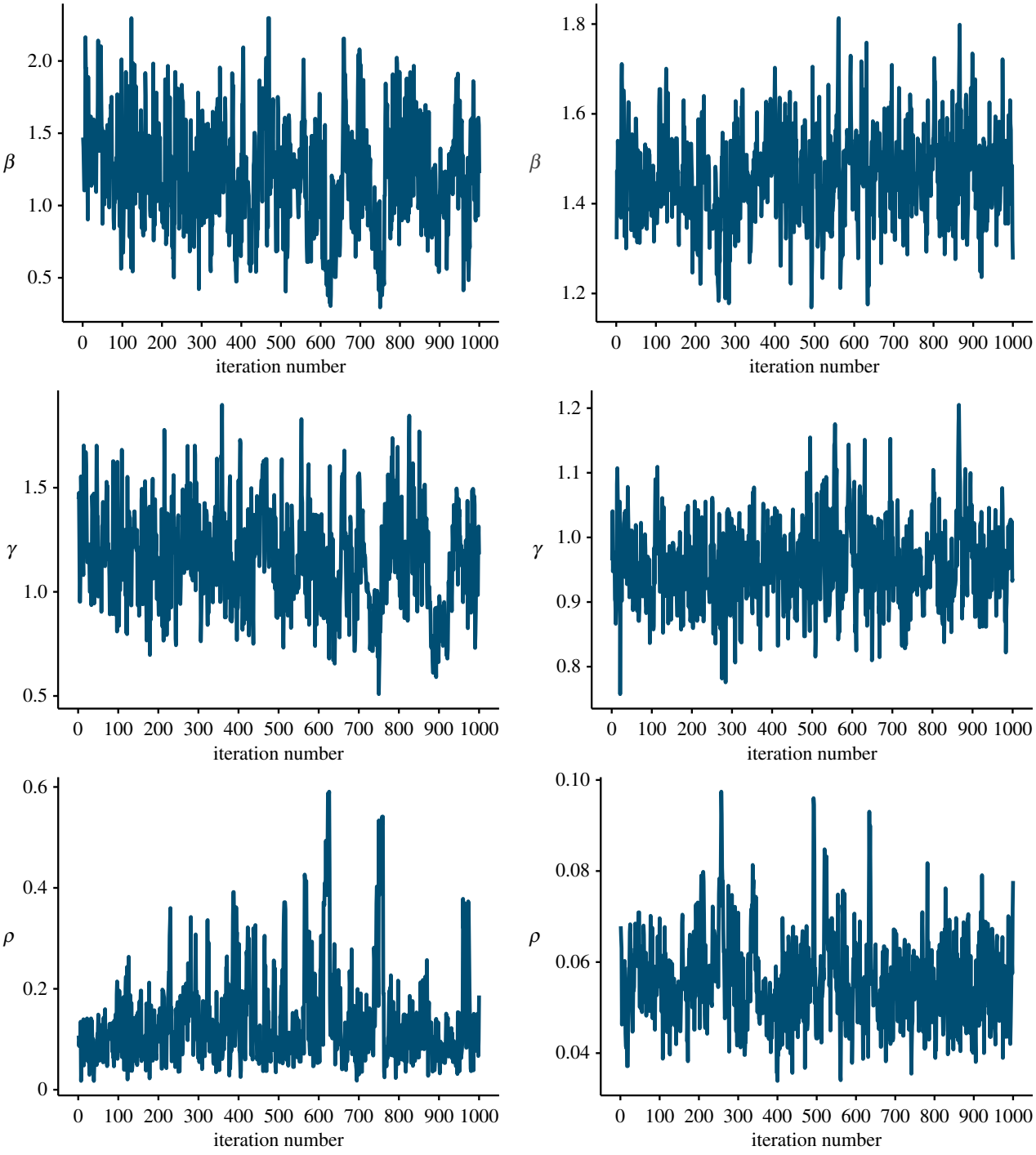

Figure 11. The thinned trace of a single Markov chain in the MCMC implementation of Method 3 for the smaller cut-off time ($T = 3$). Separate panels are shown for each of the parameters β, γ, and ρ. The left-hand panels correspond to $n = 10^2$, and the right-hand panels correspond to $n = 10^3$. The true parameter values are $\beta = 1.5$, $\gamma = 1$ and $\rho = 0.05$ (parameter setting θ_3). (Online version in colour.)

heavily truncated data, the MSE in table 3 are higher than those in table 2. Also, the sharp decrease in MSE with increasing n in table 2 is less pronounced in table 3. Nevertheless, the estimates obtained are still quite accurate. Also, the MSE for Method 3 are slightly better than those of Method 1 or 2. Interestingly, the parameter ρ is accurately estimated by Method 3.

In figures 9 and 10, we show the posterior distributions of the Method 3 estimators of β, γ and β based on the SDS likelihood. To avoid repetition, we show only two posterior plots: figure 9 shows results for the parameter setting θ_1 under the smaller cut-off time, and figure 10 shows results for the parameter setting θ_2 under the larger cut-off time. As shown in tables 3 and 2, the variances of the posterior distributions shrink drastically as we increase n from 10^2 to 10^3. We do not show the posterior distributions for the $n = 10^4$ case because it does not provide any additional insights into the quality of the inference procedure except for the fact that the posterior variance further reduces.

Finally, figure 11 shows additional diagnostic statistics for the MCMC implementation of Method 3. We show the thinned trace of a single Markov chain for $n = 10^2$ and 10^3. As expected, the chain mixes faster when $n = 10^3$ than when $n = 10^2$ because Method 3 is based on an LLN of the scaled Poisson processes keeping track of the population counts. As before, we omit the $n = 10^4$ case. For completeness, we consider the third parameter setting θ_3 in figure 11. The Markov chains also converge for the other parameter settings (not shown).

References

1. Enserink M. 2013 SARS: chronology of the epidemic. *Science* **339**, 1266–1271. (doi:10.1126/science.339.6125.1266)
2. Coltart CE, Lindsey B, Ghinai I, Johnson AM, Heymann DL. 2017 The Ebola outbreak, 2013–2016: old lessons for new epidemics. *Phil. Trans. R. Soc. B* **372**, 20160297. (doi:10.1098/rstb.2016.0297)
3. Meltzer MI, Atkins CY, Santibanez S, Knust B, Petersen BW, Ervin ED, Nichol ST, Damon IK, Washington ML. 2014 Estimating the future number of cases in the Ebola epidemic—Liberia and Sierra Leone, 2014–2015. *MMWR Suppl.* **63**, 1–14.
4. Handel A, Rohani P. 2015 Crossing the scale from within-host infection dynamics to between-host transmission fitness: a discussion of current assumptions and knowledge. *Phil. Trans. R. Soc. B* **370**, 20140302. (doi:10.1098/rstb.2014.0302)
5. Kermack WO, McKendrick AG. 1927 A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A* **115**, 700–721. (doi:10.1098/rspa.1927.0118)
6. Ethier SN, Kurtz TG 1986 *Markov processes: characterization and convergence*. Wiley Series in Probability and Mathematical Statistics. New York, NY: John Wiley & Sons, Inc.
7. Anderson H, Britton T 2000 *Stochastic epidemic models and their statistical analysis*. New York, NY: Springer.
8. Anderson DF, Kurtz TG 2015 *Stochastic analysis of biochemical systems*, vol. 1. Cham, Switzerland: Springer.
9. Banisch S 2016 *Markov chain aggregation for agent-based models*. Cham, Switzerland: Springer International Publishing.
10. KhudaBukhsh WR, Auddy A, Disser Y, Koeppl H. 2019 Approximate lumpability for Markovian agent-based models using local symmetries. *J. Appl. Probab.* **56**, 647–671. (doi:10.1017/jpr.2019.44)
11. Sellke T. 1983 On the asymptotic distribution of the size of a stochastic epidemic. *J. Appl. Probab.* **20**, 390–394. (doi:10.2307/3213811)
12. Kenah E. 2011 Contact intervals, survival analysis of epidemic data, and estimation of R_0. *Biostatistics* **12**, 548–566. (doi:10.1093/biostatistics/kxq068)
13. Kenah E. 2013 Non-parametric survival analysis of infectious disease data. *J. R. Stat. Soc.: Ser. B* **75**, 277–303. (doi:10.1111/j.1467-9868.2012.01042.x)
14. Aalen OO, Borgan Ø, Gjessing HK 2008 *Survival and event history analysis: a process point of view*. New York, NY: Springer Science & Business Media.
15. Fleming TR, Harrington DP 1991 *Counting processes and survival analysis*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York, NY: John Wiley & Sons, Inc.
16. Buchholz P. 1994 Exact and ordinary lumpability in finite Markov chains. *J. Appl. Probab.* **31**, 59–75. (doi:10.2307/3215235)
17. Kemeny JG, Snell JL 1983 *Finite Markov chains*. New York, NY: Springer.
18. Simon PL, Taylor M, Kiss IZ. 2011 Exact epidemic models on graphs using graph-automorphism driven lumping. *J. Math. Biol.* **62**, 479–508. (doi:10.1007/s00285-010-0344-x)
19. Kurtz TG. 1970 Solutions of ordinary differential equations as limits of pure jump Markov processes. *J. Appl. Probab.* **7**, 49–58. (doi:10.2307/3212147)
20. Kurtz TG. 1978 Strong approximation theorems for density dependent Markov chains. *Stocha. Process. Appl.* **6**, 223–240. (doi:10.1016/0304-4149(78)90020-0)
21. Dembo A, Zeitouni O 2010 *Large deviations techniques and applications*. Berlin, Germany: Springer.
22. Dupuis P, Ellis R 2011 *A weak convergence approach to the theory of large deviations*. Hoboken, NJ: John Wiley & Sons.
23. Feng J, Kurtz TG 2006 *Large deviations for stochastic processes*, vol. 131. Providence, RI: American Mathematical Society.
24. Djehiche B, Schied A. 1998 Large deviations for hierarchical systems of interacting jump processes. *J. Theor. Probab.* **11**, 1–24. (doi:10.1023/A:1021690707556)
25. Dolgoarshinnykh R. 2009 Sample path large deviations for SIRS epidemic processes. Preprint.
26. Pardoux E, Samegni-Kepgnou B. 2017 Large deviation principle for epidemic models. *J. Appl. Probab.* **54**, 905–920. (doi:10.1017/jpr.2017.41)
27. Baladron J, Fasoli D, Faugeras O, Touboul J. 2012 Mean-field description and propagation of chaos in networks of Hodgkin–Huxley and FitzHugh–Nagumo neurons. *J. Math. Neurosci.* **2**, 10. (doi:10.1186/2190-8567-2-10)
28. McDonald D. 2007 *Lecture notes on mean field convergence*. Toronto, Canada: Department of Mathematics, University of Toronto.
29. Méléard S. 1996 Asymptotic behaviour of some interacting particle systems; McKean-Vlasov and Boltzmann models. In *Probabilistic models for nonlinear partial differential equations*, pp. 42–95. Berlin, Germany: Springer.
30. Hao B-L. 1989 *Elementary symbolic dynamics and chaos in dissipative systems*. Singapore: World Scientific.
31. Kakutani S. 1972 Strictly ergodic symbolic dynamical systems. In *Proc. 6th Berkeley Symp. on Mathematical Statistics and Probability* (eds LM LeCam, J Neyman, EL Scott), pp. 319–326. Berkeley, CA: University of California Press.
32. Lind D, Marcus B 1995 *An introduction to symbolic dynamics and coding*. Cambridge, UK: Cambridge University Press.
33. Choi B, Rempala GA. 2012 Inference for discretely observed stochastic kinetic networks with applications to epidemic modeling. *Biostatistics* **13**, 153–165. (doi:10.1093/biostatistics/kxr019)
34. Smith AF, Roberts GO. 1993 Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B* **55**, 3–23. (doi:10.1111/j.2517-6161.1993.tb01466.x)
35. McRae C 2014 *Bayesian inference in nonlinear differential equation models*. Melbourne, Australia: Australian Mathematical Sciences Institute.
36. Vihola M. 2012 Robust adaptive Metropolis algorithm with coerced acceptance rate. *Stat. Comput.* **22**, 997–1008. (doi:10.1007/s11222-011-9269-5)
37. Claeskens G, Hjort NL. 2003 The focused information criterion. *J. Am. Stat. Assoc.* **98**, 900–916. (doi:10.1198/016214503000000819)
38. Schwartz EJ, Choi B, Rempala GA. 2015 Estimating epidemic parameters: application to H1N1 pandemic data. *Math. Biosci.* **270**, 198–203. (doi:10.1016/j.mbs.2015.03.007)
39. Schwartz E, Morgan M, Lapin S. 2015 Pandemic 2009 H1N1 influenza in two settings in a small community: the workplace and the university campus. *Epidemiol. Infect.* **143**, 1606–1609. (doi:10.1017/S0950268814002684)
40. Marchbanks TL. 2011 An outbreak of 2009 pandemic influenza A (H1N1) virus infection in an elementary school in Pennsylvania. *Clin. Infect. Dis.* **52**, S154–S160. (doi:10.1093/cid/ciq058)
41. Vaidya N, Morgan M, Jones T, Miller L, Lapin S, Schwartz E. 2015 Modelling the epidemic spread of an H1N1 influenza outbreak in a rural university town. *Epidemiol. Infect.* **143**, 1610–1620. (doi:10.1017/S0950268814002568)
42. Hoffmann P, Terplan K 2005 *Intelligence support systems: technologies for lawful intercepts*. Boca Raton, FL: CRC Press.
43. KhudaBukhsh WR, Choi B, Kenah E, Rempała GA. SDS_epidemic: code implementation in R language. Available from: https://github.com/cbskust/SDS.Epidemic.
44. Althaus CL. 2014 Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. *PLoS Curr.* **6**. (doi:10.1371/currents.outbreaks.91afb5e0f279e7f29e7056095255b288)
45. Getz WM, Dougherty ER. 2018 Discrete stochastic analogs of Erlang epidemic models. *J. Biol. Dyn.* **12**, 16–38. (doi:10.1080/17513758.2017.1401677)
46. Pardoux E, Samegni-Kepgnou B. 2016 Large deviation principle for poisson driven SDEs in epidemic models. (http://arxiv.org/abs/quant-ph/1606.01619)
47. Kenah E, Robins JM. 2007 Second look at the spread of epidemics on networks. *Phys. Rev. E* **76**, 036113. (doi:10.1103/PhysRevE.76.036113)
48. Newman MEJ. 2002 Spread of epidemic disease on networks. *Phys. Rev. E* **66**, 016128. (doi:10.1103/PhysRevE.66.016128)
49. Volz E. 2008 SIR dynamics in random networks with heterogeneous connectivity. *J. Math. Biol.* **56**, 293–310. (doi:10.1007/s00285-007-0116-4)
50. Burch MG, Jacobsen KA, Tien JH, Rempała GA. 2017 Network-based analysis of a small Ebola outbreak.

Math. Biosci. Eng. **14**, 67–77. (doi:10.3934/mbe.2017005)

51. Jacobsen KA, Burch MG, Tien JH, Rempała GA. 2018 The large graph limit of a stochastic epidemic model on a dynamic multilayer network. *J. Biol. Dyn.* **12**, 746–788. (doi:10.1080/17513758.2018.1515993)
52. KhudaBukhsh WR, Woroszylo C, Rempała GA, Koeppl H. 2017 Functional central limit theorem for susceptible-infected process on configuration model graphs. (http://arxiv.org/abs/quant-ph/1703.06328)
53. Rubino G, Sericola B. 1989 On weak lumpability in Markov chains. *J. Appl. Probab.* **26**, 446–457. (doi:10.2307/3214403)