Mathematical Biosciences 375 (2024) 109265

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Original Research Article

How to *correctly* fit an SIR model to data from an SEIR model?

Wasiur R. KhudaBukhsh [a], Grzegorz A. Rempała [b,*]

[a] School of Mathematical Sciences, The University of Nottingham, University Park, Nottingham, NG7 2RD, Nottinghamshire, United Kingdom
[b] Division of Biostatistics, College of Public Health, The Ohio State University, 1841 Neil Avenue, Cunz Hall, Columbus, 43210, OH, United States of America

ARTICLE INFO

MSC:
37N25
60J28
92B05

Keywords:
SIR model
SEIR model
Dynamical survival analysis

ABSTRACT

In epidemiology, realistic disease dynamics often require Susceptible-Exposed-Infected-Recovered (SEIR)-like models because they account for incubation periods before individuals become infectious. However, for the sake of analytical tractability, simpler Susceptible-Infected-Recovered (SIR) models are commonly used, despite their lack of biological realism. Bridging these models is crucial for accurately estimating parameters and fitting models to observed data, particularly in population-level studies of infectious diseases.

This paper investigates stochastic versions of the SEIR and SIR frameworks and demonstrates that the SEIR model can be effectively approximated by a SIR model with time-dependent infection and recovery rates. The validity of this approximation is supported by the derivation of a large-population Functional Law of Large Numbers (FLLN) limit and a finite-population concentration inequality.

To apply this approximation in practice, the paper introduces a parameter inference methodology based on the Dynamic Survival Analysis (DSA) survival analysis framework. This method enables the fitting of the SIR model to data simulated from the more complex SEIR dynamics, as illustrated through simulated experiments.

1. Introduction

One of the pioneering works in communicable disease modeling is the seminal paper by Kermack and McKendrick [1]. This paper introduced a foundational epidemiological model that segments the population into three compartments: susceptible (S), infected (I), and recovered or removed (R). This model is widely known as the Susceptible-Infected-Recovered (SIR) model.

Famously, as a special case of their general framework, Kermack and McKendrick proposed a simple system of Ordinary Differential Equations (ODEs) (see (2.1) below) to describe the time evolution of population proportions in each compartment of the SIR model. However, their classical SIR model is not entirely realistic because it assumes instantaneous infectiousness upon contact, while most infectious or transmittable diseases have an incubation period.

The incubation period can be incorporated into the SIR framework by adding an additional compartment, resulting in the Susceptible-Exposed-Infected-Recovered (SEIR) system, which accounts for exposed or incubating individuals, as discussed in the next section. Although the SIR model is mathematically simpler and sometimes more convenient to analyze, it lacks epidemiological and biological realism. Conversely, realistic data from infectious disease studies often derive from SEIR-like models, which present more modeling challenges. Therefore, it is crucial to understand how to translate between these frameworks to avoid biased estimates of relevant parameters while also avoiding unnecessary computational and conceptual overhead.

The purpose of the current paper is to demonstrate that, in large populations, a biologically more realistic SEIR model can be approximated by a mathematically more convenient SIR system with time-varying infection and recovery rates. To formally introduce and justify this approximation and quantify the approximation error, we consider a stochastic Markovian setting and provide both a large-population Functional Law of Large Numbers (FLLN) limit and a finite-population concentration inequality. Additionally, we present a parameter inference methodology based on a dynamical survival model to fit the approximating SIR system to synthetic data generated from a SEIR framework.

The rest of the paper is structured as follows: In Section 2 we briefly recall the classical ODE SIR and SEIR frameworks based on differential equations. In Section 3, we describe the Continuous Time Markov Chain (CTMC)-based stochastic SEIR model and its large population description in terms of a system of ODEs from Section 2. In Section 4, we describe the approximating stochastic SIR model and furnish necessary convergence results. To illustrate our results' applicability to data analysis, we describe the so-called DSA-based parameter inference methods and given some numerical examples in Section 5 before the concluding remarks in Section 6. Additional mathematical derivations and the list of acronyms are provided in Appendices A and B.

* Corresponding author.
E-mail addresses: wasiur.khudabukhsh@nottingham.ac.uk (W.R. KhudaBukhsh), rempala.3@osu.edu (G.A. Rempała).
URLs: https://www.wasiur.xyz/ (W.R. KhudaBukhsh), https://neyman.mbi.ohio-state.edu/ (G.A. Rempała).

https://doi.org/10.1016/j.mbs.2024.109265

Available online 30 July 2024

2. Deterministic SIR and SEIR models

Consider the classical ODE SIR model of Kermack and McKendrick, where the proportions of individuals x_S, x_I, and x_R in susceptible, infected, and removed compartments satisfy the following system of ODEs:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}x_S &= -\beta x_S x_I,\\ \frac{\mathrm{d}}{\mathrm{d}t}x_I &= \beta x_S x_I - \gamma x_I,\\ \frac{\mathrm{d}}{\mathrm{d}t}x_R &= \gamma x_I,\end{aligned} \tag{2.1}$$

with $x_S(0) = 1, x_I(0) = \rho \in (0,1), x_R(0) = 0$, where $\beta > 0$ is the infection rate, and $\gamma \geq 0$ is the recovery rate. To obtain the corresponding deterministic SEIR system from (2.1) we would simply replace the middle equation by a pair of equations

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}x_S &= -\beta x_S x_I,\\ \frac{\mathrm{d}}{\mathrm{d}t}x_E &= \beta x_S x_I - \alpha x_E,\\ \frac{\mathrm{d}}{\mathrm{d}t}x_I &= \alpha x_E - \gamma x_I,\\ \frac{\mathrm{d}}{\mathrm{d}t}x_R &= \gamma x_I,\end{aligned} \tag{2.2}$$

where now x_E represents a proportion of individuals in exposed compartment, with additional initial condition $x_E(0) = 0$.

Upon elementary manipulations (shown in Appendix A for completeness), the system of ODEs in (2.1) reduces to the single ODE

$$-\frac{\mathrm{d}}{\mathrm{d}t}x_S = \beta x_S(1 - x_S) + \gamma x_S \log(x_S) + \rho\beta x_S, \text{ with } x_S(0) = 1. \tag{2.3}$$

Note that similar reduction is not possible for the system (2.2).

As is well-known in the theory of Markov jump processes, one may view the systems of ODEs in (2.1) (2.2) as the FLLN limits of the proportions of individuals in a corresponding stochastic compartmental Markovian SIR or SEIR model where the counts of susceptible, exposed, infected, and removed individuals are assumed to form a CTMC. Kurtz [2], Kurtz [3], Darling and Norris [4] provide the probabilistic justification for doing so. See also, Andersson and Britton [5, Chapter 5].

It is remarkable that an equation similar to (2.3) may be also obtained as a limit of a stochastic epidemic process evolving on a random graph with the same infection rate β, recovery rate γ and initial proportion of infected individuals ρ. Indeed, let us consider a stochastic SIR epidemic process on a Configuration Model (CM) random graph with a Poisson(μ) degree distribution, *i.e.*, the degrees are drawn from a Poisson(μ) distribution. Variants of this model have been studied in Volkening et al. [6], Decreusefond et al. [7], Janson et al. [8], Khudabukhsh et al. [9], Ball et al. [10]. The monograph by van der Hofstad [11, Part III, Chapter 7] is a great resource to learn about CM random graphs. In the limit of a large graph, the proportion of susceptible vertices $\tilde{x}_S$ in the graph can be described as the solution to the following ODE [12,13]:

$$-\frac{\mathrm{d}}{\mathrm{d}t}\tilde{x}_S = \tilde{\beta}\tilde{x}_S(1 - \tilde{x}_S) + \tilde{\gamma}\tilde{x}_S \log(\tilde{x}_S) + \rho\tilde{\beta}\tilde{x}_S, \text{ with } \tilde{x}_S(0) = 1.0, \tag{2.4}$$

where $\tilde{\beta} = \mu\beta$, and $\tilde{\gamma} = \beta + \gamma$. Notice the similarity between (2.3) and (2.4).

By a novel application of the Sellke construction [14], [5, Chapter 2], the limiting proportions of susceptible individuals x_S, and $\tilde{x}_S$ can be interpreted as the survival function for the time to infection of an initially susceptible individual chosen randomly from among a large population. That is,

$$\mathsf{P}\left(T_I > t\right) = x_S(t),$$

where the random variable T_I denotes the time to infection of a randomly chosen initially susceptible individual in the mass-action SIR model in (2.1). This forms the basis for the so called DSA approach [15–18]. Comparing the survival function $\tilde{x}_S$ in (2.4) with the survival function x_S in (2.3) reveals that, as far as the probability laws of times of infection (and subsequent times of recovery after an exponentially distributed infectious period) are concerned, an SIR model on an infinitely large CM random graph with infection rate β, and recovery rate γ is *equivalent* to a mass-action SIR model with infection rate $\tilde{\beta}$, and recovery rate $\tilde{\gamma}$. For further discussion, we refer the reader to [19,20].

The above observation about the equivalence of two seemingly different approaches prompts us to consider which other frameworks can be shown to be in some way equivalent to the simple SIR system. An obvious candidate is the SEIR model.

3. The stochastic SEIR model

The standard mass-action stochastic SEIR model keeps track of the counts of individuals with different immunological statuses. Let us define the stochastic process $X(t) := (S(t), E(t), I(t), R(t))$ where $S(t), E(t), I(t)$, and $R(t)$ are counts of susceptible, exposed, infected, and removed individuals respectively. We assume initially there are n susceptible, and m infected individuals. Under the stochastic law of mass-action, each infected individual contacts other individuals in the population following a Poisson clock with rate β. If the contacted individual is susceptible, they get infected and move into the E compartment in which they are infected but not infectious. Individuals spend an Exponential(α) amount of time in the E compartment before they move into the I compartment at which point they turn infectious. Once infectious, they attempt to infect susceptible individuals before recovering after an infectious period of length distributed according to an Exponential(γ) distribution. We assume $X(t)$ is a CTMC. The jumps of the CTMC $X(t)$ are given by

$$\begin{aligned}&\mathsf{P}\left(X(t+\Delta t) - X(t) = (-1,1,0,0) \mid X(t)\right) \approx \frac{\beta}{n}S(t)I(t)\Delta t,\\ &\mathsf{P}\left(X(t+\Delta t) - X(t) = (0,-1,1,0) \mid X(t)\right) \approx \alpha E(t)\Delta t,\\ &\mathsf{P}\left(X(t+\Delta t) - X(t) = (0,0,-1,1) \mid X(t)\right) \approx \gamma I(t)\Delta t,\end{aligned} \tag{3.5}$$

for small Δt. Then, the process X satisfies the stochastic equations:

$$\begin{aligned}S(t) &= S(0) - Q_1\left(\int_0^t \frac{\beta}{n}S(u)I(u)\mathrm{d}u\right),\\ E(t) &= E(0) + Q_1\left(\int_0^t \frac{\beta}{n}S(u)I(u)\mathrm{d}u\right) - Q_2\left(\int_0^t \alpha E(u)\mathrm{d}u\right),\\ I(t) &= I(0) + Q_2\left(\int_0^t \alpha E(u)\mathrm{d}u\right) - Q_3\left(\int_0^t \gamma I(u)\mathrm{d}u\right),\\ R(t) &= R(0) + Q_3\left(\int_0^t \gamma I(u)\mathrm{d}u\right),\end{aligned} \tag{3.6}$$

where $S(0) = n, E(0) = 0, I(0) = m(= m(n)), R(0) = 0$, and Q_1, Q_2, and Q_3 are independent unit-rate Poisson processes. This follows from the random time change representation of Poisson processes [5,21,22]. See Appendix A for more details about this representation. The trajectories can be simulated using the popular Doob–Gillespie's algorithm [22,23]. See Algorithm 1 for a pseudocode. From an individual perspective, a randomly chosen initially susceptible individual remains susceptible till time t with probability $\exp\left(-\int_0^t n^{-1}\beta I(u)\mathrm{d}u\right)$, given the history of the process.

We assume $m/n \to \rho \in (0,1)$ as $n \to \infty$. Then, following the standard results for Markov process (see Kurtz [2,3],Darling and Norris [4],Andersson and Britton [5],Ethier and Kurtz [21]), we can show the scaled stochastic process $n^{-1}X = (n^{-1}S, n^{-1}E, n^{-1}I, n^{-1}R)$ converges to the solution $x := (s,e,i,r)$ of the already familiar SEIR ODE system (2.2), which in our current notation may be written as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}s_t &= -\beta s_t i_t,\\ \frac{\mathrm{d}}{\mathrm{d}t}e_t &= \beta s_t i_t - \alpha e_t,\\ \frac{\mathrm{d}}{\mathrm{d}t}i_t &= \alpha e_t - \gamma i_t,\\ \frac{\mathrm{d}}{\mathrm{d}t}r_t &= \gamma i_t,\end{aligned} \tag{3.7}$$

Algorithm 1 Pseudocode for the Doob–Gillespie algorithm

Initialize $(S(0), E(0), I(0), R(0)) = (n, 0, m, 0)$
Assume you have the process value $(S(t), E(t), I(t), R(t))$ at t
Calculate rates $\lambda_{S\to I}(t) = \beta S(t)I(t)/n$, $\lambda_{E\to I} = \alpha E(t)$ and $\lambda_{I\to R}(t) = \gamma\, I(t)$
Set $\Lambda(t) = \lambda_{S\to I}(t) + \lambda_{E\to I}(t) + \lambda_{I\to R}(t)$
Set next transition time Δt as $\textsc{Exponential}(\Lambda(t))$
Draw a random sample u of $U \sim \textsc{Uniform}(0, 1)$
if $u \le \lambda_{S\to I}(t)/\Lambda(t)$ **then** Update

$$(S(t+\Delta t), I(t+\Delta t), R(t+\Delta t)) = (S(t) - 1, E(t) + 1, I(t), R(t))$$

else
    if $u < (\lambda_{S\to I}(t) + \lambda_{E\to I}(t))/\Lambda(t)$ **then** Update

$$(S(t+\Delta t), I(t+\Delta t), R(t+\Delta t)) = (S(t), E(t) - 1, I(t) + 1, R(t))$$

    else Update

$$(S(t+\Delta t), I(t+\Delta t), R(t+\Delta t)) = (S(t), E(t), I(t) - 1, R(t) + 1)$$

    end if
end if
Set $t = t + \Delta t$ and return to Step 2

with $s_0 = 1, e_0 = 0, i_0 = \rho$, and $r_0 = 0$. More precisely,

$$\lim_{n\to\infty} \mathsf{P}\left(\sup_{0<t\le T} \|n^{-1}X(t) - x(t)\| > \epsilon\right) = 0,$$

for any $\epsilon > 0$, and $0 < T < \infty$, where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}_+{}^4$. The system of ODEs in (3.7) is often referred to as the mean-field SEIR ODE system. This convergence result establishes the connection between the stochastic and the deterministic models.

The SEIR model described in this section is our reference model. Our goal is to approximate this model by a simpler SIR model, which we describe next.

4. An SIR approximation to the SEIR model

The SIR approximation that we propose is a time-inhomogeneous CTMC model in which the infection and recovery rates are assumed time-varying. To this end, define a new variable $v_t := e_t + i_t$ and note that v_t is the proportion of individuals who are either exposed or infected at time t. Define the new time-varying infection and recovery rates as follows:

$$\beta_t := \beta\frac{i_t}{v_t}, \quad \text{and } \gamma_t := \gamma\frac{i_t}{v_t}. \tag{4.8}$$

Then, of course, it is immediate from (3.7) that v_t satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}v_t = \beta_t s_t v_t - \gamma_t v_t,$$

with initial condition $v_0 = \rho$. Our approximating SIR model is described by the stochastic process $Y(t) := (\tilde S(t), \tilde V(t), \tilde R(t))$ where $\tilde S(t), \tilde V(t)$, and $\tilde R(t)$ are counts of susceptible, infected, and removed individuals respectively. We assume $Y(t)$ is a CTMC satisfying the trajectory equations:

$$\begin{aligned}
\tilde S(t) &= \tilde S(0) - Q_1\left(\int_0^t \frac{\beta_u}{n}\tilde S(u)\tilde V(u)\mathrm{d}u\right),\\
\tilde V(t) &= \tilde V(0) + Q_1\left(\int_0^t \frac{\beta_u}{n}\tilde S(u)\tilde V(u)\mathrm{d}u\right) - Q_3\left(\int_0^t \gamma_u\tilde V(u)\mathrm{d}u\right),\\
\tilde R(t) &= \tilde R(0) + Q_3\left(\int_0^t \gamma_u\tilde V(u)\mathrm{d}u\right),
\end{aligned} \tag{4.9}$$

with $\tilde S(0) = n, \tilde V(0) = m$, and $\tilde R(0) = 0$, where Q_1, and Q_3 are the two independent, unit-rate Poisson processes defined in (3.6). That is, the two models share the processes Q_1, and Q_3. We have chosen the same Poisson processes because it will be useful when studying the approximation error. As before, we assume $m/n \to \rho$ as $n \to \infty$. Trajectories of the stochastic process Y can be simulated adapting the Doob–Gillespie's algorithm or the next-reaction method [24,25]. It is also worth mentioning that the $\tilde S(t) + \tilde I(t) + \tilde R(t) = n + m$ for all $t \ge 0$, *i.e.*, as before, the total mass remains conserved.

4.1. Functional Law of Large Numbers

For $x := (x_1, x_2, x_3) \in \mathbb{R}^3$, let us define $\|x\|_\infty := \max\{|x_1|, |x_2|, |x_3|\}$. As with the SEIR model, we expect a deterministic limit for scaled process

$$n^{-1}Y(t) = (n^{-1}\tilde S(t), n^{-1}\tilde V(t), n^{-1}\tilde R(t))$$

in the approximating SIR model. The following FLLN establishes this limit. Note that due to the conservation laws (closed populations) both systems considered here are non-explosive over any compact time interval.

Theorem 1. *Assume* $\lim_{n\to\infty} n^{-1}Y(0) = y(0) = (1, \rho, 0)$. *Then, for any* $T > 0$,

$$\lim_{n\to\infty}\sup_{0\le t\le T}\left\|n^{-1}Y(t) - y(t)\right\|_\infty = 0 \text{ almost surely},$$

where $y(t) = (\tilde s_t, \tilde v_t, \tilde r_t)$ *is the solution to the system of ODEs:*

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\tilde s_t &= -\beta_t\tilde s_t\tilde v_t,\\
\frac{\mathrm{d}}{\mathrm{d}t}\tilde v_t &= \beta_t\tilde s_t\tilde v_t - \gamma_t\tilde v_t,\\
\frac{\mathrm{d}}{\mathrm{d}t}\tilde r_t &= \gamma_t\tilde v_t,
\end{aligned} \tag{4.10}$$

with $\tilde s_0 = 1, \tilde v_0 = \rho$, *and* $\tilde r_0 = 0$.

Albeit time-inhomogeneity of the rates, the proof of Theorem 1 follows from the standard use of the FLLN for Poisson processes, which states that

$$\lim_{n\to\infty}\sup_{0<t\le T}|n^{-1}Q(nt) - t| = 0 \text{ almost surely},$$

for a unit-rate Poisson process Q, and an application of the Grönwall's inequality. See Andersson and Britton [5, Chapter 5] or Ethier and Kurtz [21, Chapter 11]. For the sake of completeness, we include it here.

Proof of Theorem 1. For $x := (x_1, x_2, x_3) \in \mathbb{R}^3$, and $t \in [0, T]$, let

$$\Psi_t(x) := (-\beta_t x_1 x_2, \beta_t x_1 x_2 - \gamma_t x_2, \gamma_t x_2).$$

Note that $v_0 = \rho > 0$ ensures that $\beta_t \le \beta$ and $\gamma_t \le \gamma$ for all $t \in [0, T]$, since $i_t \le v_t$ for all $t \in [0, T]$. Then, for any compact $K \subset \mathbb{R}^3$, there exists a constant C_K such that

$$\sup_{0\le t\le T}\left\|\Psi_t(x) - \Psi_t(y)\right\|_\infty \le C_K\|x - y\|_\infty, \qquad \text{for all } x, y \in K,$$

since the space $\mathbb{R}^3$ is locally compact and Ψ_t is locally Lipschitz for each $t \ge 0$. This ensures the system of ODEs in (4.10) admits unique solutions for all $t \in [0, T]$. Now, note that

$$\begin{aligned}
n^{-1}Y(t) - y(t) &= n^{-1}Y(0) - y(0) + n^{-1}(-1, 1, 0)Q_1\left(n\int_0^t \beta_u\frac{\tilde S(u)}{n}\frac{\tilde V(u)}{n}\mathrm{d}u\right)\\
&\quad + n^{-1}(0, -1, 1)Q_3\left(n\int_0^t \gamma_u\frac{\tilde V(u)}{n}\mathrm{d}u\right) - \int_0^t \Psi_s(y(s))\mathrm{d}s\\
&= n^{-1}Y(0) - y(0) + n^{-1}(-1, 1, 0)\hat Q_1\left(n\int_0^t \beta_u\frac{\tilde S(u)}{n}\frac{\tilde V(u)}{n}\mathrm{d}u\right)\\
&\quad + n^{-1}(0, -1, 1)\hat Q_3\left(n\int_0^t \gamma_u\frac{\tilde V(u)}{n}\mathrm{d}u\right)\\
&\quad + \int_0^t \left(\Psi_u(n^{-1}Y(u)) - \Psi_u(y(u))\right)\mathrm{d}u,
\end{aligned}$$

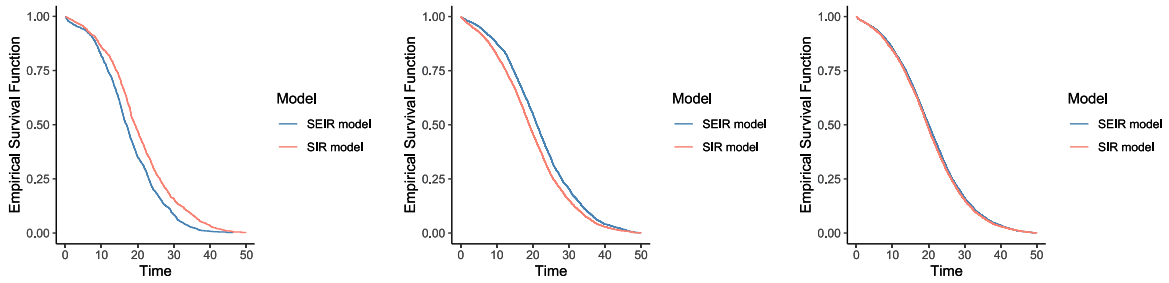

Fig. 1. Comparison of empirical survival functions of times of infection under the original SEIR model from (3.6) and the approximating SIR model from (4.9). The parameter values in this simulation are: $\alpha = 0.25, \beta = 1.5, \gamma = 0.75$, and $\rho = 0.01$. The initial numbers of susceptible, n, are 1000 (left), 5000 (middle), and 10 000 (right).

where $\hat{Q}_1(t) := Q_1(t) - t$, and $\hat{Q}_3(t) := Q_3(t) - t$ are the compensated unit-rate Poisson processes. Here, for a scalar $a \in \mathbb{R}$, and a vector $u := (u_1, u_2, u_3) \in \mathbb{R}^3$, the product au is defined as the vector (au_1, au_2, au_3). Therefore,

$$\begin{aligned}\left\|n^{-1}Y(t) - y(t)\right\|_\infty \le \left\|n^{-1}Y(0) - y(0)\right\|_\infty &+ \sup_{u\le t} n^{-1}|\hat{Q}_1(nK_\beta u)| \\ &+ \sup_{u\le t} n^{-1}|\hat{Q}_3(nK_\gamma u)| \\ &+ \int_0^t C_K \left\|n^{-1}Y(u) - y(u)\right\|_\infty \mathrm{d}u,\end{aligned}$$

where K_β, and K_γ are constants dependent on K (and β, and γ respectively), but independent of n. Therefore, by Grönwall's inequality, we have

$$\left\|n^{-1}Y(t) - y(t)\right\|_\infty \le (A_n + B_n(t))\exp(C_K t), \tag{4.11}$$

where

$$A_n := \left\|n^{-1}Y(0) - y(0)\right\|_\infty,$$
$$B_n(t) := \sup_{u\le t} n^{-1}|\hat{Q}_1(nK_\beta u)| + \sup_{u\le t} n^{-1}|\hat{Q}_3(nK_\gamma u)|.$$

Taking supremum on both sides of the above inequality, we have

$$\sup_{0\le t\le T}\left\|n^{-1}Y(t) - y(t)\right\|_\infty \le (A_n + B_n(T))\exp(C_K T).$$

Now, $\lim A_n = 0$ by our assumption and $\lim B_n(T) = 0$ almost surely as $n \to \infty$ by the FLLN for Poisson processes. Therefore,

$$\lim_{n\to\infty}\sup_{0\le t\le T}\left\|n^{-1}Y(t) - y(t)\right\|_\infty = 0,$$

almost surely, for all $T > 0$. □

In Fig. 1, we show that the empirical survival functions of the times of infection under the original SEIR model from (3.6) and the approximating SIR model from (4.9) are close to each other suggesting the approximation error is small. In Fig. 2, we show that the empirical densities of the times of infection and the recovery times under the two models are close to each other. As Fig. 2 shows, the empirical densities under the two models are close to each other, again suggesting a small approximation error. We expect the distance (in some appropriate sense) between the original SEIR model from (3.6) and the approximating SIR model from (4.9) to vanish in the limit as n tends to infinity. In the next section, we provide a concentration inequality giving an estimate on the error of approximation for finite n and show that the magnitude of error indeed goes to zero as n tends to infinity.

4.2. Approximation error

In practice, especially when n is not large enough, the SIR model may not be a good approximation to the SEIR model. Therefore, it is important to understand the error in the approximation. A concentration inequality is useful in quantifying the approximation error.

Before presenting our concentration inequality, we state two basic lemmas that will be useful in proving the inequality. The first one is a concentration inequality for compensated Poisson processes.

Lemma 1. *Let Q be a unit rate Poisson process. Then, for all $\epsilon > 0$, and $T > 0$, we have*

$$\mathsf{P}\left(\sup_{s\le T}|n^{-1}Q(ns) - s| \ge \epsilon\right) \le 6\exp(CT - \frac{\epsilon}{3}\sqrt{n}), \tag{4.12}$$

where C is a positive constant (independent of n).

Proof. First, note that

$$\mathsf{E}\left[\exp(\theta(Q(t) - t))\right] = \exp(t(\exp(\theta) - 1 - \theta)) \le \exp(Ct\theta^2),$$

for some positive constant C and $|\theta| < 1$. The proof of (4.12) then follows from a simple algebraic manipulation of the Etemadi's inequality [26] for a unit-rate Poisson process:

$$\mathsf{P}\left(\sup_{s\le T}|n^{-1}Q(ns) - s| \ge 3\epsilon\right) \le 3\sup_{s\le T}\mathsf{P}\left(|n^{-1}Q(ns) - s| \ge \epsilon\right).$$

Indeed, for any $T > 0$, and $|\theta| < 1$, we have

$$\begin{aligned}\sup_{s\le T}\mathsf{P}\left(n^{-1}Q(ns) - s \ge \epsilon\right) &\le \sup_{s\le T}\mathsf{P}\left(\exp(\theta(Q(ns) - ns)) \ge \exp(\theta n\epsilon)\right) \\ &\le \sup_{s\le T}\exp(-\theta n\epsilon)\mathsf{E}\left[\exp(\theta(Q(ns) - ns))\right] \\ &\le \exp(nCT\theta^2 - \theta n\epsilon).\end{aligned}$$

Choosing $\theta = \frac{1}{\sqrt{n}}$, gives us

$$\sup_{s\le T}\mathsf{P}\left(n^{-1}Q(ns) - s \ge \epsilon\right) \le \exp(CT - \epsilon\sqrt{n}),$$

Similarly, again using the Markov inequality, we get

$$\sup_{s\le T}\mathsf{P}\left(-(n^{-1}Q(ns) - s) \ge \epsilon\right) \le \exp(CT - \epsilon\sqrt{n})$$

The proof completes by combining the two inequalities above and replacing ϵ with $\epsilon/3$ in Etemadi's inequality. □

Next we show that the solutions to the SEIR ODEs in (3.7) can be retrieved from the solutions to the approximating SIR ODEs in (4.10). To this end, recall that $x := (s, e, i, r)$ satisfies the system of ODEs in (3.7), and define

$$\tilde{x}(t) := (\tilde{s}_t, (1 - \frac{i_t}{v_t})\tilde{v}_t, \frac{i_t}{v_t}\tilde{v}_t, \tilde{r}_t),$$

where i_t, v_t satisfy (3.7), and $(\tilde{s}_t, \tilde{v}_t, \tilde{r}_t)$ satisfy (4.10). Let $F := [0, 1 + \rho]^4 \subset \mathbb{R}^4$ and define $\Phi : F \mapsto \mathbb{R}^4$ as follows

$$\Phi(u) := (-\beta u_1 u_3, \beta u_1 u_3 - \alpha u_2, \alpha u_2 - \gamma u_3, \gamma u_3) \tag{4.13}$$

for $u := (u_1, u_2, u_3, u_4) \in F$. For all $u, v \in F$, we have

$$\|\Phi(u) - \Phi(v)\|_\infty \le K\|u - v\|_\infty,$$

where the constant K could be chosen as $K = 3(1 + \rho)\max\{\alpha, \beta, \rho\}$.

Lemma 2. *Let $x := (s, e, i, r)$ be the solution to the system of ODEs in (3.7), and $\tilde{x}(t) := (\tilde{s}_t, (1 - \frac{i_t}{v_t})\tilde{v}_t, \frac{i_t}{v_t}\tilde{v}_t, \tilde{r}_t)$, where $(\tilde{s}_t, \tilde{v}_t, \tilde{r}_t)$ solves the system of ODEs in (4.10). If $x(0) = \tilde{x}(0) = (1, 0, \rho, 0)$, then $x(t) = \tilde{x}(t)$ for all $t \in [0, T]$.*

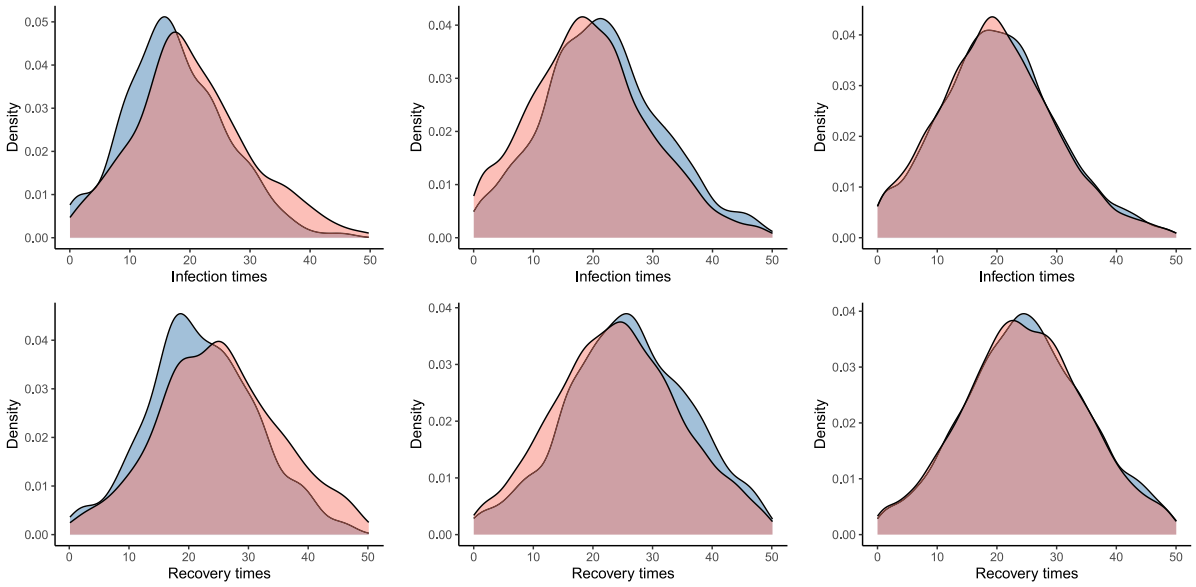

Fig. 2. Comparison of empirical densities of times of infection and recovery times under the original SEIR model from (3.6) and the approximating SIR model from (4.9). The parameter values in this simulation are: $\alpha = 0.25, \beta = 1.5, \gamma = 0.75$, and $\rho = 0.01$. The initial numbers of susceptible, n, are 1000 (left), 5000 (middle), and 10 000 (right).

Proof of Lemma 2. Note that if $x(0) \in F$, and x satisfies the system of ODEs in (3.7), then $x(t) \in F$ for all $t \in [0, T]$ by the conservation law. Now, associate to the system of ODEs in (3.7) we have the following ODE for v_t:

$$\frac{\mathrm{d}}{\mathrm{d}t} v_t = \beta s_t i_t - \gamma i_t, \qquad \text{with } v_0 = \rho.$$

Then, dividing this equation by $\frac{\mathrm{d}}{\mathrm{d}t} s_t$ and then solving partially yields

$$v_t = (1 + \rho - s_t) + \frac{\gamma}{\beta} \log(s_t),$$

which suggests if $v_t = 0$ for some t, $s_u = s_\infty$ for all $u \geq t$, where s_∞ is the unique solution to the equation $s_\infty = 1 + \rho + \frac{\gamma}{\beta} \log(s_\infty)$. Moreover, $s_\infty \in (0, 1)$, and $s_t \to s_\infty$ from above as $t \to \infty$. From the continuity of the functions s_t, i_t, and v_t we can choose a uniform lower bound A, strictly bounded away from zero, such that $s_t, i_t, v_t > A$ in bounded time interval $[0, T]$, provided $i_0 = v_0 = \rho > 0$.

Now, note that $\tilde{x}$ satisfies

$$\tilde{x}(t) = \tilde{x}(0) + \int_0^t \tilde{\Psi}_s(\tilde{x}(s))\mathrm{d}s,$$

with $\tilde{\Psi}_t(\tilde{x}(t)) := (\tilde{\Psi}_t^{(1)}(\tilde{x}(t)), \tilde{\Psi}_t^{(2)}(\tilde{x}(t)), \tilde{\Psi}_t^{(3)}(\tilde{x}(t)), \tilde{\Psi}_t^{(4)}(\tilde{x}(t)))$ where

$$\begin{aligned}
\tilde{\Psi}_t^{(1)}(\tilde{x}(t)) &= -\beta \frac{i_t}{v_t} \tilde{s}_t \tilde{v}_t, \\
\tilde{\Psi}_t^{(2)}(\tilde{x}(t)) &= \beta \frac{i_t}{v_t} \tilde{s}_t \tilde{v}_t - \gamma \frac{i_t}{v_t} \tilde{v}_t - \frac{\tilde{v}_t}{v_t^2} \left(\alpha e_t v_t - \gamma i_t v_t - \beta s_t i_t^2 + \gamma i_t^2 + i_t^2 (\beta \tilde{s}_t - \gamma) \right), \\
&= \beta \frac{i_t}{v_t} \tilde{s}_t \tilde{v}_t - \alpha e_t \frac{\tilde{v}_t}{v_t} + \frac{i_t^2 \tilde{v}_t}{v_t^2} \beta (s_t - \tilde{s}_t), \\
\tilde{\Psi}_t^{(3)}(\tilde{x}(t)) &= \frac{\tilde{v}_t}{v_t^2} \left(\alpha e_t v_t - \gamma i_t v_t - \beta s_t i_t^2 + \gamma i_t^2 + i_t^2 (\beta \tilde{s}_t - \gamma) \right), \\
\tilde{\Psi}_t^{(4)}(\tilde{x}(t)) &= \gamma \frac{i_t}{v_t} \tilde{v}_t.
\end{aligned}$$

Now, note that $\tilde{\Psi}_t$ is Lipschitz continuous on F. Moreover, if $\tilde{x}(0) \in F$, then $\tilde{x}(t) \in F$ for all $t \in [0, T]$. Therefore, by the Picard-Lindelöf theorem, the solutions to $\frac{\mathrm{d}}{\mathrm{d}t} \tilde{x}(t) = \tilde{\Psi}_t(\tilde{x}(t))$ with $\tilde{x}(0) = (1, 0, \rho, 0)$ exist and are unique.

It is easy to verify that $\tilde{\Psi}_t(\tilde{x}(t)) = \Phi(x(t))$ if $\tilde{x}(t) = x(t) \in F$, where Φ is defined in (4.13). Therefore, if $x(t) = \tilde{x}(t)$ for some t, then $x(u) = \tilde{x}(u)$ for all $u \geq t$. Because we have the same initial condition $\tilde{x}(0) = x(0) = (1, 0, \rho, 0)$ with $\rho > 0$, the conclusion follows from the uniqueness of solutions. □

In order to show that the approximating SIR model is a good approximation of the SEIR model, we need to show that the difference between the two models is small. To this end, we define the following process $\Delta_n(t) := (\Delta_S^{(n)}(t), \Delta_E^{(n)}(t), \Delta_I^{(n)}(t), \Delta_R^{(n)}(t))$ where

$$\begin{aligned}
\Delta_S^{(n)}(t) &= n^{-1} \left(S(t) - \tilde{S}(t) \right), \\
\Delta_E^{(n)}(t) &= n^{-1} \left(E(t) - (1 - \frac{i_t}{v_t}) \tilde{V}(t) \right), \\
\Delta_I^{(n)}(t) &= n^{-1} \left(I(t) - \frac{i_t}{v_t} \tilde{V}(t) \right), \\
\Delta_R^{(n)}(t) &= n^{-1} \left(R(t) - \tilde{R}(t) \right),
\end{aligned} \tag{4.14}$$

where $(S(t), E(t), I(t), R(t))$ evolve according to the trajectory equations (in terms of Poisson processes) in (3.6), and $(\tilde{S}(t), \tilde{V}(t), \tilde{R}(t))$, according to (4.9). We show that the process $\Delta_n(t)$ converges to zero in probability as $n \to \infty$, and in fact, satisfies a concentration inequality, which we present next.

Theorem 2. *Assume the initial conditions $S(0) = n, E(0) = 0, I(0) = m = [\rho n], R(0) = 0$, and $\tilde{S}(0) = n, \tilde{V}(0) = m = [\rho n], \tilde{R}(0) = 0$, i.e., $\Delta_n(0) = (0, 0, 0, 0)$ for all n, with $[\rho n]$ denoting the integer part of ρn. Then, for all $\epsilon > 0$, and $T > 0$, we have*

$$\mathrm{P}\left(\sup_{t \in [0,T]} \|\Delta_n(t)\|_\infty \geq \epsilon \right) \leq 2 \times 1_{\frac{[\rho n]}{n} > \frac{\delta}{4}} + 36 \exp\left(CLT - \frac{\delta}{36} \sqrt{n} \right), \tag{4.15}$$

where $\delta := \epsilon \exp(-KT)$ for some constants C, K, and L that depend on the parameters of the SEIR model, and $\{u\} = u - [u]$ denotes the fractional part of u.

Proof of Theorem 2. Note that the map Φ defined in (4.13) is Lipschitz continuous on F with constant K that could be chosen as $K = 3(1 + \rho) \max\{\alpha, \beta, \rho\}$. Recall that $X(t) := (S(t), E(t), I(t), R(t))$ satisfies the stochastic Eq. (3.6), and $x := (s, e, i, r)$ satisfies (3.7). Then, following calculations similar to those in the proof of the FLLN in Theorem 1, we obtain the following inequality for $n^{-1} X(t) - x(t)$:

$$\sup_{t \leq T} \left\| n^{-1} X(t) - x(t) \right\|_\infty \leq (A_n + B_n(T)) e^{KT},$$

where

$$A_n := \left\| n^{-1} X(0) - x(0) \right\|_\infty = \left| \frac{[\rho n]}{n} - \rho \right| = \frac{\{\rho n\}}{n} \to 0,$$

and

$$B_n(T) := \sup_{t\le T}|\hat{Q}_1(nK_\beta t)| + \sup_{t\le T}|\hat{Q}_2(nK_\alpha t)| + \sup_{t\le T}|\hat{Q}_3(nK_\gamma t)|,$$

where $\hat{Q}_1, \hat{Q}_2, \hat{Q}_3$ are the compensated Poisson processes corresponding to Q_1, Q_2, Q_3, respectively in (3.6) and $K_\beta, K_\alpha, K_\gamma$ are the constants dependent on β, α, γ but independent of n. Then, we have

$$\begin{aligned}\mathsf{P}\left(\sup_{t\le T}\left\|n^{-1}X(t)-x(t)\right\|_\infty > \epsilon\right) &\le 1_{\frac{\{\rho n\}}{n}>\frac{\epsilon}{2}\exp(-KT)} \\ &\quad + \mathsf{P}\left(B_n(T) > \frac{\epsilon}{2}\exp(-KT)\right) \\ &\le 1_{\frac{\{\rho n\}}{n}>\frac{\epsilon}{2}\exp(-KT)} \\ &\quad + 3\mathsf{P}\left(\sup_{t\le T}|\hat{Q}_1(nLt)| > \frac{\epsilon}{6}\exp(-KT)\right),\end{aligned}$$

where $L = \max\{K_\beta, K_\alpha, K_\gamma\}$. By Lemma 1, we have

$$\mathsf{P}\left(\sup_{t\le T}|\hat{Q}_1(nLt)| > \frac{\epsilon}{6}\exp(-KT)\right) \le 6\exp\left(CLT - \frac{\epsilon}{18}\exp(-KT)\sqrt{n}\right),$$

Therefore, we have an exponential estimate:

$$\mathsf{P}\left(\sup_{t\le T}\left\|n^{-1}X(t)-x(t)\right\|_\infty > \epsilon\right) \le 1_{\frac{\{\rho n\}}{n}>\frac{\delta}{2}} + 18\exp\left(CLT - \frac{\delta}{18}\sqrt{n}\right), \tag{4.16}$$

with $\delta := \epsilon\exp(-KT)$.

Now, let us consider the stochastic process

$$\tilde{X}(t) := (\tilde{S}(t), (1-\frac{i_t}{v_t})\tilde{V}(t), \frac{i_t}{v_t}\tilde{V}(t), \tilde{R}(t))$$

where $(\tilde{S}(t), \tilde{V}(t), \tilde{R}(t))$ evolves according to the trajectory equations in (4.9). From the FLLN for $n^{-1}Y$, it is easy to see that $n^{-1}\tilde{X}(t)$ converges to $\tilde{x}(t) := (\tilde{s}_t, (1-\frac{i_t}{v_t})\tilde{v}_t, \frac{i_t}{v_t}\tilde{v}_t, \tilde{r}_t)$ as $n\to\infty$. We derive a concentration inequality for the process $n^{-1}\tilde{X}$. Given continuous functions $x = (s, e, i, r)$ satisfying the SEIR system of ODEs in (3.7), define the operator Γ_x as follows:

$$\Gamma_x y(t) = (y_1(t), (1-\frac{i_t}{v_t})y_2(t), \frac{i_t}{v_t}y_2(t), y_3(t)),$$

for $y(t) = (y_1(t), y_2(t), y_3(t))$. Then, it is easy to see

$$\tilde{X} = \Gamma_x Y, \text{ and } \tilde{x} = \Gamma_x y$$

where Y evolves according to the trajectory equations in (4.9), $x = (s, e, i, r)$ satisfies (3.7) with initial condition $s_0 = 1, e_0 = 0, i_0 = \rho, r_0 = 0$, and $y = (\tilde{s}, \tilde{v}, \tilde{r})$ satisfies (4.10) with the initial condition $\tilde{s}_0 = 1, \tilde{v}_0 = \rho, \tilde{r}_0 = 0$. Moreover, we have

$$\left\|\Gamma_x y(t) - \Gamma_x z(t)\right\|_\infty \le \|y(t) - z(t)\|_\infty,$$

at each $t \ge 0$ for two functions $y(t) := (y_1(t), y_2(t), y_3(t))$, and $z(t) := (z_1(t), z_2(t), z_3(t))$.

Therefore, we have

$$\begin{aligned}\mathsf{P}\left(\sup_{t\le T}\left\|n^{-1}\tilde{X}(t)-\tilde{x}(t)\right\|_\infty > \epsilon\right) &= \mathsf{P}\left(\sup_{t\le T}\left\|\Gamma_x n^{-1}Y(t) - \Gamma_x y(t)\right\|_\infty > \epsilon\right) \\ &\le \mathsf{P}\left(\sup_{t\le T}\left\|n^{-1}Y(t)-y(t)\right\|_\infty > \epsilon\right) \\ &\le \mathsf{P}\left((A_n + B_n(T)) > \epsilon\exp(-KT)\right) \\ &= \mathsf{P}\left((A_n + B_n(T)) > \delta\right)\end{aligned}$$

where

$$A_n = \left\|n^{-1}Y(0) - y(0)\right\|_\infty = \frac{\{\rho n\}}{n},$$
$$B_n(T) = \sup_{t\le T} n^{-1}|\hat{Q}_1(nK_\beta T)| + \sup_{t\le T} n^{-1}|\hat{Q}_1(nK_\gamma T)|,$$

from (4.11). Here, we have replaced the constant C_K in (4.11) by K, which is justified by restricting Ψ_t to $[0, 1+\rho]^3$. Therefore, we have

$$\begin{aligned}\mathsf{P}\left(\sup_{t\le T}\left\|n^{-1}\tilde{X}(t)-\tilde{x}(t)\right\|_\infty > \epsilon\right) &< 1_{\frac{\{\rho n\}}{n}>\frac{\delta}{2}} + 2\mathsf{P}\left(\sup_{t\le T} n^{-1}|\hat{Q}(nLt)| > \frac{\delta}{4}\right) \\ &< 1_{\frac{\{\rho n\}}{n}>\frac{\delta}{2}} + 12\exp\left(CLT - \frac{\delta}{12}\sqrt{n}\right).\end{aligned} \tag{4.17}$$

Finally, we note that

$$\Delta_n(t) = n^{-1}X(t) - n^{-1}\tilde{X}(t) = (n^{-1}X(t) - x(t)) - (n^{-1}\tilde{X}(t) - \tilde{x}(t)),$$

because $x(t) = \tilde{x}(t)$ for all $t \in [0, T]$ according to Lemma 2. Therefore, we have

$$\|\Delta_n(t)\|_\infty \le \left\|n^{-1}X(t) - x(t)\right\|_\infty + \left\|n^{-1}\tilde{X}(t) - \tilde{x}(t)\right\|_\infty,$$

whence it follows that

$$\begin{aligned}\mathsf{P}\left(\sup_{t\le T}\|\Delta_n(t)\|_\infty > \epsilon\right) &\le \mathsf{P}\left(\sup_{t\le T}\left\|n^{-1}X(t)-x(t)\right\|_\infty > \frac{\epsilon}{2}\right) \\ &\quad + \mathsf{P}\left(\sup_{t\le T}\left\|n^{-1}\tilde{X}(t)-\tilde{x}(t)\right\|_\infty > \frac{\epsilon}{2}\right) \\ &\le 2\times 1_{\frac{\{\rho n\}}{n}>\frac{\delta}{4}} + 18\exp(CLT - \frac{\delta}{36}\sqrt{n}) \\ &\quad + 12\exp(CLT - \frac{\delta}{24}\sqrt{n}) \\ &\le 2\times 1_{\frac{\{\rho n\}}{n}>\frac{\delta}{4}} + 36\exp\left(CLT - \frac{\delta}{36}\sqrt{n}\right).\end{aligned}$$

This completes the proof. □

Corollary 1. *The stochastic process $\Delta_n(t)$ converges to zero in probability as $n\to\infty$. That is, for any $\epsilon > 0$,*

$$\lim_{n\to\infty}\mathsf{P}\left(\sup_{t\in[0,T]}\|\Delta_n(t)\|_\infty \ge \epsilon\right) = 0.$$

Proof. Follows directly from Theorem 2 by taking the limit of $n\to\infty$ in (4.15). □

Remark 1. The implication of Corollary 1 is that the distance between the original SEIR model $n^{-1}X(t) := (S(t), E(t), I(t), R(t))$ satisfying the stochastic Eqs. (3.6), and $\tilde{X}(t) := (\tilde{S}(t), (1-\frac{i_t}{v_t})\tilde{V}(t), \frac{i_t}{v_t}\tilde{V}(t), \tilde{R}(t))$ where $(\tilde{S}(t), \tilde{V}(t), \tilde{R}(t))$ evolve according to the trajectory equations in (4.9), vanishes in the limit as $n\to\infty$. We could have, of course, established such as a convergence result as a corollary to Theorem 1 and the standard FLLN for the mass-action SEIR model. However, the purpose of Theorem 2 is to give an explicit estimate of the error in the approximation.

5. Parameter inference

Having shown that the time-varying SIR model is a good approximation to the SEIR model, we now turn to the problem of parameter inference and model fitting. For this purpose, we consider the DSA approach [12,15–17], which is a survival analysis based approach designed for dynamical systems. In this approach, one interprets the limiting mean-field (FLLN) equations as probabilistic quantities, such as survival functions and densities corresponding to certain time-to-event random variables, as opposed to proportions or concentrations. For instance, in the context of SIR-type models of infectious disease epidemiology, the function s_t denoting the limiting proportion of susceptible individuals is interpreted as a survival function describing the time to infection of an initially susceptible individual. This change in perspective in the DSA approach has crucial several advantages. First, one does not need the size of the total population. In fact, DSA is able to estimate what is called an effective population size. Second, DSA does not require full epidemic trajectories. Instead, it only requires a random sample of times of infection, and times of recovery, if available. Third, the method is quite flexible and easily applicable to a wide range of models and data availability scenarios. This includes non-Markovian models under mass action; for example, see Di Lauro et al. [16]. Network-based models, such as those explored by KhudaBukhsh

et al. [12] and Kiss et al. [27], are also considered. In terms of data availability scenarios, these models have been applied to single snapshot data [28], spatio-temporal data [29], testing and repeated testing data [30,31], and wastewater surveillance data [32].

5.1. DSA-likelihood based on the approximate SIR model

Let $\theta := (\alpha, \beta, \gamma, \rho)$ denote the vector of unknown parameters. We begin by noting that the system of ODEs in (4.10) can be reduced to

$$-\frac{\mathrm{d}}{\mathrm{d}t}\tilde{s}_t = \beta_t \tilde{s}_t(1-\tilde{s}_t) + \gamma_t \tilde{s}_t \log(\tilde{s}_t) + \rho\beta_t \tilde{s}_t, \text{ with } \tilde{s}_0 = 1.$$

The derivation can be done almost exactly the same way as we have shown in Appendix A for the system of ODEs in (2.1). Then, following the DSA approach, we interpret the function $\tilde{s}_t$ as the improper survival function corresponding to the random variable T_I^{SIR} denoting the time of infection of an initially susceptible individual in an infinitely large population. That is,

$$\mathsf{P}\left(T_I^{SIR} > t\right) = \tilde{s}_t.$$

Note that the random variable T_I^{SIR} does not have finite mean. Indeed, the expectation

$$\mathsf{E}\left[T_I^{SIR}\right] = \int_0^\infty \mathsf{P}\left(T_I^{SIR} > t\right)\mathrm{d}t = \int_0^\infty \tilde{s}_t \mathrm{d}t$$

diverges because $\tilde{s}_\infty := \mathsf{P}\left(T_I^{SIR} = \infty\right) = \lim_{t\to\infty}\tilde{s}_t > 0$. However, if we observe an epidemic till a final observation time $T > 0$, we can condition on the event $\{T_I^{SIR} \le T\}$ to get the conditional Probability Density Function (PDF) of T_I^{SIR} as

$$\tilde{f}_T(t) = -\left(\frac{1}{1-\tilde{s}_T}\right)\frac{\mathrm{d}}{\mathrm{d}t}\tilde{s}_t = \frac{\beta_t \tilde{s}_t \tilde{v}_t}{1-\tilde{s}_T}.$$

Given a random sample $t_1, t_2, \ldots, t_{l_1}$ of infection times T_I^{SIR} under the approximating SIR model, their contribution to the likelihood function is

$$\tilde{\ell}_I(\theta \mid t_1, t_2, \ldots, t_{l_1}) = \prod_{j=1}^{l_1} \tilde{f}_T(t_j). \tag{5.18}$$

In general, parameter inference can be conducted using only a random sample of infection times. However, when samples of other transition times (e.g., recovery times) are available, they should be utilized to enhance the quality of inference. Therefore, let us assume we also have access to a random sample $r_1, r_2, \ldots, r_{l_2}$ of recovery times T_R^{SIR}. The PDF of T_R^{SIR} is given by

$$\tilde{h}_T(t) = \frac{\int_0^t \tilde{f}_T(u) r_u(t-u)\mathrm{d}u}{\int_0^T \int_0^t \tilde{f}_T(u) r_u(t-u)\mathrm{d}u\mathrm{d}t},$$

where

$$r_u(s) := \gamma_{u+s} \exp\left(-\int_u^{u+s} \gamma_v \mathrm{d}v\right),$$

for $u, s \ge 0$. The PDF $\tilde{h}_T$ is a convolution of two PDFs because the distribution of the infectious period of an individual who got infected at time u is described by the hazard function $\gamma_{u+(\cdot)}$ and the infectious periods are independent of the time of infection. See [15, Section 2], or [16, Section 3] for a detailed derivation of such densities under the DSA approach. Therefore, the likelihood contribution of the random sample $r_1, r_2, \ldots, r_{l_2}$ of the random variable T_R^{SIR} describing the times of recovery is

$$\tilde{\ell}_R(\theta \mid r_1, r_2, \ldots, r_{l_2}) = \prod_{j=1}^{l_2} \tilde{h}_T(r_j).$$

Finally, the DSA likelihood function based on a random sample $t_1, t_2, \ldots, t_{l_1}$ of infection times T_I^{SIR} and a random sample $r_1, r_2, \ldots, r_{l_2}$ of the recovery times T_R^{SIR} is given by

$$\begin{aligned}\ell_{SIR}(\theta \mid t_1, t_2, \ldots, t_{l_1}, r_1, r_2, \ldots, r_{l_2}) := \tilde{\ell}_I(\theta \mid t_1, t_2, \ldots, t_{l_1}) \\ \times \tilde{\ell}_R(\theta \mid r_1, r_2, \ldots, r_{l_2}).\end{aligned} \tag{5.19}$$

The likelihood function ℓ_{SIR} in (5.19) can be used for parameter inference in various statistical ways. For instance, one could maximize it to obtain the Maximum Likelihood Estimates (MLEs) of the unknown parameters θ. Obtaining closed-form expressions for the maximizers of the likelihood function ℓ_{SIR} does not seem feasible. However, numerical methods can be employed. In this paper, we will follow a Bayesian approach instead.

We note a few practical points here: (1) It is often much easier to work with the log-likelihood function, i.e., the logarithm of the likelihood in (5.19). This is precisely what we do in our implementation. (2) The densities $\tilde{h}_T(.)$ are computationally expensive to compute. Calculating the derivatives with respect to t and then solving them as ODEs appears to be a much less expensive approach. This method is also employed in our implementation. (3) We do *not* assume that the random samples of infection and recovery times come from the same individuals. Since l_1 need not equal l_2, this point is perhaps clear. However, this ability to consider cross-sectional data is, in our opinion, one of the crucial practical advantages of the DSA approach.

In order to illustrate the method, we first simulate a single trajectory of X satisfying (3.6) using the Doob–Gillespie's algorithm 1. From that trajectory, we take random samples of infection times, and recovery times. We follow a Bayesian approach to parameter inference. First, we apply the DSA likelihood to estimate the parameters using the random sample of only infection times, *i.e.*, using the likelihood function $\tilde{\ell}_I$. We do this by drawing posterior samples of the parameters using the Hamiltonian Monte Carlo (HMC) method under uninformative flat priors. The estimated posterior distributions are shown in Fig. 3. The posterior distributions are unimodal, and the method is able to recover the true parameters with remarkable accuracy. Next, we perform the DSA method using samples of both infection and recovery times using the full DSA likelihood in (5.19). Again, we follow the same approach here: We draw posterior samples of the parameters using the HMC method under uninformative flat priors. The estimated posterior distributions are shown in Fig. 4. As we can see in Fig. 4, the posterior distributions are unimodal. The point estimates (means of the posterior distributions) are fairly accurate.

Comparing Fig. 3 with Fig. 4, we see that the DSA method is able to recover the true parameters fairly accurately in both cases as the means of the posterior distributions (shown as red triangles in plots) are close to the true parameter values. However, using the recovery times appears to improve the quality of the inference in that the posterior samples are more concentrated around the true parameter values. In both Figs. 3 and 4, we followed a HMC-based Bayesian approach implemented in `R` [33] using the package `CmdStanR` [34]. The `R` script is available at https://github.com/wasiur/SIRapproximatingSEIR.

6. Summary and conclusions

Although there has been a multitude of works on compartmental epidemic SIR and SEIR models, formal treatment of the relationships between the two appears to have received less attention. In this paper, we have formally demonstrated how, in large populations, a biologically more realistic SEIR model can be approximated by a mathematically more convenient SIR system with time-varying infection and recovery rates. To introduce and justify this approximation and quantify the approximation error, we considered a stochastic Markovian setting and provided both a large-population FLLN limit and a finite-population concentration inequality, showing that the approximation is effective, with an error of the order $\exp(-\sqrt{n})$ that decays rapidly. Additionally, we presented a parameter inference methodology based on a dynamical survival model, demonstrating how to use the so-called DSA approach to fit an approximating SIR system to synthetic data generated from an SEIR framework. Based on our method of analysis as well as some recent discussions in [35], it appears that a general result may also be established on the asymptotic equivalence between a wide class of Markovian compartmental models and their counterparts with

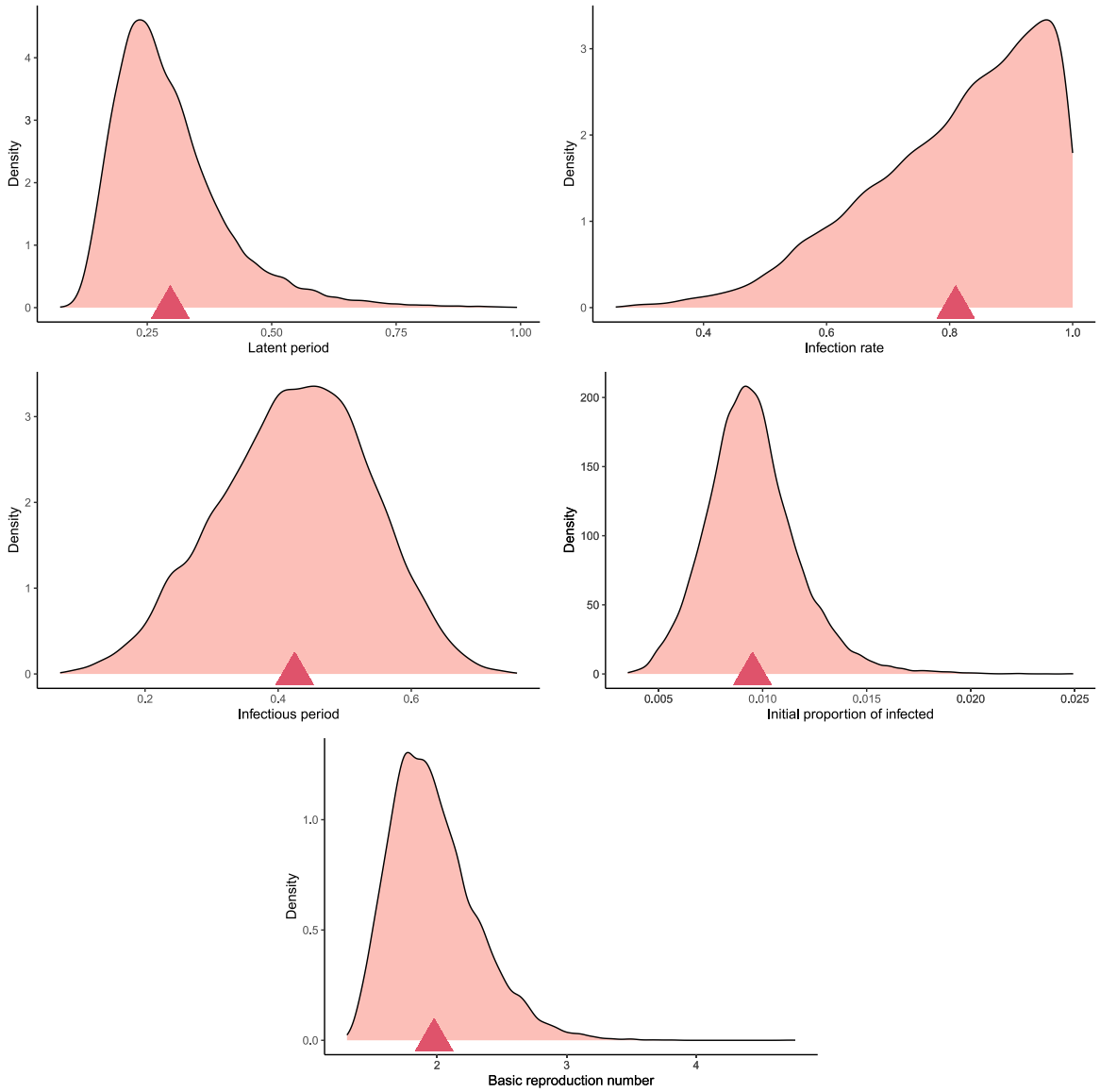

Fig. 3. Posterior distributions of the parameters $(\alpha, \beta, \gamma, \rho)$ and R_0 based on the partial DSA-likelihood (5.18) and a random sample of only infection times under the approximating SIR model. We used uninformative flat priors. The true parameter values are: $\alpha = 0.25$, $\beta = 0.8$, $\gamma = 0.4$, $\rho = 0.01$, and $R_0 = 2.0$. Here, $n = 50\,000$ and random samples of size 5000 of infection were taken for the purpose of inference. The red triangles represent the means of the posterior samples, often employed as substitutes for the Bayesian point estimates.

a reduced number of compartments but with time-dependent rates. We leave the establishment of such a general result for future research.

We have used only synthetic data, which can be generated using the software package.

CRediT authorship contribution statement

Wasiur R. KhudaBukhsh: Writing – original draft, Software, Conceptualization. **Grzegorz A. Rempała:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

The authors declare no conflicts of interest.

Data availability

The R software package is freely available at KhudaBukhsh and Rempała [37]. In this work only synthetic data has been used, which can be generated using the provided software package.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors took assistance of ChatGPT [36], an AI language model developed by OpenAI, in generating the R [33] simulation code used in the data examples. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

WRKB extends sincere thanks to the organizers of the Workshop on "Dynamical Systems in Life Sciences" (July 13–15, 2023) held at the Ohio State University, where discussions pivotal to this manuscript took place. The Workshop celebrated the scientific achievements of Professors Marty Golubitsky and Avner Friedman. During WRKB's tenure as a postdoctoral researcher at the Mathematical Biosciences Institute (MBI) at Ohio State University (OSU), he had the privilege of interacting with

both professors and is deeply grateful for their mentorship. GAR also expresses his gratitude to Professors Golubitsky and Friedman for their longstanding contributions to the mathematical biology community and MBI, where he served as Deputy Director for over a decade.

Funding

WRKB was partially supported by an International Collaboration Fund awarded by the Faculty of Science, University of Nottingham (UoN), United Kingdom, and the Engineering and Physical Sciences Research Council (EPSRC), United Kingdom [grant number EP/Y027795/1 awarded to WRKB]. The Workshop on "Dynamical Systems in Life Sciences" (July 13–15, 2023) at the Ohio State University was partially funded by the National Science Foundation, United States of America award DMS-2310816 to GAR.

Appendix A. Additional mathematical background

A.1. Continuous Time Markov Chains

As the name suggests, a CTMC $(X(t), t \geq 0)$ is a pure-jump Markov process taking values in a countable state space $\mathcal{X}$ and satisfying the Markov property

$$\mathsf{P}\left(X(t) \in B \mid \mathcal{F}_s\right) = \mathsf{P}\left(X(t) \in B \mid X(s)\right)$$

for measurable subsets B of $\mathcal{X}$, where σ-field $\mathcal{F}_s$ denotes the history of the process up to time s. We refer the readers to [38, Chapter 2] for learning about CTMCs beyond their random time change representations, which we discuss here. A standard approach to describe a CTMC

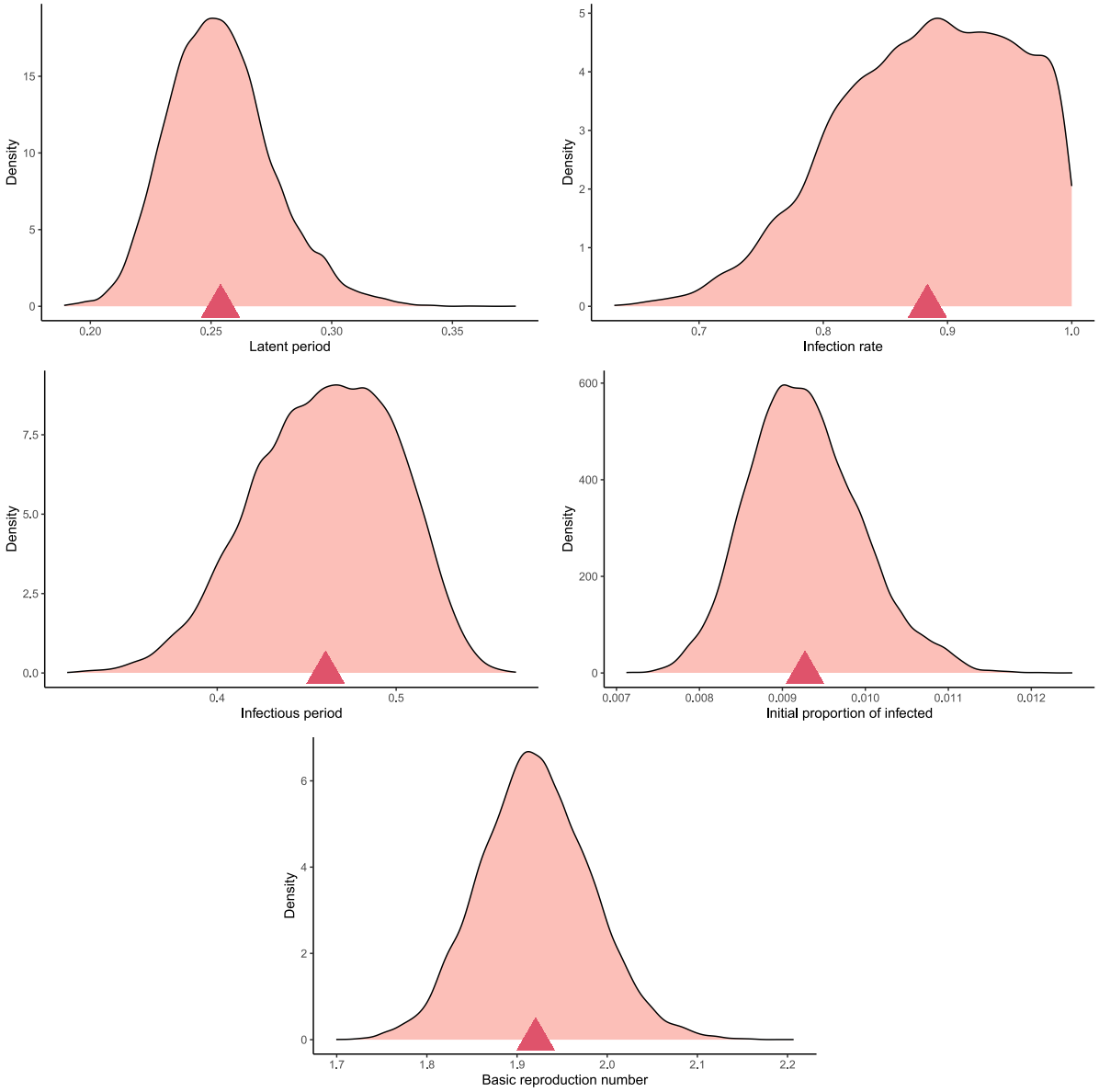

Fig. 4. Posterior distributions of the parameters $(\alpha, \beta, \gamma, \rho)$ and R_0 obtained based on the complete DSA-likelihood (5.19) and a random sample of infection times and recovery times under the approximate SIR model. Uninformative flat priors are used. The true parameter values are: $\alpha = 0.25$, $\beta = 0.8$, $\gamma = 0.4$, $\rho = 0.01$, and $R_0 = 2.0$. Here, $n = 50\,000$ and random samples of size 5000 of infection and recovery times were taken for the purpose of inference. The red triangles indicate the means of the posterior samples.

is via its generator G defined as

$$Gf(x) := \sum_l r_l(x)\left(f(x+e_l) - f(x)\right), \tag{A.1}$$

for bounded functions $f : \mathcal{X} \mapsto \mathbb{R}$, where $r_l : \mathcal{X} \mapsto \mathbb{R}_+$ is the instantaneous jump rate (also called "intensity") in the direction $e_l \in \mathcal{X}$ such that

$$\mathsf{P}\left(X(t+h) = x + e_l \mid X(t) = x\right) \approx r_l(x)h$$

for small $h > 0$. It is a standard result in the theory of Markov processes that if X is a CTMC with generator G, then the stochastic process

$$M_f(t) := f(X(t)) - f(X(0)) - \int_0^t Gf(S(s))\mathrm{d}s \tag{A.2}$$

is an $\mathcal{F}_t$-martingale for bounded functions $f : \mathcal{X} \mapsto \mathbb{R}$. They are sometimes called the Dynkin's martingales in the literature. On the other hand, if a stochastic process $(X(t), t \geq 0)$ is such that the stochastic process $M_f(t)$ defined in (A.2) is an $\mathcal{F}_t$-martingale for every bounded function f, then we say that the stochastic process X is a solution to the martingale problem for the operator G, and it can be shown that the stochastic process X is, in fact, a Markov process with generator G. That is, the martingale problem provides a means to characterize Markov processes.

In practice, it is often easier to view the CTMC as a solution to some stochastic equation. The following theorem from [22, Theorem 1.22] justifies the approach we have taken in Section 3 to describe the CTMC keeping track of the counts $(S(t), E(t), I(t), R(t))$ in terms of randomly time changed Poisson processes in (3.6).

Theorem 3. *Assume $r_l(x) > 0$ implies $x + e_l \in \mathcal{X}$, i.e., there are no transitions that take the process outside $\mathcal{X}$. Also, assume $\sum_l r_l(x) < \infty$ for all $x \in \mathcal{X}$, and $\lim_{|x|\to\infty} Gf(x) = 0$ for functions f with finite support in $\mathcal{X}$. Then, the solution of the stochastic equation*

$$X(t) = X(0) + \sum_l e_l Y_l\left(\int_0^t r_l(X(s))\mathrm{d}s\right)$$

with $X(t) = \Delta$ for $t \geq J_\infty$ is the unique minimal solution to the martingale problem for G, where $Y_1, Y_2, \ldots$ are independent unit-rate Poisson processes, and $J_\infty := \lim_{K\to\infty} J_K$, and $J_K = \inf\{t : |X(t)| > K\}$.

A.2. Derivation of (2.3)

Consider the system of ODEs in (2.1) with the initial condition $x_S(0) = 1, x_I(0) = \rho$, and $x_R(0) = 0$. Now, dividing the equation for $\frac{\mathrm{d}}{\mathrm{d}t}x_I$ by $\frac{\mathrm{d}}{\mathrm{d}t}x_S$, which is nonzero, yields

$$\frac{\mathrm{d}x_I}{\mathrm{d}x_S} = -1 + \frac{\gamma}{\beta}\frac{1}{x_S},$$

solving which along with the initial conditions gives the relation

$$x_I = -x_S + \frac{\gamma}{\beta}\log(x_S) + 1 + \rho.$$

Plugging the above solution back into the equation for $\frac{\mathrm{d}}{\mathrm{d}t}x_S$ gives us

$$-\frac{\mathrm{d}}{\mathrm{d}t}x_S = \beta x_S(1 - x_S) + \gamma\log(x_S) + \beta\rho x_S.$$

Appendix B. Acronyms

ABM	Agent-based Model
BA	Barabási-Albert
CDC	Centers for Disease Control and Prevention
CDF	Cumulative Distribution Function
CLT	Central Limit Theorem
CM	Configuration Model
CME	Chemical Master Equation
CRM	Conditional Random Measure
CRN	Chemical Reaction Network
CTBN	Continuous Time Bayesian Network
CTMC	Continuous Time Markov Chain
DSA	Dynamic Survival Analysis
DTMC	Discrete Time Markov Chain
DRC	Democratic Republic of Congo
ER	Erdös-Rényi
ESI	Enzyme-Substrate-Inhibitor
FCLT	Functional Central Limit Theorem
FLLN	Functional Law of Large Numbers
FPT	First Passage Time
GP	Gaussian Process
HJB	Hamilton–Jacobi–Bellman
HMC	Hamiltonian Monte Carlo
iid	independent and identically distributed
IPS	Interacting Particle System
KL	Kullback–Leibler
LDP	Large Deviations Principle
LLN	Law of Large Numbers
LNA	Linear Noise Approximation
MAPK	Mitogen-activated Protein Kinase
MCMC	Markov Chain Monte Carlo
MFPT	Mean First Passage Time
MGF	Moment Generating Function
MLE	Maximum Likelihood Estimate
MM	Michaelis–Menten
MPI	Message Passing Interface
MSE	Mean Squared Error
ODE	Ordinary Differential Equation
PDE	Partial Differential Equation
PDF	Probability Density Function
PGF	Probability Generating Function
PMF	Probability Mass Function
psd	positive semi-definite
PT	Poisson-type

QSSA	Quasi-Steady State Approximation
rQSSA	reversible QSSA
SD	Standard Deviation
SEIR	Susceptible-Exposed-Infected-Recovered
SI	Susceptible-Infected
SIR	Susceptible-Infected-Recovered
SIS	Susceptible-Infected-Susceptible
sQSSA	standard QSSA
tQSSA	total QSSA
TK	Togashi–Kaneko
WS	Watts–Strogatz
whp	with high probability

References

[1] W.O. Kermack, A.G. McKendrick, A contribution to the mathematical theory of epidemics, Proc. R. Soc. A 115 (1927).

[2] T.G. Kurtz, Strong approximation theorems for density dependent markov chains, Stoch. Process. Appl. 6 (1978) 223–240.

[3] T.G. Kurtz, Solutions of ordinary differential equations as limits of pure jump markov processes, J. Appl. Probab. 7 (1970) 49–58, http://dx.doi.org/10.2307/3212147.

[4] R. Darling, J. Norris, Differential equation approximations for markov chains, Probab. Surv. 5 (2008) http://dx.doi.org/10.1214/07-ps121.

[5] H. Andersson, T. Britton, Stochastic Epidemic Models and their Statistical Analysis, vol. 151, Springer-Verlag New York, 2000, http://dx.doi.org/10.1007/978-1-4612-1158-7.

[6] A. Volkening, D.F. Linder, M.A. Porter, G.A. Rempala, Forecasting elections using compartmental models of infection, SIAM Rev. 62 (2020) 837–865, http://dx.doi.org/10.1137/19M1306658.

[7] L. Decreusefond, J.S. Dhersin, P. Moyal, V.C. Tran, Large graph limit for an SIR process in random network with heterogeneous connectivity, Ann. Appl. Probab. 22 (2012) http://dx.doi.org/10.1214/11-aap773.

[8] S. Janson, M. Luczak, P. Windridge, Law of large numbers for the SIR epidemic on a random graph with given degrees, Random Structures Algorithms 45 (2014) 726–763, http://dx.doi.org/10.1002/rsa.20575.

[9] W.R. Khudabukhsh, C. Woroszylo, G.A. Rempała, H. Koeppl, A functional central limit theorem for SI processes on configuration model graphs, Adv. in Appl. Probab. 54 (2022) 880–912, http://dx.doi.org/10.1017/apr.2022.52.

[10] F. Ball, T. Britton, K.Y. Leung, D. Sirl, A stochastic SIR network epidemic model with preventive dropping of edges, J. Math. Biol. 78 (2019) 1875–1951, http://dx.doi.org/10.1007/s00285-019-01329-4.

[11] R. van der Hofstad, Random Graphs and Complex Networks, Cambridge University Press, 2016, http://dx.doi.org/10.1017/9781316779422.

[12] W.R. KhudaBukhsh, C.D. Bastian, M. Wascher, C. Klaus, S.Y. Sahai, M.H. Weir, E. Kenah, E. Root, J.H. Tien, G.A. Rempała, Projecting COVID-19 cases and hospital burden in Ohio, J. Theoret. Biol. 561 (2023) 111404, http://dx.doi.org/10.1016/j.jtbi.2022.111404.

[13] K.A. Jacobsen, M.G. Burch, J.H. Tien, G.A. Rempała, The large graph limit of a stochastic epidemic model on a dynamic multilayer network, J. Biol. Dyn. 12 (2018) 746–788, http://dx.doi.org/10.1080/17513758.2018.1515993.

[14] T. Sellke, On the asymptotic distribution of the size of a stochastic epidemic, J. Appl. Probab. 20 (1983) 390–394, http://dx.doi.org/10.2307/3213811.

[15] W.R. KhudaBukhsh, B. Choi, E. Kenah, G.A. Rempał, Survival dynamical systems: individual-level survival analysis from population-level epidemic models, Interface Focus 10 (2020) 20190048, http://dx.doi.org/10.1098/rsfs.2019.0048.

[16] F. Di Lauro, W.R. KhudaBukhsh, I.Z. Kiss, E. Kenah, M. Jensen, G.A. Rempała, Dynamic survival analysis for non-markovian epidemic models, J. R. Soc. Interface 19 (2022) 20220124, http://dx.doi.org/10.1098/rsif.2022.0124.

[17] C.D. Bastian, G.A. Rempala, Throwing stones and collecting bones: Looking for Poisson-like random measures, Math. Methods Appl. Sci. (2020) http://dx.doi.org/10.1002/mma.6224.

[18] G.A. Rempała, W.R. KhudaBukhsh, Dynamical survival analysis for epidemic modeling, in: Handbook of Visual, Experimental and Computational Mathematics, Springer International Publishing, 2023, pp. 1–17, http://dx.doi.org/10.1007/978-3-030-93954-0_31-1.

[19] I.Z. Kiss, E. Kenah, G.A. Rempała, Necessary and sufficient conditions for exact closures of epidemic equations on configuration model networks, J. Math. Biol. 87 (2023) 36, http://dx.doi.org/10.1007/s00285-023-01967-9.

[20] G.A. Rempała, Equivalence of mass action and Poisson network SIR epidemic models, BIOMATH 12 (2023) 2311237, http://dx.doi.org/10.55630/j.biomath.2023.11.237.

[21] S.N. Ethier, T.G. Kurtz, Markov Processes: Characterization and Convergence, vol. 282, John Wiley & Wiley, 1986.

[22] D.F. Anderson, T.G. Kurtz, Continuous time markov chain models for chemical reaction networks, in: Design and Analysis of Biomolecular Circuits, Springer, 2011, pp. 3–42.

[23] D.J. Wilkinson, Stochastic Modelling for Systems Biology, Chapman and Hall/CRC, 2018.

[24] M.A. Gibson, J. Bruck, Efficient exact stochastic simulation of chemical systems with many species and many channels, J. Phys. Chem. A 104 (2000) 1876–1889, http://dx.doi.org/10.1021/jp993732q.

[25] D.F. Anderson, A modified next reaction method for simulating chemical systems with time dependent propensities and delays, J. Chem. Phys. 127 (2007) http://dx.doi.org/10.1063/1.2799998.

[26] F. Kühn, R.L. Schilling, Maximal inequalities and some applications, Probab. Surv. 20 (2023) http://dx.doi.org/10.1214/23-ps17.

[27] I.Z. Kiss, L. Berthouze, W.R. KhudaBukhsh, Towards inferring network properties from epidemic data, Bull. Math. Biol. 86 (2023) http://dx.doi.org/10.1007/s11538-023-01235-3.

[28] W.R. KhudaBukhsh, S.K. Khalsa, E. Kenah, G.A. Rempała, J.H. Tien, COVID-19 dynamics in an ohio prison, Front. Public Health 11 (2023) http://dx.doi.org/10.3389/fpubh.2023.1087698.

[29] H. Vossler, P. Akilimali, Y. Pan, W.R. KhudaBukhsh, E. Kenah, G.A. Rempała, Analysis of individual-level data from 2018–2020 ebola outbreak in democratic republic of the congo, Sci. Rep. 12 (2022) 5534, http://dx.doi.org/10.1038/s41598-022-09564-4.

[30] M. Wascher, P.M. Schnell, W.R. KhudaBukhsh, M.B.M. Quam, J.H. Tien, G.A. Rempała, Estimating disease transmission in a closed population under repeated testing, J. R. Stat. Soc. Ser. C. Appl. Stat. (2024) qlae021, http://dx.doi.org/10.1093/jrsssc/qlae021, arXiv:https://academic.oup.com/jrsssc/advance-article-pdf/doi/10.1093/jrsssc/qlae021/57417654/qlae021.

[31] I. Somekh, W.R. KhudaBukhsh, E.D. Root, L.K. Boker, G. Rempala, E. Simões, Quantifying the population-level effect of the COVID-19 mass vaccination campaign in israel: A modeling study, Open Forum Infect. Dis. 9 (2022) http://dx.doi.org/10.1093/ofid/ofac087.

[32] T. Smith, R.H. Holm, R.J. Keith, A.R. Amraotkar, C.R. Alvarado, K. Banecki, B. Choi, I.C. Santisteban, A.M. Bushau-Sprinkle, K.T. Kitterman, J. Fuqua, K.T. Hamorsky, K.E. Palmer, J.M. Brick, G.A. Rempala, A. Bhatnagar, Quantifying the relationship between sub-population wastewater samples and community-wide SARS-CoV-2 seroprevalence, Sci. Total Environ. 853 (2022) 158567, http://dx.doi.org/10.1016/j.scitotenv.2022.158567.

[33] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2023, https://www.R-project.org/.

[34] J. Gabry, R. Češnovar, Cmdstanr: R interface to 'CmdStan', 2022, https://mc-stan.org/cmdstanr/, https://discourse.mc-stan.org.

[35] D. Cappelletti, G.A. Rempala, Individual molecules dynamics in reaction network models, SIAM J. Appl. Dyn. Syst. 22 (2023) 1344–1382, http://dx.doi.org/10.1137/21M1459563.

[36] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020, arXiv:2010.11929.

[37] W.R. KhudaBukhsh, G.A. Rempała, 2023, `R software package`, https://github.com/wasiur/SIRapproximatingSEIR.

[38] J.R. Norris, Markov Chains, Cambridge University Press, 1997, http://dx.doi.org/10.1017/cbo9780511810633.