

**Estimating time-varying transmission rates of the SIR model**

**Sang Woo Park**

supervised by

Dr. Ben Bolker

A thesis presented for Math 4P06

Department of Mathematics & Statistics  
McMaster University  
April 1, 2019

## **Abstract**

### **Acknowledgements**

I would like to thank Dr. Ben Bolker and Dr. Jonathan Dushoff for their guidance throughout my undergraduate studies.

I would also like to thank Mac-Theobio group.

Finally, I would like to thank my family for supporting me.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Methods</b>	<b>5</b>
2.1	SIR model . . . . .	5
2.2	Epidemic time series . . . . .	6
2.3	Fitting methods . . . . .	7
2.4	Splines . . . . .	11
<b>3</b>	<b>Results</b>	<b>11</b>
3.1	Types of time series . . . . .	11
3.2	Susceptible reconstruction . . . . .	13
3.3	Density dependence . . . . .	14
3.4	Simulation study: comparison of statistical properties . . . . .	16
3.5	Simulation study: estimating time-varying transmission rates . .	17
<b>4</b>	<b>Discussion</b>	<b>19</b>
<b>5</b>	<b>Appendix</b>	<b>19</b>
5.1	Equivalence of incidence, mortality, and prevalence . . . . .	19
<b>6</b>	<b>Fitting continuous models</b>	<b>19</b>
6.1	Gradient matching . . . . .	19
6.2	Spectral matching . . . . .	19
<b>7</b>	<b>Appendix</b>	<b>23</b>
7.1	Derivation of growth rate in discrete-time SIR model . . . . .	23
7.2	Probability of infection . . . . .	23
7.3	Infectious period . . . . .	24

# 1 Introduction

Over the last few decades, a plethora of computational tools have been developed to study irregular patterns observed in epidemic time series, such as those of measles. Central to these studies are mechanistic models, which are sets of equations that represent our understanding of the underlying biological system (Bretó et al., 2009). Mechanistic modeling of disease dynamics allows us to study the qualitative dynamics of a system and infer underlying process that drove the observed patterns in the data. For example, we now know that the complex dynamics of measles outbreaks are driven by seasonal transmission patterns (Earn et al., 2000; Dalziel et al., 2016).

Major challenges in epidemiological modeling come from developing the actual model and parameterizing it. One must make an appropriate set of assumptions to balance generality, realism, and precision (Levins, 1966). These assumptions may depend on a system as well as an inferential tool. Fortunately, a basic epidemic model, such as the Susceptible-Infected-Recovered (SIR) model can describe the measles dynamics sufficiently well (Earn et al., 2000; Krylova and Earn, 2013; Hempel and Earn, 2015; Dalziel et al., 2016). Focusing on a simple, well-understood model allows us to make a fair comparison between different statistical methods in their ability to estimate underlying parameters; If any changes have to be made to the model to apply a statistical tool, they will be relatively easy to understand and control. Here, we focus on estimating the time-varying transmission rate of the SIR model.

One way to estimate the transmission rate is by matching the deterministic trajectory of the system with the observed data by minimizing their “distance”, often measured by the sum of squares difference or negative log-likelihood (Riley et al., 2003; Chowell et al., 2004). This method of fitting a deterministic model (often referred to as trajectory matching) assumes that *all* variation in data can be explained by observation error alone (Bolker, 2008). Likewise, it is possible to match gradients of the system (1) with the observed gradient, which can be estimated by fitting a smooth curve to a time series, such as piecewise polynomials, and taking its derivative (Ellner et al., 2002). These methods are computationally efficient and easy to implement.

Monte Carlo methods are computationally demanding but allow us to model *all* sources of variation (e.g., observation, demographic, and environmental stochasticity) explicitly. In particular, Sequential Monte Carlo methods (also known as particle filtering) has been successfully applied in mechanistic analyses of epidemic time series (He et al., 2009, 2011; Didelot et al., 2017); its ability to approximate the likelihood through a series of simulations provides a powerful technique for testing hypotheses and inferring underlying parameters (Ionides et al., 2006; Bretó et al., 2009; King et al., 2015). Approximate likelihood obtained by particle filtering can be maximized by iterated filtering (Ionides et al., 2011, 2015) or can be incorporated into a Bayesian scheme as a particle Monte Carlo Markov Chain (MCMC) (Andrieu et al., 2010).

Unlike any of the previously discussed methods, which can be used to estimate parameters of a wide range of systems, the time series SIR (TSIR) method

seeks to estimate time-varying transmission rates of fully-immunizing childhood infections, specifically (Finkenstadt and Grenfell, 2000). The TSIR method first estimates the dynamics of the susceptible population based on the assumption that all individuals eventually become infected; conditional on the estimated susceptible dynamics, transmission rates can be estimated using a regression. Even though the TSIR method is designed to tackle a specific kind of problems, it has been consistently used since its introduction. So what makes it work so well?

**[SWP: Come back]** The main purpose of this project is to compare the TSIR method against other methods. We begin by discussing types of epidemic time series (incidence, prevalence, mortality) and necessary assumptions behind each method, including TSIR. Then, we compare their ability to estimate transmission rates by testing them against simulations. Finally, we provide practical guidance for estimating parameters of the SIR model.

## 2 Methods

### 2.1 SIR model

The Susceptible-Infected-Recovered (SIR) model is one of the most basic epidemic models (Kermack and McKendrick, 1927). It describes how disease spreads in a homogeneously mixing population. Here, we present the model as a set of ordinary differential equations:

$$\begin{aligned} \frac{dS}{dt} &= b(t) - \left( \beta(t) \frac{I}{N} + \mu \right) S \\ \frac{dI}{dt} &= \beta(t) S \frac{I}{N} - (\gamma + \mu) I \\ \frac{dR}{dt} &= \gamma I - \mu R \end{aligned} \tag{1}$$

where  $S$  is the number of susceptible individuals,  $I$  is the number of infected individuals, and  $R$  is the number of recovered (or removed) individuals. We assume that individuals enter the susceptible pool at a rate  $b(t)$ . Susceptible individuals become infected at a rate of  $\beta(t)$  by contacting infected individuals. Infected individuals recover at a rate of  $\gamma$ ; in other words, the infectious period is exponentially distributed with mean  $1/\gamma$  time units. Finally, all individuals experience natural mortality at a rate of  $\mu$ .

Here, we mainly consider two sets of stochastic simulations: (1) a single outbreak with constant transmission rate (and no movement in and out of the susceptible population) and (2) recurrent outbreaks driven by sinusoidal transmission rates. To simulate a single outbreak, we use the following parameterization:  $\beta(t) = 2$ ,  $\gamma = 1$ ,  $\mu = b(t) = 0$ ,  $N = 100000$ . Initial conditions are specified as follows:  $I(0) = 10$  and  $S(0) = N - I(0)$ . To simulate recurrent outbreaks, we use the following parameterization:  $\beta(t) = b_0(1 + b_1 \cos(t/26 \times 2\pi))$ ,  $b_0 = 500/26$ ,  $b_1 = 0.15$ ,  $\gamma = 1$ ,  $\mu = 1/(50 \times 26)$ ,  $N = 500000$ . Birth rate is specified

to keep the total population size constant:  $b(t) = \mu N$ . Initial conditions are specified as follows:  $I(0) = 1 \times 10^{-1} \times N$  and  $S(0) = 0.05 \times N$ .

Stochastic simulation of the SIR model is performed using the Gillespie algorithm (Gillespie, 1976).

## 2.2 Epidemic time series

In order to estimate the parameters of the SIR model, it is important to know what kinds of data are available. Epidemic time series can be classified into three categories: incidence, mortality, and prevalence. Incidence is defined as the number of newly infected individuals generated over a reporting period. Then, true incidence of the SIR model (1) between time  $t$  and  $t - t_{\text{rep}}$ , where  $t_{\text{rep}}$  is the length of reporting time step, can be obtained by integrating total infection rate over the reporting period:

$$i_t = \int_{t-t_{\text{rep}}}^t \beta(s)S \frac{I}{N} ds. \quad (2)$$

Similarly, true mortality case, which is defined as the number of individuals that died during a reporting period, can be obtained by integrating total death rate over the reporting period:

$$m_t = \int_{t-t_{\text{rep}}}^t \gamma I ds. \quad (3)$$

Finally, prevalence is defined as the number of infected individuals that are present in the population and corresponds to the state variable  $I$ ; to distinguish the continuous dynamics from the discrete reports, we write  $p_t$  to denote true prevalence at time  $t$ . Incidence, mortality, and prevalence for a stochastic simulation is defined in the same manner as described above.

Reported incidence, mortality, and prevalence cases are drawn from a beta-binomial distribution with reporting rate  $\rho$  and dispersion parameter  $\theta_{\text{rep}}$  (Morris, 1997). We assume that cases are reported instantaneously but a more realistic model can incorporate a delay between infection time and reporting time (e.g., infected individuals are likely to report after symptom onset but before recovery). We do not explore these ideas here.

Instead, we focus on how different assumptions about an epidemic time series (whether it represents incidence, mortality, or prevalence) affect our inference. While this question is initially motivated by the TSIR framework, where incidence is assumed to be equal to prevalence when reporting time is equal to mean generation time, the actual problem of identifying the correct representation of an epidemic time series is far more common. For example, incidence has been assumed to be measured either upon transition from latent to infectious stage (Hooker et al., 2010; Althaus, 2014) or upon recovery (He et al., 2009), where the latter assumption is equivalent to our definition of mortality reports. Understanding these differences is crucial in order to correctly estimate underlying parameters as well as their associated uncertainties. For all other purposes, we use incidence time series throughout this study.

## 2.3 Fitting methods

### Time-series SIR model

The time-series SIR (TSIR) method was developed by Bjørnstad et al. (2002) to estimate seasonally varying transmission rates of measles. The TSIR method begins by discretizing the infection process of the SIR model, assuming that disease generation time is equal to reporting interval:

$$\begin{aligned} S_{t+1} &= B_t + S_t - i_{t+1} \\ i_{t+1} &= \beta_t S_t \frac{i_t^\alpha}{N_t} \end{aligned} \quad (4)$$

where  $\alpha$  is often referred to as a conversion factor from a continuous-time model to a discrete-time model or a heterogeneity parameter (Glass et al., 2003). Taking log on both sides, estimation of the transmission rates  $\beta_t$  becomes a regression problem:

$$\log i_{t+1} = \alpha \log \beta_t + \log S_t + \alpha \log i_t - \log N_t + \epsilon_t. \quad (5)$$

A key assumption behind the TSIR model is that all, if not most, individuals become infected eventually. In other words, the total number of births should be equal to the total number of cases over a long period of time. Rewriting true incidence  $i_t$  as a product of reported cases  $i_{t,\text{rep}}$  and reciprocal of reporting rate  $1/\rho_t$  and iteratively solving (4), we get

$$\sum_{t=1}^N B_t = \sum_{t=1}^N \frac{i_{t,\text{rep}}}{\rho_{t+1}} + Z_{t+1} - Z_1, \quad (6)$$

where  $Z_t = S_t - \bar{S}$ . Under this assumption, fitting a regression model to cumulative birth as a function of cumulative cases allows us to estimate time-varying reporting rate  $\rho_t$  from the reciprocal of the slope and susceptible dynamics  $Z_t$  from the residuals (Finkenstadt and Grenfell, 2000). Conditional on the known susceptible dynamics, we can apply the TSIR regression to epidemic time series.

### Trajectory matching

Trajectory matching is a way of estimating underlying parameters of an ODE model by matching the observed time series and a deterministic solution of the model. Given observed incidence  $i_{t,\text{rep}}$  (or equivalently, observed prevalence or mortality cases), we can define its mean dynamics using the solution of the SIR model:

$$\mu_t = \frac{i_t}{\rho} = \frac{1}{\rho} \int_{t-t_{\text{rep}}}^t \beta(s) S \frac{I}{N} ds. \quad (7)$$

Since it is difficult to integrate of the state-space, we include another state variable  $C$  that represents cumulative incidence where  $dC/dt = \beta SI/N$  and numerically solve for  $C$  by integrating alongside the SIR model; taking the

difference in  $C$  between two consecutive reporting periods allows us to obtain true  $i_t$ . Finally, the likelihood of observing the data can be defined as a deviation from this mean trajectory:

$$i_{t,\text{rep}} \sim \text{NegBin}(\mu_t, \phi_{\text{obs}}), \quad (8)$$

where  $\phi_{\text{obs}}$  is the negative binomial dispersion parameter.

### Gradient matching

Gradient matching is a way of estimating underlying parameters of an ODE model by matching the gradients of a system, rather than the deterministic solution of the system (Ellner et al., 2002). Gradient matching provides a flexible way of modeling a system as it does not require the parametric form of the rate equations to be pre-specified (e.g., we do not need to that per-capita rate of transmission is a linear function of  $I$ ). Instead, parameters can be estimated by fitting a non-parametric regression model to the “observed gradients” as a function of state variables that are relevant. Therefore, all state variables must be observed in order to apply gradient matching.

In practice, it is impossible to observe all state variables of an epidemic model, even for simple models such as the SIR model. When either incidence or mortality cases are reported, *no* state variables are observed. Nonetheless, it is still possible to estimate the transmission rates of the SIR model using gradient matching by making similar assumptions as the TSIR model. Here, we briefly outline the estimation process; note that the purpose of section is to introduce the method, rather than to provide a rigorous justification of the statistical model.

First, we begin by assuming that the mean generation time is equal to the length of reporting period and that incidence is proportional to prevalence, i.e.,  $i_t = aI_t$  for some  $a > 0$ . Note that this is a weaker assumption than the condition for the TSIR method. We then fit a smooth curve  $f(t)$  to log reported incidence  $\log i_{t,\text{rep}}$ :

$$\log i_{t,\text{rep}} \sim \text{Normal}(f(t), \sigma_{\text{smooth}}^2), \quad (9)$$

where natural cubic splines are used to model the function  $f$  [**SWP: CITE spline package**]. Basis knots are placed every 6 generations.

The smooth curve  $f(t)$  represents the dynamics of the log of prevalence ( $\log I_t$ ) up to a constant. Taking the derivative of the curve  $f(t_i)$ , we have an estimate of the derivative of the log of prevalence:

$$E\left[\frac{df(t)}{dt}\right] = \beta S - (\gamma + \mu). \quad (10)$$

There are several ways we can estimate the transmission rate using this equation. First, we can model transmission rate as a function of susceptibles  $S$  by fitting a non-parametric regression by treating known values of  $\gamma$  and  $\mu$  as offset terms:

$$\frac{df(t)}{dt} \sim \text{Normal}(g(S_t) - (\gamma + \mu), \sigma^2), \quad (11)$$

where  $g$  is a non-parametric function that represents the transmission rate and the unobserved susceptible population  $S_t$  can be estimated from susceptible reconstruction. When transmission varies over time, we can either (1) model transmission rate as an interaction term between a factor variable that represents time of the year and a smoothing term (i.e., fit a separate nonparametric curve  $g$  for each factor level) or try to separate transmission rate from the susceptible dynamics using a log link function:

$$\frac{df(t)}{dt} + (\gamma + \mu) \sim \text{Normal}(\exp(\log \beta_t + \log S_t), \sigma^2), \quad (12)$$

where log number of susceptible individuals  $\log S_t$  is treated as an offset term and the log transmission rate can be estimated is estimated as a function of time of year using a nonparametric regression.

### Sequential Monte Carlo

Calculating the exact likelihood for a nonlinear stochastic system is practically impossible. Instead, Sequential Monte Carlo (particle filtering) allows the likelihood to be approximated via simulations. Let  $X_t = \{\mathbf{x}_{t,i} \mid i = 1, 2, \dots, n\}$  be the set of states (particles) at time  $t$  and  $h$  be the process model. For the SIR model, each particle  $\mathbf{x}_{t,i} \in X$  consists of number of susceptible individuals  $S_t$  and number of infected individuals  $I_t$ . Then, we can predict the states at next time step  $t + 1$  for each particle using the process model:  $h(\mathbf{x}_{t,i})$ . For each prediction  $h(\mathbf{x}_{t,i})$ , we can calculate the likelihood  $w_{t,i}$  of the observation at time  $t + 1$  and define marginal likelihood by taking the average:  $\bar{w}_t$ . Finally, we draw  $n$  weighted samples from the set  $\{h(\mathbf{x}_{t,i}) \mid i = 1, 2, \dots, n\}$ , where the weight of each  $h(\mathbf{x}_{t,i})$  is proportional to  $w_i$ , to obtain the next set of particles  $X_{t+1}$ . This algorithm is repeated until the final time step. Then, complete likelihood of the data is given by the product of the marginal likelihood at each time step:  $\prod \bar{w}_t$ .

While Sequential Monte Carlo provides a powerful technique for approximating the likelihood, it is still impractical to fit a continuous-time stochastic model. For example, it can take more than [SWP: 10] hours to simulate an 20-year-long measles time series with gillespie algorithm. Computing a likelihood with Sequential Monte Carlo would take more than a few days, let alone maximizing the likelihood.

Instead of fitting a continuous-time model, we can assume that an epidemic process is a partially observed discrete-time Markov process and discretize the system by using a hazard approximation of the transition probabilities between

states:

$$\begin{aligned}
B(t) &\sim \text{Poisson}(b(t)\Delta t) \\
F_S(t) &\sim \text{Binomial}(S_t, 1 - \exp(-(r_1 + r_2)\Delta t)) \\
F_I(t) &\sim \text{Binomial}(I_t, 1 - \exp(-(r_3 + r_4)\Delta t)) \\
F_{S,I}(t) &\sim \text{Binomial}\left(F_S(t), \frac{r_1}{r_1 + r_2}\right) \\
S(t + \Delta t) &= B(t) + S(t) - F_S(t) \\
I(t + \Delta t) &= F_S(t) + I(t) - F_I(t) \\
C(t + \Delta t) &= C(t) + F_S(t)
\end{aligned} \tag{13}$$

where  $r_1 = \beta I/N$ ,  $r_3 = \gamma$ , and  $r_2 = r_4 = \mu$ . Random variables  $F_S(t)$  and  $F_I(t)$  represent the total number of individuals that move out from susceptible  $S$  and infected  $I$  states. Likewise,  $F_{S,I}(t)$  represent the total number of individual move from susceptible state  $S$  to infected state  $I$ . Finally, the state variable  $C(t)$  keeps track of cumulative incidence. Again, observation likelihood is defined using a negative-binomial distribution:

$$i_{t,\text{rep}} \sim \text{NegBin}(\rho(C(t) - C(t - t_{\text{rep}})), \phi_{\text{obs}}). \tag{14}$$

### Renewal equation

Previously discussed models assume that the shape of generation-interval distribution is exactly known. The continuous time SIR model assumes that generation intervals are exponentially distributed. The discrete-time SIR model represented as a partially observed Markov process assumes that generation intervals are geometrically distributed. Finally, the TSIR assumes that generation intervals are fixed. All these assumptions are unrealistic; we expect a realistic generation-interval distribution of measles to be narrow but not fixed.

It is possible to relax the assumption on generation-interval distribution of the SIR model by including a latent period and allowing for infectious and latent periods to have multiple stages. However, the number of stages (compartments) in each state must be predetermined before parameter estimation is performed. Instead, we can model true incidence as a function of previous incidence and generation-interval distributions explicitly via the renewal equation:

$$i(t) = \mathcal{R}_0 S(t) \int_0^t g(t-s)i(s)ds, \tag{15}$$

where  $i(t)$  represents incidence,  $S(t)$  represents number of susceptible individuals,  $g(t)$  represents generation-interval distribution and  $\mathcal{R}_0$  represents the basic reproductive number. Because this model generalizes the compartmental models with Erlang distributed latent and infectious periods, including the standard SIR model (Champredon et al., 2018), we expect this model to perform better in analyzing time series of outbreaks.

Previous studies used a discrete time version of the renewal equation model to analyze Ebola epidemics but assumed that the simulation time step was equal

to the reporting time step (Li et al., 2018; Champredon et al., 2018). Here, we extend their model by simulating the epidemic on a finer scale and allowing for natural birth and death in the susceptible population:

$$\begin{aligned}
B(t) &\sim \text{Poisson}(b(t)\Delta t) \\
\phi(t) &= \mathcal{R}(t) \sum_{k=1}^{\ell} i(t - k\Delta t)g(k\Delta t) \\
F_S(t) &\sim \text{Binomial}\left(S(t), 1 - e^{-(\phi(t)+\mu\Delta t)}\right) \\
i(t + \Delta t) &\sim \text{Binomial}\left(F_S(t), \frac{\phi(t)}{\phi(t) + \mu\Delta t}\right) \\
S(t + \Delta t) &= B(t) + S(t) - F_S(t)
\end{aligned} \tag{16}$$

## 2.4 Splines

In order to estimate time-varying transmission rates as a smooth function of time, we model transmission rate  $\beta(t)$  using a spline function (Hooker et al., 2010):

$$\beta(t) = \sum_{i=1}^k \phi_i(t)b_i, \tag{17}$$

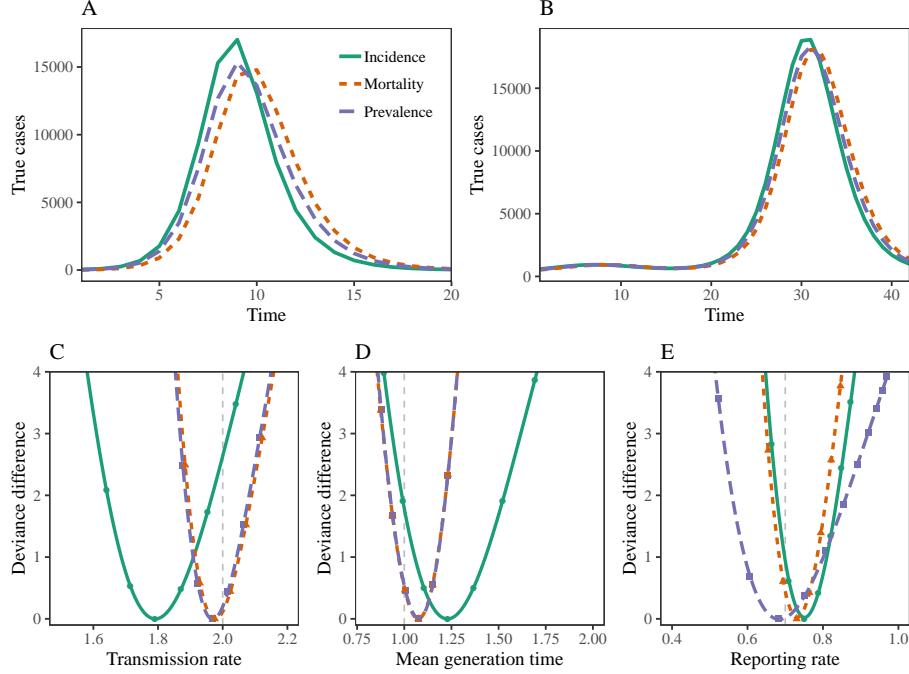
where each  $\phi_i$  is a cyclic cubic B-spline basis and  $b_i$  is the corresponding coefficient. Here, we fix knot locations at every two biweeks (1/13 years); this parameterization results in almost a two-fold reduction in the number of parameters to be estimated compared to the regular TSIR procedure, which estimates a separate transmission rate for every biweek. For regression-based methods (TSIR and gradient matching), we allow the number of knots to be estimated within the regression method by the Restricted Maximum Likelihood (REML) approach (Wood, 2012).

## 3 Results

### 3.1 Types of time series

When reporting period is equal to mean generation time, incidence and prevalence are often assumed to be equivalent (Fig. 1). Indeed, for the simple SIR model (1), it can be shown that incidence, prevalence, and mortality are similar to each other under Euler approximation when reporting period is equal to mean generation time (see Appendix). Their deterministic dynamics are sufficiently similar that we expect these three reports to be nearly indistinguishable from each other when demographic stochasticity and observation error is introduced. However, this is not necessarily true for more complicated models. For a more realistic model with exposed period (SEIR model), incidence is similar to prevalence when reporting period is equal to mean *generation time* whereas mortality

is similar to prevalence when reporting period is equal to mean *infectious period*.  
**[SWP: TODO: confirm]**



**Figure 1: Comparison of incidence, prevalence, and mortality from dynamical and estimation perspective.** A: dynamics of incidence, mortality and prevalence of the SIR model for a single outbreak (no natural birth/death and fixed transmission rate) when mean generation time ie equal to reporting period. The parameters and initial conditions are  $\beta = 2$ ,  $\gamma = 1$ ,  $N = 1 \times 10^5$ ,  $I(0) = 10$ , and  $S(0) = N - 10$ . B: dynamics of incidence, mortality and prevalence of the SIR model with sinusoidal transmission rate ( $\beta(t) = b_0(1 + b_1 \cos(2\pi t/26))$ ) when mean generation time ie equal to reporting period. The parameters and initial conditions are  $b_0 = 500/26$ ,  $b_1 = 0.15$ ,  $\gamma = 1$ ,  $\mu = 1/(50 \times 26)$ ,  $N = 5 \times 10^6$ ,  $I(0) = 0.0001N$ ,  $S(0) = 0.05N$ . C, D, E: deviance difference (profile likelihood minus the maximum likelihood) of each parameters when incidence, mortality, and prevalence are fit to same time series using trajectory matching. Grey dashed line represents the true value.

Even if incidence, mortality, and prevalence are expected to be exhibit similar dynamics other when reporting period is equal to mean generation time, it is important to distinguish them when we are estimating underlying parameters of the SIR model, especially when mean generation time  $1/\gamma$  is not exactly known. To demonstrate this idea, we first simulate an epidemic time series representing prevalence over time by solving the SIR model numerically and

calculating prevalence,  $I(t)$ , for 20 reporting periods, where reporting period is equal to the mean generation time. Then, we draw a beta-binomial random variable at each time step with a reporting rate of 70% and an overdispersion parameter of 10. Using this time series, we try to estimate four parameters – transmission rate  $\beta$ , recovery rate  $\gamma$ , reporting rate  $\rho$ , and an initial number of infected individuals  $I(0)$  – by treating the same time series as if it were incidence, mortality, and prevalence report (Fig. 1).

We find that fitting prevalence and mortality curves to the same time series yields almost identical estimates of the transmission rate  $\beta$  and the recovery rate  $\gamma$  (as well as related uncertainty in the estimates) whereas fitting incidence and mortality curves to the same time series yields consistent estimates of the reporting rate  $\rho$  (as well as related uncertainty in the estimates). As mortality is approximately a lagged reflection of prevalence regardless of reporting period (reported mortality case at time  $t$  is approximately equal to  $\rho\gamma I(t - t_{\text{rep}})t_{\text{rep}}$ ), it is possible to generate a mortality curve that matches its corresponding prevalence curve by adjusting the reporting rate  $\rho$ . This explains why estimates of dynamical parameters,  $\beta$  and  $\gamma$ , are similar whether we treat the same time series as prevalence or mortality. On the other hand, incidence report no longer matches prevalence report when reporting period is not equal to the mean generation time (i.e., when  $\gamma$  is being estimated); therefore, fitting incidence curve yields a different estimate of the dynamical parameters,  $\beta$  and  $\gamma$ , from fitting mortality or prevalence curves. Fitting incidence and mortality curve yields similar estimates of the reporting rate  $\rho$  because they both contain equivalent information about the final size of an epidemic: total *true* incidence and total *true* mortality is equal to the final size of an epidemic.

Even when mean generation time is exactly known, we find that estimates of transmission rate  $\beta$  and reporting rate  $\rho$  depends on our assumptions about the data. We find that fitting mortality and prevalence yields consistent estimates of both  $\beta$  and  $\rho$  in this case, whereas fitting incidence yields similar but still different estimates. Treating incidence report as prevalence report (and vice versa) should be avoided.

[SWP: *Say something more*]

### 3.2 Susceptible reconstruction

One of the main assumptions behind the TSIR model is that the dynamics of susceptible population  $S_t$  must be known. Assuming that all newborn infants become infected eventually, it is possible to recover the reporting rate,  $\rho_t$ , as well as the dynamics of the susceptible population as a deviation from its mean,  $Z_t = S_t - \bar{S}$ , by fitting a regression model to between cumulative births and cumulative cases:

$$\sum_{t=1}^N B_t = \sum_{t=1}^N \frac{C_{t+1}}{\rho_{t+1}} + Z_{t+1} - Z_1, \quad (18)$$

where  $C_t$  is the number of observed cases. Then, the mean susceptible population  $\bar{S}$  can be estimated using profile likelihood with the TSIR regression.

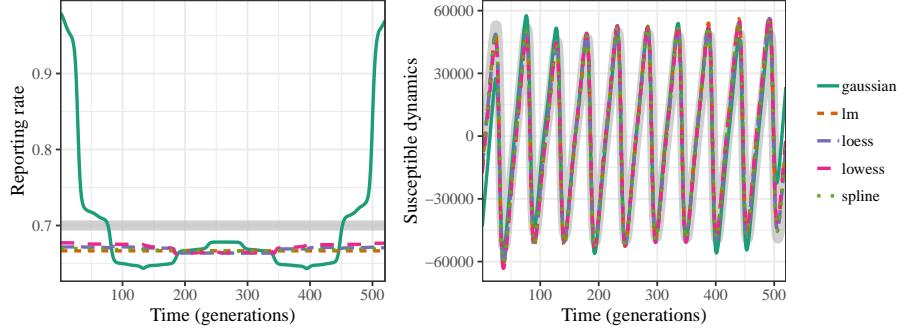


Figure 2: [SWP: TODO]

While this method of reconstructing the susceptible dynamics has been successfully used with the TSIR method, several questions remain to be answered. First, how sensitive is our inference to regression methods? Second, does  $Z_t$  represent deviation from the overall population mean  $\bar{S}$  or from some sort of moving average (assuming that population changes over time)? Likewise, how does changes in population size or reporting rate over time affect our estimate of  $\rho_{t+1}$  or  $Z_t$ ? Finally, how do assumptions about the population size affect estimate of the transmission rate?

We first simulate an epidemic using the continuous deterministic SIR model for 10 years and apply 5 different regression models available from the `tsiR` package to test their ability to infer reporting rate and susceptible dynamics (Fig. ??). We find that estimates of transmission rates are highly sensitive to regression methods. The gaussian regression is particularly inaccurate in boundaries. Other methods provide similar estimates. In general, estimates of susceptible dynamics appear to be robust regardless of the regression method.

Then, we compare estimates of susceptible dynamics and reporting rate under different scenarios (Fig. 3). We allow for population size to stay constant or increase over time and also allow for reporting rate to stay constant or increase over time; these combinations yields 4 different scenarios. We find that estimating time-varying reporting rate is difficult. We also find that  $Z_t$  estimated from regression better matches  $S_t - \bar{S}$  rather than  $S_t - \sigma N$ , i.e., any long term dynamics in the susceptible population is captured in  $Z_t$ .

### 3.3 Density dependence

In a continuous time SIR model, density dependence in the infection process is captured in the nonlinear infection term  $\beta SI/N$ . As the number of infected individuals increases, competition among infected individuals to find a susceptible host also increases, which results in saturation in the probability of infection. The most common way to account for density dependence is to approximate the probability of infection a hazard function:  $1 - \exp(-\phi \Delta t)$  where a constant force

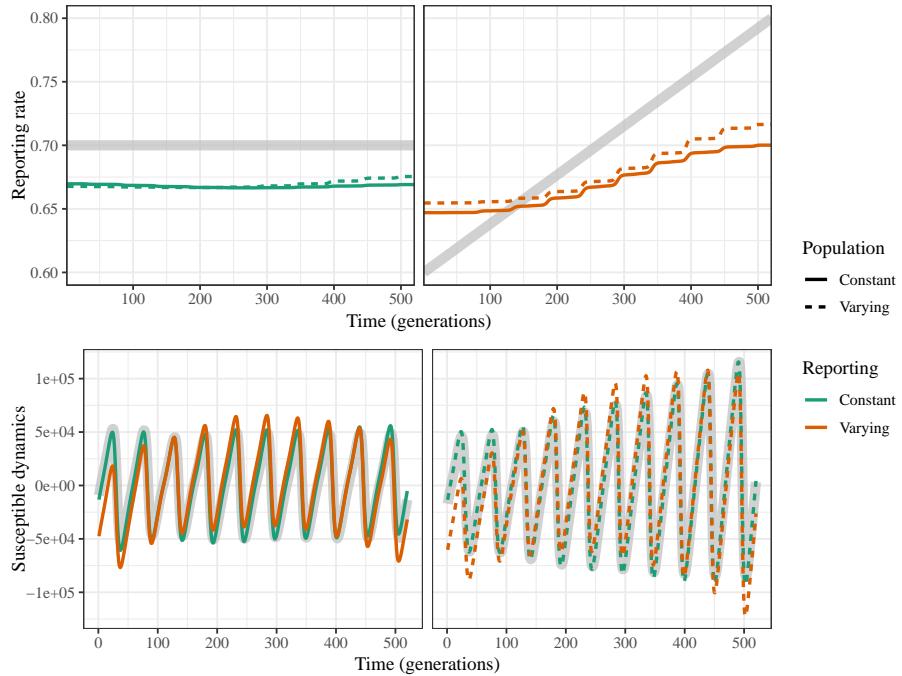


Figure 3: [SWP: TODO]

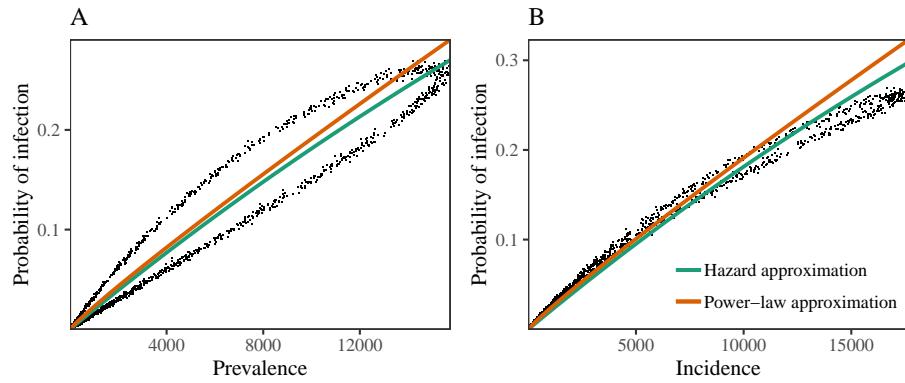


Figure 4: **Density dependence of the infection process.** A: probability of infection over a generation as a function of prevalence. B: probability of infection over a generation as a function of incidence. Probability of infection is calculated from 100 gillespie simulations. Hazard approximation curve is calculated using the true values. Power-law approximation is calculated by fitting the TSIR model to 100 simulations.

of infection  $\phi = \beta I/N$  is assumed over  $\Delta t$ . While this approximation captures the *average* probability of infection fairly well, the *realized* probability of infection deviates from this approximation (Fig. 4A). When the epidemic is growing, the force of infection increases over time and the probability of infection is higher than on average for a given value of prevalence  $I$ . When the epidemic is dying out, the force of infection decreases over time and the probability of infection is lower than on average for a given value of prevalence  $I$ . Modeling infection this way is still useful because it provides a probabilistic interpretation of the transition between states, which can be extended to a stochastic model.

On the other hand, the TSIR model assumes that the probability of infection follows a power function of previous incidence:  $\hat{\beta}i^\alpha/N$ , where  $\alpha$  has been referred to as the correction factor for heterogeneity or discretization of a continuous model (Glass et al., 2003). Glass et al. (2003) further derived an analytical expression for the optimal value of  $\alpha$  as a function of birth rate and the basic reproductive number by comparing the second order Taylor expansion of the TSIR model to that of a continuous time model. In the absence of susceptible recruitment to the population, the optimal value of  $\alpha$  based on their approximation is unity, which has been known to be the condition for homogeneous mixing. This is false.

Instead, we propose that  $\alpha$  should be interpreted as the strength of density dependence. Smaller values of  $\alpha$  captures stronger density dependence and faster saturation in the probability of infection. Under this interpretation, it is clear that  $\alpha$  should always be less than unity even in a homogeneously mixing population. Fitting the TSIR model to a simulated epidemic confirms this prediction (Fig. 4B). For a single epidemic with  $\mathcal{R}_0 = 2$ , we estimate the optimal value for  $\alpha$  to be 0.92 and  $\hat{\beta}$  to be 3.9.

Finally, high estimate of  $\hat{\beta}$ , in comparison to the true basic reproductive number  $\mathcal{R}_0$ , suggests that the raw estimate of the transmission rate  $\hat{\beta}$  from the TSIR regression should not be taken as an estimate of the reproductive number. Instead, we can translate this estimate to a transmission rate of a continuous-time SIR model by matching the estimated probability infection from the TSIR model,  $\hat{\beta}i^\alpha/N$ , with the hazard function,  $1 - \exp(-\beta I/N)$ , at mean incidence,  $\bar{i}$ , assuming that incidence and prevalence are equivalent:

$$\beta_{\text{continuous}} \approx -\frac{N}{\bar{i}} \log \left( 1 - \hat{\beta} \bar{i}^\alpha / N \right). \quad (19)$$

Then, we get  $\beta_{\text{continuous}} = 2.1$ , which is much closer to the true value.

### 3.4 Simulation study: comparison of statistical properties

We compare performance of each method by conducting a simulation study. We simulate 100 stochastic epidemic time series with fixed transmission rate without natural birth or death. [SWP: Parameter and rationale]

Results of each method is summarized into three summary statistics: coverage probability, bias, and root mean squared error (RMSE). Coverage probability is the proportion of confidence intervals that contain the true value over

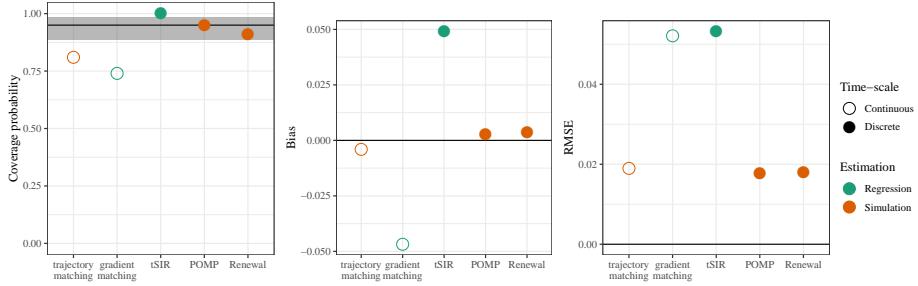


Figure 5: add me.

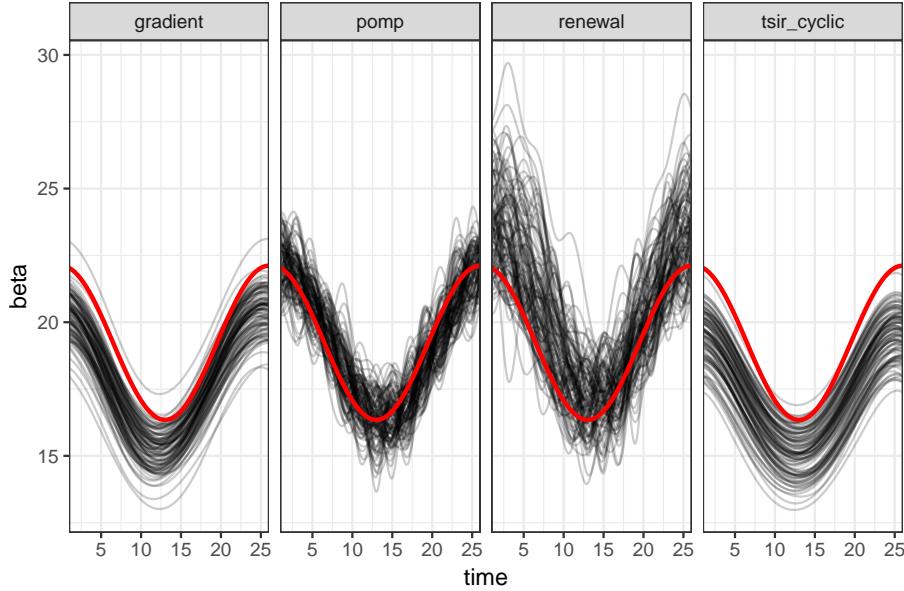
repeated experiments. For example, the 95% confidence interval should contain the true value 95% of the time. Bias and RMSE are computed on a log scale.

We find that the Monte Carlo methods (whether the model is represented as a compartmental model or a renewal equation) performs best (Fig. 5). Estimates of the transmission rates based on Sequential Monte Carlo attains nominal coverage and has least bias and RMSE. However, in order for the renewal equation model to attain nominal coverage, the simulation time step has to be as fine as one twentieth of the mean generation length; simulating an epidemic at one tenth of the mean generation length results in [SWP: 70%] coverage ([SWP: see Appendix]). While trajectory matching allows us to fit a continuous-time model without having to rely on a discretization process, it attains a slightly lower coverage than the nominal probability.

Both regression-based methods (gradient matching and TSIR regression) perform similarly in terms of RMSE. Gradient matching and TSIR regression give slightly biased answers in opposite directions but they seem to be negligible. Surprisingly, the TSIR regression attains 100% coverage if we assume that we know the exact susceptible dynamics. However, the true coverage is likely to be lower because susceptible reconstruction method will not be able to recover the *exact* dynamics. Low coverage of the gradient matching method is likely driven by not accounting for uncertainty during the initial smoothing process; we do not explore this idea here. Finally, we note the the summary statistics for the TSIR method are based on corrected estimates; using raw estimates of the transmission rate obtained from the TSIR regression, without correcting for discretization, yields an estimate with strong bias and 0% coverage.

### 3.5 Simulation study: estimating time-varying transmission rates

We simulate 100 stochastic epidemic time series with sinusoidal transmission rate and test each method's ability to recover the true shape of the transmission rate. We excluded trajectory matching from this analysis because fitting a deterministic model to a time series that is driven by highly nonlinear and



stochastic system is difficult and not likely to result in useful inference. We find that both regression-based methods are able to recover the true shape of the transmission rate reasonably well but give biased estimates. TSIR method yields a slightly more variable estimates; this is likely to be driven by inaccuracy in the estimates of the reporting rate.

Monte Carlo methods yield more variable estimates. Estimates follow a sinusoidal shape (higher transmission rates near the boundaries and lower transmission rates near the middle) but are too wiggly. Fitting a renewal equation yields much more variable estimates of the transmission rates because the model allows for extra uncertainty in the system by trying to estimate the variance of the generation-interval distribution. On one hand, these results suggest that small changes in the transmission rate are unidentifiable in the presence of process error; in fact, fitting a TSIR model without using splines also estimates a wiggly transmission rate. On the other hand, we simply may not have enough power to estimate the true shape of the transmission rates without constraining the shape.

## 4 Discussion

## 5 Appendix

### 5.1 Equivalence of incidence, mortality, and prevalence

First, recall that mortality case can be written as

$$\int_{t-t_{\text{rep}}}^t \gamma I ds. \quad (20)$$

When reporting period is equal to mean generation time, i.e.,  $t_{\text{rep}} = 1/\gamma$ , it follows that:

$$\int_{t-t_{\text{rep}}}^t \gamma I ds \approx \gamma I(t - t_{\text{rep}}) t_{\text{rep}} = I(t - t_{\text{rep}}) \quad (21)$$

Likewise, we can derive a similar relation for incidence:

$$\begin{aligned} I(t + t_{\text{rep}}) &= I(t) + \int_t^{t+t_{\text{rep}}} \frac{dI(s)}{ds} ds \\ &= I(t) + \int_t^{t+t_{\text{rep}}} \beta(s) S \frac{I}{N} ds - \int_t^{t+t_{\text{rep}}} \\ &\approx \int_t^{t+t_{\text{rep}}} \beta(s) S \frac{I}{N} ds \end{aligned} \quad (22)$$

Here, we assumed that individuals leaving infected class  $I$  from natural mortality is negligible over the reporting period.

[SWP: TODO]

## 6 Fitting continuous models

### 6.1 Gradient matching

Conditions:

- See Jost and Ellner (???)
- sampled sufficiently
- not too much error
- Brunel Parameter estimation of ODE's via nonparametric estimators

### 6.2 Spectral matching

What is this? Read Reuman et al. 2006

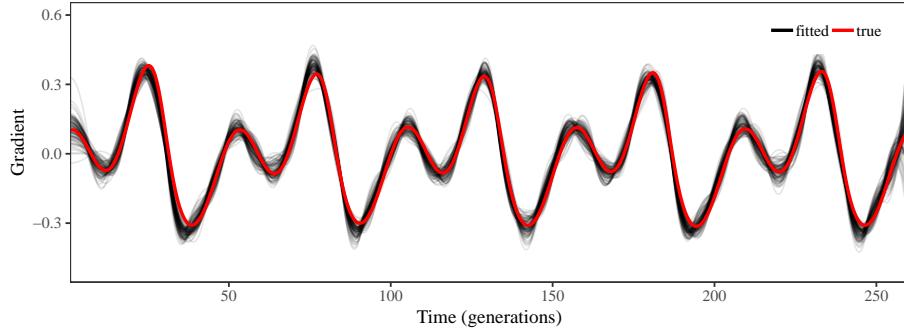


Figure 6: This is not a caption.

## References

- Althaus, C. L. (2014). Estimating the reproduction number of ebola virus (ebov) during the 2014 outbreak in west africa. *PLoS currents* 6.
- Andrieu, C., A. Doucet, and R. Holenstein (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(3), 269–342.
- Bjørnstad, O. N., B. F. Finkenstädt, and B. T. Grenfell (2002). Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs* 72(2), 169–184.
- Bolker, B. M. (2008). *Ecological models and data in R*. Princeton University Press.
- Bretó, C., D. He, E. L. Ionides, A. A. King, et al. (2009). Time series analysis via mechanistic models. *The Annals of Applied Statistics* 3(1), 319–348.
- Champredon, D., J. Dushoff, and D. Earn (2018). Equivalence of the Erlang SEIR epidemic model and the renewal equation. *BioRxiv*, 319574.
- Champredon, D., M. Li, B. M. Bolker, and J. Dushoff (2018). Two approaches to forecast Ebola synthetic epidemics. *Epidemics* 22, 36–42.
- Chowell, G., N. W. Hengartner, C. Castillo-Chavez, P. W. Fenimore, and J. M. Hyman (2004). The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *Journal of theoretical biology* 229(1), 119–126.
- Dalziel, B. D., O. N. Bjørnstad, W. G. van Panhuis, D. S. Burke, C. J. E. Metcalf, and B. T. Grenfell (2016). Persistent chaos of measles epidemics in the prevaccination United States caused by a small change in seasonal transmission patterns. *PLoS computational biology* 12(2), e1004655.

- Didelot, X., L. K. Whittles, and I. Hall (2017). Model-based analysis of an outbreak of bubonic plague in Cairo in 1801. *Journal of The Royal Society Interface* 14(131), 20170160.
- Earn, D. J., P. Rohani, B. M. Bolker, and B. T. Grenfell (2000). A simple model for complex dynamical transitions in epidemics. *Science* 287(5453), 667–670.
- Ellner, S. P., Y. Seifu, and R. H. Smith (2002). Fitting population dynamic models to time-series data by gradient matching. *Ecology* 83(8), 2256–2270.
- Finkenstadt, B. F. and B. T. Grenfell (2000). Time series modelling of childhood diseases: a dynamical systems approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 49(2), 187–205.
- Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics* 22(4), 403–434.
- Glass, K., Y. Xia, and B. Grenfell (2003). Interpreting time-series analyses for continuous-time biological models—measles as a case study. *Journal of theoretical biology* 223(1), 19–25.
- He, D., J. Dushoff, T. Day, J. Ma, and D. J. Earn (2011). Mechanistic modelling of the three waves of the 1918 influenza pandemic. *Theoretical Ecology* 4(2), 283–288.
- He, D., E. L. Ionides, and A. A. King (2009). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of the Royal Society Interface*.
- Hempel, K. and D. J. Earn (2015). A century of transitions in New York City’s measles dynamics. *Journal of The Royal Society Interface* 12(106), 20150024.
- Hooker, G., S. P. Ellner, L. D. V. Roditi, and D. J. Earn (2010). Parameterizing state-space models for infectious disease dynamics by generalized profiling: measles in Ontario. *Journal of The Royal Society Interface* 8(60), 961–974.
- Ionides, E. L., A. Bhadra, Y. Atchadé, A. King, et al. (2011). Iterated filtering. *The Annals of Statistics* 39(3), 1776–1802.
- Ionides, E. L., C. Bretó, and A. A. King (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* 103(49), 18438–18443.
- Ionides, E. L., D. Nguyen, Y. Atchadé, S. Stoev, and A. A. King (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences* 112(3), 719–724.

- Kermack, W. O. and A. G. McKendrick (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115(772), 700–721.
- King, A. A., D. Nguyen, and E. L. Ionides (2015). Statistical inference for partially observed markov processes via the r package pomp. *arXiv preprint arXiv:1509.00503*.
- Krylova, O. and D. J. Earn (2013). Effects of the infectious period distribution on predicted transitions in childhood disease dynamics. *Journal of The Royal Society Interface* 10(84), 20130098.
- Levins, R. (1966). The strategy of model building in population biology. *American scientist* 54(4), 421–431.
- Li, M., J. Dushoff, and B. M. Bolker (2018). Fitting mechanistic epidemic models to data: a comparison of simple Markov chain Monte Carlo approaches. *Statistical methods in medical research* 27(7), 1956–1967.
- Morris, W. F. (1997). Disentangling effects of induced plant defenses and food quantity on herbivores by fitting nonlinear models. *The American Naturalist* 150(3), 299–327.
- Riley, S., C. Fraser, C. A. Donnelly, A. C. Ghani, L. J. Abu-Raddad, A. J. Hедley, G. M. Leung, L.-M. Ho, T.-H. Lam, T. Q. Thach, et al. (2003). Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science* 300(5627), 1961–1966.
- Wood, S. (2012). mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation.

## 7 Appendix

### 7.1 Derivation of growth rate in discrete-time SIR model

Derivation:

$$\begin{aligned} S_{t+\Delta t} &= S_t - S_t(1 - \exp(-\beta I_t \Delta t / N)) \\ I_{t+\Delta t} &= S_t(1 - \exp(-\beta I_t \Delta t / N)) + I_t - I_t(1 - \exp(-\gamma \Delta t)) \end{aligned} \quad (23)$$

Assuming that  $S_t$  is approximately equal to  $N$ , we get

$$I_{t+\Delta t} = N_t(1 - \exp(-\beta I_t \Delta t / N)) + I_t \exp(-\gamma \Delta t) \quad (24)$$

When  $I_t \approx 0$ ,  $\exp(-\beta I_t \Delta t / N) \approx 1 - \beta I_t \Delta t / N$  and so

$$\begin{aligned} I_{t+\Delta t} &\approx I_t \beta \Delta t + I_t \exp(-\gamma \Delta t) \\ &= I_t(\beta \Delta t + \exp(-\gamma \Delta t)) \end{aligned} \quad (25)$$

Substituting  $\hat{\gamma}$ , we get

$$I_{t+\Delta t} = I_t \left(1 + \beta \Delta t - \frac{\Delta t}{\mu}\right), \quad (26)$$

Therefore,

$$I_t = I_0 \left(1 + \beta \Delta t - \frac{\Delta t}{\mu}\right)^{t/\Delta t} \quad (27)$$

and the initial growth rate is given by

$$r = \frac{1}{\Delta t} \log \left(1 + \beta \Delta t - \frac{\Delta t}{\mu}\right). \quad (28)$$

### 7.2 Probability of infection

In both the TSIR model and the hazard-based model, transition from the susceptible compartment,  $S$ , to the infected compartment,  $I$ , is represented as a product of number of susceptible individuals and probability of infection between two time steps  $t$  and  $t + \Delta t$ . The TSIR model assumes that the probability of infection is a linear function of prevalence  $I_t$ . This formulation is based on the Euler approximation to the solution of the ordinary differential equation. On the other hand, the hazard-based model assumes that the probability of infection is an inverse exponential function of prevalence  $I_t$ . Besides their differences in the relationship between prevalence and probability of infection, there are two more assumptions that we need to consider: (1) force of infection remains constant over two time steps and (2) new susceptible individuals have zero probability of infection between the two time steps.

### 7.3 Infectious period

Geometric distribution with probability of  $1 - \exp(-\gamma\Delta t)$  and unit of  $\Delta t$ . Then, mean infectious period and generation interval is given by

$$\frac{\Delta t}{1 - \exp(-\gamma\Delta t)}. \quad (29)$$

An important but often overlooked component is the variance of the distribution. Squared coefficient of variation for this distribution is equal to  $\exp(-\gamma\Delta t)$ , which necessarily depends on pre-specified time-step  $\Delta t$ . If we want to match the mean of the distribution to a fixed value  $\mu$  regardless of our choices of  $\Delta t$ , we obtain

$$\hat{\gamma} = -\frac{1}{\Delta t} \log \left( 1 - \frac{\Delta t}{\mu} \right) \quad (30)$$

Then,

$$\text{CV}_{\text{Infectious period}}(\Delta t)^2 = 1 - \frac{\Delta t}{\mu}. \quad (31)$$

We expect changes in variation in infectious period to affect variation in stochastic realizations.

This means that the relationship between  $r$  and  $\mathcal{R}$  changes. For this generation-interval distribution, we have

$$\mathcal{R} = 1 + \frac{\exp(r) - 1}{(1 - \exp(-\gamma\Delta t))} \quad (32)$$

Substituting

$$r = \frac{1}{\Delta t} \log \left( 1 + \beta\Delta t - \frac{\Delta t}{\mu} \right), \quad (33)$$

as well as  $\hat{\gamma}$ , we get

$$\mathcal{R} = 1 + \frac{\mu}{\Delta t} \left( 1 + \beta\Delta t - \frac{\Delta t}{\mu} \right)^{1/\Delta t} \quad (34)$$

Then, we can obtain a relationship between  $\mathcal{R}$  of a discrete-time system and that of a corresponding continuous time system, assuming that contact rate  $\beta$  and mean generation-time  $\mu$  are the same:

$$\mathcal{R}_{\text{discrete}} = 1 + \frac{\mu}{\Delta t} \left( 1 + \frac{(\mathcal{R}_{\text{continuous}} - 1)\Delta t}{\mu} \right)^{1/\Delta t} \quad (35)$$

On the other hand, we may want to match choose  $\beta$  and  $\mu$  to match true  $r$  and  $\mathcal{R}$ .