Thesis

Sang Woo Park

February 11, 2019

1 Introduction

Fitting models. Too many methods and too many models. What should we do?

2 SIR model

The susceptible-infected-recovered (SIR) model is one of the simplest epidemic models which describe how disease spreads in a homogeneously mixing population. In general, the SIR model can be written as

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

where S, I, and R describe the number of susceptible, infected, and recovered individuals. $b(t)$ represents recruitment rate to the susceptible population, μ represents natural mortality rate, and γ represents per-capita recovery rate. $\phi(t)$ represents force of infection, which is defined as the rate at which a susceptible individual acquires infection. We write $\phi(t) = \beta(t)I/N$ to represent *frequency*-dependent action or $\phi(t) = \beta I$ to represent *density*-dependent action, where $N = S + I + R$ is the total population size. For this study, we focus on $\phi(t) = \beta(t)I/N$.

Typically, data suffers from under-reporting and we may want to understand the dynamics of ρI instead, where ρ is the reporting rate. Then, we can write

$$\begin{aligned} \frac{dS}{dt} &= b(t) - \beta S\frac{\hat{I}}{\rho N} \\ \frac{d\hat{I}}{dt} &= \beta S\frac{\hat{I}}{N} - \gamma\hat{I} \end{aligned} \tag{2}$$

2.1 Incidence, prevalence, and mortality

We want to distinguish among different kinds of data: prevalence, incidence, and mortality. Prevalence is defined as the total number of infected individuals present in the population and corresponds to $\hat{I}(t)$. Incidence is defined as the number of newly infected cases. Hence, incidence report at time t corresponds to the integral of the total infection rate from the previous reporting period $t - \Delta t$ and current reporting period t:

$$\int_{t-\Delta t}^{t} \beta(s)S\frac{\hat{I}}{N}ds. \tag{3}$$

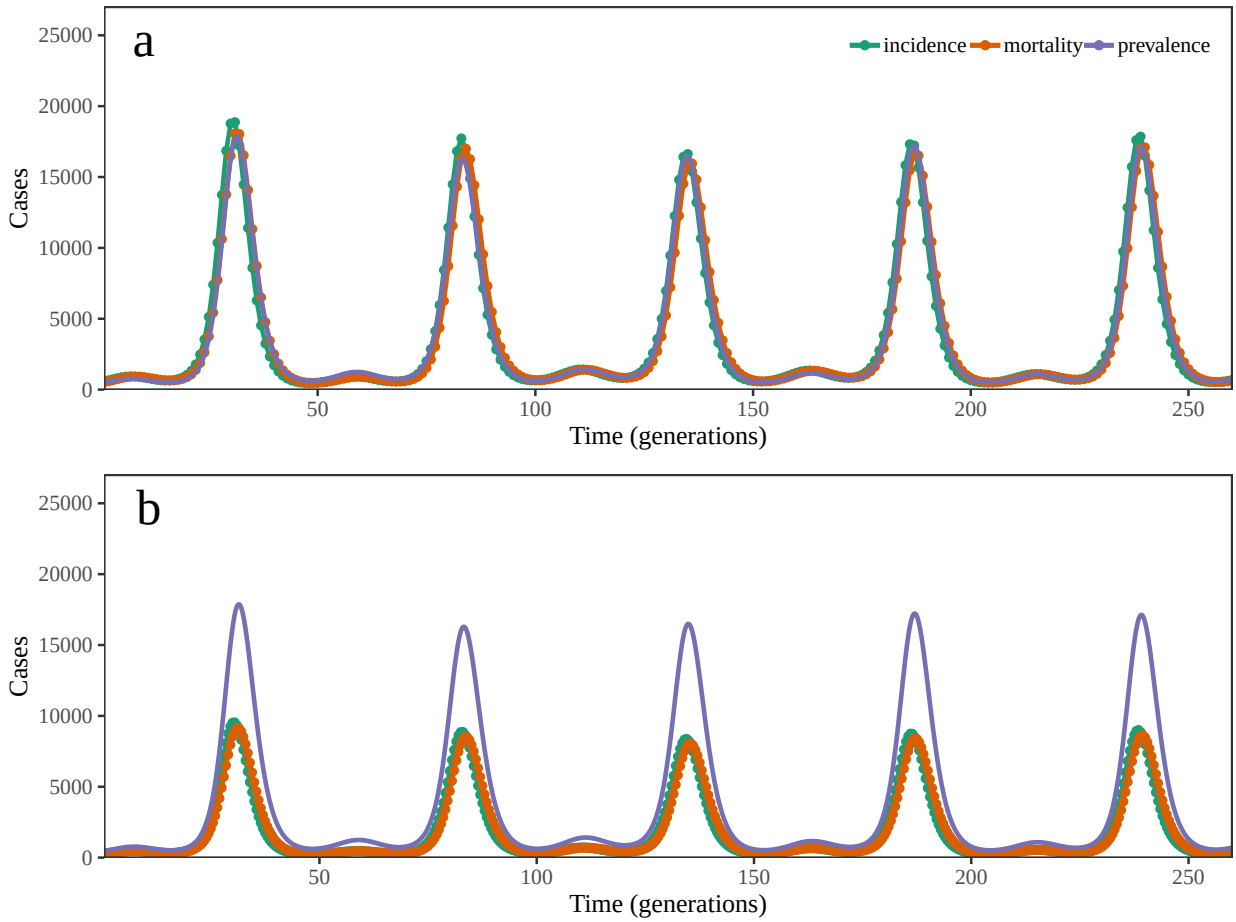


Figure 1: This is not a caption.

Finally, mortality is defined as the number of individuals that died over a time period. Like incidence, mortality can be written as the integral of the total death rate:

$$\int_{t-\Delta t}^{t} \gamma \hat{I} ds. \tag{4}$$

The main difference among the three types is that prevalence counts provide a direct observation of a state variable whereas incidence and mortality counts do not. As a result,

In this article, we focus on fitting to incidence data only. In particular, we consider two kinds of models: continuous deterministic models and discrete stochastic models. We do not consider fitting continuous stochastic models because they are computationally too expensive.

3 Fitting continuous models

3.1 Trajectory matching

Trajectory matching is one of the simplest fitting methods. As its name suggests, the goal is to find a set of parameters such that the simulated trajectory, from the ODE model, best matches the observed data. This method assumes that there is only measurement error and no process error. The measurement model can

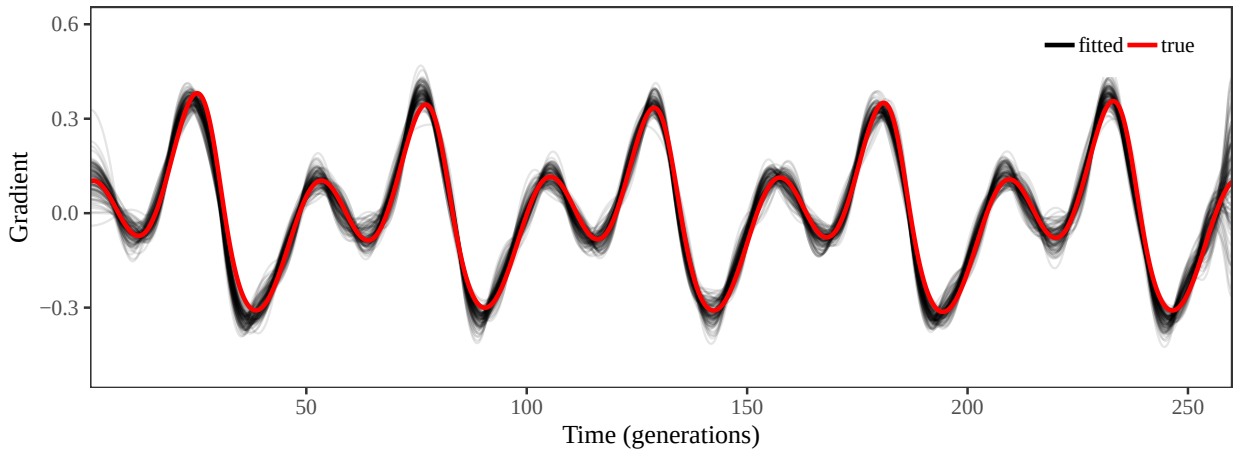


Figure 2: This is not a caption.

be written as:

$$y_t \sim \text{NegBinomial}\left(\mu_t = \int_{t-\Delta t}^{t} \beta(s) S \frac{\hat{I}}{N} ds, \phi\right), \tag{5}$$

where ρ is reporting rate and ϕ is dispersion parameter. We prefer this model compared to the binomial or beta-binomial model because fitting binomial or beta-binomial models may not work if there are any irregular patterns in the data.

3.2 Gradient matching

When *all* state variables are observed, we can fit smooth curves to the data and obtain an estimate of the gradients of the ODE by taking the derivatives of the estimated smooth functions. Gradient matching methods seek to estimate the rate equation using nonparametric regression. However, for most epidemics, it is impossible to have data of all state variables. When only incidence is provided, we have no direct observations of any state variables. If we can reconstruct the dynamics of the susceptible population and assume that incidence is sufficiently similar to prevalence, we can still be able to apply gradient mathing methods.

Observe that

$$\frac{d \log \hat{I}}{dt} = \beta(t) \frac{S}{N} - (\gamma + \mu), \tag{6}$$

If we can obtain a reasonable estimate of the gradient $Z_i = d \log \hat{I}/dt(t_i)$, we can write

$$Z_i + (\gamma + \mu) = \exp\left(\log \beta(t_i) + \log S_i - \log N\right) + \epsilon_i \tag{7}$$

and estimate β, given that we can reconstruct the susceptible dynamics S_i. Estimate of β can be done using nonparametric regression by treating $\log S_i$ and $\log N$ as offset terms.

First, we want to test whether gradients can be reliably estimated from incidence data alone by simulating a deterministic SIR model with measurement error. We use natural cubic spline with knots placed every 6 biweeks. It works generally well but there is systematic bias. When gradients are at their local maxima (and minima), we tend to overestimate (and underestimate) the gradients. This is because incidence precedes prevalence???

If we simply construct confidence intervals based on equation ?? we underestimate uncertainty significantly by ignoring the uncertainty in the estimated gradients Z_i.

Conditions:

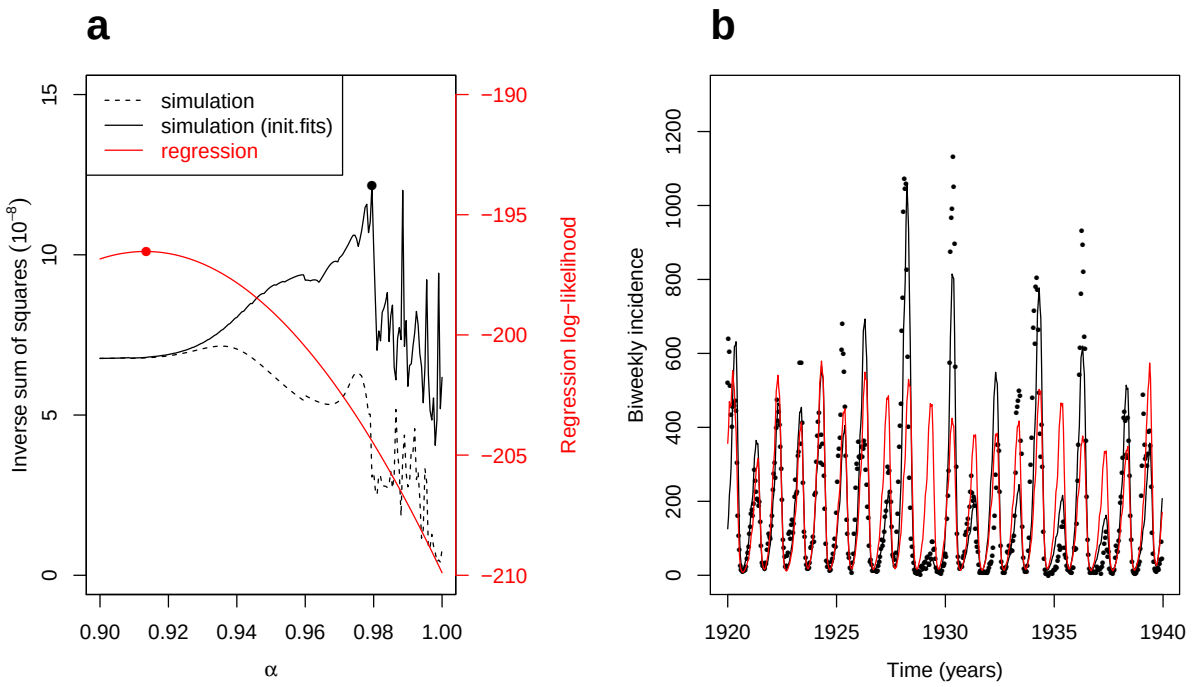


Figure 3: This is not a caption.

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

3.3 Generalized profiling

Generalized profiling combines trajectory matching with gradient matching.

3.4 Spectral matching

What is this? Read Reuman et al. 2006

4 Fitting discrete time models

4.1 Time-series SIR

The time-series SIR (TSIR) method was introduced by Bjørnstad et al. (2002) to estimate seasonally varying transmission rates of measles. It relies on reconstructing the susceptible dynamics from birth and case reports and using linear regression, conditional on knowing "true susceptible dynamics", to estimate transmission rates. Initial conditions are often estimated to improve goodness of fit.

The TSIR method begins by discretizing the infection process, assuming that disease generation-time is equal to reporting interval:

$$S_{t+1} = B_t + S_t - I_{t+1}I_{t+1} \quad = \beta_t S_t \frac{I_t^\alpha}{N_t} \tag{8}$$

where α is often referred to as a conversion factor from continuous-time model to discrete-time model or a heterogeneity parameter [CITE]. True incidence, I_t, is inferred from the observed incidence, y_t, and estimated reporting rate ρ_t: $I_t = y_t/\rho_t$. Taking log on both sides, estimation of transmission rates β_t becomes a regression problem:

$$\log I_{t+1} = \log \beta_t + \log S_t + \alpha \log I_t - \log N_t + \epsilon_t. \tag{9}$$

The estimation process is deterministic, but the TSIR method is analogous to fitting a stochastic model: equation ?? models step-ahead prediction as a function of previous state and process error, ϵ_t. Coupled with a deterministic susceptible reconstruction and estimation of reporting rate, the TSIR method attributes all errors to one-step process noise and oversmooths the likelihood surface (figure ?). Several studies have associated dynamics of the deterministic model (equation ??) with the inferred parameters from the regression but they are intrinsically different models.

4.2 Particle filtering

Particle filtering provides a way of approximating the likelihood for stochastic models. In this paper, we focus on mif2 method.

$$\begin{aligned} S_{t+\Delta t} &= B_t + S_t - S_t \left(1 - \exp\left(-\beta_t \frac{I_t}{N_t}\Delta t\right)\right) \\ I_{t+\Delta t} &= S_t \left(1 - \exp\left(-\beta_t \frac{I_t}{N_t}\Delta t\right)\right) + I_t - I_t(1 - \exp(-(\gamma + \mu)\Delta t)) \end{aligned} \tag{10}$$

Tries to account for variation in ... but still ... When are these assumptions appropriate?

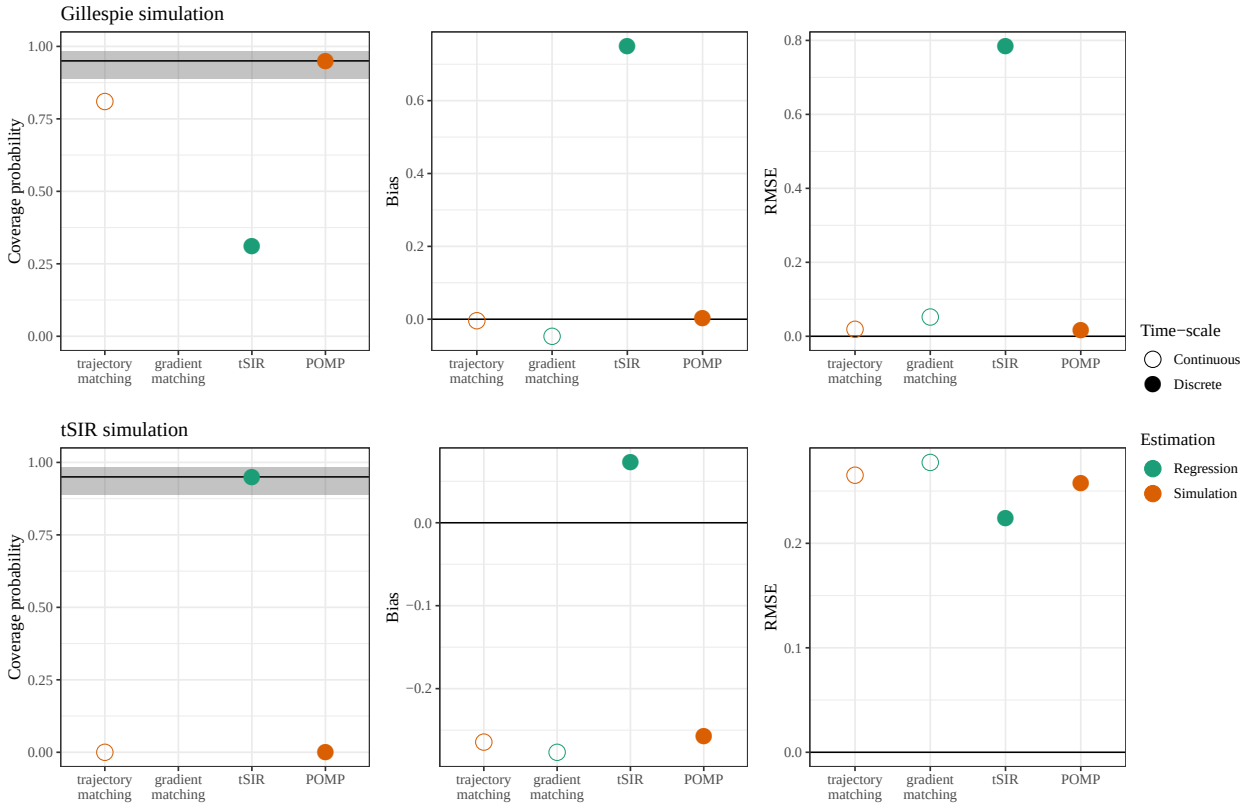
4.3 MCMC

Li et al. (2018) compared incidence-based models across several Bayesian platforms. In our case, we can write a similar model as:

$$\begin{aligned} S_{t+\Delta t} &= B_t + S_t - I_{t+\Delta t} \\ \phi_t &= \sum_{i=1}^{\ell} k(i) I_{t-\ell+i} \\ I_{t+\Delta t} &\sim \text{BetaBin}(1 - e^{-\phi_t}, S_t, \delta_P) \\ \text{Obs}_t &\sim \text{NegBin}(P_{rep}, \sum I_t, \delta_{obs}) \end{aligned} \tag{11}$$

When reporting period is same as the mean generation time, we want to simulate the model on a finer scale.

Disadvantage: unable to model mechanisms; difficult to deal with prevalence or mortality data. Advantage: flexible generation time distribution.



4.4 Probe-based methods

5 Results

5.1 Statistical inference

References

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.

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.

6 Appendix

6.1 Derivation of growth rate in discrete-time SIR model

Derivation:

$$\begin{aligned} S_{t+\Delta t} &= S_t - S_t(1 - \exp(-\beta I_t \Delta t/N)) \\ I_{t+\Delta t} &= S_t(1 - \exp(-\beta I_t \Delta t/N)) + I_t - I_t(1 - \exp(-\gamma \Delta t)) \end{aligned} \tag{12}$$

Assuming tat S_t is approximately equal to N, we get

$$I_{t+\Delta t} = N_t(1 - \exp(-\beta I_t \Delta t/N)) + I_t \exp(-\gamma \Delta t) \tag{13}$$

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

$$\begin{aligned} I_{t+\Delta t} &\approx I_t \beta \Delta t + I_t \exp(-\gamma \Delta t) \\ &= I_t(\beta \Delta t + \exp(-\gamma \Delta t)) \end{aligned} \tag{14}$$

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

$$I_{t+\Delta t} = I_t \left(1 + \beta \Delta t - \frac{\Delta t}{\mu}\right), \tag{15}$$

Therefore,

$$I_t = I_0 \left(1 + \beta \Delta t - \frac{\Delta t}{\mu}\right)^{t/\Delta t} \tag{16}$$

and the initial growth rate is given by

$$r = \frac{1}{\Delta t} \log \left(1 + \beta \Delta t - \frac{\Delta t}{\mu}\right). \tag{17}$$

6.2 Probability of infection

In both the TSIR model and the hazard-based model, transition from the susceptible comparment, 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.

6.3 Infectious period

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

$$\frac{\Delta t}{1 - \exp(-\gamma \Delta t)}. \tag{18}$$

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 Δt. If we want to match the mean of the distribution to a fixed value μ regardless of our choices of Δt, we obtain

$$\hat{\gamma} = -\frac{1}{\Delta t} \log \left(1 - \frac{\Delta t}{\mu}\right) \tag{19}$$

Then,

$$\mathrm{CV}_{\text{Infectious period}}(\Delta t)^2 = 1 - \frac{\Delta t}{\mu}. \tag{20}$$

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

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

$$\mathcal{R} = 1 + \frac{\exp(r) - 1}{(1 - \exp(-\gamma \Delta t))} \tag{21}$$

Substituting

$$r = \frac{1}{\Delta t} \log\left(1 + \beta \Delta t - \frac{\Delta t}{\mu}\right), \tag{22}$$

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

$$\mathcal{R} = 1 + \frac{\mu}{\Delta t}\left(1 + \beta \Delta t - \frac{\Delta t}{\mu}\right)^{1/\Delta t} \tag{23}$$

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 β and mean generation-time μ are the same:

$$\mathcal{R}_{\text{discrete}} = 1 + \frac{\mu}{\Delta t}\left(1 + \frac{(\mathcal{R}_{\text{continuous}} - 1)\Delta t}{\mu}\right)^{1/\Delta t} \tag{24}$$

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