Inferring generation interval distributions from contact tracing data

Sang Woo Park, David Champredon and Jonathan Dushoff

September 3, 2018

1 Introduction

2 Results

An epidemic can be characterized by its speed (exponential growth rate, r) and its strength (reproductive number, $\mathcal{R}$). Reproductive number, defined as the average number of secondary cases arising from a primary case, is of particular interest as it provides information about the final size of an epidemic [CITE]. However, directly measuring the reproductive number often requires knowledge of entire disease history and may not be feasible early in an epidemic [CITE]. Instead, the reproductive number can be *indirectly* estimated from exponential grwoth rate, which can easily be estimated from incidence data [CITE]. These two quantities are linked by generation interval distributions (Wallinga and Lipsitch, 2007).

Generation interval is defined as the time between when a person becomes infected and when that person infects another person. As each individual experiences different course of infection, *individual* generation interval distribution varies among infectors [CITE sven]. Hence, the *intrinsic* generation interval distribution, which provides link between r and $\mathcal{R}$, is a pooled distribution across all potential infections.

Due to individual variation in infection time, the observed generation interval distribution can change depending on when and how it is measured [CITE champ]. Generation interval can be measured *forward* in time by looking at all infections that were caused by infectors that were infected at the same period of time. Early in the epidemic when depletion of susceptible is negligible, we expect the forward generation interval distribution to be similar to the intrinsic generation interval distribution. As epidemic progresses, an infector is less likely to infect individuals later in time due to decrease in susceptibles and the forward generation interval will be shorter, overall.

On the other hand, generation intervals can be measured *backward* in time by considering all infectees that were infected at the same time period and comparing when their infectors were infected. When epidemic is growing, a

susceptible individual is more likely to be infected by a newly infected individual and the backward generation interval will be short. When epidemic is subsiding, most infections are caused by the remaining infectors, rather than new infectors, and the backward generation interval will be long.

we can measure generation intervals from infector's point of view at a given time.

3 Methods

First, we start by

4 Generation interval

Let $K(t)$ be the infection kernel. Then, the reproduction number is defined as

$$\mathcal{R} = \int_0^\infty K(t).$$

The intrinsic generation interval distributions [CITE] can be written as

$$g(t) = \frac{K(t)}{\mathcal{R}}.$$

We can think of intrinsic generation interval distribution as an intrinsic characteristic of an infector, as suggested by its name. Here, we look at spatio-temporal components associated with generation interval distribution.

5 Contact tracing

Generation and serial intervals are often sampled through contact tracing. For simplicity, suppose contact tracing is performed during an outbreak from the beginning of an epidemic to some time point. Following Champredon and Dushoff (2015), we can write the number of infection occuring at time t caused by infectors who were themselves infected at time s as

$$i_s(t) = K(t-s)i(s)S(t) \tag{1}$$

Writing the kernel as the product of the intrinsic genreation distributions and $\mathcal{R}$, we get

$$i_s(t) = \mathcal{R}g(t-s)i(s)S(t) \tag{2}$$

We define generation intervals measured through contact tracing as the *right-censored generation interval distributions*. Note that number of infection occuring at time s caused by infectors who were themselves infected at time $s-\tau$ is given by

$$i_{s-\tau}(s) = \mathcal{R}i(s-\tau)g(\tau)S(s) \tag{3}$$

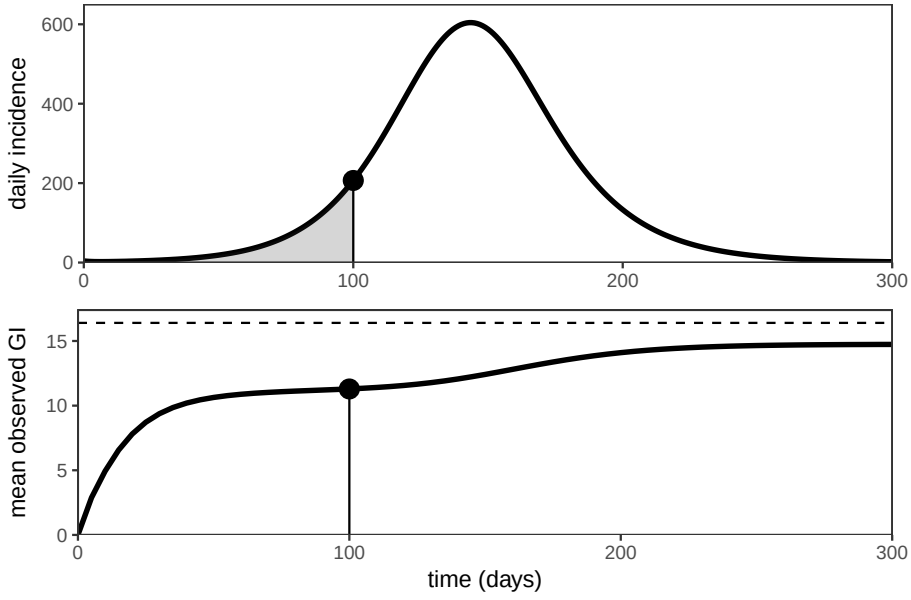

Figure 1: Fill this out.

Note we are interested in total number of secondary infections that are τ time steps apart and occur before time t:

$$\mathcal{R} \int_{\tau}^{t} i(s-\tau)g(\tau)S(s)ds. \tag{4}$$

Then, the censored generation interval at time t is given by

$$c_t(\tau) = \frac{\mathcal{R} \int_{\tau}^{t} i(s-\tau)g(\tau)S(s)ds}{\mathcal{R} \int_{0}^{t} \int_{x}^{t} i(s-x)g(x)S(s)dsdx}. \tag{5}$$

We note that the expression in the denominater is equivalent to cumulative incidence at time t. The intuition behind this is that we are normalizing acrosss all incidence before time t. Then, we have

$$c_t(\tau) = \frac{\mathcal{R} \int_{\tau}^{t} i(s-\tau)g(\tau)S(s)ds}{\int_{0}^{t} i(s)ds}. \tag{6}$$

For convenience, we ignore normalizing constants and write

$$c_t(\tau) \propto g(\tau) \int_{0}^{t} i(s-\tau)S(s)ds. \tag{7}$$

Note that the observed mean generation interval through contact tracing will always be shorter be shorter than intrinsic mean generation interval (Figure 1). There are two reasons for this. First, due to right censoring, we do

not observe generation intervals that are longer than the time at which contact tracing was performed. Second, number of susceptibles decrease over the course of an epidemic and any infector is less likely to infect someone through long generation intervals (Champredon and Dushoff, 2015). Due to the second reason, even if we could contact trace through an entire epidemic, we will still underestimate mean generation interval. In the following sections, we introduce different ways to recover intrinsic generation interval distributions from the right-censored genration interval samples.

5.1 Non-parametric approach

During an exponential growth period, we can write $i(\tau) \propto \exp(rt)$. Assuming that $S(t) \approx 1$, censored generation interval distribution can be written as follows:

$$g_{\text{obs}}(\tau) = \mathcal{R} g(\tau) \exp(-r\tau), \tag{8}$$

This is the right-censored generation interval distributions that we expect to observe during an early outbreak (growing at rate r). Hence, to recover the intrinsic generation interval distributions, we can take the weighted distribution of the observed distribution (Fig ???):

$$g(\tau) = \frac{1}{\mathcal{R}} g_{\text{exp}}(\tau) \exp(r\tau). \tag{9}$$

Furthermore, $\mathcal{R}$ should be estimated by

$$\mathcal{R} = \int_0^\infty g_{\text{obs}}(\tau) \exp(r\tau) d\tau \tag{10}$$

This contrasts with the well-known Euler-Lotka equation:

$$\mathcal{R} = 1 \Big/ \int_0^\infty g(\tau) \exp(-r\tau) d\tau \tag{11}$$

Using the censored generation interval without correction leads to underestimation of $\mathcal{R}$ as well as mean generation interval.

5.2 Parametric approach

Conditional likelihood...

The likelihood is given by

$$\mathcal{R}^{n_e} \cdot \prod g(\tau_e) \cdot \exp\left(-\mathcal{R} \int_0^{c - t_{\text{inf}}} g(s) ds\right)$$

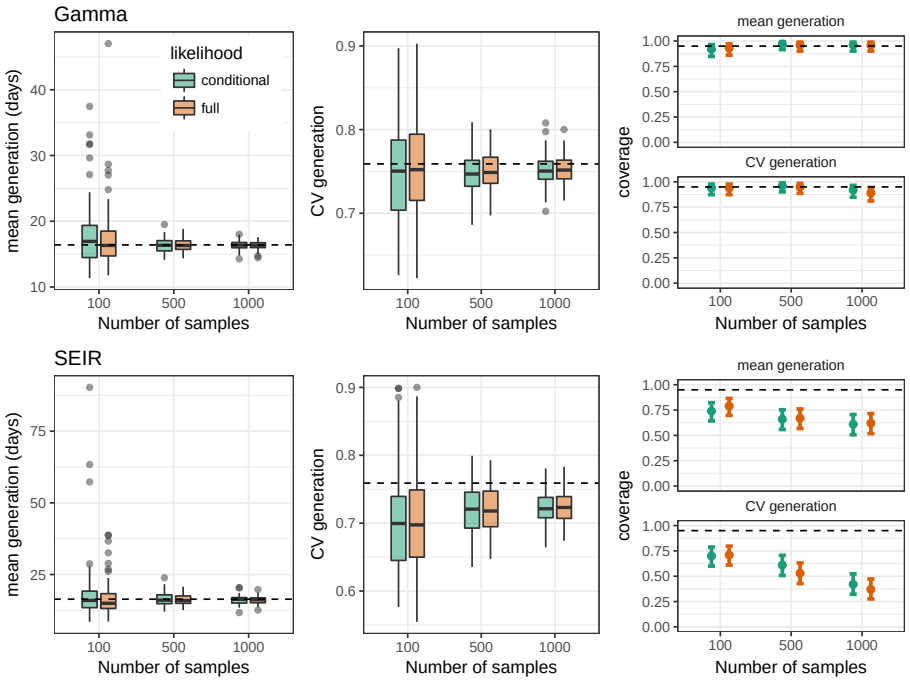

Figure 2: Fill this out as well.

6 Spatial variation - Effective generation interval

Intrinsic generation interval distribution implicitly that an infector can exert all infectious contacts without wasting any throughout the infectious period. In other words, it is conditional on the assumption that a contacted individual has not been contacted before. When the population is limited, we must take the probability that a susceptible individual can be found into account.

Let $\beta(t)$ be infectious contact rate per pair. The probability that a susceptible is still susceptible at time t is given by

$$\exp\left(-\int_0^t \beta(s)ds\right).$$

Then, the effective generation interval distribution of an infected individual is proportional to the product of intrinsic generation interval distribution and this survival probability

$$g_{\text{eff}}(\tau) \propto g(\tau)\exp\left(-\int_0^\tau \beta(s)ds\right).$$

Note that the previous formulation does not take into account presence of other potential infectors. During an outbreak, we can imagine a susceptible individual being exposed to multiple infected individuals. Since effective generation interval is conditional on the assumption that a contacted susceptible individual has not been contacted previously, we have to take this into account... Then, the previous formulation can be taken as an upper bound of the actual effective GI distribution... It is very difficult to formalize this idea but we use a numerical example to demonstrate the idea:

6.1 Numerical example

We can imagine contact structure being embedded in generation or serial intervals sampled through contact tracing. Then, when we apply temporal correction, we expect to obtain the effective generation interval distributions rather than the intrinsic generation interval distribution. In particular, we claim that we must use effective generation interval distribution in order to estimate $\mathcal{R}$.

[Example]

- Call it right censored
- time censored are different from backward but during exponential period, they are
- local correction and full correction
- 1) epidemic is growing, 2) I'm infecting people (local), 3) semi-local effect - sharing contact

- unsolved problem of stats
- relationship between R and beta
- gamma distributed
- fit at individual level (poisson level)
- we should focus on r-R
- Gamma-network simulation
- Coverage on ???
- SEmInR

7 Individual level bootstrap

Any GI that I observe, what proportion of infector were observed for at least that long

- JD observed 100 individuals - I infected Mike after 8 day - weight = 1/prop of ppl that were observed that long (whose censorship time is greater than the observed generation interval of that individual)

References

Champredon, D. and J. Dushoff (2015). Intrinsic and realized generation intervals in infectious-disease transmission. *282*(1821).

Wallinga, J. and M. Lipsitch (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society of London B: Biological Sciences 274*(1609), 599–604.