Inference of seasonal forcings on reproduction numbers during epidemics

Lu Lu, Ziyao Tian, Fengyang Wang, Yi Xiao Zhang

Abstract

Introduction

[TK: this is retained from Proposal, modify]

Exponential growth of the number of symptomatic cases is usually observed in the initial stages of infectious disease epidemics. Typically, the rate of this growth is decomposed into two quantities:

- The serial interval: time between onset of symptoms of two adjacent individuals in a transmission chain
- The reproductive number, R: the expected number of new infections a single symptomatic individual will cause

The estimation of the latter quantity is important for policy, because infectious disease outbreaks will switch from an exponential growth regime to an exponential decay regime when the reproductive number falls below 1. The goal of mitigation or suppression policies is therefore to effect change on the reproductive number R.

Many infectious diseases are believed to be seasonal, including most respiratory viruses, like the SARS-CoV-2 virus implicated in the 2019–2020 coronavirus pandemic. Wang et al. (2020) have analysed how the weather of locations within China correlates with local point estimates of the reproductive number R. However, because their analysis is frequentist in nature, their estimates of the seasonal forcing do not come with any estimate of uncertainty.

Methodology

Thompson et al. (2019) outline a Bayesian framework for estimation of $R(t)$ using symptomatic case counts, decomposed into local and imported cases. The EpiEstim R package implements ideas from this paper. A simplified summary of their approach, ignoring the imported cases:

1. Estimation of (discrete) distribution of serial interval from line-list data with intervals reflecting date of infection.
2. Estimation of time-varying R from time-bucketed symptomatic case counts.

We developed a method to estimate the strength of various seasonal forcings on R through MCMC. This involves both substantial deviations from the second part of the Thompson et al. model and also less less substantial adjustments for the nature of our data. Our method does not extend the first part of the Thompson et al. model.

To sanity-check the results of our method, we compare its conclusions against a point estimate for the seasonal forcings through linear regression with the EpiEstim estimates of R.

Incidence likelihood of the Thompson et al. model

In its estimation of time-varying R, the Thompson et al. model uses the following likelihood function for the number of new infections at time t:

$$I(t)|I(1),\cdots,I(t-1),\mathbf{w},R(t) \sim \text{Poisson}\left(R(t)\sum_{s=1}^{t} I(t-s)w_s\right)$$

where

- $I(t)$ is observed data containing the number of new infections at time t
- $R(t)$ is a parameter for the reproduction number at time t
- w_s denotes the proportion of serial intervals which take on discrete value s, i.e. $w_s = \text{P}(\text{SI} = s)$.

Because the Thompson et al. model uses a gamma-distributed prior for $R(t)$ (the conjugate prior for a Poisson likelihood) and because they chose a convenient family of serial interval distributions, the Bayesian posterior in the Thompson et al. model can be described by a closed form analytic expression.

Extension of Thompson et al. for R subject to forcings

Our Bayesian model extends the method of Thompson et al. to multiple locations, each with its own incidence time series $I_1(t),\cdots,I_n(t)$ such that

$$R_i(t) = \exp\left(\theta(t) + \lambda_i + \beta^T x_i(t) + \varepsilon_i(t)\right)$$

where:

- $x_i(t)$ is data of observed seasonal factors at time t and location i, with each feature standardized to mean 0 and variance 1
- β is a parameter describing the locally linear effect of each seasonal forcing
- $\theta(t)$ is a parameter for logarithm of country-wide R independent of modeled seasonal forcings
- λ_i is a parameter for local variation of $\log R$ independent of modeled seasonal forcings
- $\varepsilon_i(t)$ is a parameter for remaining fluctuation in $\log R$ not modeled above

Unlike the Thompson et al., our analysis fundamentally estimates the logarithm of R instead of R itself. This has some practical advantages: a linear model for R could predict negative values for sufficiently extreme climate conditions, but negative values for R do not make sense epidemiologically.

We maintain a very similar incidence likelihood function as the Thompson et al. model, but impose an additional correction for importations:

$$I_i(t)|P_i,x_i(t),I_i(1),\cdots,I_i(t-1),\mathbf{w},\theta(t),\lambda_i,\beta,\varepsilon(t),\gamma \sim \text{Poisson}\left(\gamma P_i + R_i(t)\sum_{s=1}^{t} I_i(t-s)w_s\right)$$

where in addition to the parameters described (with $R_i(t)$ given by the equation above), the following additional data and parameters are introduced:

- P_i is data giving the population of location i
- γ is a parameter estimating the nationwide rate of imports per capita per day, on which we impose prior $\gamma \sim \text{Gamma}(2, 10^{-5})$

The only substantative change to this likelihood function, beyond the decomposition of $R_i(t)$ into the various parameters influencing it, is the addition of importation risk. The reason for this adjustment is that the dataset we are using does not distinguish between importations and local cases. Thus, the Thompson likelihood would be 0 upon detection of the first case, had we not modeled the importations separately.

The most important feature of our model is the parameter β which are the coefficients of the linear model of $\log R$. This feature is assumed to be global across all locations and times, and thus models inherent properties of the contagiousness of the outbreak, as opposed to local deviations in space or time.

To avoid the model ascribing deviations in R to the seasonal forcings without widespread evidence across locations and times, three other features of our model are the parameters $\theta(t)$, λ_i, and $\varepsilon_i(t)$. The first parameter, $\theta(t)$, describes the logarithm country-wide R before adjusting for location and climate and local idiosyncratic factors. That is, it can be thought of as a baseline R value affected simply by the national circumstances at the given time; such as national policies or human behavioral changes. This will encapsulate various unmodeled behaviours such as weekends or interventions. At time zero, our prior belief is that this R will match the consensus of international estimates for R_0, which vary widely but tend to cluster between 2 or 3 [TK: Citation needed?]. Because of the great uncertainty, there is significant variance in this assumption, so the prior is quite weak (note that the $\mu \pm 2\sigma$ values, when transformed exponentially, are below 1 and above 6 respectively).

$$\theta(0) \sim \text{Normal}(\log 2.5, 0.5^2)$$

The remaining days of θ are modeled using a simple autoregressive model:

$$\theta(t+1) \sim \text{Normal}(\theta(t), \theta^\sigma)$$

where θ^σ is a parameter describing the spread of daily drifts in θ. The prior on θ^σ is weakly informative:

$$\theta^\sigma \sim \text{Gamma}(2, 0.01)$$

The location-specific deviation from the country average in $\log R$ independent of modeled seasonal forcings is encoded in λ_i. The existence of this parameter helps the model explain the large variations of apparent R across different locations, some of which appear to be affected much more so than others. The remaining parameter $\varepsilon_i(t)$ encapsulates all other unexplained deviations across both time and location. All these parameters (and β) are given the following weakly informative prior:

$$\lambda_i, \varepsilon_i(t), \beta_k \sim \text{Normal}(0, 0.5)$$

Our technique is general and can be used for any epidemic and any combination of potential seasonal forcings.

Dataset

[TK: retained from proposal, need to modify]

- The Italian government dataset at https://github.com/pcm-dpc/COVID-19 contains the disease incidence for each province. This data contains date of confirmation instead of date of onset of symptoms, and so some procedure may be necessary to correct for this. Abbott et al. (2020) outlines a technique to correct for reporting delays using an exponential distribution of reporting delay fit to line list data; we may use the same distribution.
- The `worldmet` R package enables access to temperature and humidity data, sourced from NOAA.

Results

Discussion

Directions for Future Research

[TK: retained from proposal, need to modify]

- Looking at the climate and infection data in American and Canadian cities as the situation develops further. In particular, compare this with the results in Italy, and observe if the linear assumption between seasonal factors and R holds true across different climate regions. If not, try fitting while assuming a nonlinear relationship between R and seasonal factors.
- Look into statistical simulation models which consider intercity travel and geographic proximity.

Appendix

- (PRE-PRINT) Abbott, S., Hellewell, J., Munday, J. D., Chun, J. Y., Thompson, R. N., Bosse, N. I., Chan, Y. D., Russell, T. W., Jarvis, C. I., ... & Funk, S. (2020). Temporal variation in transmission during the COVID-19 outbreak. CMMID. Available at CMMID: https://cmmid.github.io/topics/covid19/current-patterns-transmission/global-time-varying-transmission.html
- Thompson, R. N., Stockwin, J. E., van Gaalen, R. D., Polonsky, J. A., Kamvar, Z. N., Demarsh, P. A., ... & Lessler, J. (2019). Improved inference of time-varying reproduction numbers during infectious disease outbreaks. Epidemics, 29, 100356.
- Wang, J., Tang, K., Feng, K., & Lv, W. (2020). High Temperature and High Humidity Reduce the Transmission of COVID-19. Available at SSRN: https://ssrn.com/abstract=3551767 or http://dx.doi.org/10.2139/ssrn.3551767