Adaptively temporal evolution of reproduction number estimation with trend filtering

Jiaping Liu[1], Zhenglun Cai[2], Paul Gustafson[1], Daniel J. McDonald[1*]

1 Department of Statistics, The University of British Columbia, Vancouver, British Columbia, Canada
2 Centre for Health Evaluation and Outcome Sciences, The University of British Columbia, Vancouver, British Columbia, Canada

* daniel@stat.ubc.ca

Abstract

aaa

Author summary

xxx

1 Introduction

The basic reproduction number ($\mathcal{R}_0$), defined as the expected number of secondary infections caused by an infected individual in a completely susceptible population, is a fundamental indicator of epidemiological transmissibility. The reproduction number provides a threshold such that when $\mathcal{R}_0 < 1$, the infection dies out gradually, which is known as the *disease-free equilibrium*, and when $\mathcal{R}_0 > 1$, the *endemic equilibrium* is asymptotically achieved, i.e., the infection is always present ([1,4,5]). Basic reproduction numbers reflect the potential sizes of a pandemic and suggest the scale of epidemic prevention measures such as the proportion of a population that should be vaccinated. **Effective reproduction numbers** ($\mathcal{R}_t$ at time t, also called as, instantaneous reproduction numbers) differ from the basic reproduction numbers by relaxing the assumption of a completely susceptible population. Although it varies from $\mathcal{R}_0$ by a time-varying scale (x such that $\mathcal{R}_t = \mathcal{R}_0 x$) — the proportion of susceptible individuals over the population, it is able to explain more time-varying factors that may influence the spread of infectious diseases such as the transmission rates due to interventions. Therefore, the time-varying effective reproduction numbers are more interpretable in reality.

Effective reproduction number reflects an unobservable biological reality. Mathematical biologists and theoretic epidemiologists have proposed complicated mathematical models to uncover this reality by making various domain-specific assumptions using different observations such as incidence data. It, on one hand, implies that the estimated effective reproduction numbers of a pathogen can vary. Different estimation processes and results suggest different aspects and levels of policy making in epidemic control. For example, if effective reproduction numbers are assumed to be affected by human-human interaction over time in an epidemic model, controlling the parameters influencing human-human contact rate may eventually lead to a

reduction of the effective reproduction numbers in the specific model. Since some assumptions cannot be verified in practice, it is critical for an estimator to be *robust* to model misspecification. On the other hand, the estimation relies heavily on the quality of the available data. Some data may contain poor information due to the limitations of data collection. Thus, it is important to make *accurate* estimation that is *robust* to poor condition of data such as the low incidence periods.

EpiEstim [9] is an accurate method for near real-time effective reproduction number estimation. It assumes Poisson distributed incident confirmed cases and Gamma distributed serial interval functions, uses the renewal equation to measure the contagious dynamic, and the data it requires is only the observed daily infection counts. Abry et al. [10] extended this approach by introducing a (second-order) temporal regularization and a spatial regularization of reproduction rates and proposed to solve it as an optimization problem. Pascal et al. [11] further introduced robustness against outliers. EpiFilter [**?**] is a recursive Bayesian smoother based on Kalman Filter that maximizes the posteriori of reproduction number given a gamma prior and Poisson infection counts and generates robust estimates in low incidence periods. EpiNow2 [**?**] is a Bayesian latent variable framework that provides precise estimation and forecasts. A similarity of these models is that they only require the knowledge of the limited information, such as infected counts and the distribution of serial interval, to make accurate estimation. EpiLPS [**?**] is an another Bayesian framework that generates Bayesian P-splines coupled with Laplace approximations of the conditional posterior of the splines. It provides both point and interval estimations. There are other spline-based approaches such as [**?**, **?**, **?**].

2 Methods

2.1 Renewal model for incidence data

In this project, we focus on the temporal evolution of reproduction numbers and extend it from a second-order divided difference to all orders $(1, 2, 3, ...)$. We propose a locally adaptive estimator measuring the time series of reproduction numbers. Compared to existing mathematical models for reproduction number estimation, this estimator is more flexible in the order of temporal evolution of reproduction numbers and also locally adaptive so that it captures the local changes such as the initiation of effective control measures. More specifically, it regularizes the similarity among reproduction numbers across a chosen number of neighboring time points and segments the curvature of the reproduction numbers such that there are more jumpiness in some subregions and more smoothness in others. We find the proposed estimator is identical to the univariate Poisson trend filtering estimator with a slight modification. We assume the serial interval function is known and approximated it by Gamma distributions following previous studies [9–12].

2.2 Poisson trend filtering estimator

We assume that the count of observed daily new infections at time i (y_i) follows a Poisson distribution with the natural parameter being the effective reproduction number $(\mathcal{R}_i > 0)$ scaled by the weighted sum of previous daily counts, denoted by $w_i \geq 0$ for $i = 1, \cdots, n$. Let $\theta := \log(\mathcal{R}) \in \mathbb{R}^n$, and then $w \circ \mathcal{R} = w \circ e^{\theta}$, $\log(w \circ \mathcal{R}) = \log(w) + \theta$, where $e^a, \log(a)$ apply to a vector a elementwise. We further regularize the smoothness of the reproduction number using the ℓ_1 norm of the divided difference of the natural logarithm of $\mathcal{R}$, which is real-valued.

The extended Poisson trend filtering (PTF) on univariate cases is then defined as:

$$\begin{aligned}\hat{\theta} &:= \underset{\theta\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} -y_i\theta_i + w_i e^{\theta_i} + \lambda \left\| D^{(k+1)}\theta \right\|_1, \\ \hat{\mathcal{R}} &:= e^{\hat{\theta}},\end{aligned} \tag{1}$$

where $D^{k+1} \in \mathbb{Z}^{(n-k-1)\times n}$ is the k-th order divided difference matrix with $k = 0, 1, 2, 3, \cdots$. Define $D^{(k+1)}$ recursively as $D^{(k+1)} := D^{(1)}D^{(k)}$, where $D^{(1)} \in \mathbb{N}^{(n-k-1)\times(n-k)}$ and $D^{(1)}$ is a banded matrix with $(-1, 1)$ on the columns $(i, i+1)$ for each row $i = 1, ..., n-k-1$. Define $D^{(0)} := I_n$. An exponential transformation is applied to the PTF estimator to get the estimated reproduction number. For unequally spaced signals, replace $D^{(k+1)}$ by $D^{(x,k+1)}$ with weights $x \in \mathbb{R}^n$ (which are signal locations). Define $D^{(x,k+1)}$ recursively as $D^{(x,k+1)} := D^{(1)} \cdot X^k \cdot D^{(x,k)}$, where

$$X^k := \operatorname{diag}\left(\frac{k}{x_{k+1}-x_1}, \frac{k}{x_{k+2}-x_2}, \cdots, \frac{k}{x_n - x_{n-k}}\right)$$

is a diagonal matrix depending on the order k.

We use Gamma distribution to estimate the serial interval function, i.e., weights of the previous daily infections. On day i, the weights of previous counts $y_1, ..., y_{i-1}$ have corresponding coefficients (weights) $\Phi_1, .., \Phi_{i-1}$, which are probabilities of gamma distribution with prespecified parameters (α, β) corresponding to specific quartiles. The weight corresponding to $\mathcal{R}_i$ is $w_i = \sum_{j=1}^{\tau_\Phi} \Phi_j y_{i-j}$, where τ_Φ is a chosen period of infection.

2.3 Proximal Newton solver

We use proximal Newton method to solve the proximal optimization problem in Eq (1). During each outer Newton-based iteration, we use specialized ADMM [13] to solve the problem in the inner loop.

The outer Newton-based iteration solves the following problems at time $t+1$:

$$\theta_+^t = \operatorname{prox}_{W_w^t, D^{(x,k+1)}}\left(c_w^t\right), \tag{2}$$

$$\theta^{t+1} = \theta_+^t + \gamma^{t+1}\left(\theta_+^t - \theta^t\right), \tag{3}$$

where $W_w^t := \operatorname{diag}\left(w \circ e^{\theta^t}\right) \in \mathbb{R}^{n\times n}$ and $c_w^t := y^t \circ w^{-1} \circ e^{-\theta^t} + \theta^t - \mathbf{1}$.

3 Results

Implementation of the proximal Newton method is provided in the `R` package `rtestim`.

3.1 Covid-19 cases

We implement the proposed model on the Covid-19 confirmed cases in British Columbia (B.C.) as of May 18, 2023 reported by B.C. Conservation Data Centre. We choose the gamma distribution with shape 2.5 and scale 2.5 to approximate the serial interval function.

Considering the temporal evolutions of neighboring $3, 4, 5$ reproduction numbers, the estimated reproduction numbers of Covid-19 in British Columbia (displayed in the top right, bottom left, and bottom right panels in Fig 1 respectively) are always lower than 2.5, which means that two distinct infected individuals can on average infect less than

Fig 1. Covid19 daily confirmed counts between March 1st, 2020 and April 15th, 2023 in British Columbia, Canada. The top left panel displays the time trend of the observed infectious cases. The top right, bottom left and bottom right panels illustrated the estimated reproduction numbers ($\mathcal{R}_t$) using the Poisson trend filtering (in Eq (1)) with degrees $k = 1, 2, 3$ respectively.

five other individuals in the population. The three degrees of the temporal evolution (across all regularization levels λ) all yield similar results that $\hat{\mathcal{R}}_t$ achieves the highest peak around the end of 2021 and reaches the lowest trough shortly thereafter. Throughout the estimated curves, the peaks and troughs of the reproduction numbers roughly come prior to the following growths and decays of confirmed cases respectively.

The reproduction numbers are relatively unstable before April 1st, 2022. The highest peak coincides with the emergence and globally spread of the Omicron variant. The estimated reproduction numbers are apparently below the threshold 1 during two time periods – roughly from April 1st, 2021 to July 1st, 2021 and from January 1st, 2022 to April 1st, 2022. The first trough of $\hat{\mathcal{R}}_t$ coincides with the first authorization for use of Covid-19 vaccines in British Columbia. The second trough shortly after the greatest peak may credit to many aspects, including self-isolation of the infected individuals and application of the second shot of Covid-19 vaccines. Since around April 1st, 2022, the reproduction numbers stay stable (at around 1) and the infected cases stay low.

Greater regularization levels (i.e., larger λs) result in smoother estimated curves. Smoother curves (e.g., the yellow curve in the top right panel in Fig 1) suggest that the estimated reproduction numbers are around 1 during most time periods; however, they may not be appropriate to interpret the reality. More wiggly curves better reflect the fluctuation of $\mathcal{R}_t$, but sometimes fail to highlight the significant peaks or troughs. The tuning parameter λ needs to be chosen corresponding to the information in practice for a better interpretation.

4 Discussion

The proposed methodology provides a locally adaptive estimator using Poisson trend filtering on univariate data for capturing the heterogeneous smoothness of effective reproduction numbers given time series of infective observations in a given region. This is a nonparametric regression model which can be written as convex optimization(minimization) problem. Minimizing the distance (averaged KL divergence per coordinate) between the estimators and observations guarantees the data fidelity; minimizing a certain order of divided differences between each pair of neighboring parameters regularizes the smoothness. The ℓ_1 regularization introduces sparsity to the divided differences, which leads to heterogeneous smoothness within certain periods of time. The homogeneous smoothness within a time period can be either performed by a constant reproduction number, or a constant rate of changes, or a constant graphical curvature depending on the prescribed degree ($k = 0, 1, 2$ respectively).

The property of local adaptivity is useful for interpreting seasonal outbreaks. Given a properly chosen degree of polynomials, for example, the growth rate in a seasonal outbreak can be distinguished from the counterparts in un-seasonal outbreak periods at the beginning of the outbreak, which will alert epidemiologists to propose sanitary policies to prevent the progressing outbreak. The effective reproduction numbers can be estimated afterwards to check the efficiency of the sanitary policies referring to whether they are below the threshold, their tendencies of reduction, or their graphical curvatures.

A potential future work is to extend the proposed model to analyze spatio-temporal infectious data. Such data have the inherit graphical structure such that temporal

evolution within a region can be connected by lines (as time series) and spatial connection (of cross-sectional data) can be constructed by graphs where each pair of neighboring regions is linked by an edge. Moreover, the spatio-temporal evolution, i.e., the effects of previous infectious data of one region on current infections of neighboring regions, can be measured, for example, by linking the node of region a at time $t-1$ to another node of region b at time t. In this case, we can directly apply Poisson trend filtering on graphs with minor adjustment.

Our proposed model provides a natural way to deal with missing data, e.g., on weekends and holidays. Since the edge lengths of the line graphs can be adjusted, we can manually increase the length between two observations if there are missing data in between so that the influence of a previous observation is reduced correspondingly. Moreover, the ℓ_1 penalty introduces sparsity, and thus, makes the estimators less sensitive to outliers compared to ℓ_2 regularization. It is remarkable that our focus is to provide a mathematical model for epidemiologists to use, rather than to focus on a specific disease. In addition, more specialized methodologies are needed for the diseases with relatively long incubation periods (e.g., HIV and HBV).

A group of epidemiological models are compartmental models. They establish the epidemic transmission process by creating compartments with labels and connecting them by directed edges. A simple compartmental model – for example, *Susceptible-Infectious-Susceptible* (SIS) model – divides the population (N) into two compartments for susceptible cases (S) and infectious cases (I) respectively and connects them in serial as $S \to I \to S$. It only focuses on susceptible individuals. Each directed edge corresponds to a ratio of transmission (say, α, β respectively). In such models, reproduction numbers are defined as functions of the estimated transmission parameters and the numbers of compartments or population, e.g., $\hat{\mathcal{R}}_0 = \hat{\beta}N/\hat{\alpha}$ in the SIS models [5], as by-products. Compartmental models usually solve ordinary differential equations (ODE) systems for transmission numbers (e.g., α, β in the SIS model). A disadvantage of such parametric models is that they are less flexible than nonparametric models and the number of parameters to be estimated grows along with the increase of compartments in practice, which results in a growing computational complexity. Since the epidemic mechanism depends highly on the contexts, e.g., if a latency period exists or not, such models are lack of generalizability. Moreover, data of high quality are not always available for all compartments especially when there is a pandemic outbreak that results in a sudden shortage of resources in collecting daily new infections.

Supporting information

S1 Fig. Covid-19 figure. Covid-19 cases in BC and the estimated reproduction numbers.

S2 Fig. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 File. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Video. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Appendix. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Table. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

Acknowledgments

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

References

1. Diekmann O, Heesterbeek JAP, Metz JA. On the definition and the computation of the basic reproduction ratio R 0 in models for infectious diseases in heterogeneous populations. Journal of mathematical biology. 1990;28:365–382.
2. Dietz K. The estimation of the basic reproduction number for infectious diseases. Statistical methods in medical research. 1993;2(1):23–41.
3. Fine PE. Herd immunity: history, theory, practice. Epidemiologic reviews. 1993;15(2):265–302.
4. Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. Complexity of the basic reproduction number (R0). Emerging infectious diseases. 2019;25(1):1.
5. Brauer F, Castillo-Chavez C, Feng Z. Mathematical models in epidemiology. vol. 32. Springer; 2019.
6. Anderson RM, May RM. Directly transmitted infections diseases: control by vaccination. Science. 1982;215(4536):1053–1060.
7. Anderson RM, May RM. Vaccination and herd immunity to infectious diseases. Nature. 1985;318(6044):323–329.
8. Heffernan JM, Smith RJ, Wahl LM. Perspectives on the basic reproductive ratio. Journal of the Royal Society Interface. 2005;2(4):281–293.
9. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. American journal of epidemiology. 2013;178(9):1505–1512.
10. Abry P, Pustelnik N, Roux S, Jensen P, Flandrin P, Gribonval R, et al. Spatial and temporal regularization to estimate COVID-19 reproduction number R (t): Promoting piecewise smoothness via convex optimization. Plos one. 2020;15(8):e0237901.
11. Pascal B, Abry P, Pustelnik N, Roux S, Gribonval R, Flandrin P. Nonsmooth convex optimization to estimate the Covid-19 reproduction number space-time evolution with robustness against low quality data. IEEE Transactions on Signal Processing. 2022;70:2859–2868.

12. Thompson RN, Stockwin JE, van Gaalen RD, Polonsky JA, Kamvar ZN, Demarsh PA, et al. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. Epidemics. 2019;29:100356.

13. Ramdas A, Tibshirani RJ. Fast and flexible admm algorithms for trend filtering. Journal of Computational and Graphical Statistics. 2016;25(3):839–858.