RtEstim: Effective 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

* jiaping.liu@stat.ubc.ca

Abstract

To understand the transmissibility and spread of infectious diseases, epidemiologists turn to estimates of the effective reproduction number. While many estimation approaches exist, their utility may be limited. Challenges of surveillance data collection, model assumptions that are unverifiable with data alone, and computationally inefficient frameworks are critical limitations for many existing approaches. We propose a discrete spline-based approach `RtEstim` that solves a convex optimization problem—Poisson trend filtering—using the proximal Newton method. It produces a locally adaptive estimator for effective reproduction number estimation with heterogeneous smoothness. `RtEstim` remains accurate even under some process misspecifications and is computationally efficient, even for large-scale data. The implementation is easily accessible in a lightweight `R` package `rtestim`.

Author summary

Effective reproduction number estimation presents many challenges due to data collection, modelling assumptions, and computational burden. Such limitations hinder the accurate estimation of the effective reproduction number. Our motivation is to develop a model that produces accurate estimates, is robust to model misspecification, and is straightforward to use and computationally efficient, even for large counts and long time periods. We propose a convex optimization model with an ℓ_1 trend filtering penalty. It couples accurate estimation of the effective reproduction number with desired smoothness. We solve the optimization using the proximal Newton method, which converges rapidly and is numerically stable. Our software, conveniently available in the `R` package `RtEstim`, can produce estimates in seconds for incidence sequences with hundreds of observations. These estimates are produced for a sequence of tuning parameters and can be selected using a built-in cross validation procedure.

1 Introduction

The effective reproduction number is defined to be the average number of secondary infections caused by a primary infection that occured sometime in the past. Also called the instantaneous reproduction number, it is a key quantity for understanding infectious disease dynamics including the potential size of an outbreak and the required stringency

of control measures. Tracking the time series of this quantity is useful for understanding whether or not future infections are likely to increase or decrease from the current state. Let $\mathcal{R}(t)$ denote the effective reproduction number at time t. Practically, as long as $\mathcal{R}(t) < 1$, infections will decline gradually, eventually resulting in a disease-free equilibrium, whereas when $\mathcal{R}(t) > 1$, infections will continue to increase, resulting in endemic equilibrium. While $\mathcal{R}(t)$ is fundamentally a continuous time quantity, it can be related to data only at discrete points in time $t = 1, \ldots, n$. This sequence of effective reproduction numbers over time is not observable, but, nonetheless, is easily interpretable and retrospectively describes the course of an epidemic. Therefore, a number of procedures exist to estimate $\mathcal{R}_t$ from different types of observed incidence data such as cases, deaths, or hospitalizations, while relying on various domain-specific assumptions. Importantly, accurate estimation of effective reproduction numbers relies heavily on the quality of the available data, and, due to the limitations of data collection, such as underreporting and lack of standardization, estimation methodologies rely on various assumptions to compensate. Because model assumptions may not be easily verifiable from data alone, it is also critical for any estimation procedure to be robust to model misspecification.

Many existing approaches for effective reproduction number estimation are Bayesian: they estimate the posterior distribution of $\mathcal{R}_t$ conditional on the observations. One of the first such approaches is the software `EpiEstim` [1], described in [2]. This method is prospective, in that it uses only observations available up to time t in order to estimate $\mathcal{R}_t$ for each $i = 1, \ldots, t$. An advantage of `EpiEstim` is its straightforward statistical model: new incidence data follows the Poisson distribution conditional on past incidence combined with the conjugate gamma prior distribution for $\mathcal{R}_t$ with fixed hyperparameters. Additionally, the serial interval distribution, the distribution of the period between onsets of primary and secondary infections in a population, is fixed and known. For this reason, `EpiEstim` requires little domain expertise for use, and it is computationally fast. [3] modified this method to distinguish imported cases from local transmission and simultaneously estimate the serial interval distribution. [4] further extended `EpiEstim` by using "reconstructed" daily incidence data to handle irregularly spaced observations. Recently, [5] proposed a Bayesian latent variable framework, `EpiNow2` [6], which leverages incident cases, deaths or other available streams simultaneously along with allowing additional delay distributions (incubation period and onset to reporting delays) in modelling. [7] proposed an extension that handles missing data by imputation followed by a truncation adjustment. These modifications are intended to increase accuracy at the most recent (but most uncertain) timepoints, to aid policymakers. [8] also proposed a Bayesian approach, `EpiFilter` based on the (discretized) Kalman filter and smoother. `EpiFilter` also estimates the posterior of $\mathcal{R}_t$ given a Gamma prior and Poisson distributed incident cases. Compared to `EpiEstim`, however, `EpiFilter` estimates $\mathcal{R}_t$ retrospectively using all available incidence data both before and after time t, with the goal of being more robust in low-incidence periods. [9] proposed a Bayesian P-splines approach, `EpiLPS`, that assumes negative Binomial distributed observations. [10] also proposed a Bayesian model estimated with particle filtering to incorporate spatial structures. Bayesian approaches estimate the posterior distribution of the effective reproduction numbers and possess the advantage that credible intervals may be easily computed. A limitation of many Bayesian approaches, however, is that they usually require more intensive computational routines, especially when observed data sequences are long or hierarchical structures are complex. Below, we compare our method to two of the more computationally efficient Bayesian models, `EpiEstim` and `EpiLPS`.

There are also frequentist approaches for $\mathcal{R}_t$ estimation. [11] proposed regularizing the smoothness of $\mathcal{R}_t$ through penalized regression with second-order temporal

regularization, additional spatial penalties, and with Poisson loss. [12] extended this procedure by introducing another penalty on outliers. [13] proposed a spline-based model relying on the assumption of exponential-family distributed incidence. [14] estimates $\mathcal{R}_t$ while monitoring the time-varying level of overdispersion. There are other spline-based approaches such as [15, 16], autoregressive models with random effects [17] that are robust to low incidence, and generalized autoregressive moving average (GARMA) models [18] that are robust to measurement errors in incidence data.

We propose a retrospective effective reproduction number estimator called `RtEstim` that requires only incidence data. Our model makes the conditional Poisson assumption, similar to much of the prior work described above, but is empirically more robust to misspecification. This estimator is defined by a convex optimization problem with Poisson loss and ℓ_1 penalty on the temporal evolution of $\log(\mathcal{R}_t)$ to impose smoothness over time. As a result, `RtEstim` generates discrete splines, and the estimated curves (in logarithmic space) appear to be piecewise polynomials of an order selected by the user. Importantly, the estimates are locally adaptive, meaning that different time ranges may possess heterogeneous smoothness. Because we penalize the logarithm of $\mathcal{R}_t$, we naturally accommodate the positivity requirement, in contrast to related methods, can handle large or small incidence measurements, and are automatically (reasonably) robust to outliers without additional constraints. A small illustration using three years of Covid-19 case data in Canada is shown in Fig 1.

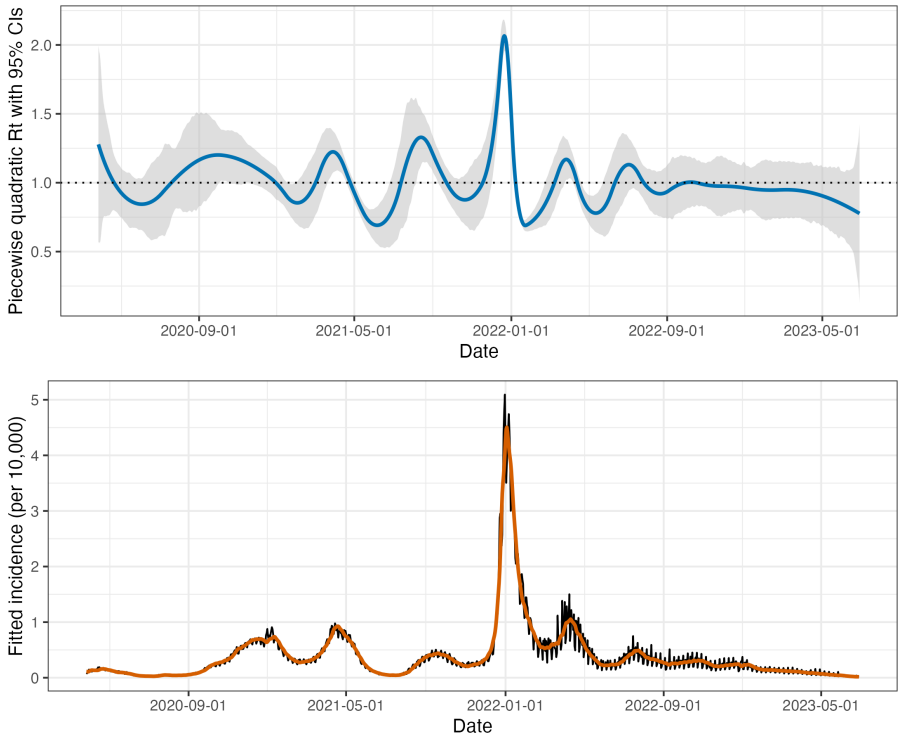

Fig 1. A demonstration of effective reproduction number estimation by `RtEstim` and the corresponding predicted incident cases for the Covid-19 epidemic in Canada during the period from March 28, 2020 to June 28, 2023. In the top panel, the blue curve is the estimated piecewise quadratic $\mathcal{R}_t$ and the gray ribbon is the corresponding 95% confidence band. The black curve in the bottom panel is the observed Covid-19 daily confirmed cases, and the orange curve is the predicted incident cases corresponding to the estimated $\mathcal{R}_t$.

While our approach is straightforward and requires little domain knowledge for implementation, we also implement a number of refinements. We use a proximal Newton method to solve the convex optimization problem along with warm starts to produce estimates efficiently, typically in a matter of seconds, even for long sequences of data. In a number of simulation experiments, we show empirically that our approach is more accurate than existing methods at estimating the true effective reproduction numbers.

The manuscript proceeds as follows. We first introduce the methodology of `RtEstim` including the renewal equation and the development of Poisson trend filtering estimator. We explain how this method could be interpreted from the Bayesian perspective, connecting it to previous work in this context. We provide illustrative experiments comparing our estimator to `EpiEstim` and `EpiLPS`. We then apply our `RtEstim` on the Covid-19 pandemic incidence in British Columbia and the 1918 influenza pandemic incidence in the United States. Finally, we conclude with a discussion of the advantages and limitations of our approach and describe practical considerations for effective reproduction number estimation.

2 Methods

2.1 Renewal model for incidence data

The effective reproduction number $\mathcal{R}(t)$ is defined to be the expected number of secondary infections at time t produced by a primary infection sometime in the past. To make this precise, denote the number of new infections at time t as $y(t)$. Then the total primary infectiousness can be written as $\eta(t) := \int_0^\infty p(i)y(t-i)\mathsf{d}i$, where $p(i)$ is the probability that a new secondary infection is the result of a primary infection that occurred i time units in the past. The reproduction number is then given as the value that equates

$$\mathbb{E}[y(t) \mid y(j),\ j < t] = \mathcal{R}(t)\eta(t) = \mathcal{R}(t)\int_0^\infty p(i)y(t-i)\mathsf{d}i, \tag{1}$$

otherwise known as the renewal equation. The period between primary and secondary infections is exactly the generation time of the disease, but given real data, observed at discrete times (say, daily) this delay distribution must be discretized into contiguous time intervals, say, $(0,1],(1,2],\ldots$, which results in the sequence $\{p_i\}_0^\infty$ corresponding to observations y_t and resulting in the discretized version of Eq (1),

$$\mathbb{E}[y_t \mid y_1,\ldots,y_{t-1}] = \mathcal{R}_t\eta_t = \mathcal{R}_t\sum_{i=1}^{\infty} p_i y_{t-i}. \tag{2}$$

Many approaches to estimating $\mathcal{R}_t$ rely on Eq (2) as motivation for their procedures, among them, `EpiEstim` [2] and `EpiFilter` [8].

In most cases, it is safe to assume that infectiousness disappears beyond τ timepoints ($p(i) = 0$ for $i > \tau$) so that the truncated integral of the generation interval distribution $\int_0^\tau p(i)\mathsf{d}i = 1$. Generation time, however, is usually unobservable and tricky to estimate, so common practice is to approximate it by the serial interval: the period between the symptom onsets of primary and secondary infections. If the infectiousness profile after symptom onset is independent of the incubation period (the period from the time of infection to the time of symptom onset), then this approximation is justifiable: the serial interval distribution and the generation interval distribution share the same mean. However, other properties may not be similarly shared, and, in general, the generation interval distribution is a convolution of the serial interval distribution with the distribution of the difference between independent draws from the delay

distribution from infection to symptom onset. See, for example, [19] for a fuller discussion of the dangers of this approximation. Nonetheless, treating these as interchangeable is common [2] and doing otherwise is beyond the scope of this work. Additionally, we assume that the generation interval (and, therefore, the serial interval), is constant over time t. That is, the probability $p(i)$ depends only on the gap between primary and secondary infections and not on the time t when the secondary infection occurs. For our methods, we will assume that the serial interval can be accurately estimated from auxiliary data (say by contact tracing, or previous epidemics) and we will take it as fixed, as is common in existing studies, e.g., [2, 11, 12].

The renewal equation in Eq (2) relates observable data streams (incident cases) occurring at different time points to the reproduction number given the serial interval. The fact that it depends only on the observed incident counts makes it reasonable to estimate $\mathcal{R}_t$. However, data collection idiosyncrasies can obscure this relationship. Diagnostic testing targets symptomatic individuals, omitting asymptomatic primary infections which can lead to future secondary infections. Testing practices, availability, and uptake can vary across space and time [20, 21]. Finally, incident cases as reported to public health are subject to delays due to laboratory confirmation, test turnaround times, and eventual submission to public health [22]. For these reasons, reported cases are lagging indicators of the course of the pandemic. Furthermore, they do not represent the actual number of new infections that occur on a given day, as indicated by exposure to the pathogen. The assumptions described above (constant serial interval distribution, homogenous mixing, similar susceptibility and social behaviours, etc.) are therefore consequential. That said, Eq (2) also provides some comfort about deviations from these assumptions. If y_t is scaled by a constant (in time) describing the reporting ratio, then it will cancel from both sides. Similar arguments mean that even if such a scaling varies in time, as long as it varies slowly relative to the set of p_i that are larger than 0, Eq (2) will be a reasonably accurate approximation, so that $\mathcal{R}_t$ can still be estimated well from reported incidence data. Finally, even a sudden change in reporting ratio, say from c_1 for $i = 1, \ldots, t_1$ to c_2 for $i > t_1$ would only result in large errors for t in the neighbourhood of t_1 (where the size of this neighbourhood is again determined by the effective support of $\{p_i\}$). This robustness to certain types of data reporting issues partially justifies using Eq (2) to calculate $\mathcal{R}_t$.

2.2 Poisson trend filtering estimator

We use the daily confirmed incident cases y_t on day t to estimate the observed infectious cases under the model that y_t given previous incident cases $y_{t-1}, \ldots, y_1$ and a constant serial interval distribution follows a Poisson distribution with mean Λ_t. That is,

$$y_t \mid y_1, \ldots, y_{t-1} \sim \text{Poisson}(\Lambda_t), \text{ where } \Lambda_t = \mathcal{R}_t \sum_{i=1}^{t-1} p_i y_{t-i} = \mathcal{R}_t \eta_t.$$

Given a history of n confirmed incident counts $\mathbf{y} = (y_1, \ldots, y_n)^\mathsf{T}$, our goal is to estimate $\mathcal{R}_t$. A natural approach is to maximize the likelihood, producing the MLE:

$$\begin{aligned} \widehat{\mathcal{R}} &= \underset{\mathcal{R} \in \mathbb{R}_+^n}{\text{argmax}}\ \mathbb{P}(\mathcal{R} \mid \mathbf{y},\ \mathbf{p}) = \underset{\mathcal{R} \in \mathbb{R}_+^n}{\text{argmax}} \prod_{t=1,\ldots,n} \frac{(\mathcal{R}_t \eta_t)^{y_t} \exp\{-\mathcal{R}_t \eta_t\}}{y_t!} \\ &= \underset{\mathcal{R} \in \mathbb{R}_+^n}{\text{argmin}}\ \frac{1}{n} \sum_{t=1}^{n} \mathcal{R}_t \eta_t - y_t \log(\mathcal{R}_t \eta_t). \end{aligned} \tag{3}$$

This optimization problem, however, is easily seen to yield a one-to-one correspondence between the observations and the estimated effective reproduction number, i.e., $\widehat{\mathcal{R}}_t = y_t / \eta_t$, so that the estimated sequence $\widehat{\mathcal{R}}$ will have no significant smoothness.

The MLE is an unbiased estimator of the true parameter $\mathcal{R}_t$, but unfortunately has high variance: changes in y_t result in proportional changes in $\widehat{\mathcal{R}}_t$. To avoid this behaviour, and to match the intuition that $\mathcal{R}_t \approx \mathcal{R}_{t-1}$, we advocate enforcing smoothness of the effective reproduction numbers. This constraint will decrease the estimation variance, and hopefully lead to more accurate estimation of $\mathcal{R}$, as long as the smoothness assumption is reasonable. Smoothness assumptions are common (see e.g., [8] or [19]), but the type of smoothness assumed is critical. [1] imposes smoothness indirectly by estimating $\mathcal{R}_t$ with moving windows of past observations. The Kalman filter procedure of [8] would enforce in ℓ_2-smoothness ($\int_0^n \left(\widehat{\mathcal{R}}''(t)\right)^2 \mathsf{d}t < C$ for some C), although the computational implementation results in $\widehat{\mathcal{R}}$ taking values over a discrete grid. [12] produces piecewise-linear $\widehat{\mathcal{R}}_t$, which turns out to be closely related to a special case of our methodology. Smoother estimated curves will provide high-level information about the entire epidemic, obscuring small local changes in $\mathcal{R}(t)$, but may also remove the ability to detect large sudden changes, such as those resulting from lockdowns or other major containment policies.

To enforce smoothness of $\widehat{\mathcal{R}}_t$, we add a trend filtering penalty to Eq (4) [23–26]. Because $\mathcal{R}_t > 0$, we explicitly penalize the divided differences (discrete derivatives) of neighbouring values of $\log(\mathcal{R}_t)$. Let $\theta := \log(\mathcal{R}) \in \mathbb{R}^n$, so that $\Lambda_t = \eta_t \exp(\theta_t)$, and $\log(\eta_t \mathcal{R}_t) = \log(\eta_t) + \theta_t$. For evenly spaced incident case, we write our estimator as the solution to the optimization problem

$$\widehat{\mathcal{R}} = \exp(\widehat{\theta}) \quad \text{where} \quad \widehat{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \; \eta^{\mathsf{T}} \exp(\theta) - \mathbf{y}^{\mathsf{T}} \theta + \lambda \| D^{(k+1)} \theta \|_1, \tag{4}$$

where $\exp(\cdot)$ applies elementwise. Here, $D^{(k+1)} \in \mathbb{Z}^{(n-k-1)\times n}$ is the $(k+1)^{\text{th}}$ order divided difference matrix for any $k \in \{0, \ldots, n-1\}$. $D^{(k+1)}$ is defined recursively as $D^{(k+1)} = D^{(1)} D^{(k)}$, where $D^{(1)} \in \{-1, 0, 1\}^{(n-k-1)\times(n-k)}$ is a sparse matrix with diagonal band:

$$D^{(1)} = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}.$$

The tuning parameter λ balances data fidelity with desired smoothness. When $\lambda = 0$, the problem in Eq (4) reduces to the MLE in Eq (3). Larger tuning parameters privilege the regularization term and yield smoother estimates. Finally, there exists $\lambda_{\max}$ such that any $\lambda \geq \lambda_{\max}$ will result in $D^{(k+1)}\widehat{\theta} = 0$ and $\widehat{\theta}$ will be the Kullback-Leibler projection of $\mathbf{y}$ onto the null space of $D^{(k+1)}$ (see subsection 2.3).

The solution to Eq (4) will result in piecewise continuous polynomials, specifically called discrete splines. For example, 0^{th}-degree discrete splines are piecewise constant, 1^{st}-degree curves are piecewise linear, and 2^{nd}-degree curves are piecewise quadratic. For $k \geq 1$, k^{th}-degree discrete splines are continuous and have continuous discrete differences up to degree $k-1$ at the knots. This penalty results in more flexibility compared to the homogeneous smoothness that is created by the squared ℓ_2 norm. Using different orders of divided differences result in estimated effective reproduction numbers with different smoothness properties.

For unevenly-spaced data, the spacing between neighboring parameters varies with the time between observations, and thus, the divided differences must be adjusted by the times that the observations occur. Given observation times $\mathbf{x} = (x_1, \ldots, x_n)^{\mathsf{T}}$, for $k \geq 1$, define a k^{th}-order diagonal matrix

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

Letting $D^{(\mathbf{x},1)} := D^{(1)}$, then for $k \geq 1$, the $(k+1)^{\text{th}}$-order divided difference matrix for unevenly spaced data can be created recursively by $D^{(\mathbf{x},k+1)} := D^{(1)} X^{(k)} D^{(\mathbf{x},k)}$. No adjustment is required for $k = 0$.

Due to the penalty structure, this estimator is locally adaptive, meaning that it can potentially capture local changes such as the initiation of control measures. [11, 12] considered only the 1^{st}-order divided difference of $\mathcal{R}_t$ rather than its logarithm. In comparison to their work, our estimator (1) allows for arbitrary degrees of temporal smoothness and (2) avoids the potential numerical issues of penalizing/estimating positive real values. Furthermore, as we will describe below, our procedure is computationally efficient for estimation over an entire sequence of penalty strengths λ and provides methods for choosing how smooth the final estimate should be.

2.3 Solving over a sequence of tuning parameters

We can solve the Poisson trend filtering estimator over an arbitrary sequence of λ that produces different levels of smoothness in the estimated curves. We consider a candidate set $\boldsymbol{\lambda} = \{\lambda_m\}_{m=1}^{M}$ that is strictly decreasing.

Let $D := D^{(k+1)}$ for simplicity in the remainder of this section. As $\lambda \to \infty$, the penalty term $\lambda \|D\theta\|_1$ dominants the Poisson objective, so that minimizing the objective is asymptotically equivalent to minimizing the penalty term, which results in $\|D\theta\|_1 = 0$. In this case, the divided differences of θ with order $k+1$ is always 0, and thus, θ must lie in the null space of D, that is, $\theta \in \mathcal{N}(D)$. The same happens for any λ beyond this threshold, so define $\lambda_{\max}$ to be the smallest λ that produces $\theta \in \mathcal{N}(D)$. It turns out that this value can be written explicitly as $\lambda_{\max} = \|\left(D^{\dagger}\right)^{\mathsf{T}} (\eta - y)\|_{\infty}$, where $D^{\dagger}$ is the (left) generalized inverse of D satisfying $D^{\dagger} D = I$. Therefore, we use $\lambda_1 = \lambda_{\max}$ and then choose the minimum λ_M to be $r\lambda_{max}$ for some $r \in (0, 1)$ (typically $r = 10^{-5}$). Given any $M \geq 3$, we generate a sequence of λ values to be equally spaced on the log-scale between λ_1 and λ_M.

To compute the sequence efficiently, the model is estimated sequentially by visiting each component of $\boldsymbol{\lambda}$ in order. The estimates produced for a larger λ are used as the initial values (warm starts) for the next smaller λ. By solving through the entire sequence of tuning parameters, we have a better chance to achieve a better trade-off between bias and variance, and accordingly, improved accuracy relative to procedures examining one fixed value of λ.

2.4 Choosing a final λ

We estimate model accuracy over the candidate set through K-fold cross validation (CV) to choose the best tuning parameter. Specifically, we divide $\mathbf{y}$ (except the first and last observations) roughly evenly and randomly into K folds, estimate $\mathcal{R}_t$ for all $\boldsymbol{\lambda}$ leaving one fold out, and then predict the held-out observations. Model accuracy can be measured by multiple metrics such as mean squared error $\text{MSE}(\widehat{y},\ y) = n^{-1}\|\widehat{y} - y\|_2^2$ or mean absolute error $\text{MAE}(\widehat{y},\ y) = n^{-1}\|\widehat{y} - y\|_1$, but we prefer to use the (average) deviance, to mimic the likelihood in Eq (3):
$D\left(y,\ \hat{y}\right) = n^{-1} \sum_{i=1}^{n} 2\left(y_i \log(y_i) - y_i \log(\hat{y}_i) - y_i + \hat{y}_i\right)$, with the convention that $0 \log(0) = 0$. Note that for any K and any M, we will end up estimating the model $(K+1)M$ times rather than once.

2.5 Approximate confidence bands

We also provide empirical confidence bands of the estimators with approximate coverage. Consider the related estimator $\widetilde{\mathcal{R}}_t$ defined as

$$\widetilde{\mathcal{R}} = \exp(\widetilde{\theta}) \quad \text{where} \quad \widetilde{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \, \eta^{\mathsf{T}} \exp(\theta) - \mathbf{y}^{\mathsf{T}}\theta + \lambda \|D\theta\|_2^2.$$

Let $\widetilde{\mathbf{y}} = \eta \circ \widetilde{\mathcal{R}}$, and then it can be shown (for example, Theorem 2 in [27]) that an estimator for $\operatorname{Var}(\widetilde{\mathbf{y}})$ is given by $\left(\operatorname{diag}(\widetilde{\mathbf{y}}^{-2}) + \lambda D^{\mathsf{T}} D\right)^{\dagger}$. Finally, an application of the delta method shows that $\operatorname{Var}(\widetilde{\mathbf{y}}_t)/\eta_t^2$ is an estimator for $\operatorname{Var}(\widetilde{\mathcal{R}}_t)$ for each $t = 1, \ldots, n$. We therefore use $\left(\operatorname{diag}(\widehat{\mathbf{y}}^{-2}) + \lambda D^{\mathsf{T}} D\right)_t^{\dagger}/\eta_t^2$ as an estimator for $\operatorname{Var}(\widehat{\mathcal{R}}_t)$. An approximate $(1-\alpha)\%$ confidence interval then can be written as $\widehat{\mathcal{R}}_t \pm s_t \times T_{\alpha/2, n-\mathrm{df}}$, where s_t is the square-root of $\operatorname{Var}(\widehat{\mathcal{R}}_t)$ for each $t = 1, \cdots, n$ and df is the number of changepoints in $\widehat{\theta}$ plus $k+1$.

2.6 Bayesian perspective

Unlike many other methods for $\mathcal{R}_t$ estimation, our approach is frequentist rather than Bayesian. Nonetheless, it has a corresponding Bayesian interpretation: as a state-space model with Poisson observational noise, autoregressive transition equation of degree $k \geq 0$, e.g., $\theta_{t+1} = 2\theta_t - \theta_{t-1} + \varepsilon_{t+1}$ for $k = 1$, and Laplace transition noise $\varepsilon_{t+1} \sim \text{Laplace}(0,\ 1/\lambda)$. Compared to `EpiFilter` [8], we share the same observational assumptions, but our approach has a different transition noise. `EpiFilter` estimates the posterior distribution of $\mathcal{R}_t$, and thus it can provide credible interval estimates as well. Our approach produces the maximum *a posteriori* estimate via an efficient convex optimization, obviating the need for MCMC sampling. But the associated confidence bands are created differently.

3 Results

Implementation of our approach is provided in the `R` package `rtestim`. All experiments are run in `R` version 4.3.1 on a MacBook with an Apple M1 Pro chip and 32GB RAM running under macOS Sonoma 14.0. The `R` packages used for simulation and real-data application are `EpiEstim 2.2-4`, `EpiLPS 1.2.0`, and `rtestim 0.0.4`.

3.1 Synthetic experiments

We simulate four scenarios of the time-varying effective reproduction number, intended to mimic different epidemics. The first two scenarios are rapidly controlled by intervention, where the $\mathcal{R}(t)$ consists of one discontinuity and two segments. Scenario 1 has constant $\mathcal{R}(t)$ before and after an intervention, while Scenario 2 grows exponentially, then decays. The other two scenarios are more complicated, where more waves are involved. Scenario 3 has four linear segments with three discontinuities, which reflect the effect of an intervention, resurgence to rapid transmission, and finally suppression of the epidemic. Scenario 4 involves sinusoidal waves throughout the epidemic. The first three scenarios and the last scenario are motivated by [8] and [9] respectively. We name the four scenarios as (1) 2-segment constant, (2) 2-segment exponential curve, (3) 4-segment linear line, and (4) periodic curve respectively.

In all cases, the times of observation are regular, and epidemics are of length $n = 300$. Specifically, in Scenario 1, $\mathcal{R}_t = 2, 0.8$ before and after $t = 70$. In Scenario 2,

$\mathcal{R}_t$ increases and decreases exponentially with rates $0.015, 0.005$ pre and post $t = 50$. In Scenario 3,

$$\mathcal{R}_t = blah.$$

$\mathcal{R}_t$ declines from 2.5 to 2 linearly between $t \in [1, 60]$, falls to 0.8 at $t = 61$, decreases linearly to 0.6 until $t = 110$, resurges to 1.7 at $t = 111$ and grows linearly to 2 until $t = 150$, and then drops to 0.9 at $t = 151$ and descends to 0.5 until the end. In Scenario 4, $\mathcal{R}_t$ is realization of the continuous, periodic curve generated by the function

$$\mathcal{R}(t) = 0.2\big((\sin(\pi t/12) + 1) + (2\sin(\pi t/6) + 2) + (3\sin(\pi t/1.2) + 3) \big),$$

evaluated at equally spaced points $t \in [0, 10]$. These settings are illustrated in the left column of Fig 2.

We assume that the serial interval follows the Gamma distribution with fixed shapes and scales $(3, 3)$, $(2.5, 2.5)$, $(3.5, 3.5)$ and $(3.5, 3.5)$ for Scenarios 1–4 respectively. We initialize all epidemics with $y_1 = 2$ cases and generating for $t = 2, \ldots, 300$. We compute the expected incidence Λ_t using the renewal equation, and generate the incident infections from the Poisson distribution $y_t \sim \text{Pois}(\Lambda_t)$. To verify the performance of our model under the violation of this distributional assumption, we also generate incident infections using the negative Binomial distribution with dispersion size 5, i.e., $y_t \sim \text{NB}(\Lambda_t,\ \text{size} = 5)$. For each setting, we generate 50 random samples, resulting in 400 total experiments. An example of each effective reproduction number scenario with a single corresponding Poisson and negative Binomial sample incidence sequences is displayed in Fig 2.

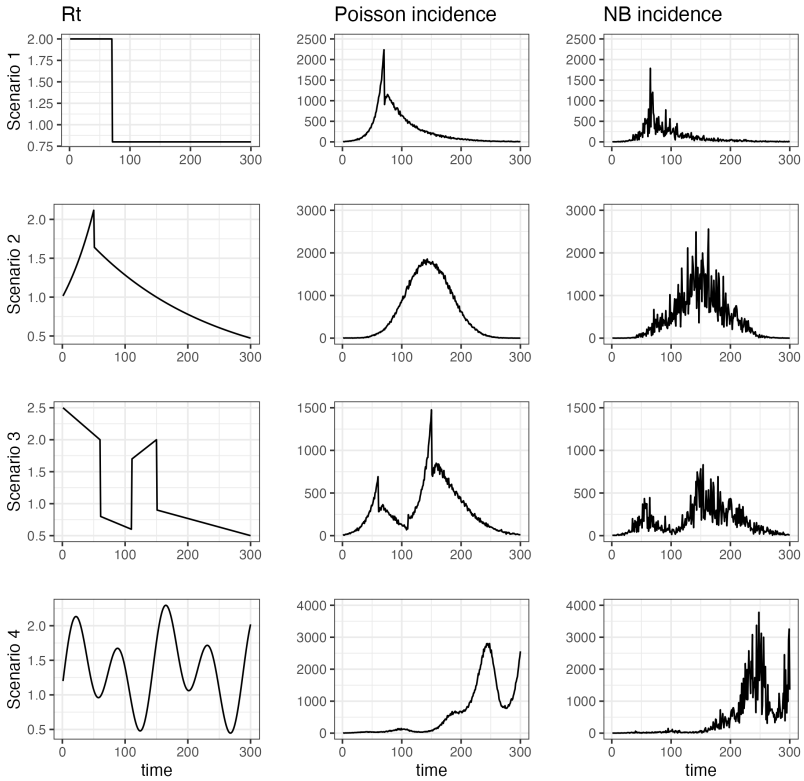

Fig 2. The effective reproduction numbers (left column) and corresponding sample incident cases drawn from a Poisson (middle column) or negative Binomial (right column) distribution. The rows correspond to the four $\mathcal{R}_t$ settings.

We compare `RtEstim` to `EpiEstim` and `EpiLPS`. Unfortunately, `EpiFilter` frequently fails to converge due to the large incident counts in these settings, so we are unable to include its results. `EpiEstim` estimates the posterior distribution of the effective reproduction number given a Gamma prior and Poisson distributed observations over a trailing window, under the assumption that the reproduction number is constant during that window. A larger window averages out more fluctuations, leading to smoother estimates, whereas, a shorter sliding window is more responsive to sudden spikes or declines. We tried the default, a weekly sliding window, as well as a monthly window. However, since neither considerably outperforms the other across all scenarios, we defer the monthly results to the Appendix. `EpiLPS` is another Bayesian approach that estimates P-splines coupled based on the Laplace approximation to the conditional posterior with negative Binomial likelihood. For `RtEstim` on the four scenarios respectively, we estimate (1) piecewise constant $k = 0$, (2) piecewise linear & cubic $k = 1, 3$, (3) piecewise linear $k = 1$ and (4) piecewise cubic polynomials $k = 3$. In each case, we examine a grid of 50 λ values, selecting the best using 10-fold cross validation. For all models and problems, we use the same serial interval distribution for estimation that was used to create the data.

To measure estimation accuracy, we compare $\widehat{\mathcal{R}}$ to $\mathcal{R}$ using the Kullback-Leibler (KL) divergence. We use the KL divergence for the Poisson distribution (averaged across all t) to measure the accuracy of the $\mathcal{R}_t$ estimates

$$D_{KL}(\mathcal{R} \parallel \widehat{\mathcal{R}}) = \sum_{t=1}^{n} w_t \left(\mathcal{R}_t \log \left(\frac{\mathcal{R}_t}{\widehat{\mathcal{R}}_t} \right) + \widehat{\mathcal{R}}_t - \mathcal{R}_t \right),$$

where $\mathcal{R} = \{\mathcal{R}_t\}_{t=1}^n$ and $w_t = \eta_t / \sum_t \eta_t$ is the rescaled total infectiousness. To fairly compare across methods, we drop the estimates during the first week because estimates from `EpiEstim` do not begin until $t = 8$ (using a weekly window). KL divergence is more appropriate for measuring accuracy because it connects directly to the Poisson likelihood used to generate the data, whereas standard measures like the mean-squared error correspond to Gaussian likelihood. Using Poisson likelihood has the effect of increasing the relative cost of mistakes when Λ_t is small. Other details of the experimental settings are deferred to the Appendix.

3.2 Results for synthetic data

`RtEstim` overall outperforms `EpiEstim` and `EpiLPS` in the experimental study. Fig 3 visualizes the KL divergence across the three models. Under both Poisson and egative Binomial distributions, `RtEstim` is easily the most accurate for Scenarios 1 and 3: the median of KL divergence is much lower and the boxes frequently fail to overlap indicating better performance than the other two methods across all 50 simulations. The advantage is less pronounced for the negative Binomial configuration, but still obvious. `RtEstim` and `EpiLPS` have similar performance in Scenarios 2 and 4. For the Poisson case, `RtEstim` and `EpiLPS` both have very small KL scores, which are very close to zero. In Scenario 4, `RtEstim` is slightly better for Poisson and `EpiLPS` is better for negative Binomial, but the boxes largely overlap each other. `EpiLPS` has a slightly lower median and a smaller IQR in Scenario 2 for the negative Binomial case. Both smoothness choices for `RtEstim` in Scenario 2 perform similarly across noise distributions, implying good performance under model misspecification. We will examine a single realization of each experiment to investigate these global conclusions in more detail.

Fig 4 shows one realization for the estimated reproduction number under the Poisson generative model for all four scenarios. Compared to `EpiEstim` and `EpiLPS`, which have rather severe difficulties at the beginning of the time series, `RtEstim` estimates are more accurate—they nearly overlap with the true values—without suffering from the

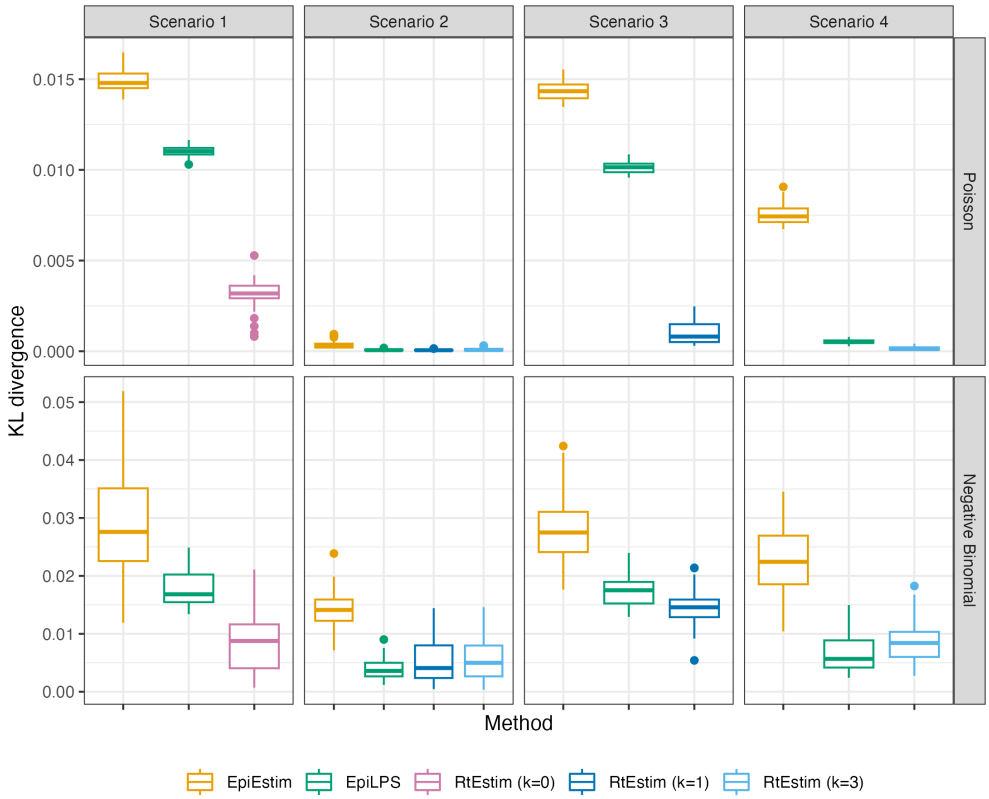

Fig 3. Boxplot of KL divergence between the estimated $\hat{\mathcal{R}}_t$ and the true $\mathcal{R}_t$ across 50 random samples for each approach given Poisson incidence *(in top panels)* and negative Binomial incidence *(in bottom panels)* respectively. Outliers are excluded.

initialization problem. Scenario 1 is the simplest case with only one knot and two constant segments. Besides the edge problem, `EpiEstim` and `EpiLPS` produce "smooth" estimated curves that are continuous at the changepoint, which results in large mistakes in that neighbourhood. Since the piecewise constant `RtEstim` estimator does not force any smoothness in $\mathcal{R}_t$, it easily captures the sharp change. Scenario 2 is relatively easy for all methods, except at the change point occurring at the end of the exponential growth. Although the truth is likely best represented with a discontinuous piecewise cubic curve, the actual curvature is so gentle that linear estimation ($k = 1$) appears potentially reasonable. However, `RtEstim` has difficulty recovering the acute rise in the growth phase because it enforces continuity at the change point.

To investigate the performance when the Poisson assumption (imposed by both `RtEstim` and `EpiEstim`) is violated, we also examine estimation accuracy with negative Binomial data. Fig 5 displays a realization, analogous to the previous case, for all methods and scenarios. `RtEstim` has more difficulty relative to the Poisson setting, especially at the beginning of the outbreak. This is most pronounced in Scenario 4, where `RtEstim` is overly smooth, except in the last wave. In Scenario 2, `RtEstim` successfully captures the changepoint, but suffers from the same discontinuity problem as in the Poisson setting. In Scenario 3, the piecewise linear version of `RtEstim` recovers the curvature of $\mathcal{R}_t$ well, but is less accurate than in the Poisson case.

Finally, it is important to provide a brief comparison of the running times of all three models across the 8 experimental settings. We find that almost all models across all experiments complete within 10 seconds. `RtEstim` generally takes the longest, due to

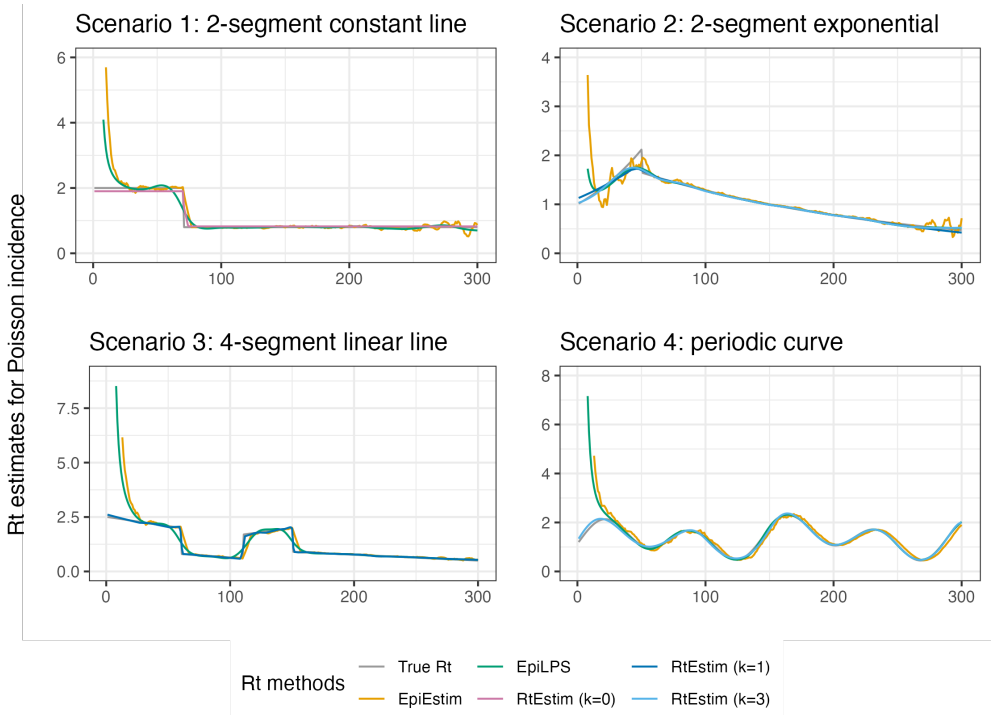

Fig 4. Example of effective reproduction number estimation for Poisson observations.

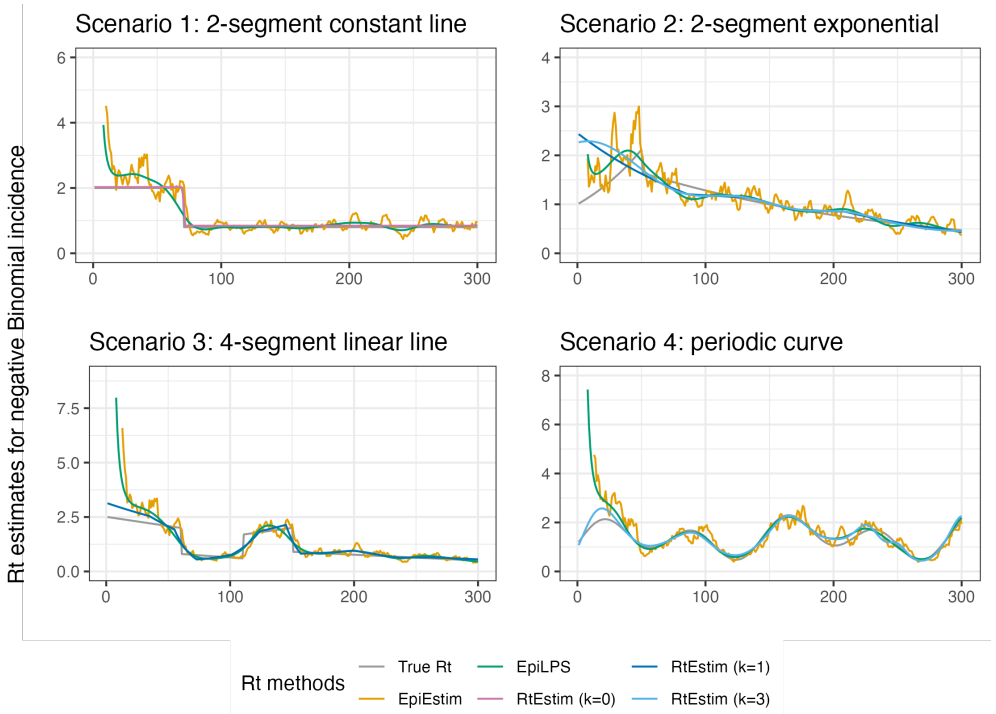

Fig 5. Example of effective reproduction number estimation for negative Binomial obsersations.

a relatively large number of estimates—50 values of λ and 10 folds of cross validation

require 500 estimates—while other models run only a single time for a fixed setting of hyperparameters per experiment. Additional results on timing comparisons are deferred to the Appendix.

3.3 Real-data results: Covid-19 incident cases in British Columbia

We implement `RtEstim` on Covid-19 confirmed incident cases in British Columbia (B.C.) as reported on May 18, 2023 (visualized in Fig 6) by the B.C. Centre for Disease Control. We use the gamma distribution with shape 2.5 and scale 2.5 to approximate the serial interval function, which is similar to empirical estimates [28].

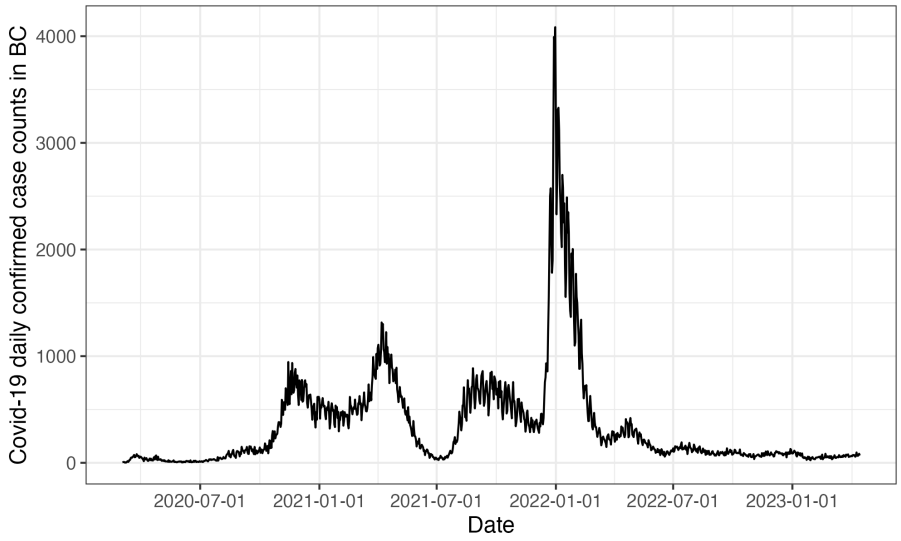

Fig 6. Covid-19 daily confirmed incident cases between March 1st, 2020 and April 15th, 2023 in British Columbia, Canada.

Considering the first, second, and third polynomial degrees, $\widehat{\mathcal{R}}_t$ for Covid-19 in British Columbia (illustrated in Fig 7) is always less than 3 except at the very early stage, which means that one distinct infected individuals on average infects less than three other individuals in the population. Examining three different settings for k, the temporal evolution of $\widehat{\mathcal{R}}$ (across all regularization levels λ) are similar near the highest peak around the end of 2021 before dropping shortly thereafter. Throughout the estimated curves, the peaks and troughs of the reproduction numbers precede the growth and decay cycles of confirmed cases, as expected. We also visualize 95% confidence bands for the point estimates with λ chosen by minimizing cross-validated KL divergence in Fig 7.

The estimated reproduction numbers are relatively unstable before April 1st, 2022. The highest peak coincides with the emergence and global spread of the Omicron variant. The estimated reproduction numbers fall below 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 coincides with the introduction of Covid-19 vaccines in British Columbia. The second trough, shortly after the largest peak may be due to variety of factors resulting in the depletion of the susceptible population such as increased self-isolation in response to media coverage of the peak or immunity incurred via recent infection. Since April 1st, 2022, the estimated reproduction number has remained relatively stable (fluctuating around one) corresponding to low reported cases,

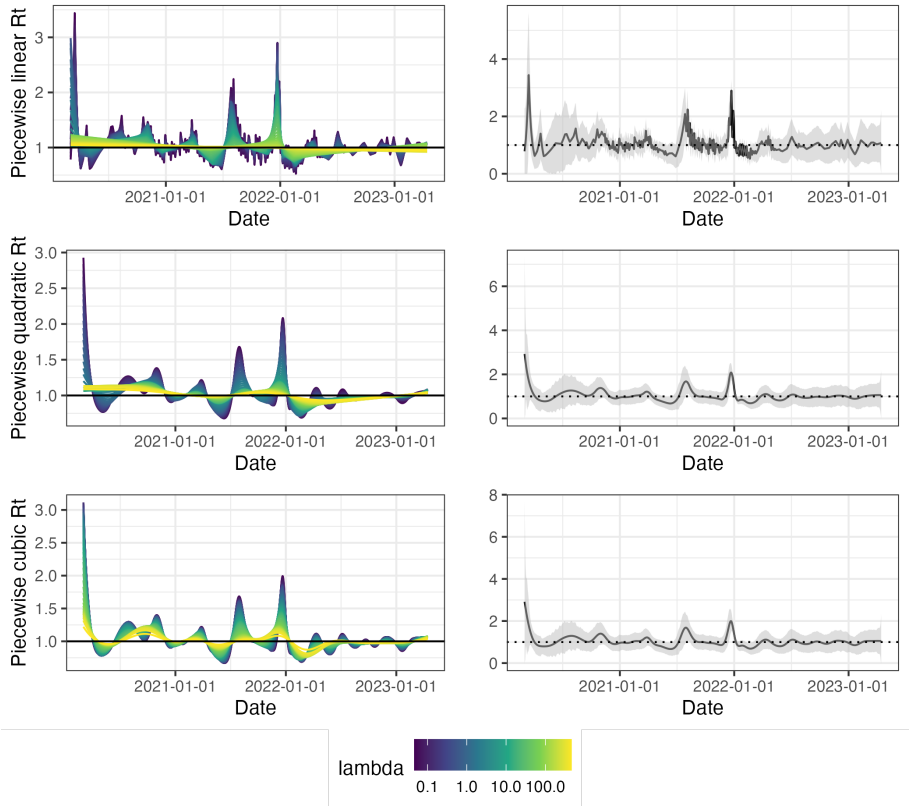

Fig 7. Estimated effective reproduction number based on Covid-19 daily confirmed incident cases between March 1st, 2020 and April 15th, 2023 in British Columbia, Canada. The left panels show estimates corresponding to 50 tuning parameters. The right panels show the CV-tuned estimate along with approximate 95% confidence bands. The top, middle and bottom panels show the estimated $\mathcal{R}_t$ using the Poisson trend filtering in Eq (4) with degrees $k = 1, 2, 3$ respectively.

though reporting behaviours also changed significantly since the Omicron wave.

3.4 Real-data results: influenza in Baltimore, Maryland, 1918

We also apply `RtEstim` to daily reported influenza cases in Baltimore, Maryland occurring during the world-wide pandemic of 1918 from September to November. The data, shown in Fig 8, is included in the `EpiEstim` `R` package. The 1918 influenza outbreak, caused by the H1N1 influenza A virus, was unprecedentedly deadly with case fatality rate over 2.5%, infecting almost one-third of the population across the world [29]. The CV-tuned piecewise cubic estimates in Fig 9 better capture the growth at the beginning of the pandemic in Fig 8. The estimated $\mathcal{R}_t$ curve suggests that the transmissibility of the pandemic grew rapidly over the first 30 days before declining below one after 50 days. However, it also suggests an increase in infectiousness toward the end of the period. With this data, it is difficult to determine if there is a second wave or a steady decline ahead. The CV-tuned piecewise constant and linear estimates in Fig 9 both suggest a steady decline. This conclusion is supported by the fact that incident cases decline to zero at the end of the period and matches $\mathcal{R}$ estimates in [2], which are all lower than one.

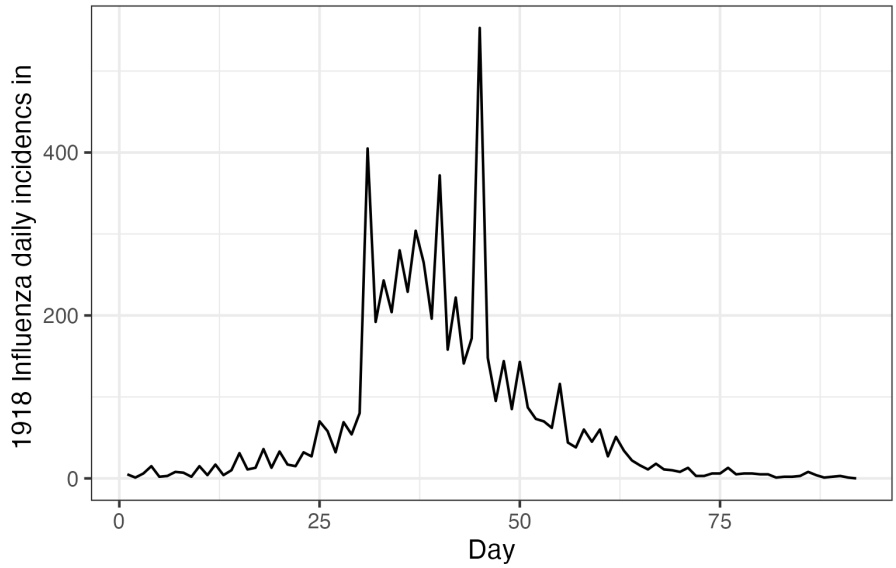

Fig 8. Daily incident influenza cases in Baltimore, Maryland between September and November in 1918.

4 Discussion

The `RtEstim` methodology provides a locally adaptive estimator using Poisson trend filtering on univariate data. It captures the heterogeneous smoothness of effective reproduction numbers given observed incidence data rather than resulting in global smoothness. This is a nonparametric regression model which can be written as a convex optimization (minimization) problem. Minimizing the distance (averaged KL divergence per coordinate) between the estimators and (functions of) observations guarantees data fidelity while the penalty on divided differences between pairs of neighbouring parameters imposes smoothness. The ℓ_1-regularization results in sparsity of the divided differences, which leads to heterogeneous smoothness across time.

The property of local adaptivity (heterogenous smoothness) is useful to automatically distinguish, for example, seasonal outbreaks from outbreaks driven by other factors (behavioural changes, foreign introduction, etc.). Given a well-chosen polynomial degree, the growth rates can be quickly detected, potentially advising public health to implement policy changes. The effective reproduction numbers can be estimated retrospectively to examine the efficacy of such policies, whether they result in $\mathcal{R}_t$ falling below 1 or the speed of their effects. The smoothness of $\mathcal{R}_t$ curves (including the polynomial degrees and tuning parameters) should be chosen based on the purpose of the study in practice, e.g., epidemic forecasting may require less smoothness while retrospective studies that solely target understanding of the pandemic may prefer a smoother estimate.

Our method `RtEstim` provides a natural way to deal with missing data, for example, on weekends and holidays or due to changes in reporting frequency. While solving the convex optimization problem, our method can be easily handle uneven spacing or irregular reporting. Computing the total primary infectiousness is also easily generalized to irregular reporting by modifying the discretization of the serial interval distribution. Additionally, because the ℓ_1 penalty introduces sparsity (operating like a median rather than a mean), this procedure is relatively insensitive to outliers compared to ℓ_2 regularization.

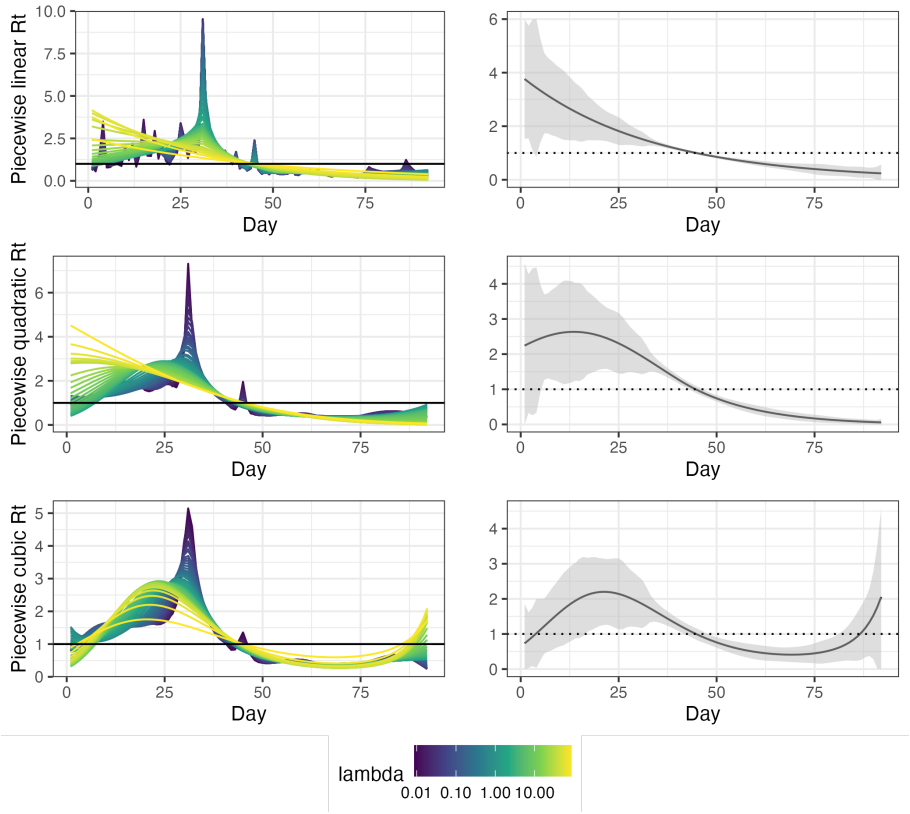

Fig 9. Estimated effective reproduction numbers for influenza in Baltimore, Maryland in 1918. The left panels show estimates for a set of 50 tuning parameters. The right column displays the CV-tuned estimates with approximate 95% confidence bands. The rows (top to bottom) show estimated reproduction numbers ($\mathcal{R}_t$) using the Poisson trend filtering in Eq (4) with $k = 1, 2, 3$ respectively.

There are a number of limitations that may influence the quality of $\mathcal{R}_t$ estimation. While our model is generic for incidence data, rather than tailored to any specific disease, it does assume that the generation interval is short relative to the period of data collection. More specialized methodologies would be required for diseases with long incubation periods such as HIV or Hepatitis. Our approach, does not explicitly model imported cases, nor distinguish between subpopulations that may have different mixing behaviour. While the Poisson assumption is common, it does not handle overdispersion (observation variance larger than the mean). The negative binomial distribution is a good alternative, but more difficult to estimate in this context. As described in section 1, the expression for $\mathcal{R}$ assumes that a relatively constant proportion of true infections is reported. However, if this proportion varies with time (say, due to changes in surveillance practices or testing recommendations), the estimates may be biased over this window. A good example is in early January 2022, during the height of the Omicron wave, British Columbia moved from testing all symptomatic individuals to testing only those in at-risk groups. The result was a sudden change that would render $\mathcal{R}_t$ estimates on either side of this time point incommensurable.

As currently implemented, `RtEstim` uses a fixed serial interval throughout the period of study, but as factors such as population immunity vary, the serial interval may vary as well [4]. Another issue relates to equating serial and generation intervals (also mentioned above). The serial interval distribution is generally wider than that of the

generation interval, because the serial interval involves the convolution of two distributions, and is unlikely to actually follow a named distribution like Gamma, though it may be reasonably well approximated by one. Our implementation allows for an arbitrary distribution to be used, but requires the user to specify the discretization explicitly, requiring more nuanced knowledge than is typically available. Pushing this analysis further, to accommodate other types of incidence data (hospitalizations or deaths), a modified generation interval distribution would be necessary, and further assumptions would be required as well. Or else, one would first need to deconvolve deaths to infection onset before using our software.

Nonetheless, our `RtEstim` estimator can be implemented easily through a lightweight `R` package `rtestim` and computed efficiently using the included proximal Newton method solver coded in `C++`, especially for large-scale data. Given available incident case data, prespecified serial interval distribution, and a choice of degree k, `RtEstim` is able to produce accurate estimates of effective reproduction number and provide efficient tuning parameter selection via cross validation.

Acknowledgments

xxx

References

1. Cori A, Cauchemez S, Ferguson NM, Fraser C, Dahlqwist E, Demarsh PA, et al.. EpiEstim: estimate time varying reproduction numbers from epidemic curves; 2020.

2. 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.

3. 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.

4. Nash RK, Bhatt S, Cori A, Nouvellet P. Estimating the epidemic reproduction number from temporally aggregated incidence data: A statistical modelling approach and software tool. PLoS Computational Biology. 2023;19(8):e1011439.

5. Abbott S, Hellewell J, Thompson RN, Sherratt K, Gibbs HP, Bosse NI, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Research. 2020;5:112.

6. Abbott S, Funk S, Hickson J, Badr HS, Monticone P, Ellis P, et al.. epiforecasts/EpiNow2: 1.4.0 release; 2023.

7. Lison A, Abbott S, Huisman J, Stadler T. Generative Bayesian modeling to nowcast the effective reproduction number from line list data with missing symptom onset dates. arXiv preprint arXiv:230813262. 2023;.

8. Parag KV. Improved estimation of time-varying reproduction numbers at low case incidence and between epidemic waves. PLoS Computational Biology. 2021;17(9):e1009347.

9. Gressani O, Wallinga J, Althaus CL, Hens N, Faes C. EpiLPS: A fast and flexible Bayesian tool for estimation of the time-varying reproduction number. PLoS Computational Biology. 2022;18(10):e1010618.

10. Trevisin C, Bertuzzo E, Pasetto D, Mari L, Miccoli S, Casagrandi R, et al. Spatially explicit effective reproduction numbers from incidence and mobility data. Proceedings of the National Academy of Sciences. 2023;120(20):e2219816120.

11. 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.

12. 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.

13. Pircalabelu E. A spline-based time-varying reproduction number for modelling epidemiological outbreaks. Journal of the Royal Statistical Society Series C: Applied Statistics. 2023;72(3):688–702.

14. Ho F, Parag KV, Adam DC, Lau EH, Cowling BJ, Tsang TK. Accounting for the Potential of Overdispersion in Estimation of the Time-varying Reproduction Number. Epidemiology. 2023;34(2):201–205.

15. Azmon A, Faes C, Hens N. On the estimation of the reproduction number based on misreported epidemic data. Statistics in Medicine. 2014;33(7):1176–1192.

16. Gressani O, Faes C, Hens N. An approximate Bayesian approach for estimation of the instantaneous reproduction number under misreported epidemic data. Biometrical Journal. 2022;65(6):2200024.

17. Jin S, Dickens BL, Lim JT, Cook AR. EpiMix: A novel method to estimate effective reproduction number. Infectious Disease Modelling. 2023;8(3):704–716.

18. Hettinger G, Rubin D, Huang J. Estimating the instantaneous reproduction number with imperfect data: a method to account for case-reporting variation and serial interval uncertainty. arXiv preprint arXiv:230212078. 2023;.

19. Gostic KM, McGough L, Baskerville EB, Abbott S, Joshi K, Tedijanto C, et al. Practical considerations for measuring the effective reproductive number, R_t. PLoS Computational Biology. 2020;16(12):e1008409.

20. Pitzer VE, Chitwood M, Havumaki J, Menzies NA, Perniciaro S, Warren JL, et al. The impact of changes in diagnostic testing practices on estimates of COVID-19 transmission in the United States. American Journal of Epidemiology. 2021;190(9):1908–1917.

21. Hitchings MD, Dean NE, García-Carreras B, Hladish TJ, Huang AT, Yang B, et al. The usefulness of the test-positive proportion of severe acute respiratory syndrome coronavirus 2 as a surveillance tool. American Journal of Epidemiology. 2021;190(7):1396–1405.

22. Pellis L, Scarabel F, Stage HB, Overton CE, Chappell LH, Fearon E, et al. Challenges in control of COVID-19: short doubling time and long delay to effect of interventions. Philosophical Transactions of the Royal Society B. 2021;376(1829):20200264.

23. Kim SJ, Koh K, Boyd S, Gorinevsky D. ℓ_1 trend filtering. SIAM Review. 2009;51(2):339–360.

24. Tibshirani RJ. Adaptive piecewise polynomial estimation via trend filtering. The Annals of Statistics. 2014;42(1):285–323.

25. Tibshirani RJ. Divided differences, falling factorials, and discrete splines: Another look at trend filtering and related problems. Foundations and Trends® in Machine Learning. 2022;15(6):694–846.

26. Sadhanala V, Bassett R, Sharpnack J, McDonald DJ. Exponential Family Trend Filtering on Lattices. arXiv preprint arXiv:220909175. 2022;.

27. Vaiter S, Deledalle C, Fadili J, Peyré G, Dossal C. The degrees of freedom of partly smooth regularizers. Annals of the Institute of Statistical Mathematics. 2017;69:791–832.

28. Lehtinen S, Ashcroft P, Bonhoeffer S. On the relationship between serial interval, infectiousness profile and generation time. Journal of the Royal Society Interface. 2021;18(174):20200756.

29. Taubenberger JK, Morens DM. 1918 Influenza: the mother of all pandemics. Emerging Infectious Diseases. 2006;17(1):69–79.