STAT 548 Qualifying Paper Report DRAFT

Justin J. Zhang

October 7, 2025

1 Summary of Paper

As the world moves forward from the Coronavirus pandemic, the question for epidemiologists and policy makers is how to best prevent the spread of future novel viruses. A major step towards epidemic prevention is the accurate statistical modelling of epidemic growth. Parag et al. (2022), henceforth denoted PTD, aim to tackle this problem by examining two popular statistics of epidemic growth: the instantaneous reproduction number, R_t, and the instantaneous growth rate, r_t. For a long time, R_t has been the preeminent statistic for health professionals, as it explicity shows when an epidemic is growing ($R_t > 1$). However, it requires model assumptions on infection time, which at best induces model error, and at worst (i.e. misspecification) incorrectly predicts epidemic growth, which has drastic effects on real-world policy, and therefore actual lives. Sceptics argue r_t is a better statistic as it does not require explicit model assumptions (Pellis et al., 2021) and is therefore more accurate at prediction. Under suitable distributional assumptions, there is a bijective link betweeen the estimates (Wallinga & Lipsitch, 2007). Morevover, PTD show that there is an even stronger relationship between R_t, r_t estimates that does not rely on model assumptions, proving that r_t has implicit assumptions. Ultimately, PTD try to reframe the question from "which statistic is better?" to "how can we use these statistics along with other metrics *in tandem* to drive our epidemic decision making?" This section covers the relevant theory and methods behind these statistics, the main contributions from PTD, and the limitations in both the paper and epidemic modelling in general.

1.1 Relevant Theory

To start, we define notation. Let $t = 1, 2, \ldots$ be a sequence of discretized time-steps, which in PTD represents days. Let I_t denote the *incidence* at time t, which is the number of new infections. Call $\{I_1, \ldots, I_T\}$ the *incidence curve*. Let $\{w_j\}_{j=0}^{\infty}$ be the *generation time distribution* (Cori et al., 2013), so that w_j denotes the probability mass that an infected individual will generate a secondary infection in exactly j days.

The *instantaneous reproduction number*, R_t, measures the mean number of secondary infections generated from each infected individual at time t, with a value greater than 1 indicating a growing epidemic. The renewal model (Fraser, 2007) in Eq. (1) is used to estimate $\hat{R}_t$. It first computes the *total infectiousness* Λ_t, the number of new infections at

time t caused by previous incidences with respect to the generation time distribution. R_t is then the multiplier of Λ_t to achieve the expected incidence number at time t, hence the *instantaneous* repreduction rate. $\hat{R}_t$ is estimated from observed infections over a period $t = 1, 2, \ldots, T$ in terms of a specific distribution for the generation time distribution using Bayesian estimation. The prior $\pi(R_t)$ is generally assumed to be from gamma distribution, and the likelihood $f(I_t \mid I_{t-1}, \ldots, I_1, w, R_t)$ assumed to be Poisson distributed. The posterior $\pi(R_t \mid I_{t-1}, \ldots, I_1, w, I_t)$ is then derived to also be a gamma distribution, and $\hat{R}_t$ is estimated based on observed incidence (Cori et al., 2013 and appendix).

$$\mathbb{E}(I_t) = \Lambda_t R_t, \qquad \Lambda_t = \sum_{j=1}^{t-1} I_{t-j} w_j \implies R_t = \frac{\mathbb{E}(I_t)}{\sum_{j=1}^{t-1} I_{t-j} w_j} \tag{1}$$

While R_t has been a mainstream measure of infectiousness for decades, it requires distributional assumptions through $\{w_j\}_{j=0}^{\infty}$ to estimate, which complicates interpretation and accuracy (Pellis et al., 2021). Accordingly, the *instantaneous growth rate*, r_t, has increased in popularity recently due its lack of distributional assumptions. r_t is derived directly from the incidence curve, along with a smooth (log-differentiable) function $\mathbb{S}_t$, as seen in the left side of Eq. (2). However, the choice of $\mathbb{S}_t$ is itself a sort of assumption as shown later.

$$r_t = \frac{\mathsf{d} \log \mathbb{S}_t}{\mathsf{d}t}, \qquad S_t = \sum_{j=(1-m)/2}^{(m-1)/2} I_{t+j} \alpha_j \tag{2}$$

An example of a smoothing function is an interpolating spline between the incidences. PTD champion the usage of the *Savitsky-Golay* (SG) filter, a local interpolation method. A SG filter of dimension m, with m odd, given in the right side of Eq. (2), performs polynomial interpolation of degree p, with $p \leq m$, on a moving window $t - \frac{1-m}{2}, \ldots, t + \frac{m-1}{2}$. It derives the coefficients $\alpha_{(1-m)/2}, \ldots, \alpha_{(m-1)/2}$ for each window by least squares estimation or using a pre-determined property (ex. standard moving average filter sets $\alpha_j = \frac{1}{m}$ for each j). The derivative is then taken with respect to the fitted polynomials.

Though no model is used, $\hat{r}_t$ is still considered an estimate in terms of a given smoothing function, which shows that the choice of $\mathbb{S}$ functions as a sort of assumption.

Under the assumption of $\{w_j\}_{j=0}^{\infty}$ for r_t there is an explicit link between $\hat{r}_t$ and $\hat{R}_t$ using its moment generating function $\mathbb{M}_w$ (Wallinga & Lipsitch, 2007). The general formulation, called the Lotka-Euler equation, as well a specfic instance where w_j is taken from a $Gamma(\alpha, \beta)$ distribution is given in Eq. (3). The equation is derived by further assuming exponential growth (or decay) for the incidence at a rate of $\hat{r}_t$, that is $I_t = I_{t-j} e^{j\hat{r}_{t-j}}$ (appendix). This gives a bijective relation to obtain $\hat{r}_t$ directly from an estimate of $\hat{R}_t$.

$$\hat{R}_t \mathbb{M}_w(-\hat{r}_t) = 1, \qquad \hat{r}_t = \beta(\hat{R}_t^{\frac{1}{\alpha}} - 1) \tag{3}$$

1.2 Main Contributions

Though this estimate $\hat{r}_t$ is model-dependent, PTD show that there is additionally a non model-dependent connection between $\hat{R}_t, \hat{r}_t$ that refutes the notion from (Pellis et al., 2021),

(Dushoff & Park, 2021) and others that the distributional assumptions on $\hat{R}_t$ minimize its usage. The novelty PTD introduce is to prove the connection between $\hat{R}_t, \hat{r}_t$, and determine their respective advantages and disadvantages. This is illustrated through a case study on the epidemic growth of the Ebola virus in West Africa (Van Kerkhove et al., 2015). $\hat{R}_t$ is estimated through the *Epifilter* package in R, but section 3 will reproduce the simulations using *rtestim*. $\hat{r}_t$ is estimated in three ways: directly using SG filters as in Eq. (2), directly using total infectiousness Λ_t as a smoothing function, and from $\hat{R}_t$ using Eq. (3). All models estimate R_t and r_t with low prediction error, though $(\hat{r}_t \mid S_t) < r_t$ consistently when I_t is small. This is because the fitted spline will further flatten the already low incidence rates, an issue more pronounced at low absolute values. The predicted $\hat{r}_t \mid S_t$ must by right shifted $\frac{m-1}{2}$ days as it requires knowledge of incidence at time $t + \frac{m-1}{2}$ and so the local spline fitted for S_t is used to predict growth rate $\hat{r}_{t+(m-1)/2}$. Likewise, $\hat{r}_t \mid \Lambda_t$ must be left shifted $\frac{m-1}{2}$ days as it requires knowledge of the incidences at $t - m, \ldots, t$ to estimate growth rate $\hat{r}_{t-(m-1)/2}$ (estimation is on midpoint). Generally, Λ_t actually requires the incidences starting at time 1, but the probability for primary transmission before $t - m$ is near 0.

The estimation of r_t using total infectiousness and SG filter as smoothing functions shows the explicit link between the model assumptions and choice of smoothing function. In fact PTD show $\Lambda_{t+\tau}$ is approximately equal to S_t for $\tau \approx \mathbb{E}(w)$. Both are functions of the daily incidence, and so under correct model specification the estimated coeffcients α_j determine it's own distribution. That is,

$$w_j \approx \alpha_{\tau-j} \implies \Lambda_{t+\tau} = \sum_{j=1}^{t+\tau-1} I_{t+\tau-j} w_j \approx \sum_{j=1}^{2\tau-1} I_{t+\tau-j}\alpha_{\tau-j} \approx \sum_{j=1-\tau}^{\tau-1} I_{t-j}\alpha_{-j} = \sum_{j=1-\tau}^{\tau-1} I_{t+j}\alpha_j \quad (4)$$

w_j will be generally near 0 for $j > 2\tau = 2\mathbb{E}(w)$ (this can be shown for example if w follows poisson). Setting $m = 2\tau - 1$ will equate the final equality in Eq. (4) with the SG filter in Eq. (2). Eq. (4) shows that generation time distirbution is functionally a smoothing filter through Λ_t, meaning the assumptions on Λ_t are used to compute both $\hat{R}_t, \hat{r}_t$ .This linkes R_t and r_t with a stronger result than Eq. (3) as this relation connects a non-model dependent $\hat{r}_t$ with $\hat{R}_t$, showing that $\hat{r}_t$ has underlying implicit model assumptions, which refutes a major advantage it has over $\hat{R}_t$ (Pellis et al., 2021). In this relation, the coefficients α_j form an arbitrary kernel for each S_t that is similar to the generation time distribution kernel. Nonetheless, Parag et al. (2022) argue there are benefits to using both statistics in conjuction to model epidemic growth and inform public policy.

One significant benefit of estimating r_t is performance when the generation time distirbution is misspecified. $\hat{R}_t$ is derived immediately from w and so naturally has high bias under misspecification, however using Eq. (3) to compute $\hat{r}_t$ from the poorly estimated $\hat{R}_t$ still recovers an estimate of r_t with low bias. Under good model specification, PTD argue $\hat{R}_t$ is a more infomrative estimate of epidemic growth since $\hat{r}_t$ is estimated from it. However, under reasonable assumptions where both estimates have low error, the two statistics can and should be used in conjunction to inform health policy. $\hat{R}_t$ quantifies the number of secondary transmissions that need to be prevented on average to slow the pandemic, which is a determining factor in epidemic control policy and vaccine coverage. $\hat{r}_t$ measures the speed

of epidemic growth, and gives metrics such as doubling time, which is a determining factor in intervention planning (ex. lockdowns). Ultimately, PTD argue that *both* $\hat{R}_t, \hat{r}_t$ are essential to understanding and implmenting informed policy for epidemics, which is of utmost importance to society as a whole.

1.3 Limitations

There are a number of limitations to the work in Parag et. al (2022), both in there analysis comparing $\hat{R}_t, \hat{r}_t$ and its applicability to accurately model epidemics. Incidence data, denoted as the time when an individual first contracts the disease is almost always earlier than the report date. Thus incidence values are right censored estimates, requiring future case numbers and incubation times, as well as other factors like susceptibility rates. In some cases, it will also be left censored if baseline infections from the start of an epidemic are not known (Fraser, 2007) due to epidemiologists not understanding the severity. Furthermore, a decent percentage of incidences are never reported, and so the true incidence is impossible to know. This adds a layer of data-dependent irreducible variance to $\hat{R}_t, \hat{r}_t$, which can have a multiplicative effect on total error, an issue PTD never mention ways to mitigate. One possible solution (Comiskey et al., 2021) is to model the distribution of incubation time, say $\{\xi_i\}_{i=1}^{\infty}$, say with gamma distribution and use observed cases C_t to back-calculate incidence, $\hat{I}_t = \sum_{i=0}^{\infty} \xi_i C_{t+i}$. This hopes that the additional variance from an extra layer of prediction is less than the reduction in variance on the main model (error due to unreported cases is still unavoidable). Another issue that is not addressed is the case of a single secondary infection coming from numerous primary sources, a common issue when larger groups of people hang out together. It is ambiguous how this should be attributed, whether it should be equal percentages to each source, attributed to a single transmitter only, or attributed in whole to each transmitter. The last option means $\hat{R}_t$ will be underestimated as there is only 1 incidence that will go towards the Λ_t calculation. The first 2 options are mathematically okay, but lead to different interpretations, which are not clarified in PTD.

There are also outside factors to consider that would impact incidence rate and hence R_t, r_t. The presence of singular events with high transmission possibility, for example festivals and concerts, will cause a jump in daily incidence. Days and seasons with good weather will have more people outdoors, leading to higher incidence as well. It may be possible to include a multiplier for total infectiousness for these effects, say $\Lambda_t = \beta_t \sum_{j=1}^{t-1} I_{t-j} w_j$, where $\lambda_t > 0$ can be estimated from a simple regression model that accounts for outside factors. λ_t should be so that values less than 1 correspond to a likely increase in infections above average (and vice versa). This is due to Λ_t being artificially lowered, and keeping the distributional assumptions on $\mathbb{E}(I_t)$ the same, R_t will be higher than the expeceted rate without accounting for outside factors. Alternatively, the times series $I_1, ..., I_t$ can incorporate these outside factors through various effect variables. This would then increase or decrease $\mathbb{E}(I_t)$ from benchmark values (i.e. no effect) accordingly, and in turn decrease or increase $\hat{R}_t$. PTD account for seasonal differences in their example, but no other outside influence addressed, though they off-handedly mention "contextual information" as something that is needed. There are also human-influenced outside factors that can shift the entire reproduction rate curve, the most prevalent being vaccine mandates and quarantines. There is a related statistic (not mentioned by PTD), the *case reproduction number*, R_t^x, which measures the average

number of secondary infections over a lifetime. This statistic will implicitly account for these changes, but can only be back-calculated as a retrospective statistic, as it requires future incidence numbers (Cori et al. 2013). There is also the *Susceptible-Infectious-Recovered* model that groups people into these three categories, but it is often an oversimplication of true epidemic dynamics (Lloyd, 2009).

Conceptually, the largest limitation is that PTD never discuss *how* to use $\hat{R}_t, \hat{r}_t$ together to understand epidemic dynamics. For a paper who's primary objectives are to clarify the understanding of these statistics, there is not any concrete applications beyond "use both". Furthermore, most of the theory is hand-waved, which is okay because they reference the background papers, but is difficult for readers to both understand their assumptions, and follow their logic. In summary, PTD do a sufficient job to compare $\hat{R}_t, \hat{r}_t$ and refute the thought that the lack of distributional assumptions in $\hat{r}_t$ is an inherent advantge, but beyond that they have not developed any novel methods or ideas that can be applied to real-world epidemiological modelling, which is subpar for an applied paper.

2 Mini-Proposals

3 Project Report

We made amazing contributions to the world of musical fractal pasta (McDonald, 2017; Tibshirani, 2013). We use Natbib, so be sure to use (Stein, 1981) for parenthetical references. Or you can say, according to Hastie et al. (2009), we should strive to balance truth and lies.

References

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Verlag.

McDonald, D. J. (2017) Minimax Density Estimation for Growing Dimension. In *Proceedings of the* 20^{th} *International Conference on Artificial Intelligence and Statistics (AISTATS)* (eds. A. Singh and J. Zhu), vol. 54, 194–203. PMLR.

Stein, C. M. (1981) Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, **9**, 1135–1151.

Tibshirani, R. J. (2013) The lasso problem and uniqueness. *Electronic Journal of Statistics*, **7**, 1456–1490.