ESTIMATING THE REPRODUCTION NUMBER IN THE PRESENCE OF SUPERSPREADING

K. JOHNSON, M. BEIGLBÖCK, M. EDER, A. GRASS, J. HERMISSON, G. PAMMER, J. POLECHOVA, D. TONEIAN, AND B. WÖLFL

Abstract. A primary quantity of interest in the study of infectious diseases is the average number of new infections that an infected person produces. This so-called reproduction number has significant implications for the disease progression. There has been increasing literature suggesting that superspreading, the significant variability in number of new infections caused by individuals, plays an important role in the spread of COVID-19. In this technical report, we consider the effect that such superspreading has on the estimation of the reproduction number and subsequent estimates of future cases. Accordingly, we employ a simple extension to models currently used in the literature to estimate the reproduction number and present a case-study of the progression of COVID-19 in Austria.

Keywords: COVID-19, reproduction number, overdispersion, superspreading.

1. Introduction

A benchmark method for the estimation of the basic reproduction number R was developed by Cori et al. [2013]. An influential framework that allows to quantify the phenomenon of *superspreading* is provided by Lloyd-Smith et al. [2005]. In this report we extend the model of Cori et al. [2013] to include the phenomenon of superspreading in the sense of Lloyd-Smith et al. [2005]. The goal is *not* to derive more accurate estimates of R but rather to better quantify the uncertainty inherent in this type of estimate.

Ultimately we are interested in the estimation of R and specifically the question whether, given current case numbers, we can claim with statistical guarantees that $R \leq 1$ or $R > 1$. Given the growing body of evidence about the existence and importance of superspreaders, we incorporate this feature into our models. We observe two important effects: first, it becomes increasingly difficult to accurately estimate the population reproduction number R in the presence of superspreading; second, models with superspreading produce prediction intervals for new cases that have approximate coverage whereas those without superpreading do not. Both of these are demonstrated in our Austrian case-study in Section 3.

In particular, the width of a credible interval for R should decrease as a function of total number of cases used during estimation and increase with the extent of superspreading. Let S be the set of days used to estimate R in the nowcasting framework presented in Section 1.1 and assume that the (average) reproduction number does not change over time. One would then expect that credible intervals have width approximately equal to

$$\frac{\text{const.}}{\sqrt{k \sum_{s \in S} I_s}}, \tag{1.1}$$

M. Beiglböck, M. Eder, A. Grass, J. Hermisson, G. Pammer, J. Polechova, D. Toneian, and B. Wölfl contributed equally.

for values of dispersion parameter k much smaller than 1; see Section 4 below. Small values of k correspond to high superspreading according to the framework of Lloyd-Smith et al. [2005].

1.1. **Nowcasting.** The goal of nowcasting is to get accurate estimates of the current state of the pandemic. Given that our observed infections are random observations from an underlying process, our goal is to understand the parameters of that process, particularly with respect to the reproduction number. In addition, we define a new time-varying parameter we call the "momentum" of the epidemic, which is a *random* realization of population infectiousness at a time-point which accounts for superspreading. This is introduced formally in Section 2

A benchmark for estimating R is the method developed in Cori et al. [2013] implemented in the software package 'EpiEstim'. An improved extension of this framework is given in Thompson et al. [2019] which accounts for variability in the generation interval (defined below). A substantial extension ('EpiNow') of the EpiEstim-package that is used to estimate R on an international level was developed by a group of researchers at the London School of Hygiene and Tropical Medicine [Abbott et al., 2020]. An important overview of the challenges involved in estimating R in the current situation is given in [Gostic et al., 2020], a comparative analysis of statistical methods to estimate R is given in [O'Driscoll et al., 2020].

If the epidemic is at an early stage, the reproduction number R and the rate of exponential growth are connected by the Euler-Lotka equation, see for instance [Wallinga and Lipsitch, 2007, Ma, 2020] for a discussion in the context of epidemiology.

As we follow the framework of Cori et al. [2013], we briefly describe their basic model. Let I_0 be the number of initial infections and $I_1, I_2, \dots$ be the number of new infections on days $1, 2, \dots$. By $(w_n)_{n\geq 1}$ we denote the *generation interval distribution.* If D_m denotes the number of people infected by a specific person on the m-th day after this person got infected, then we have for $m \in \mathbb{N}$

$$w_m = \frac{\mathbb{E}[D_m]}{\sum_{l=1}^{\infty} \mathbb{E}[D_l]}.$$

We assume that a newly infected individual does not cause secondary cases on the same day, corresponding to $w_0 = 0$.

The basic model of Cori et al. [2013] assumes that the stochastic process $(I_t)_{t\in\mathbb{Z}}$ satisfies

$$I_t \sim \text{Poisson}\left(R_t \sum_{m=1}^{\infty} I_{t-m} w_m\right), \tag{1.2}$$

for a sequence of numbers R_t, $t \in \mathbb{Z}$, and where we put $I_m = 0$ for $m < 0$. The latter convention could be interpreted as assuming that no cases occurred before time 0. In practice it is often assumed that the generation interval distribution is given as a Gamma distribution that has been discretized in such a way that $w_m = 0$ for all m larger than some cut-off number ν. As a result, the sum in (1.2) will only have $\nu \in \mathbb{N}$ summands and to make assertions about I_t we only have to consider the case numbers $I_{t-1}, \dots, I_{t-\nu}$. As ν is a parameter that can vary between diseases, this term is kept and used throughout our model description in Section 2.

When estimating the time-varying reproduction number, Cori et al. [2013] assume that the reproductive number has stayed constant over a window of τ days. Thus (1.2) simplifies to

$$I_t \sim \text{Poisson}\left(R \sum_{m=1}^{\nu} I_{t-m} w_m\right). \tag{1.3}$$

Note that the reproduction number in the sense of (1.3) does not denote number of people that actually have been infected by a given individual, rather describes what one expect in an "average" evolution of the epidemic. Furthermore, while $R_t = R$ is assumed to be constant over the window of width τ, as this window moves through time the method produces *estimates* of R_t that slowly vary over time.

1.2. **Heterogeneity in Reproduction Numbers.** The motivation for our hierarchical Bayes approach follows the framework of superspreading provided in Lloyd-Smith et al. [2005]. Even if the reproduction number $R_t = R$ is constant through time, it might vary between different individuals. We consider the reproduction number of a specific person with index x to be drawn randomly as

$$r_x \sim \text{Gamma}(k,\, k/R).$$

This distribution has mean R and variance R^2/k. The degenerate case $k = \infty$ corresponds to the deterministic case where $r_x = R$ for all individuals. Given $r_x = r$, this person causes $\text{Pois}(r)$ new infections. If one integrates out the Poisson parameter r, one is left with the unconditional number of descendants which follows a negative binomial distribution with mean R and variance $R + R^2/k$ as in Section 4.

Apparently, (1.3) corresponds to the case where each individual has the same basic reproduction number R, i.e., $k = \infty$. A basic extension of (1.3) that follows the concept of random individual reproduction numbers in the sense of Lloyd-Smith et al. [2005] is to assign, on day t, the individual reproduction numbers $r_1^t, \dots, r_{I_t}^t$ to the I_t individuals that got infected on this day. This leads to the recursion

$$I_t \sim \text{Poisson}\left(\sum_{m=1}^{\nu} w_m \sum_{x=1}^{I_{t-m}} r_x^{t-m}\right), \tag{1.4}$$

where the individual reproduction numbers r_x^m are i.i.d. according to a Gamma distribution with mean R and dispersion parameter k. Note that for the degenerate case $k = \infty$, (1.4) recovers (1.3). This forms the foundation of the model explained in detail in Section 2.

The theme of the present technical report is close to that of [Donnat and Holmes, 2020], in which heterogeneity in R between *groups* is explicitly modelled. While the high-level descriptions of these models sound nearly identical, those models are relevantly different than ours. In particular, [Donnat and Holmes, 2020] are interested in estimating group-specific or time-varying reproduction numbers for different geographical regions and age groups. On one hand, with sufficient group-specific data, this provides tools of a much broader scope than we present here; while on the other hand, it is assumed that within-group variability is negligibly small. Instead, we focus on aggregate data from a *single* geographical region but do *not* assume that individual variability is negligible. Rather, this is precisely the variability we are interested in modelling. Similarly, our critiques of the estimability of the reproduction number transfers to their setting as well: if within-group variability exists, group-specific reproduction numbers are more difficult to estimate than previously acknowledged.

2. The "Momentum" Model

As mentioned in the introduction, we identify an unobserved random variable which we term the "momentum" of the pandemic. This follows from a simple notational change in 1.4 according to the observation that a sum of i.i.d. Gamma random variables is also Gamma distributed with the same dispersion parameter. We rewrite 1.4 as

$$I_t \sim \text{Poisson}\left(\sum_{m=1}^{\nu} w_m \theta_{t-m}\right), \tag{2.1}$$

where

$$\theta_t = \sum_{x=1}^{I_t} r_x^t \sim \text{Gamma}(I_t k,\, k/R). \tag{2.2}$$

The terms $(\theta_t)_{t\geq 0}$ are collectively referred to as the "momentum" of the disease and will be treated as a set of nuisance parameters of the offspring distribution as our primary interest lies in estimating the hyperparameter R. Equation (2.1) describes the distribution of I_t conditioned on its whole past, i.e., $I_s, \theta_s,\ s < t$. Analogously, equation (2.2) is understood given the whole past, which here means $I_{s+1}, \theta_s,\ s < t$. The difference in what we understand as the relative past originates from θ_t being conceptually determined "after" I_t.

For increased clarity of the form of the model and the estimation methods required, we recast our model as a Bayesian Poisson regression using vector notation. This is made painfully explicit by using an arrow as in $\vec{I}$ for vectors. As our model is estimated over a set of τ days as in Cori et al. [2013], we specify the regression function over this window. To simplify notation, we use $[l]$, for $l \in \mathbb{N}$, to be the vector $(1, 2, \ldots, l)$. Similarly, $[l,\, m]$ for $l,\, m \in \mathbb{N}$ is shorthand for the vector $(l, l{+}1, \ldots, m)$, i.e., $[l] = [1, l]$. This notation will primarily be used for vector indices. Furthermore, the indices of our vectors increase in time. As such, our generation interval truncated to ν days can be condensely written as $\vec{w}_{[\nu]} = (w_1, \ldots, w_\nu)$. Similarly, we have $\vec{I}_{[t-\tau+1,\, t]} = (I_{t-\tau+1}, \ldots, I_t)$.

This regression formulation is important as it highlights the latent variables $\vec{\theta}$ that are required to fully determine the generative model. It also focuses attention on which observations are conditioned upon and which are treated as random, i.e., the τ observations to which we fit the model. This is relevant as more than τ nuisance parameters are present, namely $\nu + \tau - 1$. Observe that the earliest data point is $I_{t-\tau+1}$, which itself requires a history of ν momentum values of $\vec{\theta}$ to determine.

While we also think of individual reproduction numbers as changing over time due to factors such as changes in social restrictions, the assumption of constant R over a period renders this moot. For now, the dispersion parameter k is assumed constant, and we will check the distribution and estimability of R at various values of k. In general, however, a contact tracing data set would be used to estimate this parameter as in Laxminarayan et al. [2020], which identified $k = .072$ using Indian data. This is also within the range of parameter values identified in Endo et al. [2020] using Chinese data.

2.1. **Likelihood.** As we have parameterized the gamma prior on θ_t to have mean $I_t R$, it is transparent below that R has an inverse-gamma distribution. Hence we use an Inv-Gamma$(\alpha,\, \beta)$ hyperprior where these hyperparameters are set such that R has mean 2.2 and standard deviation 2 as in Abbott et al. [2020]. This yields $\alpha = 3.69$ and $\beta = 6.994$. We are suppressing notation for conditioning on all observations before time $t - \tau + 1$. Furthermore, given $\vec{\theta}_{[t-1]}$, I_t is independent of $\vec{I}_{[t-1]}$.

A full derivation of the likelihood function and posterior distribution of the pair $R, \vec{\theta}_{[t]}$ given $\vec{I}_{[t]}$ is given the Appendix. We obtain as posterior

$$
\begin{aligned}
&p(R, \vec{\theta}_{[t-\tau+1,\,t-1]} | \vec{I}_{[t-\tau-\nu+1,\,t]}) \\
&\propto p(\vec{I}_{[t-\tau+1,\,t]}, \vec{\theta}_{[t-\tau+1,\,t-1]} | \vec{\theta}_{[t-\tau-\nu+1,\,t-\tau]}, \vec{I}_{[t-\tau-\nu+1,\,t-\tau]}, R) p(\vec{\theta}_{[t-\tau]}, R | \vec{I}_{[t-\tau]}) \\
&\propto \left(\prod_{s=t-\tau+1}^{t} \left(\sum_{m<s} w_{s-m}\theta_m \right)^{I_s} e^{-\sum_{m<s} w_{s-m}\theta_m} \right) \left(\prod_{s=t-\tau+1}^{t-1} \frac{k^{I_s k}}{\Gamma(I_s k) R^{I_s k}} \theta_s^{I_s k - 1} e^{-\frac{k}{R}\theta_s} \right) \\
&\quad \cdot \left(\prod_{s=t-\nu-\tau+1}^{t-\tau} \frac{k^{I_s k}}{\Gamma(I_s k) R^{I_s k}} \theta_s^{I_s k - 1} e^{-\frac{k}{R}\theta_s} \right) \left(R^{-\alpha-1} e^{-\beta/R} \right),
\end{aligned}
$$

where the last line in the display corresponds to the chosen prior for R the momentum values $\vec{\theta}_{[t-\tau-\nu+1,\,t-\tau]}$ given the corresponding infections $\vec{I}_{[t-\tau-\nu+1,\,t-\tau]}$. In particular, we treat $\vec{I}_{[t-\tau-\nu+1,\,t-\tau]}$ as constant so that the model does not further specify nuisance parameters before time $t - \tau - \nu + 1$. This prevents an infinite recursion of in historical observations. Hence we need for a prior not only for R, but also for $\vec{\theta}_{[t-\tau-\nu+1,\,t-\tau]}$.

In order to condense notation for summations in exponents, let S_1 and S_2 be the index sets for the first and second product, respectively; i.e., $S_1 = \{t-\tau+1, t-\tau+2, \ldots, t\}$ and $S_2 = \{t-\nu-\tau+1, t-\nu-\tau+2, \ldots, t-1\}$. The additional shorthand below drops "$s \in$" from $s \in S_i$, $i \in \{1, 2\}$. With this notation, the posterior distribution of R given $\vec{\theta}$ and $\vec{I}$ is

$$
p(R|\vec{\theta}_{[t-1]}, \vec{I}_{[t]}) \propto R^{-k\sum_{S_2} I_s - \alpha - 1} e^{(-k\sum_{S_2}\theta_s - \beta)R^{-1}},
$$

which is Inv-Gamma($k\sum_{S_2} I_s + \alpha,\ k\sum_{S_2}\theta_s + \beta$). A perhaps counter-intuitive observation is that the posterior distribution of R does not depend on the generation interval $\vec{w}_{[\nu]}$. This is the result of considering $\vec{\theta}$ fixed versus integrating out it out as done in Lloyd-Smith et al. [2005]. In our case, this approach is infeasible as the dependence on $\vec{\theta}$ is too complex. If we truly know population infectiousness, i.e., the pandemic momentum at all points in time, then $\vec{w}_{[\nu]}$ is irrelevant for estimating R, because $\vec{w}_{[\nu]}$ just determines *how we learn* about $\vec{\theta}$.

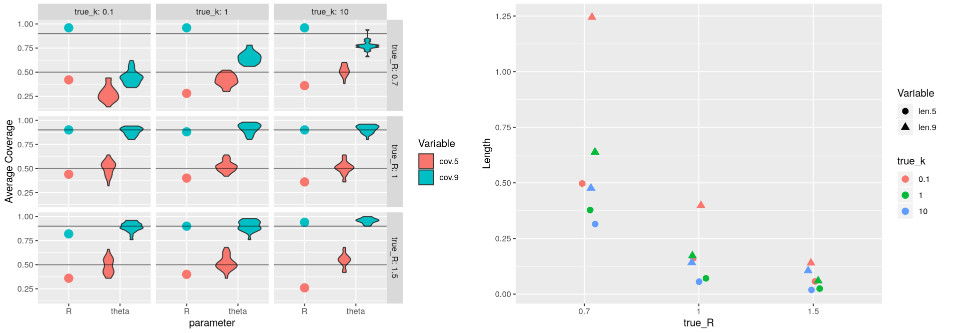
The posterior expectation and variance of R are

$$
\begin{aligned}
\mathbb{E}[R|\vec{\theta}, \vec{I}] &= \frac{k\sum_{S_2}\theta_s + \beta}{k\sum_{S_2} I_s + \alpha - 1} \qquad \text{and} \\
\mathrm{Var}[R|\vec{\theta}, \vec{I}] &= \frac{(k\sum_{S_2}\theta_s + \beta)^2}{(k\sum_{S_2} I_s + \alpha - 1)^2 (k\sum_{S_2} I_s + \alpha - 2)}.
\end{aligned}
$$

The denominator of the variance picks up an additional k term, making credible intervals wider when k is small. The dependence on $\vec{\theta}$ is difficult to remove in this general setting. Section 4 considers a simpler setting in which $\vec{\theta}$ can be integrated out in order to derive a transparent function for credible interval width.

To estimate this model, we alternate between a Gibbs-step to sample R and a Metropolis-Hastings step to sample $\vec{\theta}$. As $\mathbb{E}[\theta_s|I_s, R] = I_s R$, we can initialize reasonable starting values for $\vec{\theta}$ using various values of R such that we require little burn-in. We find total chain length to be the more important tuning parameter for accurate credible intervals. In all models presented in this paper, we set $\nu = \tau = 13$ and $\vec{w}_{[\nu]}$ a discretized gamma distribution with mean 4.46 and standard deviation 2.63.

2.2. **Model Validation.** Here we summarize estimation results for simulated data in order to more precisely show the effect of superspreading in a setting in which true parameters are known. The coverage and length of intervals are shown in



(A)

(B)

Figure 1. Illustration of average credible interval coverage and length on simulated data. As there is a single R parameter but 25 elements of $\vec{\theta}$, the coverage of the latter are summarized via a violin plot.

Figure 1. All simulations were run using an initial sequence of cases that had constant value of 50. Simulations were repeated 20 times in order to asses coverage probabilities.

Of greatest initial import is verifying that the 90% credible intervals for R indeed cover the true value with approximately nominal probability. The case $R = 1$ is of primary importance, as it represents the bright-line between the epidemic growing or shrinking. That we have nearly exact coverage in this setting is indication that our credible intervals do not achieve coverage merely by being extremely wide. Furthermore, the intervals for $\vec{\theta}$ also cover the true values with the specified probability in the $R = 1$ and $R = 1.5$ cases. With our initial sequence of cases and $R = .7$, the epidemic sometimes dies out, which can be missed by the model. As such, the $R = .7$ cases has worse coverage over the momentum parameters $\vec{\theta}$.

After establishing coverage, our motivation for modeling superspreading is verified by looking at the lengths of the credible intervals: for k small, our intervals need to be extremely wide. In fact, the interval for $k = .1$ is approximately 2.5 times longer than the interval for $k = 10$ for both $R = .7$ and $R = 1$. For $R = 1.5$, the estimation problem becomes relatively easy as case numbers grow substantially. This leads to very small credible intervals.

As the explicit conditional distribution of the momentum parameters $\vec{\theta}$ is intractable, we present a summary of the samples observed through the MCMC simulation in Figure 2. This includes all 25 momentum parameters required when $\tau = \nu = 13$ as well as R. As $R = 1.5$ in this setting, you can observe the scale increase for θ_s as s increases. It is clear that the parameters vary widely through MCMC estimation, even though the are initialized at the marginal MLE: $\hat{\theta}_s = I_s\hat{R}$. Multiple chains are run, each with a separate initial value for $\hat{R}$. When k is small, variability in $\vec{\theta}$ is large, requiring long chains to be simulated in order to overcome high auto-correlation in the MCMC draws of $\vec{\theta}$.

3. Case Study of COVID-19 in Austria

This section primarily focuses on understanding the evolution of the reproduction number in Austria between approximately mid-April and mid-September. The initial date corresponds roughly to the end of the majority of COVID restrictions in Austria. Our goals are three-fold: to demonstrate the increase in estimated variability in R due to superspreading, to provide valid prediction intervals for new

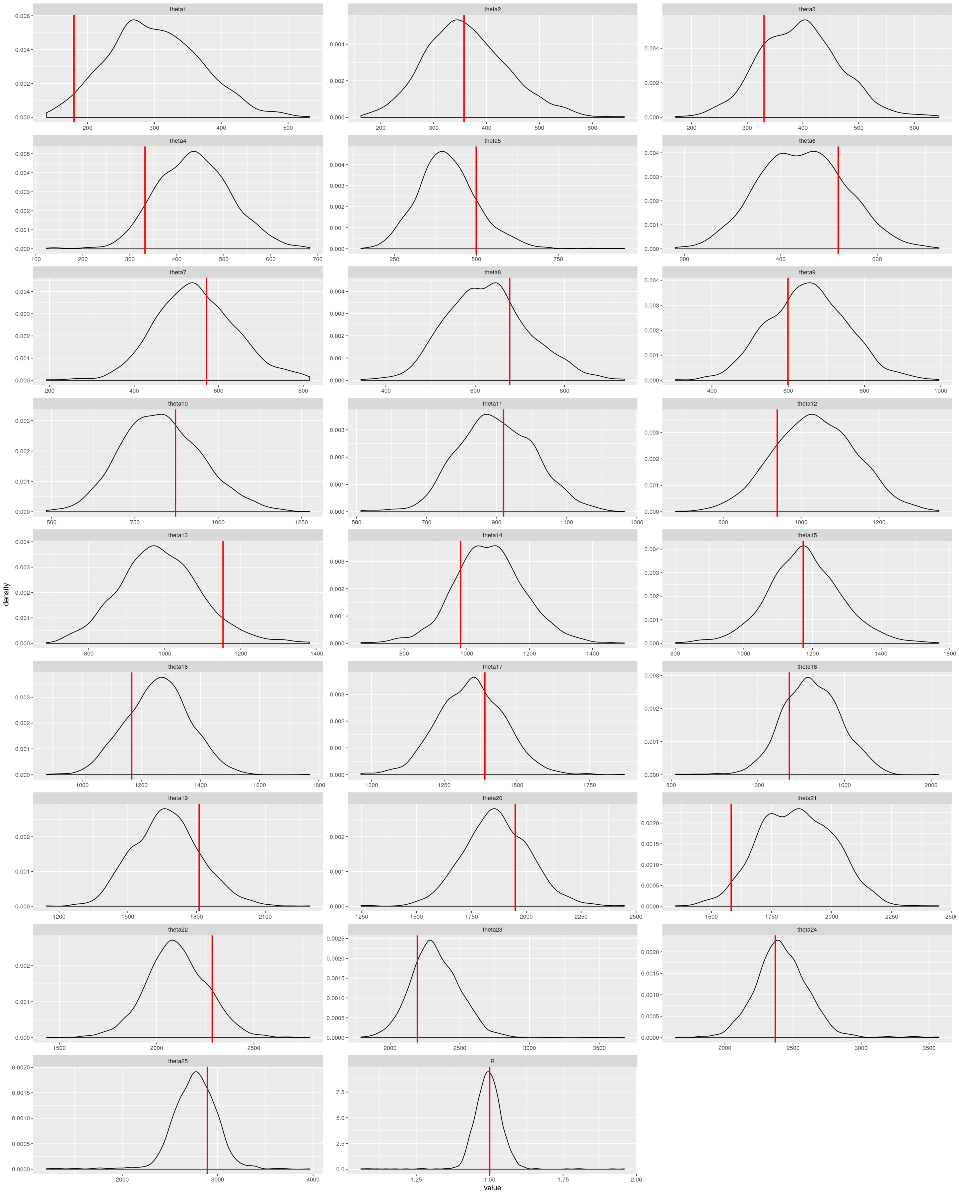


Figure 2. Samples of MCMC draws of parameters. The vertical, red lines indicate true values.

cases, and to compare to similar models without superspreading. Some results will be shown for other countries to help establish the validity of our method, but the focus is on Austrian data.

An important component of the estimation on real data is to account for the delay distribution between infections and observed cases as discussed in Gostic et al. [2020]. If a delay of length d occurs between infection and observation, then an infection included in I_t actually occurred on day $t-d$. As input for our model, we use the samples of potential infection histories as described in Abbott et al. [2020]. Austrian data is shown in Figure **??** and includes curves for the raw infection data as reported by the European Centre for Disease Prevention and Control (Raw), the 7-day moving average of Raw (Raw (MA)), each sampled infection history (Sampled

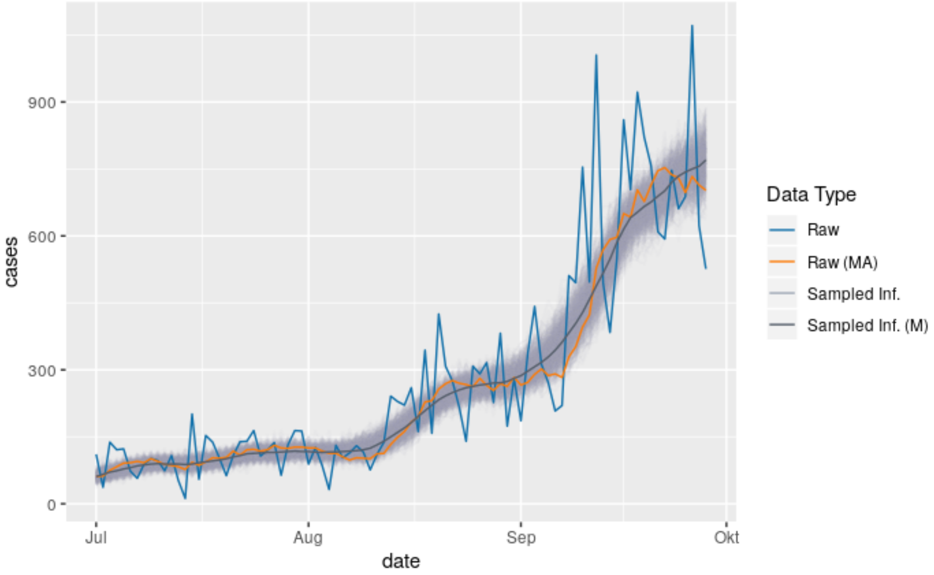


Figure 3. Summary of new cases of COVID-19 in Austria: raw infection data (Raw), the 7-day moving average of Raw (Raw (MA)), each sampled infection history (Sampled Inf.), and the daily median of the sampled infection histories (Sampled Inf. (M))

Inf.), and the daily median of the sampled infection histories (Sampled Inf. (M)). Observe that the boundary of the "band" created by the sampled infection histories is not smooth, as it is created from 1,000 separate faded lines.

For each infection history, we estimate our model using MCMC and keep 1,000 total draws from the markov chain after thinning and burn in. As this is done for each of the 1,000 potential infection histories, we are left with 1 million total samples which comprises our distributional estimates of R and momentum vector $\vec{\theta}$. To forecast future cases, we use an individual sample of parameters and run the momentum model for a specified period of time. Our graphs show results for the total number of new cases over the following week. It is important to note that before plotting we do no additional smoothing either with rolling averages or post-processing. As our goal is to cover the observed number of new infections *as reported*, such smoothing would make the estimation problem significantly easier. Between this and only plotting one observation per week, our graphs may be "rougher" or contain more spikes than commonly seen in other sources; this is by design.

In all of the following graphs, we plot predictions and intervals from three models: the momentum model with $k = .072$, the momentum model with $k = 10$, and the EpiEstim model of Cori et al. [2013]. The momentum model with $k = 10$ is included to demonstrate that our model effectively reduces to the EpiEstim model for "large" k, where $k = 10$ is sufficient as increasing k further yields no noticeable differences in parameter estimation. Practically speaking, estimating the momentum model with k significantly larger than 10 also poses numerical problems as the markov chain used moves incredibly slowly due to the implied near-identity relating $\vec{\theta}$ and $\vec{I}$ to R. We label the EpiEstim model "Epi*", because it contains a mixture of components from the EpiEstim and EpiNow modeling frameworks so as to be the appropriate competitor to our model. Most importantly, there is no superspreading but we

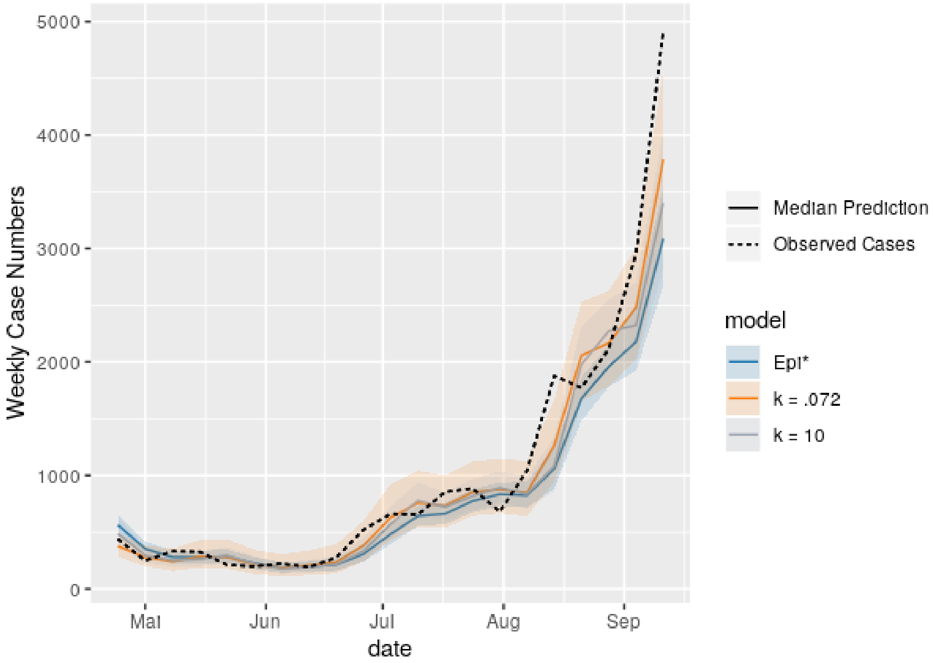


FIGURE 4. Prediction intervals for forecast of the total new cases in the following week with Austrian Data.

estimate the model for multiple samples from the delay distribution as described in the previous paragraph. As in Cori et al. [2013], we fix a generation interval, as opposed to taking samples of a generation interval estimated from a separate data source as in Thompson et al. [2019]. As a result, we are not comparing to perhaps the "best in class" model within the EpiEstim/EpiNow framework, but with a model of corresponding complexity to the momentum model. Other improvements to the modeling framework could then be built on top of the momentum model as they have been for the model of Cori et al. [2013].

To estimate the model of Cori et al. [2013], we estimate the parameters of the Cori et al. [2013] posterior distribution directly from the infection data:

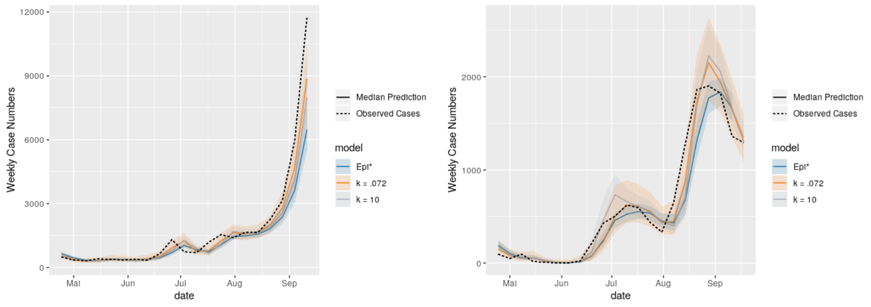
$$p(R_t|I_{[t]}) = \text{Gamma}\left(a + \sum_{s=t-\tau+1}^{t} I_s,\, b + \sum_{s=t-\tau+1}^{t} \sum_{m=1}^{\nu} w_m I_{s-m}\right),$$

where a and b are the shape and rate parameter of the gamma prior distribution on R. For each sampled infection history $I_{[t]}$, we estimate this posterior distribution, draw samples for R, and run the momentum model with $k = \infty$ for the required number of days.

Figure 4 shows the difference between models with and without superspreading. All of the prediction curves track the observed cases similarly as forecasts are only produced for the coming week. The difference in their prediction intervals, however, is significant. Most notably, the intervals for the momentum model with $k = .072$ are much wider than the others. Table 1 shows, for each method, the proportion of true weekly total new cases that fall within the prediction intervals over the prediction period. Coverage is shown for the 50% and 90% prediction intervals for both the raw infection data as well as the smoothed infection data Sampled Inf. For both the smoothed and raw data, the momentum model with $k = .072$ provides approximate coverage, while those without superspreading severely under cover.

Table 1. Table of coverage of prediction intervals for number of total cases in the following week.

Country	Data-Type	Model	Cov: 50%	Cov: 90%
Austria	Sampled Inf.	$k = .072$	.64	.95
		$k = 10$	.27	.73
		Epi*	.27	.59
	Raw	$k = .072$	.43	.90
		$k = 10$	.05	.57
		Epi*	.10	.38
Czech Republic	Sampled Inf.	$k = .072$	.59	.95
		$k = 10$	.045	.55
		Epi*	.18	.41
	Raw	$k = .072$	.38	.81
		$k = 10$	.24	.43
		Epi*	.19	.33
Croatia	Sampled Inf.	$k = .072$	.48	1.0
		$k = 10$	.17	.57
		Epi*	.32	.41
	Raw	$k = .072$	.32	.86
		$k = 10$	.23	.41
		Epi*	.10	.33



(a) (b)

Figure 5. Prediction intervals for forecast of the total new cases in the following week with Czech Data (A) and Croatian Data (B).

Figures 5a and 5b show the same graphs but for the Czech Republic and Croatia. The disease progression in the Czech Republic is similar to that of Austria over the shown period. Croatia is a common Austrian tourist destination and the disease progression is markedly different there than in Austria. The estimated coverage probabilities of the prediction intervals are also shown in Table 1. The story remains the same as before: coverage is far better for the momentum model with superpreading than without. The coverage on the raw data for the momentum model with superspreading is not as accurate as for the Austrian data, but is still a significant improvement over the models without it.

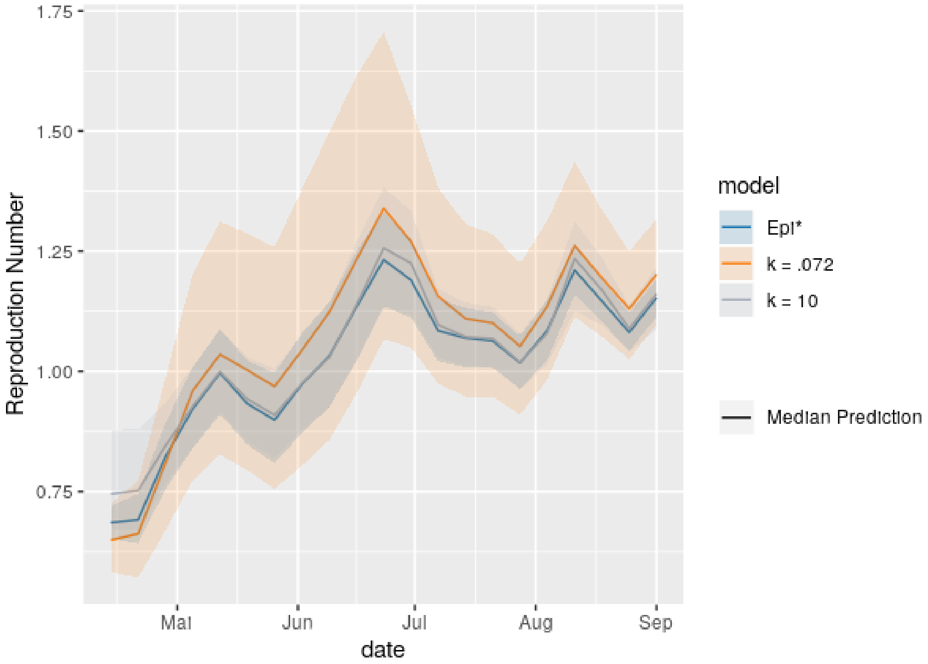
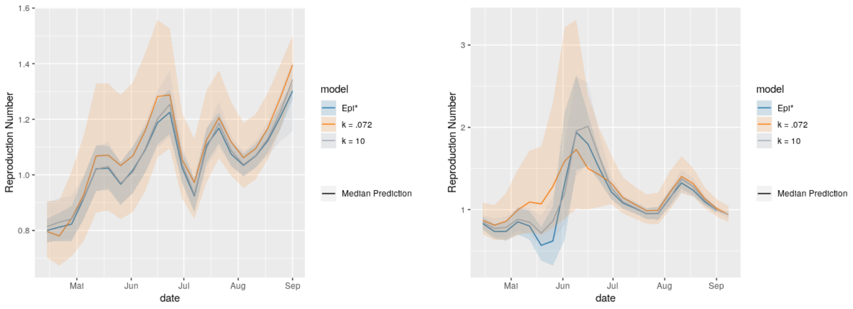


FIGURE 6. Credible intervals for R in Austria.

As the reproduction number is unobserved, we are unable to compare our predictions within a supervised setting as we compared our model forecasts. Given the previous discussion though, we believe our model performs well in Austria and thus can be used to infer about the reproduction number R. Figure 6 shows the 90% credible intervals for R derived by the momentum and Epi* models. The figure clearly demonstrates that the intervals for R are drastically different: with superspreading, intervals for R are roughly 2-3 *times* as wide as those without. Coupled with the results on raw data, this increased width appears necessary to achieve prediction intervals which provide coverage on both the smoothed and raw data. This could have potentially large implications for policy making as we know that relatively small changes in the size of R can lead to large differences in the number of new cases if the disease is allowed to progress unchecked.

At the beginning of our estimation period, at the time when restrictions were being relaxed in Austria, it quickly becomes infeasible to claim that the reproduction number is below 1; i.e., the credible interval starting in the beginning of May includes the value 1. While there have been some fluctuations over the summer, the majority of the period was worsening though remained indistinguishable from $R = 1$. As of August however, we estimate that R is strictly greater than 1, even with our comparatively wide credible intervals.

Figures 7a and 7b show the corresponding graphs but for the Czech Republic and Croatia. The story is again primarily the same, in which the momentum model with superspreading produces much wider credible intervals. One obvious feature of the Croatian data, however, is a large spike around mid June. This corresponds to a large increse and subsequent decreases in cases as seen in Figure 5b. Both models without superspreading estimate that R increases to approximately 2 with an interval estimate of roughly [1.5, 2.5]. Alternatively, in the same period, the momentum model with superspreading provides noticeably lower estimates but with an increadibly wide interval. Further exploration of the feature is warranted,



(A)

(B)

Figure 7. Credible intervals for R in the Czech Republic (A) and Croatia (B).

though it appears justifiable to say that a large spike and drop in cases over a small window window of time should produce significant more uncertainty in the value of the underlying parameter. This is still being explored.

4. Trivial Generation Interval

In order to directly relate the dispersion parameter k to the width of the credible interval, we consider the trivial generation interval in which an infected person is only infectious for a single day. Furthermore, such a simple formula comes at the cost of making several approximations as explained below. As such, results in this section should be considered heuristic as opposed to concrete.

When the generation interval w is of this form, $\vec{w}_{[1]} = (1)'$, the model is purely Markovian and the data follow a Galton-Watson process. Recall that a Poisson(λ)-distributed random variable Y where λ is a hyperparameter distributed according to Gamma(α, β), follows a negative binomial distribution as discussed in Lloyd-Smith et al. [2005]:

$$Y \sim NB\left(\alpha, \frac{1}{1+\beta}\right), \quad p(Y) = \frac{\Gamma(Y+\alpha)}{Y!\Gamma(\alpha)}\left(\frac{\beta}{1+\beta}\right)^{\alpha}\left(\frac{1}{1+\beta}\right)^{Y}. \tag{4.1}$$

Applying (4.1) to the model from Section (2.1) yields for the infections I_t on day t given the past infections and the hyperparameters R and k that

$$I_t|\vec{I}_{[t-1]}, R, k \sim NB\left(kI_{t-1}, \frac{R}{R+k}\right), \tag{4.2}$$

$$p(I_t|\vec{I}_{[t-1]}, R, k) = \frac{\Gamma(I_t + kI_{t-1})}{I_t!\Gamma(kI_{t-1})}\left(\frac{k}{R+k}\right)^{kI_{t-1}}\left(\frac{R}{R+k}\right)^{I_t}. \tag{4.3}$$

Then due to the tower property, the joint distribution of $\vec{I}_{[t-\tau+1:t]}|\vec{I}_{t-\tau}, R, k$ decomposes into a product of factors of the form (4.3). We have

$$\begin{aligned} p(\vec{I}_{[t-\tau+1:t]}|\vec{I}_{[t-\tau]}, R, k) &= \prod_{s=t-\tau+1}^{t} p(I_s|\vec{I}_{[s-1]}, R, k) \\ &= \prod_{s=t-\tau+1}^{t} \frac{\Gamma(I_s + kI_{s-1})}{I_s!\Gamma(kI_{s-1})}\left(\frac{k}{R+k}\right)^{kI_{s-1}}\left(\frac{R}{R+k}\right)^{I_s}. \end{aligned}$$

The special structure of the likelihood function makes it natural to consider estimating $R/(R+k)$ instead of R. When treating $\vec{I}_{[t-\tau]}$ and k as constant, we

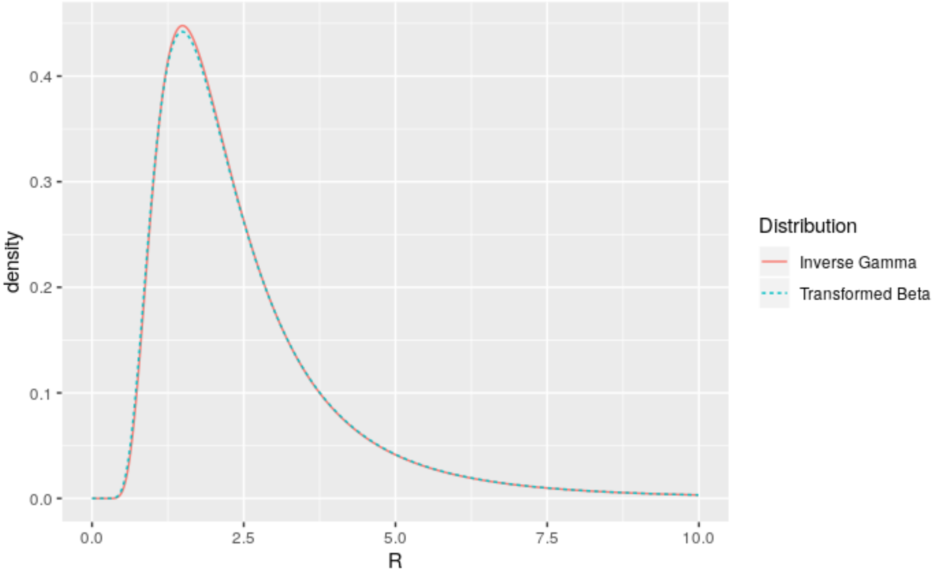


FIGURE 8. Comparison of priors on R and $R/(R+k)$.

can derive due to Bayes theorem the posterior distribution of $R/(R+k)$ given the observations $\vec{I}_{[t-\tau+1:t]}$,

$$p\left(\frac{R}{R+k}\middle|\vec{I}_{[t]},k\right) \propto \left(\frac{k}{R+k}\right)^{k\sum_{s=t-\tau}^{t-1} I_s}\left(\frac{R}{R+k}\right)^{\sum_{s=t-\tau+1}^{t} I_s} p\left(\frac{R}{R+k}\middle|I_{[t-\tau]},k\right). \tag{4.4}$$

Given this functional form, it is natural to put a beta prior on $R/(R+k)$. As shown in Figure 8, this corresponds to putting an appropriate inverse-gamma prior on R. Therefore, to mimic the $R \sim$ Inf-Gamma$(3.69, 6.994)$ prior distribution used in Section 2 we can use a Beta$(\tilde{\alpha} = 71.63, \tilde{\beta} = 3.755)$ prior on $R/(R+k)$. The posterior distribution of $R/(R+k)$ has a Beta distribution with parameters

$$\alpha = \tilde{\alpha} + \sum_{s=t-\tau+1}^{t} I_s, \quad \beta = \tilde{\beta} + k\sum_{s=t-\tau}^{t-1} I_s.$$

By a change of variables from $R(R+k)$ back to R, we derive the posterior distribution

$$p\left(R\middle|\vec{I}_{[t]},k\right) \propto \frac{k}{(R+k)^2}\left(\frac{R}{R+k}\right)^{\alpha-1}\left(\frac{k}{R+k}\right)^{\beta-1}. \tag{4.5}$$

Next, we compute the normal approximation of the posterior [Gelman et al., 2004, Section 4.1]. To this end, the first and second derivatives of the log-posterior density are

$$\frac{d}{dR}\log p(R|\vec{I}_{[t],k}) = \frac{\alpha-1}{R} - \frac{\alpha+\beta}{R+k},$$
$$\frac{d^2}{dR^2}\log p(R|\vec{I}_{[t]},k) = -\frac{\alpha-1}{R^2} + \frac{\alpha+\beta}{(k+R)^2}.$$

Thus, the mode of the posterior is

$$\hat{R} = \frac{k(\alpha-1)}{\beta+1},$$

and the variance estimate is

$$\left(-\frac{d^2}{dR^2}p(R|\vec{I}_{[t]},k)(\hat{R})\right)^{-1}=\frac{k^2(\alpha+\beta)(\alpha-1)}{\alpha+\beta}.$$

Thus, the normal approximation of the posterior is given by

$$p(R|\vec{I}_{[t]},k)\approx N\left(\frac{k(\alpha-1)}{\beta+1},\frac{k^2(\alpha+\beta)(\alpha-1)}{(\beta+1)^3}\right).$$

Consider the common setting in which $\sum_{s=t-\tau+1}^{t} I_s$ and $k\sum_{s=t-\tau}^{t-1} I_s$ are significantly larger than 1 and where we have $\beta \approx k * \alpha$ As the terms in these two sums almost entirely overlap, they will be approximately equal so long as the estimation window τ is not extremely small, e.g., 1 or 2. This yields the following simplification of the variance of the normal approximation:

$$\frac{k^2(\alpha+\beta)(\alpha-1)}{(\beta+1)^3}\approx\frac{k^2\alpha^2(k+1)}{k^3\alpha^3}=\frac{k+1}{k\alpha}\approx\frac{k+1}{k\sum_{s=t-\tau+1}^{t} I_s}.$$

Hence, the approximate length of a credible interval for R behaves like

$$\frac{const}{\sqrt{k\sum_{s=t-\tau+1}^{t} I_s}}.$$

5. Conclusion

In this technical report, we provide a simple extension of the Cori et al. [2013] model to account for superspreading. This "momentum" model, incorporates unobserved random variables which drive the process of new infections. Even if case numbers and R are relatively small, the presence of superspreaders can increase the momentum of the disease beyond what would be expected if all individuals have the same infectiousness R. The momentum model produces wider credible intervals and wider posterior predictive intervals. We find that these wider intervals are necessary to achieve approximate coverage of the observed number of new cases.

References

S. Abbott, J. Hellewell, R. N. Thompson, K. Sherratt, H. P. Gibbs, N. I. Bosse, J. D. Munday, S. Meakin, E. L. Doughty, J. Y. Chun, et al. Estimating the time-varying reproduction number of sars-cov-2 using national and subnational case counts. *Wellcome Open Research*, 5(112):112, 2020.

A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology*, 178(9):1505–1512, 2013.

C. Donnat and S. Holmes. Modeling the heterogeneity in covid-19's reproductive number and its impact on predictive scenarios. *arXiv preprint arXiv:2004.05272*, 2020.

A. Endo, S. Abbott, A. Kucharski, and S. Funk. Estimating the overdispersion in covid-19 transmission using outbreak sizes outside china. *Wellcome Open Research*, 5:67, 07 2020. doi: 10.12688/wellcomeopenres.15842.2.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman and Hall/CRC, 2nd ed. edition, 2004.

K. M. Gostic, L. McGough, E. Baskerville, S. Abbott, K. Joshi, C. Tedijanto, R. Kahn, R. Niehus, J. A. Hay, P. M. De Salazar, et al. Practical considerations for measuring the effective reproductive number, rt. *medRxiv*, 2020.

R. Laxminarayan, B. Wahl, S. R. Dudala, K. Gopal, C. Mohan, S. Neelima, K. S. J. Reddy, J. Radhakrishnan, and J. Lewnard. Epidemiology and transmission dynamics of covid-19 in two indian states. *medRxiv*, 2020. doi: 10.1101/

2020.07.14.20153643. URL https://www.medrxiv.org/content/early/2020/07/17/2020.07.14.20153643.

J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066): 355–359, 2005.

J. Ma. Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling*, 5:129–141, 2020.

M. O'Driscoll, C. Harry, C. A. Donnelly, A. Cori, and I. Dorigatti. A comparative analysis of statistical methods to estimate the reproduction number in emerging epidemics with implications for the current covid-19 pandemic. *medRxiv*, 2020.

R. Thompson, J. Stockwin, R. D. van Gaalen, J. Polonsky, Z. Kamvar, P. Demarsh, E. Dahlqwist, S. Li, E. Miguel, T. Jombart, et al. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics*, 29:100356, 2019.

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

6. Appendix

Let ν denote the "length" of the generation interval by which me mean the last day where there is positive probability of infection someone, i.e. $\omega = (\omega_1, \dots, \omega_\nu)$ and all other ω_t are zero. Let τ denote the number of days for which we assume the reproductive number R to be constant.

We know the following:

$$\begin{aligned}
I_t|\theta_{t-\nu:t-1} &\sim Pois\left(\sum_{s=1}^{\nu}\omega_s\theta_{t-s}\right), \text{ i.e.}\\
p(I_t|\theta_{t-\nu:t-1}) &= \frac{1}{I_t!}\left(\sum_{s=1}^{\nu}\omega_s\theta_{t-s}\right)^{I_t}\exp\{-\sum_{s=1}^{\nu}\omega_s\theta_{t-s}\}, \text{ and}\\
\theta_s|R, I_s &\sim \Gamma\left(I_s k, \frac{k}{R}\right), \text{ i.e.}\\
p(\theta_s|R, I_s) &= \frac{\left(\frac{k}{R}\right)^{I_s k}}{\Gamma(I_s k)}\theta_s^{I_s k-1}\exp\left\{-\frac{\theta_s k}{R}\right\}.
\end{aligned}$$

We want to consider $I_{1:\tau}$ and for this we need $\theta_{1-\nu:\tau-1}$ (for which again we need R and $I_{1-\nu:\tau-1}$).

Let us consider the joint distribution

$$\begin{aligned}
p(R, \theta_{1-\nu:0}, I_1|I_{1-\nu:0}) &= p(R|I_{1-\nu:0})p(\theta_{1-\nu:0}, I_1|R, I_{1-\nu:0})\\
&= p(R|I_{1-\nu:0})p(I_1|R, I_{1-\nu:0}, \theta_{1-\nu:0})p(\theta_{1-\nu:0}|R, I_{1-\nu:0})\\
&= p(R|I_{1-\nu:0})p(I_1|\theta_{1-\nu:0})\prod_{s=1-\nu}^{0}p(\theta_s|R, I_{1-\nu:0})\\
&= p(R|I_{1-\nu:0})p(I_1|\theta_{1-\nu:0})\prod_{s=1-\nu}^{0}p(\theta_s|R, I_s)
\end{aligned}$$

where we used that I_1 is independent of $I_{1-\nu:0}$ (as well as R) given $\theta_{1-\nu:0}$, the $\theta_{1-\nu:0}$ are conditionally independent given $I_{1-\nu:0}$, and that θ_s given I_s is independent of all other I.

Then, inductively

$$
\begin{aligned}
&p(R,\theta_{1-\nu:\tau-1},I_{1:\tau}|I_{1-\nu:0})\\
&=p(R,\theta_{1-\nu:\tau-2},I_{1:\tau-1}|I_{1-\nu:0})p(\theta_{\tau-1},I_\tau|R,\theta_{1-\nu:\tau-2},I_{1-\nu:\tau-1})\\
&=p(R,\theta_{1-\nu:\tau-2},I_{1:\tau-1}|I_{1-\nu:0})p(I_\tau|R,\theta_{1-\nu:\tau-1},I_{1-\nu:\tau-1})p(\theta_{\tau-1}|R,\theta_{1-\nu:\tau-2},I_{1-\nu:\tau-1})\\
&=p(R,\theta_{1-\nu:\tau-2},I_{1:\tau-1}|I_{1-\nu:0})p(I_\tau|R,\theta_{1-\nu:\tau-1})p(\theta_{\tau-1}|R,I_{1-\nu:\tau-1})\\
&=p(R,\theta_{1-\nu:\tau-2},I_{1:\tau-1}|I_{1-\nu:0})p(I_\tau|\theta_{1-\nu:\tau-1})p(\theta_{\tau-1}|R,I_{\tau-1})\\
&=p(R,\theta_{1-\nu:0},I_1|I_{1-\nu:0})\prod_{t=2}^{\tau}p(I_t|\theta_{1-\nu:t-1})p(\theta_{t-1}|R,I_{t-1})\\
&=p(R|I_{1-\nu:0})p(I_1|\theta_{1-\nu:0})\prod_{s=1-\nu}^{0}p(\theta_s|R,I_s)\prod_{t=2}^{\tau}p(I_t|\theta_{1-\nu:t-1})p(\theta_{t-1}|R,I_{t-1})\\
&=p(R|I_{1-\nu:0})\prod_{s=1-\nu}^{\tau-1}p(\theta_s|R,I_s)\prod_{t=1}^{\tau}p(I_t|\theta_{1-\nu:t-1})\\
&\propto p(R,\theta_{1-\nu:\tau-1}|I_{1-\nu:\tau})
\end{aligned}
$$

So for the precise likelihoods we then have

$$
\begin{aligned}
&p(R,\theta_{1-\nu:\tau-1}|I_{1-\nu:\tau})p(R|I_{1-\nu:0})\\
&\propto\prod_{s=1-\nu}^{\tau-1}p(\theta_s|R,I_s)\prod_{t=1}^{\tau}p(I_t|\theta_{1-\nu:t-1})\\
&=p(R|I_{1-\nu:0})\prod_{s=1-\nu}^{\tau-1}\frac{\left(\frac{k}{R}\right)^{I_sk}}{\Gamma(I_sk)}\theta_s^{I_sk-1}\exp\left\{-\frac{\theta_sk}{R}\right\}\prod_{t=1}^{\tau}\frac{1}{I_t!}\left(\sum_{s=1}^{\nu}\omega_s\theta_{t-s}\right)^{I_t}\exp\left\{-\sum_{s=1}^{\nu}\omega_s\theta_{t-s}\right\}\\
&=p(R|I_{1-\nu:0})\prod_{s=1-\nu}^{\tau-1}R^{-I_sk}\theta_s^{I_sk-1}\exp\left\{-\frac{\theta_sk}{R}\right\}\prod_{t=1}^{\tau}\left(\sum_{s=1}^{\nu}\omega_s\theta_{t-s}\right)^{I_t}\exp\left\{-\sum_{s=1}^{\nu}\omega_s\theta_{t-s}\right\}
\end{aligned}
$$

The main text merely separates the product over θ_s into two groups.