Inference on the dynamics of COVID-19 in the USA

Satarupa Bhattacharjee[1], Shuting Liao[2], Debashis Paul[1], and Sanjay Chaudhuri[3,*]

[1]Department of Statistics, University of California, Davis, 95616, USA
[2]Graduate Group in BioStatistics, University of California, Davis, 95616, USA
[3]Department of Statistics and Applied Probability, National University of Singapore, 117546, Singapore
[*]corresponding author: stasc@nus.edu.sg

ABSTRACT

The evolution of the COVID-19 pandemic is described through a time-dependent stochastic dynamic model in discrete time. The proposed multi-compartment model is expressed through a system of difference equations. Information on the social distancing measures and diagnostic testing rates are incorporated to characterize the dynamics of the various compartments of the model. In contrast with conventional epidemiological models, the proposed model involves interpretable temporally static and dynamic epidemiological rate parameters. A model fitting strategy built upon nonparametric smoothing is employed for estimating the time-varying parameters, while profiling over the time-independent parameters. Confidence bands of the parameters are obtained through a residual bootstrap procedure. A key feature of the methodology is its ability to estimate latent unobservable compartments such as the number of asymptomatic but infected individuals who are known to be the key vectors of COVID-19 spread. The nature of the disease dynamics is further quantified by relevant epidemiological markers that make use of the estimates of latent compartments. The methodology is applied to understand the true extent and dynamics of the pandemic in various states within the USA.

1 Introduction

The novel coronavirus has been ravaging the world since early 2020. First identified in Wuhan, Hubei Province, China, the epidemic has since spread to every corner of the world. As of February 5, 2021[1], more than 105 million people have been infected, out of which more than 2.1 million have died of the disease. The World Health Organization declared the situation a pandemic on March 11, 2020. Since then, various parts of the world have gone through multiple waves surges in the number of new infections. The pandemic has severely affected the world economy. Repeated lockdowns, travel restrictions, and other measures of containment have severely impacted the economy of many countries, stretched healthcare systems to the extreme, and caused mental health crisis for large chunks of the populations.

The new pathogen (SARS-CoV-2) that causes the disease[2] is mostly unknown in terms of its infectivity and clinical profile. It is well-known that the infection primarily spreads through infected but asymptomatic people[3–5]. The number of such people remains unknown. The reported number is based on symptomatic or positively tested persons, which grossly underestimates of the true value. Because of the undetermined denominator effect, important epidemiological markers like the death rate, hospitalization rate etc remains non-determinable from the observed data. Various estimates[6–10] of these markers have been postulated by many authors. Mathematical modelling and quantification of the epidemiological parameters[11–16] of the pandemic have been crucial in understanding and interpreting the transmission dynamics from the perspective of public health researchers and policymakers around the globe[17–20]. The dynamics of COVID-19 in various states of the United States has been studied by several authors[21,22]. We analyze such publicly available state-wise COVID-19 data from USA using the proposed methodology.

A number of popular compartmental epidemiological models, such as SIR (Susceptible-Infectious-Recovered) model, SEIR (Susceptible-Exposed-Infectious-Recovered) model and SIRD (Susceptible-Infectious-Recovered-Deceased) model, have been employed to describe the dynamics of COVID-19[23–26]. Such models yield estimates of epidemilogical markers such as the basic reproduction number (R_0), and various doubling and case fatality rates that are indicators of the disease growth pattern[27,28]. Prediction of epidemiological characteristics and transmission patterns in this context have also attracted major attention[29–32]. Advanced statistical methods have been employed in forecasting the number of cases worldwide[33] or quantifying the effects of prevention mechanisms like social distancing[34–39], public gathering, and travel restrictions[40–42] for various countries. Due to the difference in analytical methods and assumptions, the parameter estimates describing COVID-19 dynamics vary widely. This variability is also reflected in the estimates of the effectiveness of public health interventions implemented worldwide. Most epidemiological models of disease transmission are simplistic and use time-invariant transmission rates. However, in reality, due to mitigation efforts and the evolving nature of the infection mechanism, such rates become temporally dynamic. Furthermore, most SEIR-type models exclude the effects of testing and subsequent quarantining, and occasionally, even hospitalization. Such practices fail to adequately account for the size of the susceptible population and therefore tend to provide unreliable estimates of the number of asymptomatic persons infected by COVID-19 in the population.

We propose a detailed discrete-time semiparametric stochastic dynamic model for COVID-19 spread. The model is expressed through a system of difference equations connecting various interpretable compartments in the disease dynamics such as individuals who are susceptible, asymptomatic but infected, quarantined, hospitalized, dead, and have recovered from the disease. The model has interpretable time-varying parameters that reflect various temporally dynamic rates. The model also includes available information on the number of tests. On the other hand, the proposed model does not make restrictive and often untestable distributional assumptions about compartments or parameters that are commonplace in various probablistic models for the epidemiological dynamics.

We employ nonlinear nonparametric regression techniques through a profiling-based estimation procedure to estimate the model parameters and the number of people in different compartments. Using residual bootstrap based techniques, we also provide point-wise confidence intervals (bands) for the time-invariant (time-varying) parameters. The proposed model and estimation procedure relies on linear kernel weighting and fairly low dimensional optimization, thus avoiding Markov Chain Monte Carlo and other computationally expensive methods employed by Bayesian inference schemes for standard epidemiological models. Therefore, the estimates can be obtained almost instantaneously. Another key feature of this method is the ability of identifying and estimating unobservable quantities such as the actual number of asymptomatic but infected people at any given time. The estimated trajectory of the infected but asymptomatic population over time, its doubling rate, the true case fatality rate, and an analogue of the basic reproduction rate are crucial in interpreting the time-dynamics of the pandemic. They have important implications for policy decisions regarding appropriate mitigation strategies.

The contributions here are significant for the following reasons. Since the number of infected but asymptomatic individuals is unknown, conventional epidemiological models of disease spread do not readily apply to the COVID-19 dynamics. The adaption of these models to COVID-19 spread necessitates strong assumptions and costly numerical computations. Our proposed model provides a computationally inexpensive method for estimating several unobserved states as well as relevant parameters governing the spread of the disease. Various epidemiological markers based on these estimates are introduced to reveal the true extent of the pandemic in the United States.

2 A Multi-compartment Model for Disease Spread

Throughout, a closed population without emmigration or immigration is assumed. The model describes the spread of the Covid-19 pandemic in terms of various observable and partially or totally unobservable compartments.

Suppose at time t, C_t, D_t, T_t, respectively, denote the number of confirmed cases, number of deaths due to the disease and the number of tests performed up to time t. These variables are non-decreasing cumulative counts and are generally fully observed. The number of hospitalized persons due to COVID-19 infection at time t (denoted H_t) is also generally observed (see Section 4.2 for more detail). Furthermore, we observe Q_t, the number of asymptomatic individuals who are in quarantine at time t. These individuals have been tested positive, but show no significant symptoms requiring hospitalization.

The most crucial unobserved compartment is A_t, i.e., the number of infected but asymptomatic individuals at time t. It is well-known that the people in this group are primary spreaders of the disease. Furthermore, due to under-reporting, the number of confirmed cases would be a fraction of A_t. Since we do not observe how many in the population are currently infected, the number of susceptible individuals at time t, (denoted S_t) is also unobserved.

The number of recovered individuals (denoted R_t) up to time t can be partially observed. To understand this, note that the recoveries from quarantine centres and hospitals, (denoted R_t^Q and R_t^H respectively) are reported, though not necessarily separately (see Supplement Section S2., for the case when R_t^Q and R_t^H are reported separately). But since A_t is unobserved, the number of asymptomatic but infected people who recover without being quarantined or hospitalized (denoted R_t^A) cannot be observed. That is, even though $R_t^{reported} = R_t^Q + R_t^H$ is available from the data, the total recovery R_t is not.

The proposed disease propagation model are based of the following assumptions:

A1 Only an asymptomatic individual who is not either in quarantine or in hospital can transmit the disease to a susceptible individual.

A2 People who recover from the disease are immune from subsequent infection.

A3 False positive rate for the test is negligible, so that if somebody is confirmed to be positive, then he/she is assumed to be infected.

A4 Anybody who shows significant symptoms, whether being in quarantine or not, is immediately hospitalized, and is tested to be positive.

A5 There is no effective treatment regime for the asymptomatic individuals, and so they recover or turn symptomatic at the same rate regardless of whether they are tested positive (and hence quarantined) or not.

A graphical representation of the proposed disease propagation model is presented in Figure 1 below. The assumptions A1-A5 are quite general and concur to the observed dynamics of COVID-19 pandemic so far, even though a relatively tiny fraction of people do get infected by prolonged exposure to symptomatic patients, typically in hospitals. However, this small violation of assumption A1 is unlikely to have a significant influence on the overall dynamics, and in any case, the requisite data to account for this violation is practically unavailable. The number of reported reinfection after recovery is negligible, so are the false positive rates of both RTPCR and antigen tests (estimated to be less than 5%[43–46]). If necessary, the assumptions A2 and A3 can be generalised by adding a fraction of the recovered people in the susceptible category. Assumption A5 implies that the rate of transfer from compartment A_t to R_t^A is same as that of transfer from the compartments Q_t to R_t^Q and the rate of transfer from the compartments A_t and Q_t to H_t are equal.

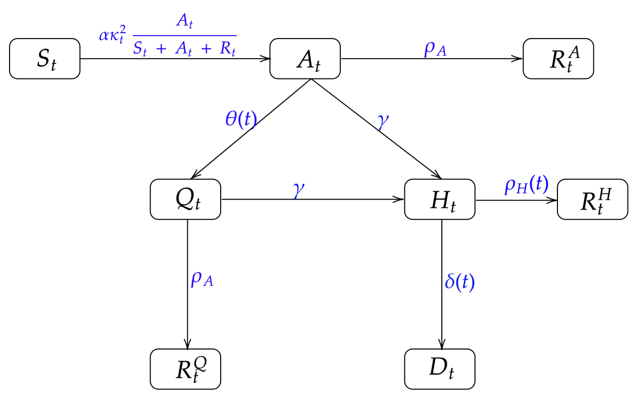


Figure 1. A graphical representation of the disease propagation model. S_t, A_t, H_t, Q_t, D_t are the number of susceptible, infected, hospitalized, quarantined, and deceased people at time t respectively. R_t^Q, R_t^H, R_t^A represent the recovered population from quarantined, hospitalization, and infected but asymptotic stages respectively. The rate parameters are as described in Section 2.1.

2.1 Disease Propagation Model

We assume an underlying Poisson process model for describing the disease dynamics. Let $\Delta C_t = C_{t+1} - C_t$ be the increments in the number of observed confirmed cases in day $t+1$. The increments ΔA_t, etc. are defined similarly. Under our model, conditionally on the current values of different compartments (collectively denoted by $\mathscr{F}_t$), the above increments follow Poisson distributions with their mean depending on $\mathscr{F}_t$ and a set of rate parameter. Based on our assumptions, the evolution model are expressed as follows:

$$\mathbb{E}[\Delta S_t|\mathscr{F}_t] = -\left(\frac{S_t}{S_t + A_t + R_t}\right)\alpha\kappa_t^2 A_t, \tag{1}$$

$$\mathbb{E}[\Delta A_t|\mathscr{F}_t] = -(\theta(t) + \gamma + \rho_A)A_t + \left(\frac{S_t}{S_t + A_t + R_t}\right)\alpha\kappa_t^2 A_t, \tag{2}$$

$$\mathbb{E}[\Delta Q_t|\mathscr{F}_t] = \theta(t)A_t - (\gamma + \rho_A)Q_t, \tag{3}$$

$$\mathbb{E}[\Delta H_t|\mathscr{F}_t] = \gamma(A_t + Q_t) - (\rho_H(t) + \delta(t))H_t,\ \mathbb{E}[\Delta D_t|\mathscr{F}_t] = \delta(t)H_t, \tag{4}$$

$$\mathbb{E}[\Delta C_t|\mathscr{F}_t] = (\theta(t) + \gamma)A_t, \tag{5}$$

$$\mathbb{E}[\Delta R_t^A|\mathscr{F}_t] = \rho_A A_t,\ \mathbb{E}[\Delta R_t^Q|\mathscr{F}_t] = \rho_A Q_t,\ \mathbb{E}[\Delta R_t^H|\mathscr{F}_t] = \rho_H(t)H_t, \tag{6}$$

$$\mathbb{E}[\Delta R_t|\mathscr{F}_t] = \mathbb{E}[\Delta R_t^H|\mathscr{F}_t] + \mathbb{E}[\Delta R_t^Q|\mathscr{F}_t] + \mathbb{E}[\Delta R_t^A|\mathscr{F}_t]. \tag{7}$$

A schematic diagram of the proposed model can be found in Figure 1. All parameters in the proposed model are non-negative. The parameter α is the baseline infection rate, in the absence of any social distancing. This means, α is the average number of susceptible individuals who may be infected on any given day by an asymptomatic but infected individual. The rate of daily recovery directly from the asymptomatic compartment is denoted by ρ_A. By assumption A5, this is also the daily rate at which a quarantined individual directly recovers. We use γ to describe the rate at which an asymptomatic individual may become symptomatic on a given day. By assumption A5, this rate is the same whether the individual is free or in quarantine. The symbols, $\rho_H(t)$ and $\delta(t)$, respectively, denote is the rate at which people recover and die from the hospitalized compartment. We assume both these rates to be time-varying to reflect the changing levels of effectiveness of treatment regimes over time. We emphasize that Poisson distributions for the increments of various compartments is only a working assumption which guides our estimation strategy (e.g., by formulating appropriate transformations of variables). In Supplement Sections S6. and S7., we carry out a detailed numerical simulation under the Poisson model to validate the statistical performance of the proposed estimation procedure.

Information about daily tests is included in the model using the function $\theta(t)$. We call it the *confirmed fraction (CF)*, i.e. the fraction of currently asymptomatic individuals who are detected through testing. Parameter $\theta(t)$ would depend on the daily number of tests, as well as the efficiency of the testing strategy in identifying the infected and asymptomatic individuals. It can also be viewed as a intervention parameter, controlling the overall testing rate per hospitalization. The contact tracing strategies were introduced by many states[47,48] with varying success. In many parts of the world, people in close contact of hospitalized patients are routinely tested. This strategy is closely connected to cluster sampling, where a cluster is defined by the contacts of a hospitalized person.

Guided by the above consideration, we reformulate the parameter $\theta(t)$ by expressing it as follows:

$$\theta(t) = \phi(t)\frac{\Delta T_t}{H_t}, \tag{8}$$

where $\phi(t)$ is interpreted as the *testing efficiency (TE)* since it measures the fraction of confirmed asymptomatic cases per test, per (currently) hospitalized patient. We use $\Delta T_t/H_t$ as a surrogate for the contact tracing intensity, since this quantity literally represents the number of new test on day $t+1$, per hospitalized (and hence severely symptomatic) patients. Clearly, the value of $\theta(t)$ is modulated by this ratio, while the factor $\phi(t)$ implicitly quantifies the extent of positivity among those tested, after accounting for the testing intensity, which justifies the nomenclature “testing efficiency”. Hypothetically, one may aim to estimate $\theta(t)$ in terms of the number of people who have been tested positive. However, in most countries (including the USA) contact tracing was limited, making estimation of $\theta(t)$ difficult. A realistic alternative is to estimate $\phi(t)$ instead, which provides an estimate of $\theta(t)$ via (8) and makes our model interpretable and more flexible. Note that, we don’t assume $\phi(t)$ to be known. It is estimated from data (see Section 3).

In addition, our model (see equation (1)) depends on a variable κ_t, which is the current state of the level of interaction among individuals. Expressed as a fraction, taking value 1 for normal activity, and 0 for complete lockdown, this variable measures the prevalent social distancing in the population. In general κ_t is not observable. However, often a surrogate variable based on data collected by internet service providers such as Google from the usage of smartphones[49] can be used[50–52].

From equation (1) the variable $\alpha\kappa_t^2$ approximately measures the daily rate at which a susceptible individual turns asymptomatic-infected. In the early stage of the epidemic, the fraction $S_t/(S_t+A_t+R_t) \approx 1$. Furthermore, rather than waiting for herd immunity to be achieved, mitigation measures are implemented in most affected places or countries to contain the spread of the disease. As a consequence, at any given time, the number of non-susceptible people is much lower as compared to the susceptible population. So $S_t/(S_t+A_t+R_t)$ has remained quite close to 1 for almost the duration of the pandemic until this point, due to the absence of mass-scale vaccination.

Notice that equation (5), provides a connection between the daily reported confirmed cases ΔC_t and the number of asymptomatic-infected individuals A_t in the population. In our model, an asymptomatic-infected person can be discovered either through a positive test and subsequent quarantining, or through hospitalization upon showing severe symptoms. Therefore, once the estimates of $\theta(t)$ and γ are available, equation (5) allows us to estimate the unknown A_t from the observed C_t. It is also clear that, due to unavoidable severe under-reporting, ΔC_t will only be a fraction of the number of total infected individuals at any time point.

2.2 Some relevant epidemiological markers

The proposed model is more realistic than the traditional such as SIR model, SEIR model etc., and allows us to estimate different epidemiological markers which can measure the dynamics of disease spread. Our focus here is on estimating epidemiological markers related to the number of asymptomatic but infected persons (i.e. A_t) in the population. It is well-known that the disease is mostly spread through persons in that group. Thus the proposed epidemiological markers reveal more fundamental trends of disease dynamics, than what can be obtained only by the confirmed case counts. In particular, we define the following epidemiological markers:

2.2.1 Relative Change in Confirmed Fraction (RCCF)

The relative change in confirmed fraction measures the change in the fraction of currently asymptomatic-infected individuals who are caught in the quarantine net through testing relative to the total fraction of currently infected individuals are either quarantined or hospitalized. From Section 2.1 we get:

$$RCCF(t) = \frac{\Delta\theta(t)}{\theta(t)+\gamma}. \tag{9}$$

The above equation is obtained by applying the difference operator on both sides of the equation $\Delta C_t = (\theta(t)+\gamma)A_t$ (see (16)), and subsequently dividing both sides by ΔC_t. The marker $RCCF(t)$ measures the dynamics of the efficacy of the testing regime to isolate the asymptomatic but infected individuals from the population into quarantine. From equation (8), this marker is directly controlled by the prevalent testing strategy and efficiency.

2.2.2 Crude infection rate (CIR) and Net Infection Rate (NIR)

The crude infection rate is defined as the fraction of change in the daily confirmed cases on a day to the number of confirmed cases on that day. In our notation, it follows that:

$$CIR(t) = \frac{\Delta^2 C_t}{\Delta C_t}. \tag{10}$$

Since CIR suffers from the under-representation inherent in the reported number of confirmed cases, we define a model-based estimate for the infection rate, denoted Net Infection Rate (NIR), which is the ratio of the daily change in the number of asymptomatic-infected persons to the number of the asymptomatic-infected persons. In our notations, from (9), (16), and (17) simple algebraic manipulations yield:

$$NIR(t) = \frac{\Delta A_t}{A_t} = \frac{CIR(t) - RCCF(t)}{1 + RCCF(t)}. \tag{11}$$

2.2.3 Daily New Infections (NI)

From our model and assumptions, the daily number of new infections are given by the number of susceptible population who turn asymptomatic-infected on that day. From equation (1) we define this marker as:

$$NI(t) = \alpha\kappa_t^2 \left(\frac{S_t}{S_t + A_t + R_t}\right) A_t. \tag{12}$$

The cumulative number of new infections up to time t can be defined as $CNI(t) = \sum_{i=1}^{t} NI(i)$.

2.2.4 Doubling Times and Rates:

The doubling time at time t, denoted $t_d(t)$ measures how much longer it would take for the number of infected up to time t to double. The doubling rate at time t, $\tilde{\xi}(t)$ is given by the inverse of the doubling time. A higher doubling rate reflects the faster spread of infection. This rate is often used to measure the effect of social distancing campaigns, improved hygiene and case tracking.

The doubling time for C_t computed using the relationship $C_{t+t_d(t)}/C_t = 2$. A first order approximation (see Supplement Section S4.) yields $t_d(t) \approx \left[\frac{d}{dt}\log C_t\right]^{-1}$. That is the doubling rate $\tilde{\xi}(t) = t_d(t)^{-1} = \frac{d}{dt}\log C_t$. Doubling rates for other compartments can be computed similarly.

2.2.5 Crude and Net Case Fatality Rates

In general a *case fatality rate* at time t is given by the ratio of the total death count and the total case count at that time. Depending on whether the reported case counts or the actual case counts are used, we can define two different case fatality rates. The *crude case fatality rate* (CFR) is defined as:

$$CFR(t) = \frac{D_t}{C_t} \times 100, \tag{13}$$

whereas the *net case fatality rate* is given by

$$NFR(t) = \frac{D_t}{CNI(t)} \times 100. \tag{14}$$

2.2.6 The Basic Reproduction Rate

In the conventional SIR or SEIR models, basic reproduction rate (R_0), which measures the expected number of cases directly generated by one case in a population where all individuals are susceptible to infection[53], is used to determine the nature, rate of growth and possible measures for controlling the pandemic[27,28].

Our model is more detailed and allows for time varying parameters and as a result, the conventional R_0 cannot be directly estimated from our model. The closest epidemiological quantity we can observe is the background infection rate, α, measuring the average number of susceptible individuals who may be infected on any given day by an asymptomatic but infected individual. However, an analogue of the basic reproduction rate for the compartment A_t can be computed[54,55].

By focusing on the compartment A_t, under our assumptions from equation (2) new infections arrive at the compartment at the rate of $\alpha\kappa_t^2 S_t/(S_t+A_t+R_t)$ and leave at the rate of $(\theta(t)+\gamma+\rho_A)$. There is no other pathway for disease spread. Thus we can define an analogue of the basic reproduction rate as:

$$\tilde{R}_0(t) = \frac{\alpha\kappa_t^2}{\theta(t)+\gamma+\rho_A}\left(\frac{S_t}{S_t+A_t+R_t}\right). \tag{15}$$

Note that, the proposed $\tilde{R}_0(t)$ can be interpreted in the same way as the conventional basic reproduction rate. By construction $\tilde{R}_0(t) < 1$ indicates negative growth of the number of asymptomatic-infected persons, whereas $\tilde{R}_0(t) > 1$ indicates its positive growth. However, temporal variation of $\tilde{R}_0(t)$ is more complex. Assuming that, $S_t/(S_t+A_t+R_t) \approx 1$, $\tilde{R}_0$ can decrease with time either due to reduction in κ_t, that is the current state of interaction among individuals, or due to an increase in the confirmed fraction $\theta(t)$. That is, the proposed $\tilde{R}_0(t)$ is directly influenced by the mitigation efforts such as social distancing, adherence to use of masks, increased testing and subsequent quarantining, hospitalization of symptomatic patients etc.

Most epidemiological models such as SIR, SEIR etc., assume fixed doubling rate parameters. In reality, however, the doubling time is a dynamic quantity, which changes continuously due to mitigation efforts and the inherently changing nature of virus-spreading mechanisms. It is then vital that policymakers and researchers have access to frequent and up-to-date estimates of doubling time[56]. For example, fixed-in-time estimates of epidemic parameters of COVID-19 (e.g. growth rate, doubling time, basic reproduction number, case detection rate) during the first 50 days of onset in China is provided[57]. In recent work[58] the basic reproduction number and doubling time have been studied in a dynamic manner by considering a varying coefficient model with daily new cases as the response and time as a predictor. A related approach focused on the real-time estimation of case fatality rates using Poisson mixture models can be found in[59].

3 Methods : Parameter and compartment Estimation

The core of our estimation strategy is to utilize equations (1)–(7) to formulate appropriate regression problems. Our estimation procedure is based on the availability of the compartments C_t, D_t, H_t, Q_t, T_t and $R_t^{reported}$ only. We do not assume that data on the social distancing factor κ_t is available. Described crudely, the proposed estimation method uses local regression (linear or nonlinear) methods for estimating the time-varying parameters, while *profiling* over the time-independent ones.

In the absence of data on κ_t, the parameter α in equation (1) is not identifiable. We first describe how the product $\alpha\kappa_t^2$ can be estimated. Notice that, ignoring the stochasticity, we may rewrite equation (5) as

$$\Delta C_t = (\theta(t)+\gamma)A_t. \tag{16}$$

Defining $\eta(t) = \theta(t)+\gamma$, and applying the difference operator on both sides of equation (16), and finally dividing both sides by ΔC_t, we obtain

$$\frac{\Delta^2 C_t}{\Delta C_t} = \left(1+\frac{\Delta\eta(t)}{\eta(t)}\right)\frac{\Delta A_t}{A_t} + \frac{\Delta\eta(t)}{\eta(t)}. \tag{17}$$

Now, ignoring the second order factor $(\Delta\eta(t)\Delta A_t)/(\eta(t)A_t)$, from equation (2), at the onset of the epidemic (i.e. $S_t/(S_t+A_t+R_t) \approx 1$), we have the approximate relationship:

$$\frac{\Delta^2 C_t}{\Delta C_t} \approx \frac{\Delta\eta(t)}{\eta(t)} - \eta(t) - \rho_A + \alpha\kappa_t^2. \tag{18}$$

Note that equation (18) establishes an approximate linear relationship, between the observable quantity $\Delta^2 C_t/\Delta C_t$ and the product $\alpha\kappa_t^2$. Below we show that, the other parameters in equation (18) can be estimated, from the available data. These estimates can be plugged in to get an estimate of $\alpha\kappa_t^2$.

3.1 Point Estimates

Broadly speaking, the estimation strategy consists of separating the time-dependent and time-independent parameters, into vectors $\beta_t = (\phi(t), \rho_H(t), \delta(t))$ and $\zeta = (\gamma, \rho_A)$ respectively. First the vector ζ is kept fixed and for each t the time-dependent parameter β_t is estimated (denoted $\widehat{\beta}_t^h(\zeta)$) by minimizing the "conditional" local loss function $\widetilde{L}_t^h(\beta_t|\zeta)$ (described below) with respect to β_t, subject to appropriate constraints on the parameters (non-negativity as well as certain upper bounds). The optimal local conditional loss is then combined across different time points to obtain the *profile loss* function for ζ, which is given by

$$L^h(\zeta) = \sum_t \widetilde{L}_t^h(\widehat{\beta}_t^h(\zeta)|\zeta). \tag{19}$$

The estimate $\widehat{\zeta}^h$ of ζ is obtained by minimizing $L^h(\zeta)$ under appropriate constraints. We update the estimates of β_t as $\widehat{\beta}_t^h = (\widehat{\phi}(t), \widehat{\rho}_H(t), \widehat{\delta}(t)) = \widehat{\beta}_t^h(\widehat{\zeta}^h)$.

In order to define the conditional loss function, let $K(\cdot)$ be a nonnegative kernel integrating to one. Now, for a bandwidth parameter $h > 0$, the *local weighted conditional loss* function of β_t, given ζ is defined as:

$$\widetilde{L}_t^h(\beta_t|\zeta) = \sum_s \frac{1}{h} K\left(\frac{t-s}{h}\right) d_s(\beta_t|\zeta) \tag{20}$$

where

$$\begin{aligned} d_s(\beta_t|\zeta) \quad &= \quad \left|\sqrt{\Delta H_s + \Delta D_s + \Delta R_s^{reported}} - \sqrt{(\rho_A + \gamma)Q_s + \frac{\gamma \Delta C_s}{\phi(t) F_s + \gamma}}\right|^2 \\ &\quad + \left|\sqrt{\Delta R_s^{reported}} - \sqrt{\rho_A Q_s + \rho_H(t) H_s}\right|^2 + \left|\sqrt{\Delta D_s} - \sqrt{\delta(t) H_s}\right|^2. \end{aligned} \tag{21}$$

Note that the RHS of equation (21) only uses the observed data. The first addendum originates from equations (4), (5) and (6). The second and the third term use equations (6) and (4) respectively. The square-root transformation of the responses are used as a variance stabilising transformation, which is driven by the assumed Poissonian characteristics of the responses. Also by construction, the estimate of $\delta(t)$ does not depend on ζ.

Estimated values of the parameters readily yields estimates of the key compartments of the model. In particular, from the definition of $\theta(t)$, equations (16) and (17) we get:

$$\widehat{\theta}(t) = \widehat{\phi}(t)\frac{\Delta T_t}{H_t}, \quad \hat{A}_t = \frac{\Delta C_t}{\widehat{\theta}_t + \hat{\gamma}}, \quad \widehat{\left(\frac{\Delta A_t}{A_t}\right)} = \frac{\frac{\Delta^2 C_t}{\Delta C_t} - \frac{\Delta \hat{\theta}(t)}{\hat{\theta}(t)+\hat{\gamma}}}{1 + \frac{\Delta \hat{\theta}(t)}{\hat{\theta}(t)+\hat{\gamma}}}, \quad \widehat{\Delta A_t} = \widehat{\left(\frac{\Delta A_t}{A_t}\right)}\hat{A}_t.$$

Now, by plugging in $\hat{\gamma}$, $\widehat{\theta}(t)$, $\hat{A}_t$ and $\hat{\delta}(t)$ in equation (4) we get an updated estimator of $\rho_H(t)$ as

$$\widehat{\rho}_H(t) = \frac{\Delta H_t - \hat{\gamma}(\hat{A}_t + Q_t) + \hat{\delta}(t) H_t}{H_t}.$$

Finally, using equation (17) an estimate of $\alpha \kappa_t^2$ can be obtained as:

$$\widehat{\alpha \kappa_t^2} = \widehat{\left(\frac{\Delta A_t}{A_t}\right)} + \hat{\theta}(t) + \hat{\gamma} + \hat{\rho}_A.$$

The rest of the compartments can be estimated by plugging in the appropriate parameter or compartment estimates in equations (1) - (7) (see the Supplement Sections S1. and S3.).

The tuning parameter h in equation (20) is obtained by minimising a standardised L_1 distance between the fitted and model based estimates of various compartments through a cross-validation strategy. The actual minimisation is achieved by a grid-search. Details can be found in the Supplement Sections S1. and S2.

3.2 Confidence Intervals

We employ residual bootstrap[60–62] to compute the confidence intervals for our parameter and compartment estimates. Briefly put, the technique adds resampled residuals to the fitted values to create several "resampled" datasets. The point estimation technique described above is applied to each of these resampled datasets to create a new set of parameter and compartment estimates. The empirical distribution of these estimates are then used to construct the confidence interval. The details of the algorithm can be found in the Supplement Section S5. The theoretical validity of the residual bootstrap method is well justified in existing literature[63,64].

4 Results : Application to COVID-19 data from the United States (US)

4.1 Data Preparation

We consider the dynamics of the spread of COVID-19 in various states of United States (US) for a tentative time window of late April to mid December, 2020. The proposed model is based on the observed state-wise daily counts of confirmed infections, deaths, hospitalizations and reported recoveries from the hospitals and quarantining facilities. Daily counts of the confirmed COVID-19 cases in various states were obtained from the COVID-19 Data Repository maintained by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. This is publicly available at `https://github.com/CSSEGISandData/COVID-19` and was accessed on December 15, 2020. The state-wise daily counts of positive and negative COVID-19 test results, current hospitalization, and recovery per day and state, were obtained from the CDC data repository - the COVID Tracking Project and are publicly available at `https://COVIDtracking.com/` (accessed on December 15, 2020.)

The collected noisy data used is pre-processed and cleaned, removing the irregularities present in the recording and maintenance of the data repositories. Any missing or evidently wrong (e.g. negative counts) observations were replaced by the average of the data from adjacent five days. Inherent noise present in the daily counts was removed by pre-smoothing the trajectories using a *Lowess* method[65–67] with bandwidth $1/16$.

4.2 Results

Unfortunately, a continuous record on hospitalization and recovery information were not available for many states. For example, most counties in California are not reporting recovery information. Data on hospitalization is found to be updated once a week in Massachusetts and Florida. New York, on the other hand, started documenting the hospitalization information only after the initial surge of the pandemic was over for the state. In our analysis we only consider the states for which daily observations on C_t, D_t, $R_t^{reported}$, Q_t, and H_t are available throughout the time window under consideration. Any missing/negative values are replaced by the average of the adjacent five days' data. For a few states e.g. Alabama, the available data turned out to be too unreliable. We present results for fifteen states in US which demonstrate the efficacy of the proposed model and the estimation methods. For succinct representation, the results from only one state i.e. Utah are presented in details below. The results for other fourteen states can be found in the Supplement Section S8.

4.2.1 Case study for the state Utah

We present our results for the state *Utah* for the time window between 7th May, 2020 to 4th December, 2020. The time interval includes the Thanksgiving weekend (27th -28th November, 2020), when due to the long holiday, the reported data may be unreliable. In Figure 2 plots of various time-varying compartments and epidemiological markers defined in Section 2.2. The plots of the parameters with their residual bootstrap confidence intervals can be found in Figure 3. Due to unreliable reporting around the Thanksgiving holiday, the estimates values after 21st November, 2020 should be interpreted with caution.

The curves in Figure 2a compare the observed and the fitted number of daily number of people in the hospitals. It can be seen that, the fitted values obtained from the model closely follow the observed values. This validates our proposed model and the estimation procedure. From the data and the fit two waves of infection can be identified. It seems the first wave starts at the end of May, 2020 stabilises and begins to die down around 7th August, 2020. The daily number of people in hospitals starts increasing again around the end of August, 2020.

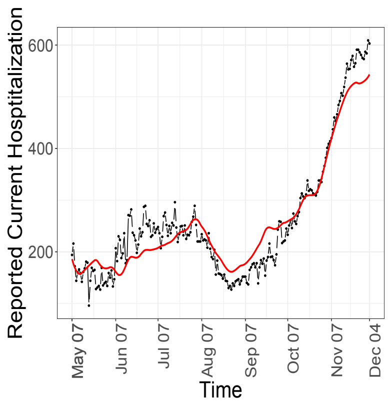
4.2.2 Estimation of latent compartment

The estimated number of infected asymptomatic people (Figure 2d) shows a similar pattern. From a high point around the beginning of August it dips to a low value at the end of August. The number remains stable for a few weeks and starts growing again at the end of September. Estimation of such latent trajectory is a key feature of our proposed methodology which cannot naturally be obtained from the conventional epi-models. The projections from IHME[29] which employ a more complex but less robust parametric estimation method based on a SEIR model provide an estimate of a "pre-symptomatic" population. Members of this compartment can be considered asymptomatic. We use the term in a more general sense.

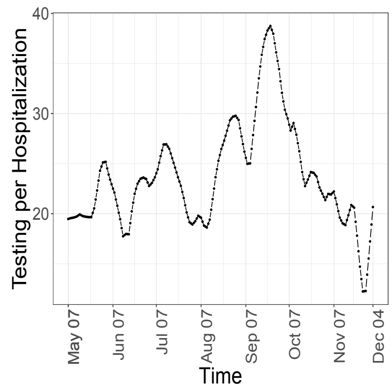
4.2.3 Analogue of Basic Reproduction Rate

The phenomenon of two waves is clearly observed from the plot of the proposed analogue of the basic reproduction rate $\tilde{R}_0$ (the solid red curve in Figure 3a) - the estimated $\tilde{R}_0$ was larger than 1 in two sub-intervals, namely from middle May to middle of July and then from end of August to beginning of November. Our estimate is compared with three other relevant sources viz., the generative COVID-model considered by the Systrom et. al[68] (blue, "longdash" line in Figure 3a), the SEIR model based COVID-19 projections using machine learning from Youyang Gu[69] (in green), and SEIR model used by the IHME team in[29] (in magenta).

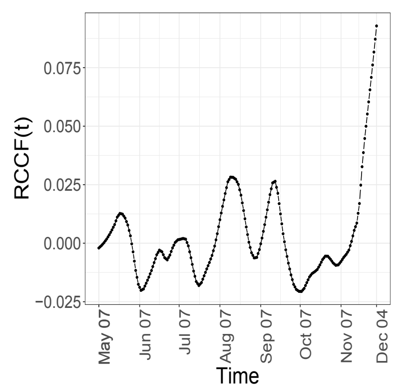
Our analogue of this epidemiological marker seems more realistic, since it tallies with the other observed and estimated compartments. For example, around August 7, 2020, the cumulative new infections (CNI), both observed and estimated,



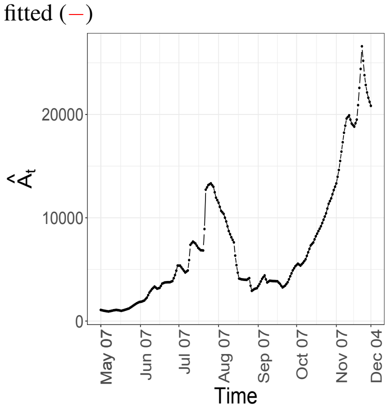
(a) Daily hospitalizations, observed (−), fitted (−)



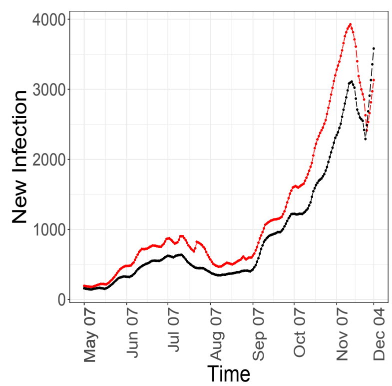
(b) $\Delta T_t/H_t$



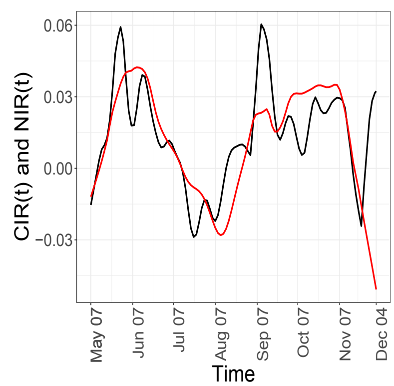
(c) RCCF



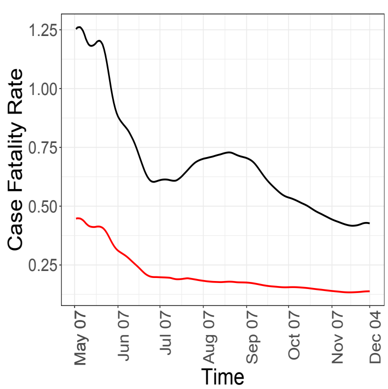
(d) $\hat{A}_t$



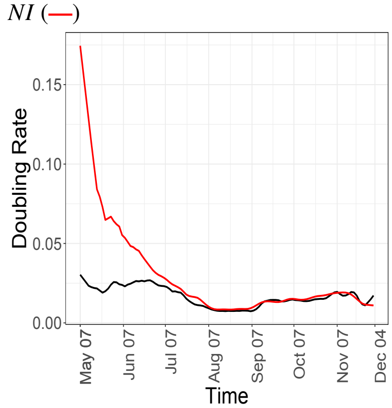
(e) Daily New Infections, ΔC_t (—) and NI (—)



(f) CIR (—) and NIR (—)

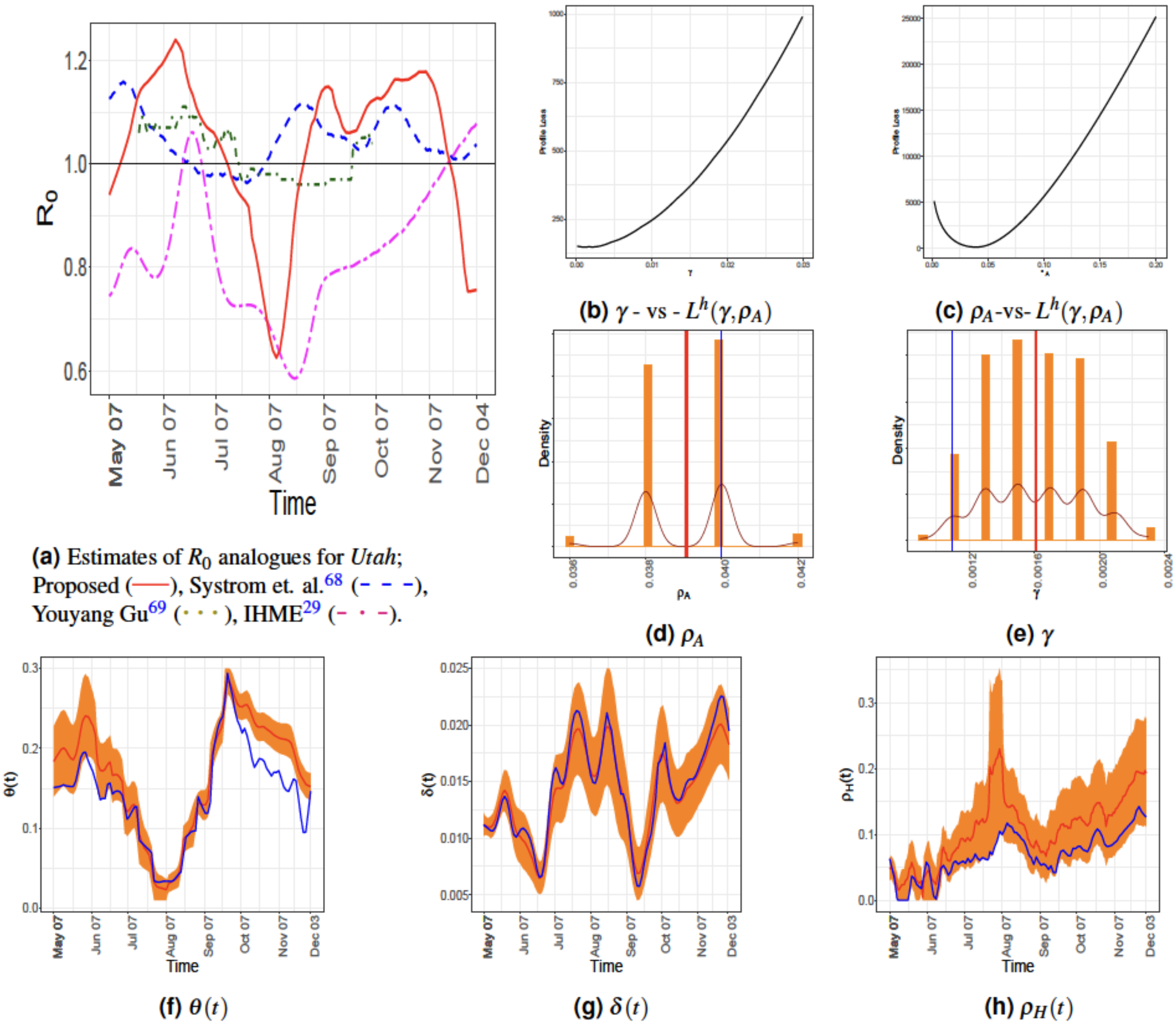


(g) CFR (—) and NFR (—)



(h) Doubling rate, C_t (—) and CNI (—)

Figure 2. Temporal patterns of some compartments and epidemiological markers for Utah.



(a) Estimates of R_0 analogues for *Utah*; Proposed (——), Systrom et. al.[68] (– – –), Youyang Gu[69] (· · ·), IHME[29] (– · –).

(b) γ - vs - $L^h(\gamma, \rho_A)$

(c) ρ_A-vs- $L^h(\gamma, \rho_A)$

(d) ρ_A

(e) γ

(f) $\theta(t)$

(g) $\delta(t)$

(h) $\rho_H(t)$

Figure 3. 3a Estimates of R_0 analogues. 3b to 3h estimates and residual bootstrap based confidence intervals for time invariant and time-varying parameters for the state of Utah. The estimate from the data is in blue. The 95% confidence band is in yellow and the mean of the bootstrap estimates are presented in red.

hospitalization and the asymptomatic population (estimated) were quite low and almost constant over a period of time (see Supplement Section *S*7). The estimated social mobility index $\alpha\kappa_t^2$ also experienced a sharp decline around that time (see Figure 10 in Supplement Section *S*7), which all give evidence to the fact that the spread of the pandemic was indeed contained around mid-July to mid- August in Utah. This is clearly resonated in our version of reproduction rate, but is not so well captured by the two other models considered above. The estimate released by IHME[29] seems to follow our estimate in August, however, it hardly gets higher than 1, not even in October, when the number of new infections were high. From this it seems that the IHME estimate does not qualitatively reflect the real nature of COVID spread.

The plot of the number of daily new and daily reported infections (Figure 2e) shows a local maximum near the middle of November. However, we cannot rule out the boundary effect as its cause.

4.2.4 Model Parameter Estimates

The estimate of $\delta(t)$ in Figure 3g seems to remain stable throughout the time period under consideration. The $\widehat{\rho_H}(t)$ shows an overall increasing trend. On the other hand, the estimate of $\theta(t)$ decreases to a near-zero value at the end of the first wave (7th August, 2020) it then increases to its maximum value at the end of September and starts to decrease again. The parameters (γ, ρ_A) are estimated based on minimization of the profile loss using a grid search algorithm with grid size 0.0001. In Figures 3d and 3e the estimates from residual bootstrap samples take discrete values, resulting in a discrete histogram counts. In Table 1

	Estimate	95% Confidence Interval	Mean	s.d.
γ	0.0011	[0.0011, 0.0021]	0.0016	0.0003
ρ_A	0.0400	[0.0360, 0.0420]	0.0391	0.0012

Table 1. Estimates, and the residual bootstrap Confidence intervals, mean and standard deviations for the time-invariant parameters for Utah. The latter three are computed based on 1000 bootstrap resamples.

we present the estimates, 95% residual bootstrap confidence intervals, the residual bootstrap mean and standard deviations of the above parameters.

4.2.5 Transmission rates

The plots of CIR and NIR seem to be similar (Figure 2f). In fact, the observed doubling rate obtained from C_t and that estimated from *CNI* seems to be very close in the second wave of the pandemic (see Figure 2h). This implies that in the second wave the reporting kept pace with the spread of the disease. Figure 2g shows the crude and net fatality rates. Due to the denominator effect, naturally the crude fatality rate is much larger than the net fatality rate. However, our estimate of NFR is mostly below 0.25%, which complies with widely held beliefs[29,70–72].

	CNI/Population	Recovery/Population	Average Seroprevalence
Period 1 (July 27 - August 13)	1.78 1.144, 3.597)	1.62 (1.595, 2.152)	3.2 (1.20, 5.03)
Period 2 (August 10 - August 27)	2.00 (1.258, 4.090)	1.93 (1.922, 2.493)	5.5 (2.94, 8.71)
Period 3 (August 24 - September 10)	2.27 (1.420, 4.500)	2.17 (2.163, 2.733)	4.9 (2.82, 7.67)
Period 4 (September 7 - September 24)	2.41 (1.748, 5.105)	2.44 (2.440, 3.020)	5.1 (3.29, 7.90)

Table 2. Table comparing the seroprevalence estimates for the state *Utah*.

4.2.6 Seroprevalence:

Seroprevalence studies to estimate the prevalence of persons with SARS-CoV-2 antibodies have been of immense interest. Seroprevalence is calculated as the number of reactive specimens divided by the number of specimens tested[73]. Even though our model cannot explicitly compute it, analogues of such estimates can be found from the ratios such as percentage of cumulative new infections in the population and the percentage of total recovery (from quarantine, hospitalization, or asymptomatic states). The estimates of such seroprevalence analogue for the state *Utah* are illustrated in Table 2 and the 95% residual bootstrap confidence intervals mostly overlap with the 95% confidence intervals provided in[73] for all four periods of time considered.

4.2.7 Testing and hospitalization

The daily number of tests and its effect in quarantining asymptomatic but infected people can be judged from the Figures 2b and 2c. The state of Utah increased its testing capacity by public-private partnership. Empirical comparison of the Figures 2a and 2b seems to reveal that although the number of daily tests could keep pace with daily number of hospitalized patients up to the third week of September, growing number of hospitalized people ultimately outpaced the number of daily tests. Note that estimated $\theta(t)$ increases at the onset of the second wave (see Figure 3f between 7th, August and 21st, September), however, from Figure 2d, $\hat{A}_t$ remains more or less constant. Thus, growth in the number of new infections could be due to the increase in κ_t, that is due to more interaction among individuals and less social distancing.

4.2.8 Impact of testing in the disease control

From a public policy perspective, our model-based simulation provides strong quantitative evidence on the significant role of testing rate in controlling the spread of the pandemic. This could be the key to mitigate the explosive nature of the epidemic even before any intervention strategies are put into practice. Numerical simulation of the pandemic based on the estimates obtained from our model shows explicitly that, with all the time-invariant and time- varying rate parameters remaining the same, a higher testing rate leads to a suppression and eventual decline in the number of infected individuals as well as hospitalizations and deaths (see Supplement Section *S*7). For example, Figure 4 shows that the curves are clearly flattened when the confirmed

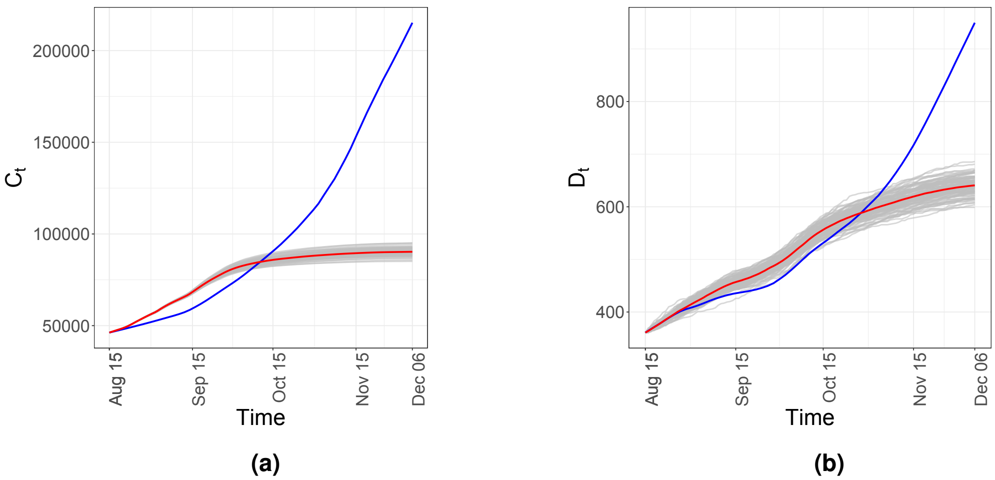


Figure 4. The 100 simulated trajectories when testing rate is increased by 30% are in gray while the mean of them are presented in red. The observed C_t and D_t in Utah are marked in blue.

fraction, $\theta(t)$ is increased by 30%. Non-increasing patterns shown in the cumulative compartments, C_t and D_t indicate a containment of the disease.

4.2.9 Summary of Results for Other States

We present a summary of the results obtained from applications of proposed method on the data procured from fifteen other states in United States (US). The estimated parameters are in Table 3. The time-varying parameters, ($\phi(t)$, $\rho_H(t)$, $\delta(t)$), are summarised by their means. The computed $\hat{\gamma}$, that is, the rate for an asymptomatic person turning symptomatic on a particular day is the smallest in Arizona and largest in Tennessee. This estimate is smaller than 0.001 for Arizona and Idaho. Minnesota, has by far the highest recovery rate for an asymptomatic person without needing hospitalization on a particular day (i.e. $\widehat{\rho}_A$). For Iowa, Nebraska, Pennsylvania, and Utah this rate is comparable and reasonably high, whereas Arizona, Delaware and Idaho have their $\widehat{\rho}_A$ value below 0.01. The average confirmed fraction $\overline{\hat{\theta}}$ is larger than 0.1 in Delaware, Tennessee and Utah. It is lowest in Texas. This can be associated with better estimates obtained for these states due to the availability of more reliable data, whereas for Idaho, South Dakota, and Texas, a lower value of there epi-markers tend to give evidence for a more relaxed testing paradigm. More testing is required for isolating the confirmed cases to contain the disease faster, which can be reflected in the numbers for these states. The detailed results and bootstrap confidence regions for these additional states can be found in the Supplement Section S8.

Among the states not included in Table 3, many, such as California did not report all the required compartments. For many states such Alabama, Colorado, Maryland, Massachusetts, North Carolina etc. the reported data produced monotone profile likelihoods which yielded unreliable boundary estimates. This could be due to the change in definition of many compartments over time, which violated our assumptions. Furthermore, for some states such as New York, New Jersey, Michigan etc., the pandemic started quite early and ran its course even before a proper testing protocol and other mitigation measures could be introduced. Thus the data from these states is contaminated with an inherent bias, the number of people in quarantine or symptomatic states are to too low to produce reliable estimates.

5 Discussion

We introduce a multi-compartment model for COVID-19 dynamics which can incorporate data from compartments like quarantine, hospitalization etc. Unlike the conventional SIR and similar models, the proposed model is based on interpretable time-varying parameters, which are more suitable for describing the disease dynamics in the presence of mitigating procedures. It also incorporates the information about testing and subsequent quarantining. We estimate the model parameters using profile likelihood and nonparametric regression. This provides a much faster alternative to Markov Chain Monte Carlo based Bayesian models which are commonly used in estimating SIR parameters. Using the proposed detailed and robust model one can estimate the daily number of asymptomatic but infected individuals, who are universally regarded as the key agent for the COVID-19 spread. To the best of our knowledge, no other model gives both such epi-estimates, which are important from a health policy perspective, as well as the projections for the unobservable latent quantities such as the trajectories of susceptible, asymptomatic, and recovered (from quarantine, hospitalization, or asymptomatic states) population, which are

	$\hat{\gamma}$	$\hat{\rho}_A$	$\widehat{\delta(t)}$	$\widehat{\rho_H(t)}$	$\widehat{\theta(t)}$	$\widehat{\phi(t)}$
Arizona	0.0003	0.002	0.0208	0.0023	0.0887	0.0079
Arkansas	0.0029	0.094	0.0249	0.0975	0.0809	0.0038
Delaware	0.0017	0.008	0.0159	0.0093	0.1076	0.0037
Idaho	0.0009	0.010	0.0230	0.0138	0.0289	0.0019
Iowa	0.0011	0.032	0.0263	0.0372	0.0478	0.0033
Minnesota	0.0023	0.128	0.0315	0.0654	0.0899	0.0034
Nebraska	0.0011	0.020	0.0141	0.0266	0.0394	0.0035
Ohio	0.0023	0.048	0.0180	0.0532	0.0625	0.0024
Oklahoma	0.0037	0.084	0.0122	0.1029	0.0494	0.0033
Pennsylvania	0.0013	0.026	0.0293	0.0372	0.0535	0.0033
South Dakota	0.0021	0.058	0.0190	0.0922	0.0262	0.0038
Tennessee	0.0059	0.064	0.0158	0.0413	0.1206	0.0076
Texas	0.0019	0.036	0.0207	0.0341	0.0212	0.0013
Utah	0.0011	0.040	0.0144	0.0252	0.1434	0.0061
Wisconsin	0.0017	0.068	0.0217	0.0707	0.0477	0.0026

Table 3. Mean estimated parameters for different states in the United States (US).

essential for understanding the dynamics of the pandemic. We define several epidemiological markers that uses the number of asymptomatic-infected individuals and therefore reveal the true underlying dynamics of the pandemic.

Our model only uses information on the number of confirmed infected, hospitalized, deaths and total reported recoveries from hospitals and quarantine. We don't require those numbers separately. However, such numbers are often available. In such a case, the loss function in equation (21) can be simplified a bit. The details can be found in the Supplement in Section S2.

The model parameters have been estimated assuming that no information about the mobility within the population is available. Such information identifies the parameters κ_t and α in our model. Reliable data on the compliance to social distancing, mask wearing etc. are difficult to get. Various aspects of the mobility data available from Google can be one potential surrogate for κ_t[74,75]. However, such data only look into the fraction of people going to workplace or recreation and so on, and does not collect information on the people who are the super spreaders or not wearing masks. Thus, it does not necessarily reflect the the social mobility index κ_t, as incorporated in our model. In the Supplement (see Sections S3. and S8.), we present results by using the Google mobility data as a surrogate to κ_t. In particular, information on the change in the mobility patterns, as the percentage decrease (increase) from the baseline, in different areas such as parks, residential locations, retail locations, among others during the pandemic from the Google mobility database were obtained. The publicly available data was sourced from `https://www.google.com/COVID19/mobility/` (accessed on December 15, 2020). When information on κ_t is available, the parameter α, which is the average number of susceptible individuals who may be infected in a day by an asymptomatic-infected individual is identifiable and can be estimated. The details can be found in the Supplement Section S3.

The proposed method and estimation procedure do not explicitly use the underlying assumption of a Poisson process. In the Supplement (see Section S6.− 7.), however, we use an ensemble of independent Poisson processes to simulate data from the proposed model. These aggregated data sets are then used to accurately estimate various parameters, which validate our estimation procedure. The aggregation has the effect of increasing the number of observations in the compartments and thereby improving estimation accuracy. If the number of individuals in the symptomatic or quarantined compartments are low, e.g. at the onset of the pandemic, inherent biases are introduced in the estimated trajectories. A bigger sample size is required to correct such contaminants.

In our model the compartment A_t includes the asymptomatic individuals, as well as those infected before they are quarantined, tested positive, or hospitalized. We further assume that anybody, whether quarantined or not, is immediately hospitalized, and is tested positive, upon onset of significant symptoms. In reality however, some symptomatic people might not get tested and remain in the community as spreaders. Furthermore, the rate at which a truly asymptomatic person infects a susceptible may differ from the same rate for an non-tested mildly symptomatic person. In practice, little data are available on mildly symptomatic people. Under the ideal situation we consider here, such differences should be negligible.

Because of the limited availability and relatively poor quality of detailed data, we allow no strata with respect to age or intrinsic vulnerability to the disease in our homogeneous population. Moreover, due to the presence of unobservable compartments like A_t, even when the data quality is good, there is a near lack of identifiability of the parameters if all of them are assumed to be time-dependent. In any case, for most practical situations, it is reasonable to assume a constant rate of recovery ρ_A and a constant rate γ of getting severely ill from the asymptomatic compartment. We consider dynamic models of pandemic propagation in a stratified population in a subsequent article.

Since the proposed method is non-parametric, they suffer from possible boundary effects near the end-points of the time window. It should also be noted that COVID-19 analyses based on published case and death counts, including those conducted here, are subject to the same biases which affect the accuracy of the data, primarily due to under-reporting[76] or mis-recording of the data, the degree of which varies by country[77]. The reasons for such under-reporting are many, including insufficient testing materials, political incentives, and administrative delays. If such irregularities are present even after pre-processing steps, the underlying model in (1)– (7) may not be adequate. In such cases the profile loss functions of γ and ρ_A in (19) may attain their minimums at the boundaries. This may influence other parameter estimates and their interpretations. Furthermore, our model assumes a closed population. It ignores migration between cities, states or countries which play an essential role in the propagation of the disease. We only count the deaths solely due to COVID-19 infections and as such completely ignore any competing causes of morbidity, as well as increase in population due to new births.

With this caveat in mind, the study of available data presented in this article nevertheless provides useful insights into the COVID-19 propagation and ways to control it. It clearly follows that in order to break the chain of transmission and "flatten the curve", we need extensive testing and adhere to strict social distancing protocols.

6 Data and Code Availability

All data necessary for the replication of our results is collated in
`https://github.com/Satarupa3671/COVID-19-Nonparametric-Inference`.
The data for the number of COVID cases, deaths, hospitalizations and recovery were originally collected from `https://covidtracking.com/data/download` while the social mobility data was sourced from `https://www.google.com/covid19/mobility`. All code necessary for the replication of our results is collated in `https://github.com/Satarupa3671/COVID-19-Nonparametric-Inference`.

References

1. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* **20**, 533–534, DOI: 10.1016/S1473-3099(20)30120-1 (2020).
2. Hoffmann, M. *et al.* SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271–280.e8, DOI: 10.1016/j.cell.2020.02.052 (2020).
3. Pollock, A. M. & Lancaster, J. Asymptomatic transmission of COVID-19. DOI: 10.1136/bmj.m4851 (2020).
4. Day, M. Covid-19: Identifying and isolating asymptomatic people helped eliminate virus in italian village. *British Medical Journal* **368**, DOI: 10.1136/bmj.m1165 (2020).
5. Rossman, H. *et al.* A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys. *Nature Medicine* **26**, 634–638, DOI: 10.1038/s41591-020-0857-9 (2020).
6. Mizrahi, B. *et al.* Longitudinal symptom dynamics of COVID-19 infection. *Nature Communications* **11**, 1–10, DOI: 10.1038/s41467-020-20053-y (2020).
7. Unwin, H. J. T. *et al.* State-level tracking of COVID-19 in the united states. *Nature Communications* **11**, 1–9, DOI: 10.1038/s41467-020-19652-6 (2020).
8. Godio, A., Pace, F. & Vergnano, A. SEIR modeling of the Italian epidemic of sars-cov-2 using computational swarm intelligence. *International Journal Of Environmental Research And Public Health* **17**, 3535, DOI: 10.3390/ijerph17103535 (2020).
9. Byambasuren, O. *et al.* Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. *Official Journal Of The Association Of Medical Microbiology And Infectious Disease Canada* **5**, 223–234, DOI: 10.3138/jammi-2020-0030 (2020).
10. Wang, L. *et al.* An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China. *Journal of Data Science* **18**, 409–432, DOI: 10.6339/JDS.202007_18(3).0003 (2021).
11. Tang, Y. & Wang, S. Mathematic modeling of COVID-19 in the United States. *Emerg. Microbes & Infect.* **9**, 827–829, DOI: 10.1080/22221751.2020.1760146 (2020).
12. Wang, J. Mathematical models for COVID-19: Applications, limitations, and potentials. *Journal of Public Health and Emergency* **4**, DOI: 10.21037/jphe-2020-05 (2020).

13. Mohamadou, Y., Halidou, A. & Kapen, P. T. A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19. *Applied Intelligence* **50**, 3913–3925, DOI: 10.1007/s10489-020-01770-9 (2020).

14. Kaniadakis, G. *et al.* The κ-statistics approach to epidemiology. *Scientific Reports* **10**, 1–14, DOI: 10.1038/s41598-020-76673-3 (2020).

15. Ndairou, F., Area, I., Nieto, J. J. & Torres, D. F. Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan. *Chaos, Solitons & Fractals* 109846, DOI: 10.1016/j.chaos.2020.109846 (2020).

16. Dong, Y. *et al.* Epidemiology of COVID-19 among children in China. *Pediatrics* **145**, DOI: 10.1542/peds.2020-0702 (2020).

17. McIntosh, K., Hirsch, M. & Bloom, A. Coronavirus disease 2019 (COVID-19): Epidemiology, virology, and prevention. *The Lancet Infectious Diseases* **1**, 2019–2020 (2020).

18. Brett, T. S. & Rohani, P. Transmission dynamics reveal the impracticality of COVID-19 herd immunity strategies. *Proceedings of the National Academy of Sciences* **117**, 25897–25903, DOI: 10.1073/pnas.2008087117 (2020).

19. Bertozzi, A. L., Franco, E., Mohler, G., Short, M. B. & Sledge, D. The challenges of modeling and forecasting the spread of COVID-19. *Proceedings of the National Academy of Sciences* **117**, 16732–16738, DOI: 10.1073/pnas.2006520117 (2020).

20. Hao, X. *et al.* Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature* **584**, 420–424, DOI: 10.1038/s41586-020-2554-8 (2020).

21. Cuadros, D. F., Branscum, A. J., Mukandavire, Z., Miller, F. D. & MacKinnon, N. Dynamics of the COVID-19 epidemic in urban and rural areas in the united states. *Annals of Epidemiology* **59**, 16–20, DOI: 10.1016/j.annepidem.2021.04.007 (2021).

22. James, N. & Menzies, M. COVID-19 in the United States: Trajectories and second surge behavior. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **30**, 091102, DOI: 10.1063/5.0024204 (2020).

23. Anastassopoulou, C., Russo, L., Tsakris, A. & Siettos, C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PloS One* **15**, e0230405, DOI: 10.1371/journal.pone.0230405 (2020).

24. Peng, L., Yang, W., Zhang, D., Zhuge, C. & Hong, L. Epidemic analysis of COVID-19 in China by dynamical modeling. *AXiv Preprint ArXiv:2002.06563* DOI: 10.1101/2020.02.16.20023465 (2020).

25. Ivorra, B., Ferrández, M. R., Vela-Pérez, M. & Ramos, A. Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. *Communications In Nonlinear Science And Numerical Simulation* **88**, 105303, DOI: 10.1016/j.cnsns.2020.105303 (2020).

26. Sameni, R. Mathematical modeling of epidemic diseases; a case study of the COVID-19 coronavirus. *ArXiv Preprint ArXiv:2003.11371* (2020).

27. Park, M., Cook, A. R., Lim, J. T., Sun, Y. & Dickens, B. L. A systematic review of COVID-19 epidemiology based on current evidence. *Journal Of Clinical Medicine* **9**, 967, DOI: 10.3390/jcm9040967 (2020).

28. Riccardo, F. *et al.* Epidemiological characteristics of COVID-19 cases and estimates of the reproductive numbers 1 month into the epidemic, Italy, 28 january to 31 march 2020. *Eurosurveillance* **25**, 2000790, DOI: 10.2807/1560-7917.ES.2020.25.49.2000790 (2020).

29. IHME COVID-19 forecasting team. Modeling COVID-19 scenarios for the united states. *Nature Medicine* DOI: 10.1038/s41591-020-1132-9 (2020).

30. Eker, S. Validity and usefulness of COVID-19 models. *Humanities And Social Sciences Communications* **7**, 1–5, DOI: 10.1057/s41599-020-00553-4 (2020).

31. Friedman, J. *et al.* Predictive performance of international COVID-19 mortality forecasting models. *Nature communications* **12**, 1–13, DOI: 10.1038/s41467-021-22457-w (2021).

32. Bhatnagar, M. R. COVID-19: Mathematical modeling and predictions. DOI: 10.13140/RG.2.2.29541.96488 (2020).

33. Petropoulos, F. & Makridakis, S. Forecasting the novel coronavirus COVID-19. *PloS One* **15**, e0231236, DOI: 10.1371/journal.pone.0231236 (2020).

34. Badr, H. S. *et al.* Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet Infectious Diseases* **20**, 1247–1254, DOI: 10.1016/S1473-3099(20)30553-3 (2020).

35. Giordano, G. *et al.* Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine* 1–6, DOI: 10.1038/s41591-020-0883-7 (2020).

36. Gatto, M. *et al.* Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences* **117**, 10484–10491, DOI: 10.1073/pnas.2004978117 (2020).

37. Khosrawipour, V. *et al.* Failure in initial stage containment of global COVID-19 epicenters. *Journal of Medical Virology* DOI: 10.1002/jmv.25883 (2020).

38. Wagner, A. B. *et al.* Social distancing merely stabilized COVID-19 in the United States. *Stat* **9**, e302, DOI: 10.1002/sta4.302 (2020). E302 sta4.302, https://onlinelibrary.wiley.com/doi/pdf/10.1002/sta4.302.

39. Lau, H. *et al.* The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China. *Journal of Travel Medicine* DOI: 10.1093/jtm/taaa037 (2020).

40. Linka, K., Peirlinck, M., Sahli Costabal, F. & Kuhl, E. Outbreak dynamics of COVID-19 in Europe and the effect of travel restrictions. *Computer Methods In Biomechanics And Biomedical Engineering* 1–8, DOI: 10.1080/10255842.2020.1759560 (2020).

41. Tian, H. *et al.* An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* **368**, 638–642, DOI: 10.1126/science.abb6105 (2020).

42. Li, M.-T. *et al.* Analysis of COVID-19 transmission in Shanxi Province with discrete time imported cases. *Mathematical Biosciences And Engineering* **17**, 3710, DOI: 10.3934/mbe.2020208 (2020).

43. Centers for Disease Control and Prevention. Interim guidance for antigen testing for sars-cov-2. https://www.cdc.gov/coronavirus/2019-ncov/lab/resources/antigen-tests-guidelines.html (2020). [Online; accessed 20-December-2020].

44. Watson, J., Whiting, P. F. & Brush, J. E. Interpreting a COVID-19 test result. *British Medical Association* **369**, DOI: 10.1136/bmj.m1808 (2020).

45. Healy, B., Khan, A., Metezai, H., Blyth, I. & Asad, H. The impact of false positive COVID-19 results in an area of low prevalence. *Clinical Medicine* **21**, e54, DOI: 10.7861/clinmed.2020-0839 (2021).

46. Sethuraman, N., Jeremiah, S. S. & Ryo, A. Interpreting diagnostic tests for SARS-CoV-2. *Journal Of The American Medical Association* **323**, DOI: 10.1001/jama.2020.8259 (2020).

47. Kahn, J. P. *Digital Contact Tracing For Pandemic Response: Ethics And Governance Guidance* (Johns Hopkins University Press, 2020).

48. McClain, C. & Rainie, L. The challenges of contact tracing as U.S. battles COVID-19. *Pew Research Center* (2020).

49. Google LLC. Google COVID-19 community mobility reports. https://www.google.com/covid19/mobility/. Last accessed May 18, 2020.

50. Badr, H. S. *et al.* Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet Infectious Diseases* **20**, 1247–1254, DOI: 10.1016/S1473-3099(20)30553-3 (2020).

51. Buckee, C. O. *et al.* Aggregated mobility data could help fight COVID-19. *Science (New York, NY)* **368**, 145–146, DOI: 10.1126/science.abb8021 (2020).

52. Kraemer, M. U. *et al.* The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497, DOI: 10.1126/science.abb4218 (2020).

53. You, C. *et al.* Estimation of the time-varying reproduction number of COVID-19 outbreak in China. *International Journal Of Hygiene And Environmental Health* 113555, DOI: 10.1016/j.ijheh.2020.113555 (2020).

54. Heffernan, J. M., Smith, R. J. & Wahl, L. M. Perspectives on the basic reproductive ratio. *Journal Of The Royal Society Interface* **2**, 281–293, DOI: 10.1098/rsif.2005.0042 (2005).

55. Yang, H. M. The basic reproduction number obtained from jacobian and next generation matrices–a case study of dengue transmission modelling. *Biosystems* **126**, 52–75, DOI: 10.1016/j.biosystems.2014.10.002 (2014).

56. Pellis, L. *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* **376**, 20200264, DOI: 10.1098/rstb.2020.0264 (2021).

57. Du, Z. *et al.* Risk for transportation of coronavirus disease from Wuhan to other cities in China. *Emerging Infectious Diseases* **26**, 1049, DOI: 10.3201/eid2605.200146 (2020).

58. Kucharski, A. J. *et al.* Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases* DOI: 10.1016/S1473-3099(20)30144-4 (2020).

59. Lee, P. H. Estimating the real-time case fatality rate of COVID-19 using poisson mixtures model. *MedRxiv* DOI: 10.1101/2020.04.11.20062190 (2020).

60. Hinkley, D. V. Bootstrap methods. *Journal of The Royal Statistical Society: Series B (Methodological)* **50**, 321–337, DOI: 10.1111/j.2517-6161.1988.tb01731.x (1988).

61. Carpenter, J. & Bithell, J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics In Medicine* **19**, 1141–1164, DOI: 10.1002/(sici)1097-0258(20000515)19:9<1141::aid-sim479>3.0.co;2-f (2000).

62. Zoubir, A. M. & Iskandler, D. R. Bootstrap methods and applications. *IEEE Signal Processing Magazine* **24**, 10–19, DOI: 10.1109/MSP.2007.4286560 (2007).

63. Davison, A. C. & Hinkley, D. V. *Bootstrap Methods And Their Application*. 1 (Cambridge University Press, 1997).

64. Chernick, M. R., González-Manteiga, W., Crujeiras, R. M. & Barrios, E. B. Bootstrap Methods, DOI: 10.1007/978-3-642-04898-2_150 (2011).

65. Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *Journal Of The American Statistical Association* **74**, 829–836, DOI: 10.1080/01621459.1979.10481038 (1979).

66. Cleveland, W. S. Lowess: A program for smoothing scatterplots by robust locally weighted regression. *American Statistician* **35**, 54, DOI: 10.2307/2683591 (1981).

67. Royston, P. Lowess smoothing. *Stata Technical Bulletin* **1** (1992).

68. Systrom, K., Vladek, T. & Krieger, M. Rt.live. https://github.com/rtcovidlive/covid-model (2020).

69. Gu, Y. COVID-19 Projections Using Machine Learning. https://github.com/youyanggu/covid19_projections/ (2020). [Accessed on 16-February-2021].

70. Woody, S. *et al.* Projections for first-wave COVID-19 deaths across the US using social-distancing measures derived from mobile phones. *MedRxiv* DOI: 10.1101/2020.04.16.20068163 (2020).

71. Chirico, F., Nucera, G. & Magnavita, N. Estimating case fatality ratio during COVID-19 epidemics: pitfalls and alternatives. *The Journal Of Infection In Developing Countries* **14**, 438–439, DOI: 10.3855/jidc.12787 (2020).

72. Ioannidis, J. P. Infection fatality rate of COVID-19 inferred from seroprevalence data. *Bulletin of the World Health Organization* **99**, 19, DOI: 10.2471/BLT.20.265892 (2021).

73. Bajema, K. L. *et al.* Estimated sars-cov-2 seroprevalence in the US as of September 2020. *JAMA Internal Medicine* DOI: 10.1001/jamainternmed.2020.7976 (2020).

74. Yilmazkuday, H. Stay-at-home works to fight against COVID-19: international evidence from Google mobility data. *Journal Of Human Behavior In The Social Environment* 1–11, DOI: 10.1080/10911359.2020.1845903 (2020).

75. Wang, H. & Yamamoto, N. Using a partial differential equation with Google Mobility data to predict COVID-19 in Arizona. *Mathematical Biosciences and Engineering* **17**, DOI: 10.3934/mbe.2020266 (2020).

76. Jagodnik, K. M., Ray, F., Giorgi, F. M. & Lachmann, A. Correcting under-reported covid-19 case numbers: estimating the true scale of the pandemic. *medRxiv* DOI: 10.1101/2020.03.14.20036178 (2020).

77. Krantz, S. G. & Rao, A. S. R. S. Level of under-reporting including under-diagnosis before the first peak of COVID-19 in various countries: Preliminary Retrospective Results Based on Wavelets and Deterministic Modeling. *Infection Control @AND@ Hospital Epidemiology* 1–8, DOI: 10.1017/ice.2020.116 (2020).

7 Acknowledgements

Debashis Paul : Research supported by NSF grants DMS-1713120, DMS-1811405, and DMS-1915894, and NASA-TRISH grant 19-19BRASH-2-0055. Sanjay Chaudhuri: Research supported by MOE AcRF grants R-155-000-194-114 and R-155-000-215-114.

8 Author contributions statement

S.B: original draft, data analysis, acquisition and interpretation, project administration. S.L: data analysis and interpretation. D.P: project conception, project administration and supervision, editing. S.C: project conception, administration, and supervision, editing.

9 Additional information

The authors declare no competing interests.

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.