Epidemic Modeling and Estimation

Alberto Abadie[1, 2, 4] Paolo Bertolotti[2, 4] Ben Deaner[1] Arnab Sarker[2, 4] Devavrat Shah[2, 3, 4]

[1]Department of Economics, MIT
[2]Institute for Data, Systems, and Society, MIT
[3]Department of EECS, MIT
[4]IDSS COVID-19 Collaboration (ISOLAT) Project, IDSS, MIT

The purpose of this memo is to summarize various classical and emerging approaches for epidemic modeling. The goal here is to describe the models and the methods for learning such models in a data-driven manner as well as utilizing them for various predictive tasks.

1. Background

Overview. Epidemics, such as COVID-19, spread through human interactions. Therefore, at some level, the precise nature and details of human interactions determine how epidemics grow and infection spreads. Epidemiologists have studied this phenomenon and developed remarkably simple, parsimonious models. In addition, at the time of writing this document, the ongoing pandemic of COVID-19 has resulted in various recent proposals to better estimate parameters for existing models, as well as proposals of novel models. We attempt to document such approaches from the biased view of the authors.

At their core, epidemiological methods attempt to model the growth of infections and the duration of the epidemic using data. The key information fed into these models involves the number of individuals with infections at any given point of time, the number of individuals recovered from infection, and the number of deaths. In some cases, clinical data may be used. In reality, such observations are noisy: there are delays in reporting, inaccuracies and, most importantly, there is a possibility of lack of detection.

Setup. Throughout the document, let t denote time. We shall assume, unless stated otherwise, that the unit of time is days. Let $S(t)$ be fraction (or actual number) of the population that is "susceptible" to receive infection at time t, initially $S(0) = 1$. Let $I(t)$ denote the fraction (or actual number) of the population that is actively "infected" at time t. Let $R(t)$ denote the fraction (or actual number) of the population that has recovered (or died) at time t. In addition, let $E(t)$ denote the fraction (or actual number) of the population that is "exposed" to infection at time t.

Organization. We start by describing deterministic mechanistic models from the epidemiology literature (see Hethcote 2000, for a recent review). We follow it with statistical models and approaches to learn these mechanistic models from the data. Finally, we end with a recent proposal about a non-mechanistic, non-parametric approach introduced by Sarker & Shah (2020).

2. Deterministic Mechanistic Models

The Susceptible-Infectious-Recovered (SIR) model of epidemics was introduced by Kermack & McKendrick (1927). Kermack and McKendrick model the flow of individual

agents from the sub-population who are susceptible to a disease into the sub-population who are infectious and from the infectious sub-population into the sub-population who have recovered (and are now immune and thus no longer susceptible). Since the seminal work of Kermack and McKendrick, numerous other models have been proposed to respresent the flow of agents from and into different subpopulations in response to epidemics. In what follows, we describe few primary examples of such models.

Susceptible-Infected-Recoverd (SIR) Model. The SIR model utilizes two parameters β and γ, which capture the *rate* of flow from susceptible to infected, and infected to recovered (or dead).

$$\begin{aligned}\frac{dS(t)}{dt} &= \quad \beta S(t)I(t),\\ \frac{dI(t)}{dt} &= \beta S(t)I(t) \quad \gamma I(t),\\ \frac{dR(t)}{dt} &= \gamma I(t).\end{aligned}$$

Parameter β is the "contact rate", which captures the 'interaction frequency' and 'infectiousness' of the disease. Parameter γ captures the "recovery rate" of disease.

In this model, the fraction of individuals susceptible starts with $S(0) = 1$ and shrinks as the fraction of infected individuals grows, while the fraction of individuals who are recovered (or dead) starts with $R(0) = 0$ and grows as infected individuals recover or expire. Parameters β and γ determine how the fraction of the infected population, $I(t)$, evolves over time. $I(t)$ initially increases exponentially, then reaches a plateau, and eventually shrinks to zero. By definition, $S(t) + I(t) + R(t) = 1$ for all t.

The Susceptible-Exposed-Infectious-Recovered (SEIR) Model. The SIR model can be generalized by adding an "exposed" state. Like the SIR model, it is a deterministic flow model, now with an additional pararmeter ϵ capturing the rate at which exposed individuals become infected, modeling the "speed" or "incubation rate" at which exposure leads to infection. With ϵ *!* *1* , SEIR becomes SIR. Precisely,

$$\begin{aligned}\frac{dS(t)}{dt} &= \quad \beta S(t)I(t),\\ \frac{dE(t)}{dt} &= \beta S(t)I(t) \quad \epsilon E(t),\\ \frac{dI(t)}{dt} &= \epsilon E(t) \quad \gamma I(t),\\ \frac{dR(t)}{dt} &= \gamma I(t).\end{aligned}$$

Introducing Vital Dynamics. The SIR and SEIR model can naturally incorporate exogenous changes due to natural birth and death as well. In addition, the 'loss of immunity' of recovered individuals can also be added to such dynamics. Below, we present

such a modification of the SEIR model, also known as the SEIRS model.

$$\begin{aligned}
\frac{dS(t)}{dt} &= \nu \quad \beta S(t)I(t) \quad \nu S(t) + \delta R(t),\\
\frac{dE(t)}{dt} &= \beta S(t)I(t) \quad \epsilon E(t) \quad \nu E(t),\\
\frac{dI(t)}{dt} &= \epsilon E(t) \quad \gamma I(t) \quad \nu I(t),\\
\frac{dR(t)}{dt} &= \gamma I(t) \quad \delta R(t) \quad \nu R(t).
\end{aligned}$$

Above, ν represents both death and birth rate (assumption is that in the short term, population is stable) and δ represents the rate at which recovered individuals lose immunity to become susceptible again.

Accounting for Unobserved Infection. Naturally, in practice not all infected cases are detected or reported. To address this problem, let's assume that a fixed proportion p of individuals who are infectious are reported in the data. Let $I^o(t)$ be the fraction of reported / observed infected population at time t. Then $I(t) = I^o(t)/p$. Let $R^o(t)$ be the fraction of recovered individuals that were observed to be infected. Assume that the recovery rate for detected infected and undetected infected is identical. In that case,

$$\frac{dR^o(t)}{dt} = \gamma I^o(t). \tag{2.1}$$

This implies $R^o(t) = pR(t)$. Now $I(t)$, the true infection rate, increases at rate $\beta S(t)I(t)$, and fraction p of it is observed. That is,

$$\frac{dI^o(t)}{dt} = p\beta S(t)I(t) \quad \gamma I^o(t) \;=\; \beta S(t)I^o(t) \quad \gamma I^o(t). \tag{2.2}$$

By definition, $S(t) = 1 \quad I(t) \quad R(t)$. That is, $S(t) = 1 \quad I^o(t)/p \quad R^o(t)/p$.

We discuss next some potential empirical strategies to estimate p. A fraction of recovered cases constitute death. Let $D(t)$ be the fraction of the population that has died. Assume that we observe deaths accurately. Let $\rho > 0$ be the morbidity rate for a given disease, i.e. $D(t) = \rho R(t)$. Therefore, if we observe $R^o(t)$ and $D(t)$, then $D(t) = \rho R^o(t)/p$. That is, we can infer ρ/p.

We may have access to data from different regions with different level of testing, i.e. different values of p. However, it is reasonable to assume ρ to be constant across regions. This means that the smallest value of $D(t)/R^o(t)$ across such different regional data may provide an upper bound for ρ. Information about ρ may also be obtained independently from mortality rates in environments with large proportion of testing. For example, after adjusting for the age distribution, the *Diamond Princess* data could be used to produce a plausible lower bound on ρ (if we assume that clinical care for COVID-19 patients from the *Diamond Princess* was of relatively high quality). The mortality rate for *Diamond Princess* patients was about 1.7 percent at the time this report was written. Notice, however, that this estimate is derived for a rather small sample (12 deaths out of 712 patients). Another possible lower bound for ρ is provided by Iceland, where there has been widespread testing. In Iceland, there have been 8 deaths out of 1739 cases at time of writing, corresponding to a death rate of 0.5 percent (Johns Hopkins 2020). Plausible values of ρ can then be used to produce estimates of p for each region.

Dealing with Quarantine, Self-Isolation, Time Varying Dynamics. The response to the emerging epidemic may lead to interventions that impact the dynamics of the SIR

or SEIR model. This can be accommodated by allowing for time varying parameters. That is, β, ϵ, γ become $\beta(t), \epsilon(t), \gamma(t)$ and they may obey specific models themselves. For example, $\beta(t)$ may decrease as $I(t)$ increases, capturing interventions like social distancing when infections are high.

In a recent work, Song *et al.* (2020) proposes a modification of the SIR model to allow a state of quarantine. Specifically, a fraction of the susceptible population becomes quarantined and cannot be infected. This model can be combined with a time-varying contact rate that results from voluntary self-isolation.

Non-linearities. More complex SIR, SEIR and SEIRS models may incorporate non-linearities, for example the rate of flow from susceptible into either infectious or exposed may depend non-linearly on the prevalence of infectious. This could be due to the spatial structure of the population or heterogeneous mixing in the population. Bjornstad *et al.* (2002) shows how models with this property might be estimated from time-series data.

3. Estimation of the SIR model

Overview. In this section, we describe two empirical approaches to the estimation of the SIR model. The first approach is based on panel data. The second approach is Bayesian and uses the Markov Chain Monte Carlo (MCMC) method to estimate model parameters.

Panel data estimation of the SIR model. We present here a panel data estimator of the SIR model. The goal is to produce estimates that can be used to forecast the evolution of the epidemic and to evaluate the impact of mitigation policies (e.g., lockdown). The model in this section is a variation of the ones in Cintrón-Arias *et al.* (2020) and Chen & Qiu (2020). In order to describe the estimation method in terms of quantities directly available in the data, we consider a SIR model with un-normalized variables. That is, in this section $I(t)$, $R(t)$, and $S(t)$ represent the number (not fraction) of individuals that are infected, recovered and susceptible. Let N represent the total population. To simplify the exposition, we will assume that, at the time scale of interest, total population experiences negligible variation. Otherwise, population changes can easily be incorporated in the model. The discrete dynamics of the SIR model can be written as

$$\begin{aligned} \Delta S(t) &= -\beta \frac{I(t)}{N} S(t), \\ \Delta I(t) &= \beta \frac{I(t)}{N} S(t) + \gamma I(t), \\ \Delta R(t) &= \gamma I(t). \end{aligned}$$

We will also assume that, at time $t = 0$, $R(0) = 0$ and hence $S(0) + I(0) = N$. We will initially assume that $I(0)$ is known or can be estimated/approximated in a first step. Later, we will incorporate the initial value of $I(t)$ as a parameter to be estimated.

Assume there are n states in the data ($j = 1, \ldots, n$), with available observations for $T+1$ time steps $(0, \ldots, T)$. Let the time-varying parameters be $\beta_j(t)$ and $\gamma_j(t)$ for state j and time t. The data contains information on new cases, $\Delta I_j(t) + \Delta R_t(t)$ and "recoveries" $\Delta R_j(t)$ (which include deaths). For state j, the model gives

$$\Delta I_j(t) + \Delta R_j(t) = \beta_j(t) \frac{S_j(t)}{N_j} I_j(t), \tag{3.1}$$

$$\Delta R_j(t) = \gamma_j(t) I_j(t), \tag{3.2}$$

$$S_j(t) = N_j - I_j(t) - R_j(t).$$

The simplest way to proceed now is to separately identify and estimate $\gamma_j(t)$ as the inverses of times to recovery in different states and time periods. For the contact rates, assume

$$\beta_j(t) = \exp\ \delta(t)(\boldsymbol{\theta}) + \boldsymbol{X}_j(t)^T\boldsymbol{\beta}\ , \tag{3.3}$$

where $\boldsymbol{X}_j(t)$ collects the values of a set of observed variables which may include characteristics of the state (e.g., population density) and policy variables (e.g., lockdown) for state j at time t, and $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are vectors of parameters. $\delta(t)$ could be completely unrestricted (i.e., time "fixed effects") or modeled arbitrarily (e.g., as a polynomial in t). Now, the parameters of the model can be estimated by least squares:

$$\underset{\boldsymbol{\theta};\boldsymbol{\beta}}{\text{minimize}} \sum_{j=1}^{J}\sum_{t=0}^{T-1} \Delta I_j(t) + \Delta R_j(t) \quad \frac{S_j(t)I_j(t)}{N_j}\exp\ \delta(t)(\boldsymbol{\theta}) + \boldsymbol{X}_j(t)^T\boldsymbol{\beta}\ \Big)^2 . \tag{3.4}$$

The model can be estimated with data on detected cases provided that $S(t)$ is adjusted for undetected cases, as explained above. Notice also that $I_j(0)$ could be included as a parameter to be estimated in this regression, in which case we would have to evaluate $I_j(t)$ and $S_j(t)$ as functions of $I_j(0)$. Estimates would potentially be subject to bias when the number of periods, $T+1$, in the data is small. Plausible scales of the bias for available sample sizes could be investigated via simulations. It is also possible to estimate $\gamma_j(t)$ in the same step as $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ by adding to the objective function a second sum of squares based on equation (3.2), as in Chen & Qiu (2020).

A Bayesian Approach for Parameter Estimation. In a recent work, Song *et al.* (2020) employ and expand upon methods introduced in Osthus *et al.* (2017) (who model the spread of seasonal influenza) to model the spread of COVID-19. Specifically, Song *et al.* (2020) build on the classic linear SIR model by allowing for measurement error of the prevalence of infectious and recovered individuals. They also provide two alternatives for modeling the effect of self-isolation and quarantine. They carry out Bayesian estimation of their model using Markov Chain Monte Carlo with priors over the unknown parameters based on estimates from the SARS outbreak in the early 2000s. Their work is reproducible in the sense that the implementation in R is made available.

To that end, let $Y^I(t), Y^R(t)$ be noisy measurements of $I(t), R(t)$ respectively. Given $S(t), I(t), R(t)$, the observations $Y^I(t), Y^R(t)$ are modeled to have Beta-distribution, which has support on $[0,1]$. Specifically, $Y^I(t)$ is distributed as $\mathsf{Beta}(\lambda_I I(t), \lambda_I(1 \quad I(t)))$ and $Y^R(t)$ is distributed as $\mathsf{Beta}(\lambda_R R(t), \lambda_R(1 \quad R(t)))$. Thus,

$$\mathbb{E}[Y^I(t)|I(t)] = I(t) \quad \text{and} \quad \mathbb{E}[Y^R(t)|R(t)] = R(t).$$

In addition to the measurement error, Song *et al.* (2020) and Osthus *et al.* (2017) add a stochastic component to the SIR dynamics.

In Song *et al.* (2020), self-isolation / quarantine is captured in two alternative ways. First, they replace the constant contact rate β with a time-varying rate $\pi_t\beta$, where π_t is treated as known rather than estimated from the data. They consider either π_t to be a step function that jumps downwards by some amount in response to the imposition of a quarantine or to be an exponential function of time. In an alternative formulation, Song *et al.* (2020) assume that upon the imposition of a quarantine at time t, a fixed proportion ϕ_t of the susceptible individuals enter a quarantine state which they do not leave. Again ϕ_t is treated as known.

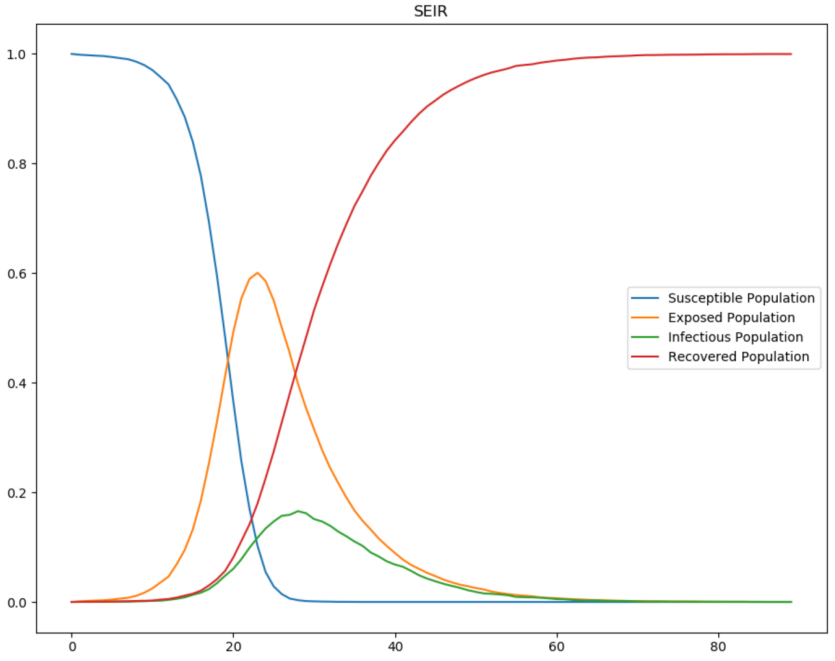

Figure 1. Typical growth and evolution under SEIR model (from Wikipedia).

4. A Non-mechanistic, Non-parametric Approach

We next describe a recent proposal for a non-mechanistic, non-parametric approach. The description below is based on promising, preliminary work by Sarker & Shah (2020).

A Challenge with the Mechanistic Models. The SIR, SEIR, and SEIRS models fundamentally assume that the entire population is going through similar phenomenon simultaneously. But as observed by Chen & Qiu (2020), the growth across geographically separated different countries may be evolving in accordance with different models (even if we assume SIR is the correct model for each country individually). In a similar manner, in any state or county the epidemic or spread of infection may be evolving with multiple growth clusters. And the number of growth clusters might change in time. To address these challenges, Sarker & Shah (2020) introduces a non-mechanistic, non-parametric model that we describe next.

Some background. The aspect of growth of an epidemic that SEIR-like models capture well is the initial exponential growth, followed by a slowdown due to saturation, followed by eventual recovery as seen in Figure 1. However, such growths are likely happening in different clusters with each growth cluster having different characteristics / scale as well as time scales at which they evolve. Now, assume each growth cluster's evolution obeys generic form of $\exp(\ \ f(t))$ where $f(\ )$ is a strongly convex function. For example, $f(t) = at^2 + bt + c$ with $a > 0$. Indeed, such form is observed for SEIR like models as well. Therefore, for the purpose of prediction, Sarker & Shah (2020) propose to utilize such a non-mechanistic, non-parametric model.

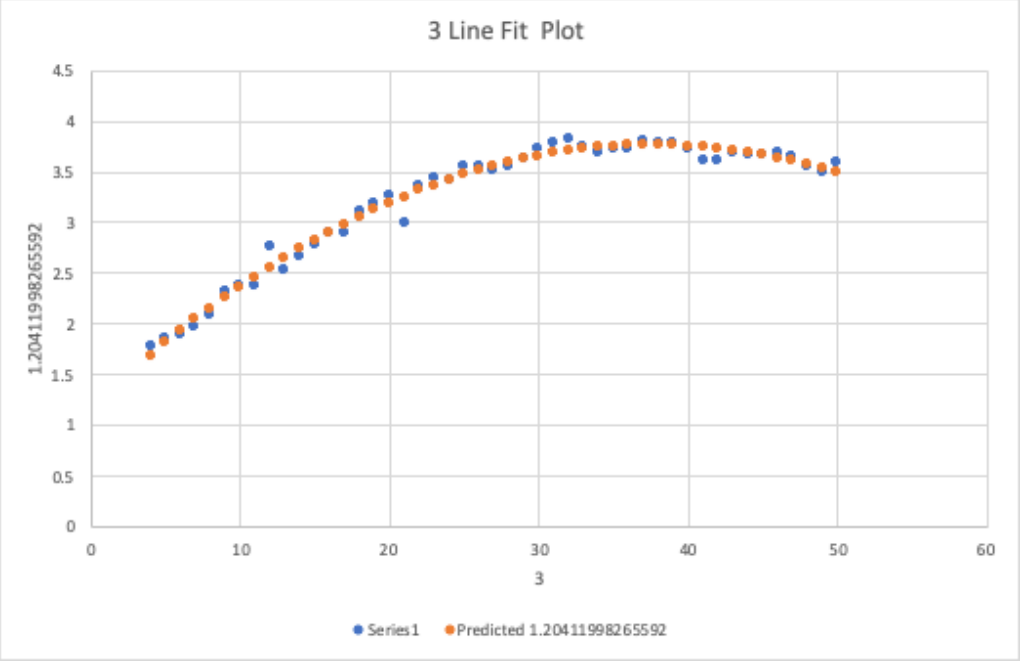

Figure 2. Model fit (R^2 = 0:99) with single component to daily COVID-19 cases in Italy (Courtesy: David Gamarnik, MIT).

A Non-mechanistic, Non-parametric Model. Let $F(t)$ $\Delta(I(t)+R(t))$ denote the new infections at time t. Then,

$$F(t) = \sum_{k=1}^{r(t)} \exp(\ a_k t^2 \ \ b_k t \ \ c_k)\ (1+\varepsilon(t)), \tag{4.1}$$

where each $a_k > 0$, $\forall\ k > 1$, $r(t) > 1$ is number of "clusters" observed till time t, $\varepsilon(t)$ is independent random variable with zero mean representing measurement error.

Preliminary results. To start with, we discuss preliminary results that provide evidence of support for the model. To that end, we utilize the data made available through the GitHub repository of *The New York Times* about number of cases reported, number of deaths at county level in United States as well as the Github repository of JHU for world-wide, global data.

Model Fit To Country-level Data: A Single Mixture. We start by verifying whether the exponential of a quadratic function is a reasonable form for capturing the growth of epidemic. To that end, we start by modeling growth in Italy through a single mixture. As shown in Figure 2, we find a remarkable fit of the model with $R^2 = 0.99$.

Model Fit To State-level Data: Multiple Mixtures. Next, we consider state-level growth data in the US. In particular, we focus on the growth data in New York State. Clearly, NYC has been a prominent growth cluster. But, in addition, there are other growth clusters. And indeed, as we fit a mixture of two clusters to NY state-level data, we find excellent fit as well as predictive power in the model. Specifically, as shown in Figure 3, we fit the model using "Orange" data which visually does not show multiple clusters, but as seen by "Green" test data, our two cluster model fit manages to predict well.

Evidence of Multiple Mixtures. The quantity $\log F(t+1)/F(t)$ (ignoring noise term),

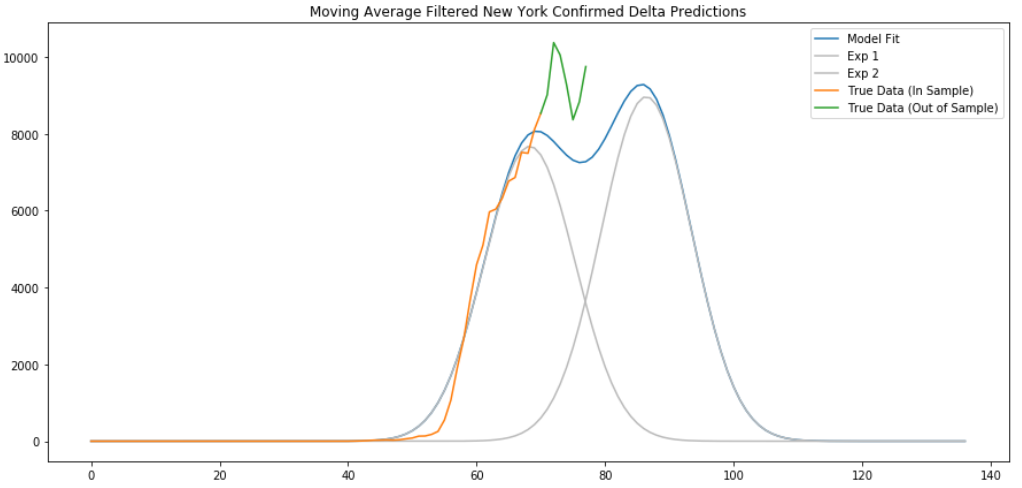

Figure 3. Model fit with two component to daily COVID-19 cases in NY State with second growth cluster predicting the future increase followed by a dip accurately.

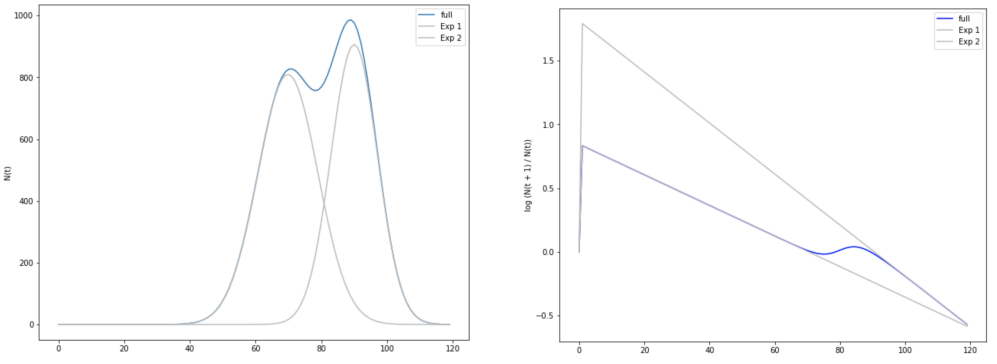

Figure 4. $F(t)$ and $\log F(t+1)/F(t)$ for a mixture of (synthetic) exponential curves.

evolves as differences of two log-exp-sums. If the growth indeed obeyed *single* mixture, then this would be linear in t. On the other hand, if it was *multiple* mixtures, then it would look more like piece-wise linear function with connections between piece-wise linear components being non-linear curves, as in Figure 4. In fact, as seen in Figure 5, such piece-wise linear curves are evident in empirical COVID-19 data.

Predicting Apex. Using this model, assuming no new growth cluster emerges, we can predict the "apex" dates for California and Louisiana as shown in Figures 6 and 7. Using the single mixture fit to the most recent growth cluster via quadratic regression technique, we can find the uncertainty band around apex. In particular, Figure 8 plots the apex estimates for various counties in US.

Parameter Fit Using Alternating Minimization. Given observations $F(t), t \in \{0, \ldots, T\}$, and a choice of the number of growth clusters, $r = r(T)$ (which could be 1 + the number of piece-wise components observed in plot of $\log F(t+1)/F(t), t \in \{0, \ldots, T\}$), we present a simple, heuristic algorithm to fit the model parameters. Specifically, we wish to find parameters $(a_k, b_k, c_k), k \leq r$ with $a_k > 0, k \leq r$. To start with, initialize each of these parameters subject to non-negativity constraints $a_k > 0, k \leq r$. Then, in each

Figure 5. Plot of $\log(F(t+1)/F(t))$ across states resembling multiple piece-wise near-linear segments, each piece effectively corresponding to different growth cluster (Courtesy: Peko Hosoi, MIT).

iteration, one-by-one, for each $k \leq r$, keeping all other (a_ℓ, b_ℓ, c_ℓ), $\ell \neq k$ fixed, find best fit for (a_k, b_k, c_k) subject to $a_k > 0$. Repeat for some large number of iterations or until parameters stop changing beyond a small threshold.

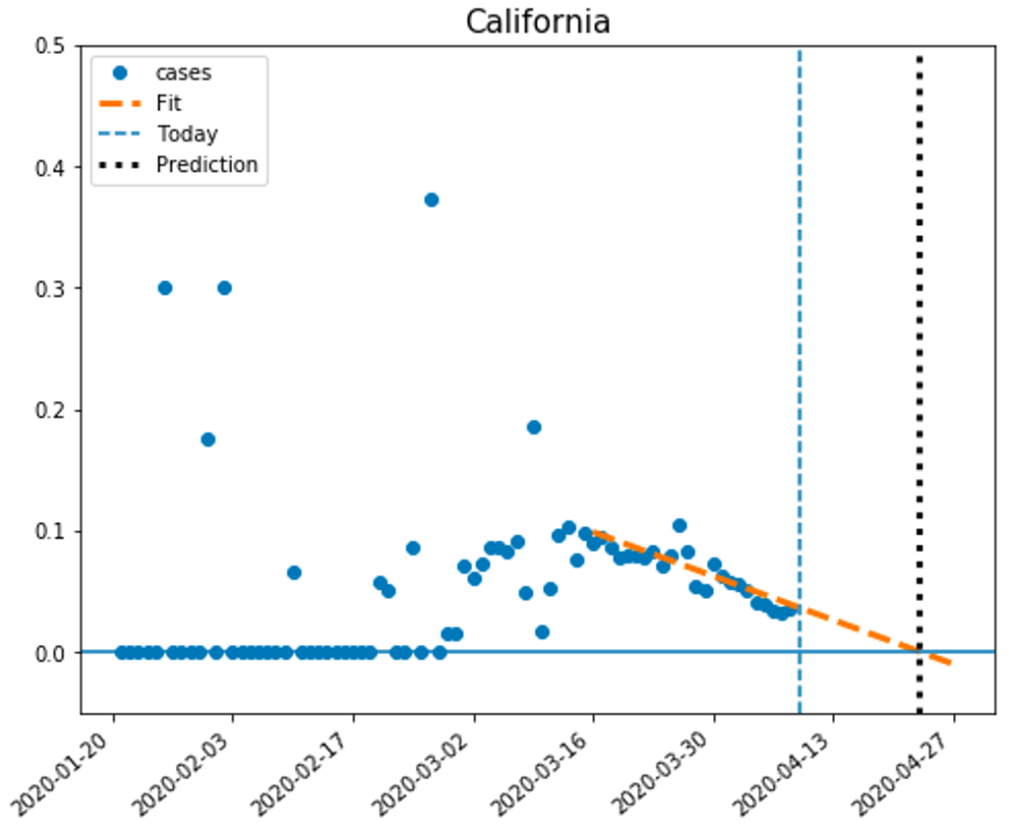

Figure 6. The estimation of apex in California using the model (Courtesy: Yash Deshpande, MIT).

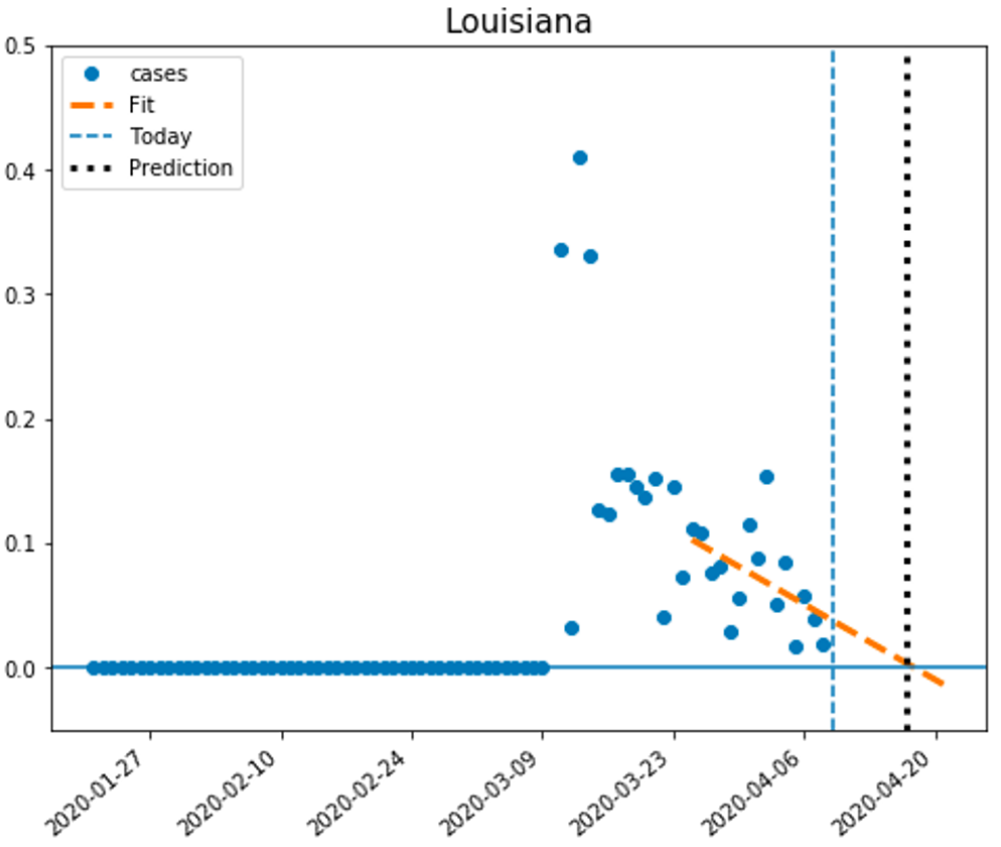

Figure 7. The estimation of apex in Louisiana using the model (Courtesy: Yash Deshpande, MIT).

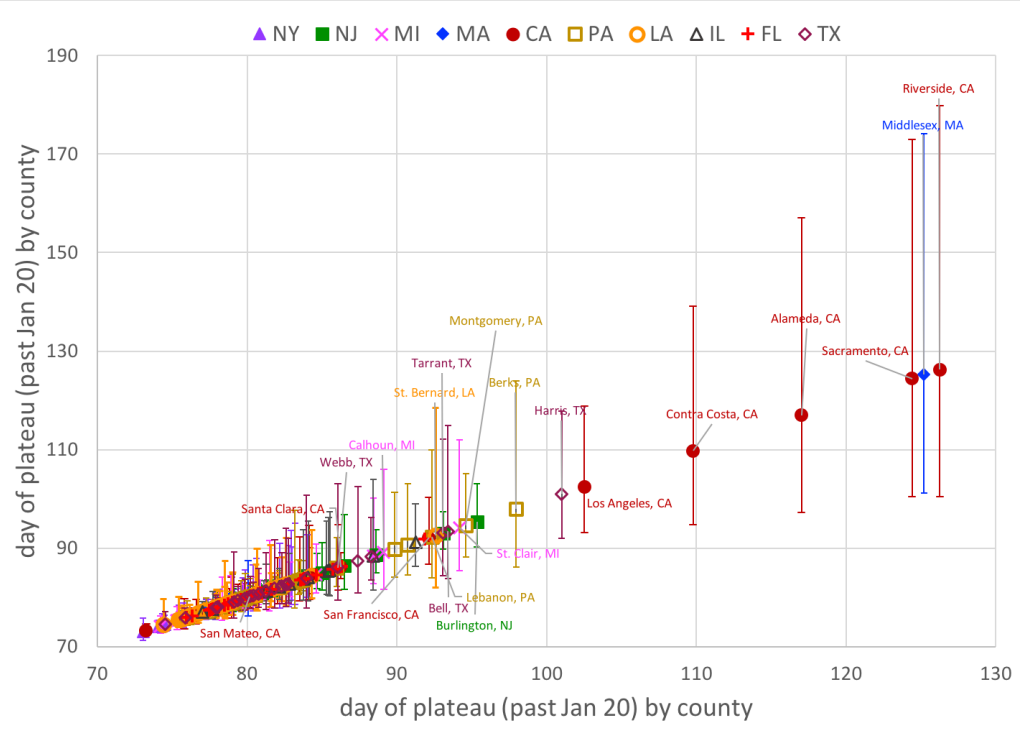

Figure 8. The estimation of apex with uncertainty in various counties in US with day 0 = January 20, 2020. Using the standard quadratic regression method, the parameter uncertainty is evaluated. The estimation of apex is given by ratio b_1=a_1 (under single mixture model assumption) and the uncertainty in apex is obtained by taking lower and upper bound of this ratio by using the 95% upper and lower bound on each of these parameters. (Courtesy: Hamsa Balakrishnan, MIT).

REFERENCES

Bjornstad, Ottar N., Finkenstadt, Barbel F. & Grenfell, Bryan T. 2002 Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs* 72, 169.

Chen, Xiaohui & Qiu, Ziyi 2020 Scenario analysis of non-pharmaceutical interventions on global COVID-19 transmissions, arXiv: 2004.04529.

Cintrón-Arias, Ariel, Castillo-Chávez, Carlos, Bettencourt, Luís M. A., Lloyd, Alun L. & Banks, H. T. 2020 The estimation of the effective reproductive number from disease outbreak data, arXiv: 2004.06827.

Hethcote, Herbert W 2000 The mathematics of infectious diseases. *SIAM review* 42 (4), 599–653.

Johns Hopkins 2020 Coronavirus resource center. coronavirus.jhu.edu/map.html.

Kermack, William Ogilvy & McKendrick, Anderson G 1927 A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115 (772), 700–721.

Osthus, Dave, Hickmann, Kyle S, Caragea, Petruţa C, Higdon, Dave & Del Valle, Sara Y 2017 Forecasting seasonal influenza with a state-space sir model. *The annals of applied statistics* 11, 202–224.

Sarker, Arnab & Shah, Devavrat 2020 A non-mechanistic, non-parametric model for epidemic. Working Paper.

Song, Peter X., Wang, Lili, Zhou, Yiwang, He, Jie, Zhu, Bin, Wang, Fei, Tang, Lu & Eisenberg, Marisa 2020 An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China. MedRxiv preprint.