EPIDEMIC MODEL GUIDED MACHINE LEARNING FOR COVID-19 FORECASTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a new epidemic model (SuEIR) for forecasting the spread of COVID-19, including numbers of confirmed and fatality cases at national and state levels in the United States. Specifically, the SuEIR model is a variant of the SEIR model by taking into account the untested/unreported cases of COVID-19, and trained by machine learning algorithms based on the reported historical data. Numerical results verify that the proposed can achieve quite good accuracy in terms of short-term (5-week) prediction.

1 INTRODUCTION

The novel coronavirus disease (COVID-19), an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Chan et al., 2020; WHO, 2020a), has emerged into a global pandemic and led to $2,581,976$ death toll in the world as of March 8, 2021 (WHO, 2020b). With the increasing availability of public data on COVID-19, more and more researches (Flaxman et al., 2020; Bendavid et al., 2020; Sutton et al., 2020; Altieri et al., 2020; Bertozzi et al., 2020; Murray et al., 2020) have been carried out to understand and prevent the spread of COVID-19 from different aspects. Among them, one important research direction is to model and forecast the spread of COVID-19, such as predicting the peak of the active cases on the virus and the size of the coronavirus outbreak. Such results can help government agencies better understand the overall impact of the disease and also facilitate policy makers in terms of pandemic preparedness and response such as allocating the medical resources.

One widely used method for modeling the spread of infectious disease is to use epidemic models such as Susceptible-Infected-Resistant (SIR) (Kermack & McKendrick, 1927) and Susceptible-Exposed-Infected-Removed (SEIR) (Hethcote, 2000). Such epidemic models are quite useful in describing the dynamics of transmission and are well-suited for predicting the peak of active cases on the virus. From the decision-making perspective, the peak prediction is able to forewarn the health system when to expect a surge in cases. Several recent works used epidemic models such as the SIR and SEIR models (Imai et al., 2020; Li et al., 2020a; Wu et al., 2020; Kucharski et al., 2020; Read et al., 2020; Tang et al., 2020; Ferguson et al., 2020) to simulate the spread of COVID-19 in different regions and were able to forecast the size and severity of such epidemic outbreak. However, it is often the case that the number of publicly reported cases (including confirmed cases, fatality cases, and recovered cases) is much less than their real numbers as many infectious cases have not been tested due to test capability and asymptomatic patients, or even possibly under-reporting (Li et al., 2020b). As a result, most of existing COVID-19 models built based on SIR or SEIR cannot accurately characterize the epidemic evolution of COVID-19 without taking such unreported cases into consideration.

The goal of this paper is to make good use of the current public data on COVID-19 to better understand the spread of the coronavirus and to facilitate informed decisions by policy makers. In order to achieve this goal, we develop a new epidemic model, called the SuEIR model, to forecast the active cases and deaths of COVID-19 by taking the untested/unreported cases into consideration. In addition, we use machine learning based methods to train our model, which enables us to train the model efficiently. Based on the proposed model, we are able to make accurate predictions on the numbers of confirmed cases and fatality cases for nations, states, and counties. Moreover, our model can also estimate the basic reproduction number ($\mathcal{R}_0$) and the effective reproduction number ($\mathcal{R}_t$) of different states in the US, which we will specify in the appendix due to the space limit.

2 METHODS

In this section, we propose a new epidemic model and a machine learning method to train this model. Due to the space limit we defer some details of the machine learning method and the calculations of reproduction numbers in the Appendix.

2.1 THE SUEIR MODEL

It is observed that COVID-19 has an incubation period ranging from 2 to 14 days (Lauer et al., 2020). It has also been observed that individuals who have been exposed to the coronavirus can also infect the susceptible group during this period. In addition, it is often the case that the number of reported cases (including confirmed cases and recovered cases) are less than their real numbers as many exposed cases have not been tested, which will not pass to the next compartment. However, such important factors cannot be characterized by the classical epidemic models such as the SIR and SEIR models. We also observe that directly applying SIR or SEIR model to fit the reported data will lead to unreasonable predictions (e.g., more than 20 million confirmed cases in the US). Therefore, we proposed a new epidemic model that takes the untested/unreported cases as well as the "silent spreaders" into consideration. We call our model the SuEIR model, and it is illustrated in Figure 1.

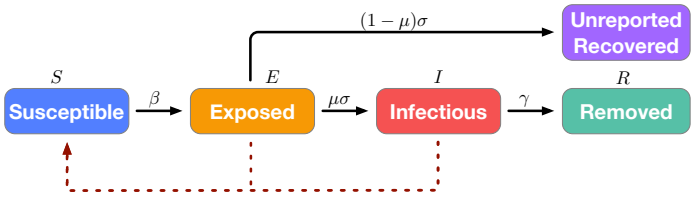


Figure 1: Illustration of the SuEIR model. Solid lines represent the transitions of individuals and dashed lines represent the routes of infection.

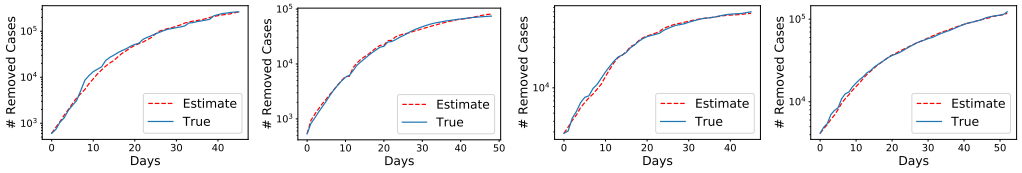
In particular, the compartment Exposed in our model is considered as the individuals that have already been infected and have not been tested. Therefore, they also have the capability to infect susceptible individuals. Moreover, some of such individuals can receive a test (typically have obvious symptoms) and be further passed to the Infectious compartment (as well as reported to the public), while the others (typically have mild or no symptoms) will recover but not appear in the publicly reported cases. Therefore, we introduce a new parameter $0 < \mu < 1$ in the evolution dynamics of I_t to characterize the ratio of the exposed cases that are confirmed and reported to the public, which we call it the *discovery rate*. This discovery rate reflects the unreported/undiscovered cases, which is an important latent factor in the dynamics of the epidemic model. As a result, we propose to use the following ordinary differential equations to describe our proposed SuEIR model:

$$\begin{aligned} \frac{dS_t}{dt} &= -\frac{\beta(I_t+E_t)S_t}{N}, \quad \frac{dE_t}{dt} = \frac{\beta(I_t+E_t)S_t}{N} - \sigma E_t, \\ \frac{dI_t}{dt} &= \mu\sigma E_t - \gamma I_t, \quad \frac{dR_t}{dt} = \gamma I_t, \end{aligned} \tag{1}$$

where β denotes the contact rate between the susceptible and "infected" groups (including both exposed and infectious compartments in Figure 1), σ is the ratio of cases in the exposed compartments that are either confirmed as infectious or dead/recovered without confirmation, μ is the discovery rate of the infected cases, and γ denotes the remove rate from the confirmed infectious group.

2.2 PARAMETER LEARNING FOR THE SUEIR MODEL

Estimation of the number of removed cases R_t. Note that I_t and R_t in our model determine the number of "current" infectious cases (a.k.a., active cases) and removed cases, i.e., the sum of recovered and fatality cases, respectively. However, most of the reported data only include the number of confirmed cases, i.e., the sum of infected cases and removed cases $I_t + R_t$. In order to train the model, we need to get I_t and R_t separately. In addition, the SuEIR model can only predict the number of removed cases, while in many cases, people are more interested in the number of



(a) US (b) China (c) France (d) Italy

Figure 2: Estimated number of removed cases for different countries using the exponential decay ratio (2) between the daily increased fatality and removed cases.

fatality cases. Therefore, in order to enable the training of the SuEIR model, as well as provide the predictions for the number of fatality cases, we have to: (1) estimate the number of removed cases; (2) determine the number of active cases in the reported data by subtracting the estimated number of removed cases. In order to do so, we propose to use the following exponential function to model the ratio between the daily increased fatality cases and the removed cases,

$$r(t) = a\exp(-bt) + c, \tag{2}$$

where $a, b, c > 0$ are parameters controlling the shape of the exponential function and t denotes the number of days since the starting date. Based on (2), we can also predict the number of fatality cases by extracting the desired numbers from the predicted number of removed cases. We demonstrate the effectiveness of this approach using the nation-level reported data, which include the numbers of both daily death and recover cases. The results are displayed in Figure 2, which clearly shows that the exponential functions can well describe the ratio between the daily increased numbers of fatality and removed cases. For states in the US, we try different choices of a, b, and c around the optimal ones obtained for the US, and pick the one with the smallest validation error.

Model training. The goal of model training is to find the optimal model parameters (including β, σ, γ, and μ) such that the estimated $\widehat{S}_t$, $\widehat{E}_t$, $\widehat{I}_t$ and $\widehat{R}_t$ are close to the observed data (i.e., S_t, E_t, I_t and R_t). Moreover, note that the ground truth data only include the numbers of confirmed cases C_t and fatality cases F_t, thus we first compute the estimated number of confirmed cases by $\widehat{C}_t = \widehat{I}_t + \widehat{R}_t$ and apply the ratio function (2) to get the estimated number of fatality cases ($\widehat{F}_t$) based on $\widehat{R}_t$. Then we propose to learn the model parameter $\widehat{\boldsymbol{\theta}} = (\widehat{\beta}, \widehat{\sigma}, \widehat{\gamma}, \widehat{\mu})$ by minimizing the following logarithmic-type mean square error (MSE):

$$L(\boldsymbol{\theta}; \mathbf{C}, \mathbf{F}) = \frac{1}{T}\sum_{t=1}^{T}\Big[\big(\log(\widehat{C}_t + p) - \log(C_t + p)\big)^2 + \big(\log(\widehat{F}_t + p) - \log(F_t + p)\big)^2\Big], \tag{3}$$

where $\mathbf{C} = \{C_t\}_{t=1}^T, \mathbf{F} = \{F_t\}_{t=1}^T$ with C_t and F_t denote the reported numbers of confirmed cases and fatalities cases at time t (i.e., date), and p is the smoothing parameter used to ensure numerical stability. The model parameter $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{C}, \mathbf{F})$ can be learnt by applying L-BFGS (Nocedal & Wright, 2006) to the loss (3) under the constraint that $\beta, \sigma, \gamma, \mu \in [0, 1]$.

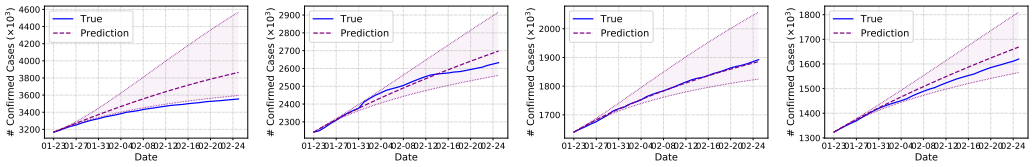
3 Results

In this section, we present the short-term (5-weeks) forecast results in some major states in the US to demonstrate the effectiveness of the proposed model and the machine learning algorithm. The experimental results of the reproduction numbers are deferred to Appendix B.

Data collection. We use the data from the Johns Hopkins University Center for Systems Science and Engineering[1] (Dong et al., 2020), including the numbers of daily confirmed cases and deaths.

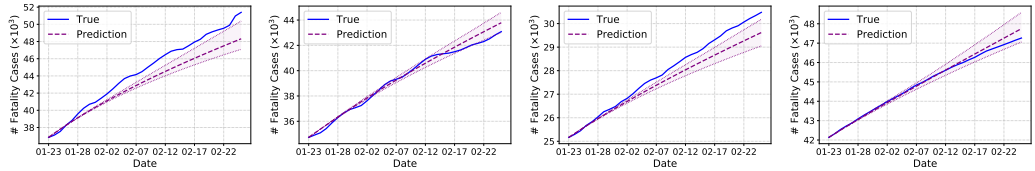
Prediction results. For the interest of space, we present the forecast results of our models for the US and some major states in the US, including New York, California, New Jersey, Texas, and Florida.

[1] https://github.com/CSSEGISandData/COVID-19.



(a) California (b) Texas (c) Florida (d) New York

Figure 3: Short-term (daily ahead) predictions of total confirmed cases in California, Texas, Florida, and New York. For each region, we present the prediction with its 95% confidence interval, and display the public reported data for comparison.



(a) California (b) Texas (c) Florida (d) New York

Figure 4: Short-term (daily ahead) predictions of total Fatality cases in California, Texas, Florida, and New York. For each region, we present the prediction with its 95% confidence interval, and display the public reported data for comparison.

Train-test splitting. We use all reported data from 2020-03-22 to 2021-02-28. In particular, all models are trained using the data from 2020-03-22 to 2021-01-16 and validated on the data from 2021-01-17 to 2021-01-23. The models are next evaluated based on the test data from 2021-01-24 to 2021-02-28.

In order to evaluate the prediction performance of our model, we present the short-term predictions of the numbers of confirmed cases and fatality cases in Figures 3 and 4, where the solid line represents the reported data, the dotted line represents the prediction, and the shadow area represents the 95% confidence interval. It can be seen that for Texas and Network the prediction results are close to the reported ones (the ground-truth data stay inside the 95% confidence interval of the prediction) for both confirmed and fatality cases, which demonstrates the effectiveness of the proposed model. For California and Florida, there exists some slight deviations between the ground-truth and predictions (particular, both confirmed cases and fatality cases predictions for California, and fatality cases prediction for Florida). These are possibly due to the inconsistency of the number of daily tests in the periods of training data and test data. In particular, the number of daily tests conducted in Florida and California keeps increasing in the period close to the end date of the training data (1.e., 2021-01-16), while turns to decrease in the period of test data.

4 Conclusion

We developed a novel epidemic model called SuEIR to infer the unreported cases of individuals contacting COVID-19. Based on this new model, we further develop a machine learning approach to forecast the numbers of confirmed and fatality cases in the US, and estimate the basic reproduction numbers as well as the effective reproduction numbers of the US and different states.

Moreover, we found that for most states, the learned discover rate (i.e., μ) is less than 0.1, which implies that a large fraction of "Exposed" individuals will finally recover/die without being tested and reported. This further suggests that the actual number of infected cases (including active, fatality and recovered cases) in the US may be more than 10 million, while most of them are not counted. This result is consistent with the recent findings by the researchers from the University of Southern California (Sood et al., 2020), which show that 4.65% (CI: $[2.8\%, 5.6\%]$) of Los Angeles residents have already contracted the COVID-19 virus, which is approximately 23 times more than the official reported numbers.

REFERENCES

Nick Altieri, Rebecca Barter, James Duncan, Raaz Dwivedi, Karl Kumbier, Xiao Li, Robert Netzorg, Briton Park, Chandan Singh, Yan Shuo Tan, et al. Curating a covid-19 data repository and forecasting county-level death counts in the united states. 2020.

Eran Bendavid, Bianca Mulaney, Neeraj Sood, Soleil Shah, Emilia Ling, Rebecca Bromley-Dulfano, Cara Lai, Zoe Weissberg, Rodrigo Saavedra, James Tedrow, et al. Covid-19 antibody seroprevalence in santa clara county, california. *medRxiv*, 2020.

Andrea L Bertozzi, Elisa Franco, George Mohler, Martin B Short, and Daniel Sledge. The challenges of modeling and forecasting the spread of covid-19. *arXiv preprint arXiv:2004.04741*, 2020.

Luis MA Bettencourt and Ruy M Ribeiro. Real time bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS One*, 3(5):e2185, 2008.

Jasper Fuk-Woo Chan, Shuofeng Yuan, Kin-Hang Kok, Kelvin Kai-Wang To, Hin Chu, Jin Yang, Fanfan Xing, Jieling Liu, Cyril Chik-Yan Yip, Rosana Wing-Shan Poon, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*, 395(10223):514–523, 2020.

Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 2020.

Neil Ferguson, Daniel Laydon, Gemma Nedjati Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, ZULMA Cucunuba Perez, Gina Cuomo-Dannenburg, et al. Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand. 2020.

Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Helen Coupland, Thomas A Mellan, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, Pablo NP Guzman, et al. Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in european countries: technical description update. *arXiv preprint arXiv:2004.11342*, 2020.

Jane M Heffernan, Robert J Smith, and Lindi M Wahl. Perspectives on the basic reproductive ratio. *Journal of the Royal Society Interface*, 2(4):281–293, 2005.

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

Natsuko Imai, Anne Cori, Ilaria Dorigatti, Marc Baguelin, Christl A Donnelly, Steven Riley, and Neil M Ferguson. Report 3: transmissibility of 2019-ncov. In *Imperial College London*. 2020.

William Ogilvy Kermack and Anderson G McKendrick. 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, 1927.

Adam J Kucharski, Timothy W Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M Eggo, Fiona Sun, Mark Jit, James D Munday, et al. Early dynamics of transmission and control of covid-19: a mathematical modelling study. *The lancet infectious diseases*, 2020.

Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 2020.

Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy SM Leung, Eric HY Lau, Jessica Y Wong, et al. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 2020a.

Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov2). *Science*, 2020b.

Junling Ma. Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling*, 2020.

Christopher JL Murray et al. Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. *medRxiv*, 2020.

Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

Jonathan M Read, Jessica RE Bridgen, Derek AT Cummings, Antonia Ho, and Chris P Jewell. Novel coronavirus 2019-ncov: early estimation of epidemiological parameters and epidemic predictions. *MedRxiv*, 2020.

Neeraj Sood, Paul Simon, Peggy Ebner, Daniel Eichner, Jeffrey Reynolds, Eran Bendavid, and Jay Bhattacharya. Seroprevalence of SARS-CoV-2–Specific Antibodies Among Adults in Los Angeles County, California, on April 10-11, 2020. *JAMA*, 05 2020.

Desmond Sutton, Karin Fuchs, Mary D'alton, and Dena Goffman. Universal screening for sars-cov-2 in women admitted for delivery. *New England Journal of Medicine*, 2020.

Biao Tang, Xia Wang, Qian Li, Nicola Luigi Bragazzi, Sanyi Tang, Yanni Xiao, and Jianhong Wu. Estimation of the transmission risk of the 2019-ncov and its implication for public health interventions. *Journal of Clinical Medicine*, 9(2):462, 2020.

WHO. Naming the coronavirus disease (covid-19) and the virus that causes it. 2020a.

WHO. Coronavirus disease 2019 (covid-19) situation report. 2020b.

Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.

Joseph T Wu, Kathy Leung, and Gabriel M Leung. Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. *The Lancet*, 395(10225):689–697, 2020.

A More Details of the Machine Learning Framework

A.1 Initialization and Train-Validation Splitting

Initialization. Regarding the initialization of the SuEIR model, we directly set $\widehat{I}_0 = I_0$ and $\widehat{R}_0 = R_0$[2]. Additionally, one can typically set $\widehat{S}_0 + \widehat{E}_0 + \widehat{I}_0 + \widehat{R}_0 = N$, where N is the total population of the region (which can be either a country or a state/county). However, since most of the states/counties in the US have already issued the stay-at-home order, the actual total number of cases in the SuEIR model will be strictly less than N. Thus we set $\widehat{S}_0 + \widehat{E}_0 + \widehat{I}_0 + \widehat{R}_0 = \widehat{N}$ for some $\widehat{N} < N$. Moreover, it is worth noting that the initialization of E, i.e., $\widehat{E}_0$, is a bit tricky since we do not know the number of infected cases before testing them. It is not reasonable to set $\widehat{E}_0 = 0$ since generally there has already existed a large number of infected cases when the local governments began to test. Therefore, we propose to use a validation set to choose the optimal initial estimates of $\widehat{N}$ and $\widehat{E}_0$ when training our model.

Validation set. To determine N_0 and the initial value $\widehat{E}_0$, we first split our data into the training data set and the validation data set. In detail, we choose the data in the most recent 7 days as the validation set, while treating the remaining as the training set. For example, suppose we have the data up to May 10, 2020, the data after May 3, 2020 will be used as the validation set, and the data up to May 3, 2020 will be used as the training set. We then do a grid search on different combinations of $\widehat{N}$ and $\widehat{E}_0$ and train different models on the training set accordingly. Finally, we choose the combination of $\widehat{N}$ and $\widehat{E}_0$ with the smallest validation loss (evaluated using the loss function (3)) along with the best model parameters (i.e., β, γ, σ, μ) to build the SuEIR model for prediction.

[2]Here we omit the numbers of removed cases and recovered cases at the initialization by setting $\widehat{I}_0$ and $\widehat{R}_0$ to be the reported numbers of confirmed cases and fatality cases.

A.2 CONFIDENCE INTERVAL

Given the initial quantities S_0, E_0, I_0, R_0, we can solve the optimization problem in (3) to obtain the model parameter $\widehat{\boldsymbol{\theta}} = (\widehat{\beta}, \widehat{\sigma}, \widehat{\gamma}, \widehat{\mu})$. To assess the confidence of our estimator, we construct the confidence interval of $\boldsymbol{\theta}$ following the previous work (Ma, 2020). More specifically, for a valid model parameter $\boldsymbol{\theta}$, we can compute the loss $L(\boldsymbol{\theta})$ in (3), and construct the test statistic as $\mathcal{T}(\boldsymbol{\theta}) = 2T\big(L(\boldsymbol{\theta}) - L(\widehat{\boldsymbol{\theta}})\big)$, which represents the loglikelihood ratio between the point estimator $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$. Note that $\boldsymbol{\theta}$ contains four free parameters (i.e., β, σ, γ and μ) while $\widehat{\theta}$ is fixed. By Wilks's Theorem (Wilks, 1938), we know that $\mathcal{T}(\boldsymbol{\theta})$ follows χ^2_4 distribution asymptotically. As a result, we can compare $\mathcal{T}(\boldsymbol{\theta})$ with the $(1-\alpha)$ quantile of the χ^2_4 distribution and determine whether $\boldsymbol{\theta}$ is in the confidence interval or not. In our experiments, we apply grid search on both sides of the point estimator $\widehat{\boldsymbol{\theta}}$ to find the boundary of the confidence interval.

A.3 COMPUTATION OF THE BASIC REPRODUCTION NUMBER $\mathcal{R}_0$

We can also compute the basic reproduction number based on our proposed SuEIR model. Note that our model has a different dynamics from that of SIR and SEIR models. Thus we cannot directly apply the standard computation method of $\mathcal{R}_0$ for the SIR or SEIR model to compute such number. Instead, we use the method proposed in Heffernan et al. (2005) to calculate $\mathcal{R}_0$ based on the next-generation matrix. In specific, let $\mathbf{x} = (x_1, \ldots, x_4)^\top$ with x_i being the number of infected individuals in the compartment i. Then we denote by function $F_i(\mathbf{x})$ the rate of new infections in compartment i, and denote by $V_i^-(\mathbf{x})$ and $V_i^+(\mathbf{x})$ the rate of individuals transferred out of the compartment i and the rate of individuals transferred into the compartment i by all other means respectively. Let $V_i(\mathbf{x}) = V_i^-(\mathbf{x}) - V_i^+(\mathbf{x})$, $F(\mathbf{x}) = (F_1(\mathbf{x}), \ldots, F_4(\mathbf{x}))^\top$ and $V(\mathbf{x}) = (V_1(\mathbf{x}), \ldots, V_4(\mathbf{x}))^\top$. The ODE (1) can be rewritten as $d\mathbf{x}/dt = F(\mathbf{x}) - V(\mathbf{x})$ with

$$F(\mathbf{x}) = \begin{bmatrix} 0 \\ \frac{\beta(x_2+x_3)x_1}{N} \\ 0 \\ 0 \end{bmatrix}, \quad V(\mathbf{x}) = \begin{bmatrix} \frac{\beta(x_2+x_3)x_1}{N} \\ \sigma x_2 \\ \gamma x_3 - \mu\sigma x_2 \\ -\gamma x_3 \end{bmatrix}.$$

Note that the disease-free equilibrium of our model is $\mathbf{x}^* = (N, 0, 0, 0)^\top$. Let $\mathbf{F}$ and $\mathbf{V}$ be the partial Jacobian matrices of functions $F(\mathbf{x})$ and $V(\mathbf{x})$ with respect to the number of individuals in the "infective" compartments (both E and I compartments in the SuEIR model), i.e., x_2 and x_3,

$$\mathbf{F} = \begin{bmatrix} \frac{\partial F_2(\mathbf{x}^*)}{\partial x_2} & \frac{\partial F_2(\mathbf{x}^*)}{\partial x_3} \\ \frac{\partial F_3(\mathbf{x}^*)}{\partial x_2} & \frac{\partial F_3(\mathbf{x}^*)}{\partial x_3} \end{bmatrix} = \begin{bmatrix} \beta & \beta \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{V} = \begin{bmatrix} \frac{\partial V_2(\mathbf{x}^*)}{\partial x_2} & \frac{\partial V_2(\mathbf{x}^*)}{\partial x_3} \\ \frac{\partial V_3(\mathbf{x}^*)}{\partial x_2} & \frac{\partial V_3(\mathbf{x}^*)}{\partial x_3} \end{bmatrix} = \begin{bmatrix} \sigma & 0 \\ -\mu\sigma & \gamma \end{bmatrix}.$$

Then the next-generation matrix $\mathbf{G} = \mathbf{F}\mathbf{V}^{-1}$ can be computed as follows:

$$\mathbf{G} = \mathbf{F}\mathbf{V}^{-1} = \begin{bmatrix} \frac{\beta}{\sigma} + \frac{\beta\mu}{\gamma} & \frac{\beta}{\gamma} \\ 0 & 0 \end{bmatrix}.$$

Note that $\mathcal{R}_0$ is given by the largest eigenvalue of next generation matrix $\mathbf{G}$ (Heffernan et al., 2005). Therefore, it is easy to show that the basic reproduction number of our proposed SuEIR model is

$$\mathcal{R}_0 = \frac{\beta}{\sigma} + \frac{\beta\mu}{\gamma}. \tag{4}$$

In contrast, the basic reproduction number for SIR and SEIR is $\mathcal{R}_0 = \beta/\gamma$.

A.4 COMPUTATION OF THE EFFECTIVE REPRODUCTION NUMBER $\mathcal{R}_t$

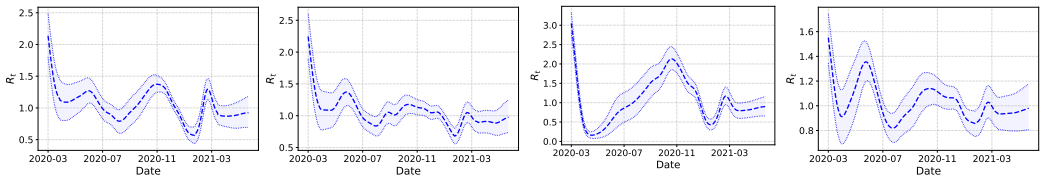
We compute the effective reproduction number $\mathcal{R}_t$ using Bayesian estimation. More specifically, we assume the Poisson distribution for new infectious cases ΔI_{t+1} at time $t+1$ as follows

$$\mathbb{P}\big(\Delta I_{t+1} = k | \lambda_{t+1}\big) = \frac{\lambda_{t+1}^k \exp(-\lambda_{t+1})}{k!},$$

where λ_{t+1} is the expected number of new infectious cases at time $t+1$ we believe we would observe. In addition, λ_{t+1} can be estimated using $\mathcal{R}_t$ and ΔI_t (Bettencourt & Ribeiro, 2008) as

Table 1: Estimated basic reproduction number $\mathcal{R}_0$ with a 95% confidence interval.

Region	US	NY	NJ	IL	MA	CA	PA	MI	FL	MD
$\mathcal{R}_0$	2.7 [2.3, 3.9]	3.5 [2.0, 5.1]	4.4 [3.7, 5.8]	3.6 [3.0, 4.7]	4.2 [3.6, 5.5]	2.2 [1.9, 2.6]	3.3 [3.0, 4.4]	2.1 [1.8, 2.7]	2.3 [2.0, 3.2]	2.9 [2.4, 3.9]



(a) California (b) Texas (c) New York (d) Florida

Figure 5: R_t curves of US, California state, Texas state, and Florida state.

$\lambda_{t+1} = \Delta I_t \exp\big(\tau(\mathcal{R}_t - 1)\big)$, where $1/\tau$ denotes the average generation time of the COVID-19. Therefore, according to the Bayesian rule, we have

$$\mathbb{P}\big(\mathcal{R}_t|\Delta I_{t+1}, \Delta I_t\big) = \frac{\mathbb{P}\big(\Delta I_{t+1}|\mathcal{R}_t, \Delta I_t\big)\mathbb{P}\big(\mathcal{R}_t|\Delta I_t\big)}{\mathbb{P}\big(\Delta I_{t+1}|\Delta I_t\big)} \propto \mathbb{P}\big(\Delta I_{t+1}|\lambda_{t+1}\big)\mathbb{P}\big(\mathcal{R}_t\big).$$

By assuming the prior distribution $\mathbb{P}\big(\mathcal{R}_t\big)$ to be a normal distribution, we can sample posterior $\mathbb{P}\big(\mathcal{R}_t|\Delta I_{t+1}, \Delta I_t\big)$ using Markov Chain Monte Carlo (MCMC) method. In our estimation, we initialize $\mathcal{R}_t$ with $\mathcal{R}_0$ computed using our SuEIR model, and for the forecast of $\mathcal{R}_t$, we use the projection of our SuEIR model.

A.5 Modeling the resurgence of COVID-19 in the US

The resurgence of COVID-19 has been observed in many states in the US since mid June or earlier. This may be caused by the careless early reopening, which makes more population be exposed to the virus and increases the contact rate between susceptible and infected populations. Therefore, the SuEIR model parameters will no longer be invariant during the entire epidemic period. In order to model the resurgence as well as describe the change of the virus spread pattern, we consider training our SuEIR in a sequential manner. In particular, we split the entire time series into to periods based on the start date of the resurgence, and assume that the susceptible population will keep increasing in the second period due to the reopen. Then we train two models for these two periods of time sequentially, where the output of the first model (estimates of the populations in S and E compartments) will be fed into the second model as the input (initial guess of the populations in S and E). Finally, we assume that the current scenario will continue and generate the forecasts of deaths and confirmed cases using the second model.

B Experiments on the Reproduction Number

Table 1 summarizes the basic reproduction number $\mathcal{R}_0$ estimated by (4) in different regions, which characterizes the spread of the virus at the beginning of the epidemic. The results vary for different states, which are consistent with the severity of the coronavirus outbreak in these regions since mid March. For example, the $\mathcal{R}_0$ values of the states in the Northeastern US (e.g., NY: 3.5, NJ: 4.4, MA:4.2) are significantly higher than those of other states (e.g. CA: 2.2, MI: 2.1, FL: 2.3). We further plot the effective reproduction number ($\mathcal{R}_t$) curves of the California, Texas, New York and Florida in Figure 5 based on the method described in Section A.4. It can be seen that the $\mathcal{R}_t$ curves of California, Florida, and Texas fluctuate more frequently than that of New York, which is consistent with the fact that there are more resurgences occurring in California, Florida, and Texas.