Journal of Data Science 18 (3), 409–432
July 2020
DOI: 10.6339/JDS.202007_18(3).0003

An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China

Lili Wang[1], Yiwang Zhou[1], Jie He[1], Bin Zhu[2], Fei Wang[3], Lu Tang[4], Michael Kleinsasser[1], Daniel Barker[1], Marisa C. Eisenberg[5], and Peter X.K. Song*[1]

[1]*Department of Biostatistics, University of Michigan, Ann Arbor, MI*
[2]*Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD*
[3]*Data Science Team, CarGurus, Cambridge, MA*
[4]*Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA*
[5]*Department of Epidemiology, University of Michigan, Ann Arbor, MI*

Abstract

We develop a health informatics toolbox that enables timely analysis and evaluation of the time-course dynamics of a range of infectious disease epidemics. As a case study, we examine the novel coronavirus (COVID-19) epidemic using the publicly available data from the China CDC. This toolbox is built upon a hierarchical epidemiological model in which two observed time series of daily proportions of infected and removed cases are generated from the underlying infection dynamics governed by a Markov Susceptible-Infectious-Removed (SIR) infectious disease process. We extend the SIR model to incorporate various types of time-varying quarantine protocols, including government-level 'macro' isolation policies and community-level 'micro' social distancing (e.g. self-isolation and self-quarantine) measures. We develop a calibration procedure for under-reported infected cases. This toolbox provides forecasts, in both online and offline forms, as well as simulating the overall dynamics of the epidemic. An R software package is made available for the public, and examples on the use of this software are illustrated. Some possible extensions of our novel epidemiological models are discussed.

Keywords *coronavirus; Infectious disease; MCMC; prediction; Runga–Kutta approximation; SIR model; turning point; under-reporting.*

1 Introduction

The outbreak of the coronavirus disease 2019 or COVID-19, originated in Wuhan, the capital city of Hubei province. From there, it spread quickly through Hubei and then to China and globally to more than 150 countries according to the WHO data available on March 2020. As of February 25, 2020, in China this large-scale epidemic, since been classified by the World Health Organization as a pandemic, has caused a total of 78,195 confirmed infections, 2,718 deaths and 30,078 recovered cases, and additionally 2,491 suspected cases still remained to be tested. Since the outbreak of the epidemic, many clinical papers (Jung et al., 2020; Chen et al., 2020; Xiang et al., 2020; Xu et al., 2020; Imai et al., 2020; Gralinski and Menachery, 2020; Luk et al., 2019; Fan et al., 2019; Hui et al., 2020; Holshue et al., 2020; Guan et al., 2020; Rothe et al., 2020; Huang et al., 2020; Zhu et al., 2020; Wang et al., 2020a) have been published to uncover limited but important knowledge of COVID-19, including that (i) COVID-19 is an infectious disease caused by SARS-CoV-2, a virus closely related to the SARS coronavirus (SARS-CoV)

*Corresponding author. Email: pxsong@umich.edu.

Received March, 2020; Accepted April, 2020

(Luk et al., 2019; Fan et al., 2019; Subissi et al., 2014); (ii) it can spread from person to person, primarily via droplet transmission (Hui et al., 2020; Holshue et al., 2020); (iii) it has a relatively high person-to-person transmission rate, especially via close contact; (iv) the median incubation time is approximately 5 days (Lauer et al., 2020), which can be as long as 24 days (Guan et al., 2020); and (v) asymptomatic person carrying SARS-CoV-2 is contagious (Rothe et al., 2020). This epidemic has been concerning not only in China but also in the rest of the world given the currently fast growing number of infected cases in South Korea, Japan, Iran, and Italy.

Quarantine or medical isolation is a key non-pharmaceutical intervention approach to stop the spreading of infectious diseases such as SARS (World Health Organization, 2020; Smith, 2006; World Health Organization, 2003) and plague (Dennis et al., 1999). The basic idea of quarantine and isolation is to separate infected cases from the susceptible population and *vice versa*. Since mid-January 2020, the Chinese government has implemented various kinds of very strict in-home isolation protocols nationwide, which have been elevated day by day through various government enforced quarantine and societally organized inspections to control the spread of COVID-19 in Hubei and other regions in China. In the meantime, the Chinese government has quickly increased the capacity of hospitals or as such that took symptomatic patients to be quarantined and treated by medical doctors and nurses.

The question of the most importance, which draws most attention, concerns when the spread of COVID-19 will end. This question has to be answered by a prediction model using the daily surveillance data from the China CDC. Unfortunately, it is extremely difficult to make accurate and precise predictions due to the limited amount of available data, which are thought to be inaccurate due to the issue of under-reporting. Additionally, predicting the peak or end of an epidemic during the exponential growth phase is well known to be highly challenging, and in many cases even potentially impossible due to parameter unidentifiability issues (Nishiura et al., 2017; Weitz and Dushoff, 2015; Kao and Eisenberg, 2018). Many prediction models (Sun et al., 2020; Li et al., 2020b; Hu et al., 2020; Rabajante, 2020; Peng et al., 2020; Zhang et al., 2020; Liu et al., 2020) have already been proposed to provide good fitting results for the publicly available data that may be potentially under-reported. Each of these models may result in different predictions of turning points, such as the dates when the daily increased or the total number of infections begin to decrease. Since such forecasting needs to extrapolate a fitted model to a relatively distant future time after the last date with observed data, whichever the chosen model is used, the model itself will dictate prediction results. In addition, data accuracy, in particular the issue of under-reporting, may cause bias in prediction, and ignoring this issue can lead to incorrect prediction of turning points. The issue of under-reporting may be attributed to the unsatisfactory sensitivity of the PCR test for SARS-CoV-2 or to the lack of enough kits for testing at the beginning of the outbreak, among other logistic and political reasons. The Chinese government tried to correct some of these issues by using a new diagnostic protocol based on clinical symptoms starting at the first week of February. However, it undermines the quality of data collected in the early phase of the epidemic.

All the above points illustrate the complexity of the impact of human interventions on the spread of COVID-19, including but not limited to in-home quarantine, hospitalization, community enforcement of wearing masks, minimizing outdoor activities, and changed diagnostic criteria by the government. The prediction model must take such features into account in order to provide meaningful analyses and hopefully reasonable predictions. However, most existing prediction models do not have the capacity to incorporate changing interventions in reality, and with no such critical component of time-varying intervention in the model, predicted turning points would be untrustworthy. Our new model and analytic toolbox aims to fill in this significant gap.

We develop an R package `eSIR` (Wang et al., 2020b) for R (R Core Team, 2020), that helps accomplish the following specific aims:

1 Utilize and calibrate publicly available data to understand the epidemiological trend of COVID-19 spread in Hubei province and the other regions of China.
2 Incorporate time-varying quarantine protocols in the model of COVID-19 infection dynamics via an extension of the classical epidemiological SIR model. This dynamic infection system necessitates the forecast of the future trend of COVID-19 spread.
3 Provide an R software package to health workers who can conveniently perform their own analyses using their substantive knowledge and perhaps better quality data from provinces in China or from other countries.

We hope to provide a data analytic toolbox to people who may have better domain-specific knowledge and experience as well as high quality data to carry out independent predictions.

Our informatics toolbox is built upon a state-space model (Zhu et al., 2012; Jørgensen et al., 1999; Song, 2000; Jøsrgensen and Song, 2007) shown in Figure 1 with an extended Markov SIR model (Kermack and McKendrick, 1927), which has the following key features: (i) Our model is specified with the temporally varying probabilities of susceptible, infected and removed (recovered and death) compartments, not directly on time series of respective counts; (ii) estimation and inference are carried out and implemented using Markov Chain Monte Carlo (MCMC); (iii) it outputs various sample draws from the posteriors of the model parameters (e.g. transmission and removal rates) and the underlying probabilities of susceptible, infected and removed compartments, as well as their credible intervals. The latter is of extreme importance to quantify prediction uncertainty. In addition, this toolbox provides predicted turning points, including (i) the date when daily increased number of infections begins to decrease or the time at which the second order derivative of the prevalence of infected compartment is zero (i.e. the turning point of infection acceleration to deceleration); and (ii) the date when daily number of removed cases is larger than that of infected cases, or the time at which the first derivative of the prevalence of infected compartment is zero (i.e. the turning point of zero infection speed). As a byproduct, the method also provides a predicted time when the COVID-19 epidemic ends.

This paper is organized as follows. Section 2 presents our new epidemiological forecast model incorporating time-varying quarantine protocols. Section 3 concerns the algorithmic implementation via Markov Chain Monte Carlo and a deliverable R software. Section 4 is devoted to the analysis of COVID-19 data within and outside Hubei, where a calibration of under-reporting is proposed. Section 5 gives some concluding remarks, and some technical details are included in the appendices.

2 State-space SIR Epidemiological Model

2.1 Basic model of coronavirus infection

We begin with a basic epidemiological model in the framework of state-space SIR models with no consideration of quarantine protocols. This framework was proposed by Osthus et al. (2017) with only one-dimensional time series of infected proportions. Refer to Chapter 9-12 of Song (2007) for an introduction to state-space models. Here we consider two time series of proportions of infected and removed cases, denoted by Y_t^I and Y_t^R at time t, respectively, where the compartment of removed R is a sum of the proportions of recovered cases and deaths at time t. We assume that the 2-dimensional time series of $(Y_t^I, Y_t^R)^\top$ follows a state-space model with the beta distributions

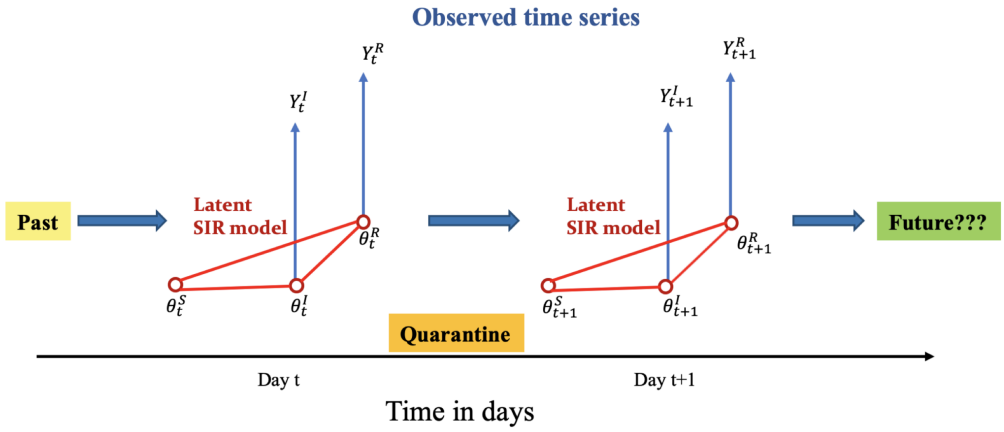

Figure 1: A conceptual framework of the proposed epidemiological state-space SIR model.

at time t:

$$Y_t^I|\boldsymbol{\theta}_t, \lambda^I \sim \text{Beta}(\lambda^I\theta_t^I, \lambda^I(1-\theta_t^I)), \tag{1}$$

$$Y_t^R|\boldsymbol{\theta}_t, \lambda^R \sim \text{Beta}(\lambda^R\theta_t^I, \lambda^R(1-\theta_t^I)), \tag{2}$$

where θ_t^I and θ_t^R are the respective probabilities of infection and removal at time t, and λ^I and λ^R are the parameters controlling the respective variances of the observed proportions (noting that the superscripts here indicate labels rather than exponents).

As shown in Figure 1, these observed time series are emitted from the underlying latent dynamics of COVID-19 infection characterized by the latent Markov process $\boldsymbol{\theta}_t$. It is easy to see that the expected proportions in both Equations (1) and (2) are equal to the prevalence of infection and the probability of removal at time t, namely $E(Y_t^I|\boldsymbol{\theta}_t) = \theta_t^I$ and $E(Y_t^R|\boldsymbol{\theta}_t) = \theta_t^R$. See Appendix B. Moreover, the latent population prevalence $\boldsymbol{\theta}_t = (\theta_t^S, \theta_t^I, \theta_t^R)^\top$ is a three-dimensional Markov process, in which θ_t^S is the probability of a person being susceptible or at risk at time t, θ_t^I is the probability of a person being infected at time t, and θ_t^R is the probability of a person being removed from the infectious system (either recovered or dead) at time t. Obviously, $\theta_t^S + \theta_t^I + \theta_t^R = 1$. We assume that this 3-dimensional probability process $\boldsymbol{\theta}_t$ is governed by the following Markov model:

$$\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\tau} \sim \text{Dirichlet}(\kappa f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)), \tag{3}$$

where parameter κ scales the variance of the Dirichlet distribution and function $f(\cdot)$ is a 3-dimensional vector that determines the mean of the Dirichlet distribution. We let all the relevant parameters be $\boldsymbol{\tau} = (\beta, \gamma, \kappa, \boldsymbol{\theta}_0, \lambda^I, \lambda^R)^\top$, where β and γ denote the transmission and removal rates of the SIR model given in (4), and $\boldsymbol{\theta}_0 = (\theta_0^S, \theta_0^I, \theta_0^R)$ are initial probabilities of the three compartments. The function f is the engine of the infection dynamics which operates according to SIR model of the form:

$$\frac{d\theta_t^S}{dt} = -\beta\theta_t^S\theta_t^I, \quad \frac{d\theta_t^I}{dt} = \beta\theta_t^S\theta_t^I - \gamma\theta_t^I, \quad \text{and} \quad \frac{d\theta_t^R}{dt} = \gamma\theta_t^I. \tag{4}$$

The ratio between the transmission and removal rates is the basic reproduction number $R_0 = \beta/\gamma$ which measures contagiousness or transmissibility of infectious agents. It provides the average secondary cases generated from one infected case when the whole population is susceptible (Fraser et al., 2009; Delamater et al., 2019). Note that the explicit solution to the above system (4) of ordinary differential equations is unavailable. Following Osthus et al. (2017), we invoke the fourth-order Runge–Kutta (RK4) approximation, resulting in an approximate of $f(\theta_{t-1}, \beta, \gamma)$ as follows:

$$f(\boldsymbol{\theta}_{t-1}, \beta, \gamma) = \begin{pmatrix} \theta^S_{t-1} + 1/6[k^{S_1}_{t-1} + 2k^{S_2}_{t-1} + 2k^{S_3}_{t-1} + k^{S_4}_{t-1}] \\ \theta^I_{t-1} + 1/6[k^{I_1}_{t-1} + 2k^{I_2}_{t-1} + 2k^{I_3}_{t-1} + k^{I_4}_{t-1}] \\ \theta^R_{t-1} + 1/6[k^{R_1}_{t-1} + 2k^{R_2}_{t-1} + 2k^{R_3}_{t-1} + k^{R_4}_{t-1}] \end{pmatrix},$$

where all these k_t terms are given in the appendix A. The set of model parameters $\boldsymbol{\tau}$ will be estimated using the MCMC method (Carlin et al., 1992).

2.2 Epidemiological model with time-varying transmission rate

The basic epidemiological model with both constant transmission and removal rates in the SIR model (4) does not reflect the reality in China, where various levels of quarantines have been enforced. Many forms of human interventions that are altering the transmission rate over time include (i) individual-level protective measures such as wearing masks and safety glasses, using hygiene, and taking in-home isolation, and (ii) community-level quarantines such as hospitalization for infected cases, city blockade, traffic control and restricted social activities, and so on. In addition, the virus itself may mutate to evolve, which may increase the potential rate of false negative in the disease diagnosis. As a result, some individual virus carriers are not captured. Thus, the transmission rate β indeed varies over time, which should be accounted in the modeling.

One extension to the above basic epidemiological model is to allow a time-varying probability that a susceptible person meets an infected person or *vice versa*. Suppose at a time t, $q^S(t) \in [0, 1]$ is the chance of an at-risk person being in-home isolation, and $q^I(t) \in [0, 1]$ is the chance of an infected person being in-hospital quarantine. Thus, the chance of disease transmission when an at-risk person meets an infected person is modified as:

$$\beta\{1 - q^S(t)\}\theta^S_t\{1 - q^I(t)\}\theta^I_t := \beta\pi(t)\theta^S_t\theta^I_t,$$

with $\pi(t) := \{1-q^S(t)\}\{1-q^I(t)\} \in [0, 1]$. In effect, this $\pi(t)$ modifies the chance of a susceptible person meeting with an infected person or *vice versa*, which is termed as a *transmission modifier* due to quarantine in this paper. Obviously, with no quarantine in place, $\pi(t) \equiv 1$ for all time. See Figure 2 Panel A. This results in a new SIR model with a time-varying transmission rate modifier:

$$\frac{d\theta^S_t}{dt} = -\beta\pi(t)\theta^S_t\theta^I_t, \quad \frac{d\theta^I_t}{dt} = \beta\pi(t)\theta^S_t\theta^I_t - \gamma\theta^I_t, \quad \text{and } \frac{d\theta^R_t}{dt} = \gamma\theta^I_t, \tag{5}$$

where the product term $\beta\pi(t)$ defines an effective transmission rate reflective to the currently enforced quarantine measures of all levels in China. Note that the above extended SIR model assumes primarily that both population-level chance of being susceptible and population-level chance of being infected remain the same, but the chance of a susceptible person meeting with an infected person is reduced by $\pi(t)$.

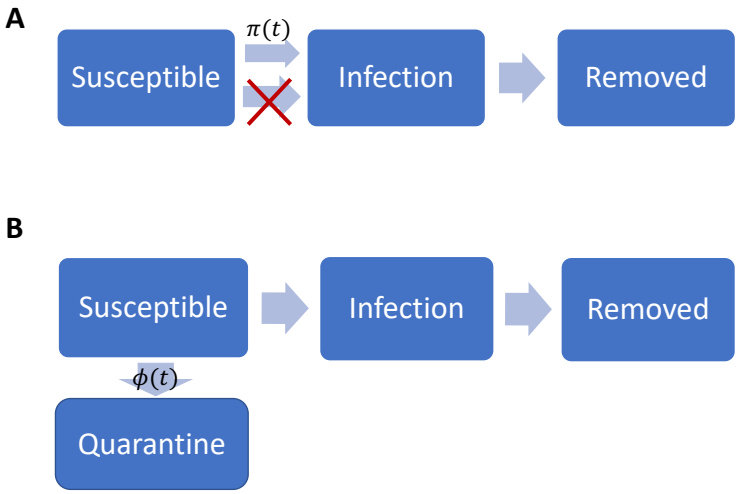

Figure 2: Extended SIR models with a time-varying transmission rate modifier $\pi(t)$ (Panel A) or a time-varying quarantine rate $\phi(t)$ (Panel B).

The transmission rate modifier $\pi(t)$ needs to be specified according to actual quarantine protocols in a given region. A possible choice of $\pi(t)$ may be a step function that reflects government-initiated macro isolation measures in Wuhan, Hubei province:

$$\pi(t) = \begin{cases} \pi_{01}, & \text{if } t \leq \text{Jan 23, no concrete quarantine protocols;} \\ \pi_{02}, & \text{if } t \in (\text{Jan 23, Feb 4], city blockade;} \\ \pi_{03}, & \text{if } t \in (\text{Feb 4, Feb 8], enhanced quarantine;} \\ \pi_{04}, & \text{if } t > \text{Feb 8, opening of new hospitals.} \end{cases}$$

When $\boldsymbol{\pi}_0 = (\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04})$ are chosen with different values, as shown in Figure 3 Panels A-C, we obtain different types of transmission rate modifiers aligned with different quarantine protocols.

Alternatively, the modifier $\pi(t)$ may be specified as a continuous function that reflects steadily increased community-level awareness and responsibility of voluntary quarantine and preventive measures, which may be regarded as a kind of micro isolation measure initiated by individuals or local self-organized inspections. For example, as shown in Figure 3 Panels D-F, we may choose the following exponential functions:

$$\pi(t) = \exp(-\lambda_0 t) \text{ or } \pi(t) = \exp\{-(\lambda_0 t)^\nu\}, \lambda_0 > 0, \nu > 0.$$

Technically, the RK's approximate of f function in Appendix A may be easily obtained by replacing β by $\beta\pi(t)$ in the specification of the latent prevalence model (3), and moreover in all quantities and steps in the MCMC implementation. See the detailed in Section 3.

2.3 Epidemiological model with quarantine compartment

An alternative way to incorporate quarantine measures into the basic epidemiological model (4) is to add a new quarantine compartment that collects quarantined individuals who would have

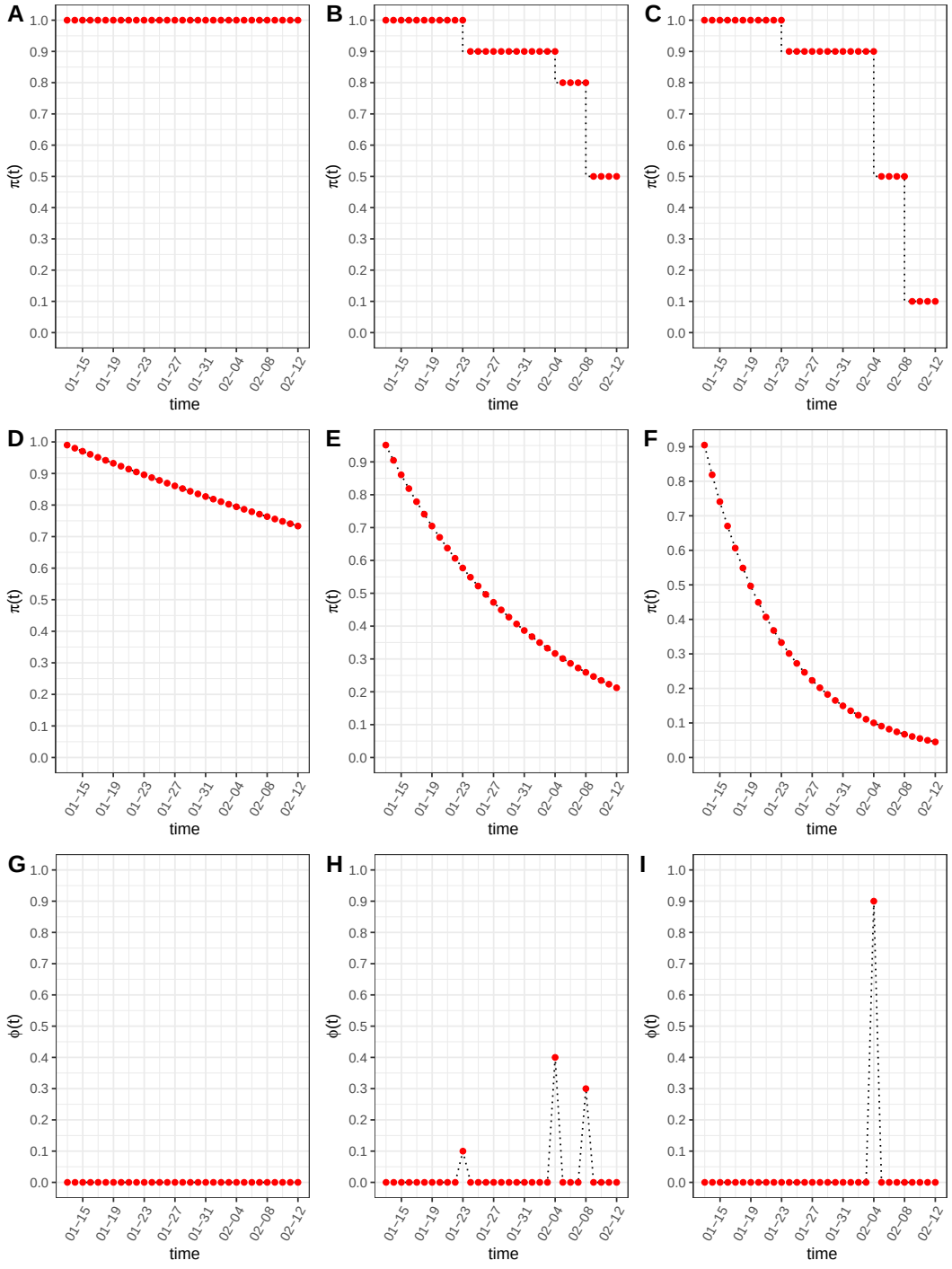

Figure 3: The functional forms of the transmission rate modifiers $\pi(t)$ and the quarantine rate $\phi(t)$: 1) Panels A-C depict step functions with $\boldsymbol{\pi}_0 = (\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04})$ equal to $(1, 1, 1, 1)$, $(1, 0.9, 0.8, 0.5)$ and $(1, 0.9, 0.5, 0.1)$ at change points (Jan 23, Feb 4, Feb 8), Panels D-F depict exponential functions under difference micro quarantine measures over time with $\lambda_0 = 0.01$, $\lambda_0 = 0.05$ and $\lambda_0 = 0.1$, and 3) Panels G-I depict multi-point instantaneous quarantine rates with $\boldsymbol{\phi}_0 = (0, 0, 0, 0)$, $\boldsymbol{\phi}_0 = (0.1, 0.4, 0.3)$ and $\boldsymbol{\phi}_0 = (0, 0.9, 0)$ at change points of (Jan 23, Feb 4, Feb 8).

no chance of meeting any infected individuals in the infection system, as shown in Figure 2 Panel B. This model allows to characterize time-varying proportions of susceptible cases due largely to the government-enforced stringent in-home isolation outside of Hubei province. The basic SIR model in equation (4) is then extended by adding a quarantine compartment with a time-varying rate of quarantine $\phi(t)$, which is the chance of a susceptible person being willing to take in-home isolation at time t. The extended SIR takes the following 4-dimensional latent process $(\theta_t^S, \theta_t^Q, \theta_t^I, \theta_t^R)^\top$:

$$\begin{aligned} \frac{d\theta_t^Q}{dt} &= \phi(t)\theta_t^S, \ \frac{d\theta_t^S}{dt} = -\beta\theta_t^S\theta_t^I - \phi(t)\theta_t^S, \\ \frac{d\theta_t^I}{dt} &= \beta\theta_t^S\theta_t^I - \gamma\theta_t^I, \ \frac{d\theta_t^R}{dt} = \gamma\theta_t^I, \end{aligned} \tag{6}$$

where $\theta_t^S + \theta_t^Q + \theta_t^I + \theta_t^R = 1$.

We suppose that the quarantine rate $\phi(t)$ is a Dirac delta function with jumps at times when major macro quarantine measures are enforced. For example, we may specify the $\phi(t)$ function as follows:

$$\phi(t) = \begin{cases} \phi_{01}, & \text{if } t = \text{Jan 23, city blockade;} \\ \phi_{02}, & \text{if } t = \text{Feb 4, enhanced quarantine;} \\ \phi_{03}, & \text{if } t = \text{Feb 8, opening of new hospitals;} \\ 0, & \text{otherwise.} \end{cases}$$

Here we show several examples of multi-point instantaneous quarantine rates in Figure 3 Panels G-H with jump sizes equal to $\boldsymbol{\phi}_0 = (\phi_{01}, \phi_{02}, \phi_{03})$ that occur respectively at dates of (Jan 23, Feb 4, Feb 8). In particular, we plot three scenarios, e.g., no intervention ($\boldsymbol{\phi}_0 = (0, 0, 0)$), multiple moderate jumps ($\boldsymbol{\phi}_0 = (0.1, 0.4, 0.3)$), and only one large jump ($\boldsymbol{\phi}_0 = (0, 0.9, 0)$). Note that at each jump, the respective proportion of the susceptible population would move to the quarantine compartment. For example, with $\boldsymbol{\phi}_0 = (0.1, 0.4, 0.3)$, the quarantine compartment will be enlarged accumulatively over three time points as $0.1\,\theta_{t_1}^S + 0.4\theta_{t_2}^S + 0.3\theta_{t_3}^S$.

The $f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)$ function determined by the above extended SIR model (6) can be solved by applying the fourth-order Runge-Kutta approximation, and the resulting solution is given in Appendix A. To deal with the Dirac delta function $\phi(t)$, we develop a two-step approximation for model (6). In brief, we first solve a continuous function without change points via the differential equations in (5), and then we directly move the mass of $\phi(t)\theta_t^S$ out of the susceptible compartment to the quarantine compartment. From our experience, this approach largely improves the approximation accuracy in the presence of discontinuities.

3 Implementation: Markov Chain Monte Carlo Algorithm

3.1 MCMC Algorithm

We implemented the MCMC algorithm to collect draws from the posterior distributions, and further obtain posterior estimates and credible intervals of the unknown parameters in the above models specified in Section 2. Because of the hierarchical structure in the state-space model considered in this paper, the posterior distributions can be obtained straightforwardly. The R package `rjags` (Plummer, 2019) can be directly applied to draw samples from the posterior distributions, so we omit the technical details. The latent Markov processes $\boldsymbol{\theta}_t$ are sampled and

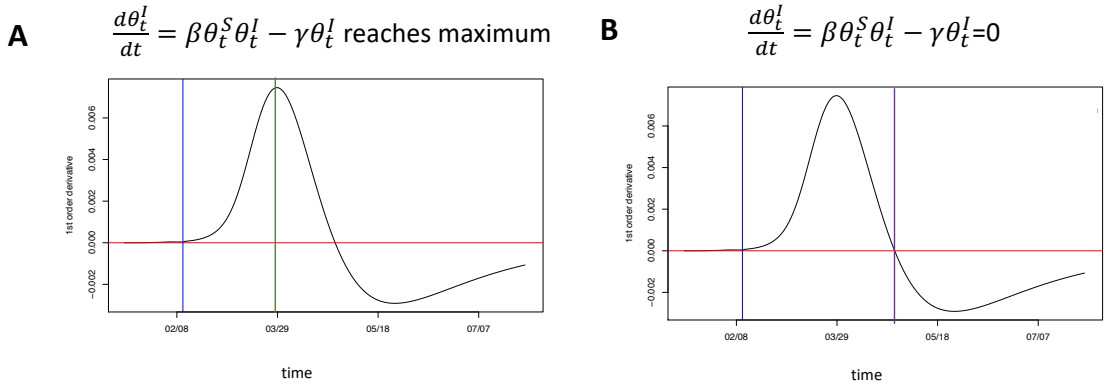

Figure 4: The first turning point in Panel A is marked by a green line, denoting the time when the estimated first-order derivative of the prevalence of infection reaches the maximum. The second turning point in Panel B is marked by a purple line, which is the time when the estimated first-order derivative of the prevalence of infection equals to zero. The vertical blue line labels the last observation day.

forecasted by the MCMC sampler, particularly for the probabilities of infection and removal, θ_t^I and θ_t^R, which enables us to determine the turning points of interest and the reproduction number R_0.

The first turning point of interest is the time when the daily number of new infected cases stops increasing. Mathematically, this corresponds to the time t at which $\ddot{\theta}_t^I = 0$ or the gradient of $\dot{\theta}_t^I$ is zero. As illustrated by Panel A in Figure 4, the peak of $\dot{\theta}_t^I$, denoted by the vertical green line, is the first turning point of interest. The second turning point is the time when the cumulative infected cases reaches its maximum, meaning $\dot{\theta}_t^I = 0$. In principle, the second turning point occurs only after the first one, as shown in Panel B in Figure 4.

The basic reproduction number R_0 of an infectious disease is estimated by the ratio $R_0 = \beta/\gamma$, where β and γ are both estimated from their posterior distributions. Because our models consider the quarantine compartment, R_0 might change according to the forms of quarantine protocols. We adopt a standard MCMC algorithm to draw samples of the latent process. Let t_0 be the current time up to which we have observed data $(Y_{0:t_0}^I, Y_{0:t_0}^R)$. To perform M draws of Y_t^I, Y_t^R for $t \in [t_0+1, T]$, we proceed as follows: for each $m = 1, \ldots, M$,

(1) draw $\boldsymbol{\theta}_t^{(m)}$ from the posterior $[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(m)}, \boldsymbol{\tau}^{(m)}]$ of the prevalence process, at $t = t_0+1, \ldots, T$;

(2) draw $(Y_t^{I(m)}, Y_t^{R(m)})$ from $[Y_t^I|\boldsymbol{\theta}_t^{(m)}, \boldsymbol{\tau}^{(m)}]$ and $[Y_t^R|\boldsymbol{\theta}_t^{(m)}, \boldsymbol{\tau}^{(m)}]$ according to the observed process, at $t = t_0+1, \ldots, T$, respectively;

The prior distributions are specified with some of the hyper-parameters being set according to the SARS data from Hong Kong (Mkhatshwa and Mummert, 2010). They are,

$$\begin{aligned}
&\boldsymbol{\theta}_0 \sim \text{Dirichlet}(1 - Y_1^I - Y_1^R, Y_1^I, Y_1^R)\\
&R_0 \sim \text{LogN}(1.099, 0.096) \text{ with } \text{E}(R_0) = 3.15, \text{SD}(R_0) = 1;\\
&\gamma \sim \text{LogN}(-2.955, 0.910) \text{ with } \text{E}(\gamma) = 0.0821, \text{SD}(\gamma) = 0.1,\ \beta = R_0\gamma;\\
&\kappa \sim \text{Gamma}(2, 0.0001),\ \lambda^I \sim \text{Gamma}(2, 0.0001),\ \lambda^R \sim \text{Gamma}(2, 0.0001).
\end{aligned}$$

Note that LogN and Gamma stand for log-normal and gamma distributions respectively, and

E and SD represent mean and standard deviation here. In the default setting the variances of the above prior distributions are set at relatively large values to reflect the fact that limited prior knowledge of these epidemiological model parameters is available. When more information becomes accessible during the course of the epidemic, smaller prior variance values may be used, leading to tighter credible intervals for the model parameters and turning points.

3.2 R software package

We deliver an R software package that is able to output the MCMC estimation, inference and prediction under the epidemiological model with two proposed extended SIR models that incorporate time-varying quarantine protocols. These new models have been discussed in detail in Sections 2.2 and 2.3. Our R package, named `eSIR`, uses daily-updated time series of infected and removed proportions as input data. The R package is available at GitHub `lilywang1988/eSIR`, and its user manual is appended as the supplementary material of this paper. The quarantine functions are predefined and will be treated as known functions of protocols/policies in the estimation and prediction steps. We let the transmission rate modifier $\pi(t)$ be either a step function or an exponential function, and let the quarantine rate $\phi(t)$ follow a Dirac delta function with pre-specified points of jump and sizes of jumps. The package provides various plots for users to visualize the MCMC results, including the estimated prevalence of infection and the estimated probability of removal, and predicted turning points of interest. Various summary statistics are listed in the output, including posterior mean estimates of the transmission and removal rates, estimate of the reproduction number, and forecasts of turning points and their 95% credible intervals. Moreover, the package gives multiple options to users who can save the entire MCMC results, including the output tables and summary plots, Gelman-Rubin convergence statistic, traceplots for MCMC quality control, and full MCMC draws for user's own summary analyses. Some illustrations on the use of this software package are given in Section 4 with sample codes in Appendix C. In addition, we developed an online R Shiny App that can automatically update the results whenever the China CDC updates the daily COVID-19 data (Kleinsasser et al., 2020).

4 Analysis of the COVID-19 Data Within and Outside Hubei

4.1 Calibration of under-reported infection data

Under-reporting of infections is a common issue in the surveillance data collection of infectious disease, especially at the beginning of an outbreak. When medical diagnostic tools become more accurate and reliable, as well the compliance of preventive measures gets improved for an exchange of voluntary in-home quarantine, certain adjustments in data typically occur. It is shown in Figure 5 that on Feb 12 the cumulative and daily added number of infected cases in Hubei had clear jumps with significantly large sizes. Such sizable jumps cannot happen within one day, rather they represent an accumulation of cases that have not been reported in previous dates prior to Feb 12. To fix this under-reporting issue, we develop a calibration procedure with details given in Appendix D. Below we briefly describe our approach for the calibration of the infected cases.

We assume an exponential growth curve for the cumulative number of infected cases in Hubei before Feb 12 of the form $y(t) = ae^{\lambda t} + b$, where parameters λ, a, b are to be estimated. Under the boundary conditions $y(t = \text{Jan } 12) = 0$ and $y(t = \text{Feb } 12) = a\exp(31\lambda) + b$, we would like to minimize the one-step ahead extrapolation error on Feb 13. The constrained optimal solution

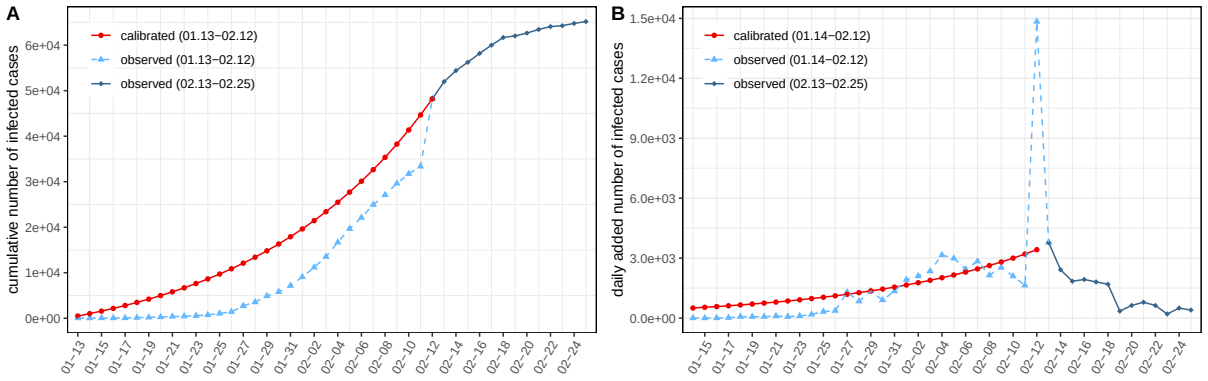

Figure 5: Under-reporting calibration of the infected cases in Hubei. (A) The cumulative number of infected cases. (B) The daily added number of infected cases. Data calibration is performed from Jan 13 to Feb 12 as is shown by the red curves.

can be obtained by the means of Lagrange Multipliers; the estimates are $\hat{\lambda} = 0.06605, \hat{a} = 7142.80, \hat{b} = -7142.80$. The resulting calibration curves for the cumulative and daily added number of infected cases are shown as the solid red curves in Figure 5. For example, on Jan 31, the reported cumulative number of infected cases is 7,153, but the calibrated number of infected cases is 17,911, with an increment of 10,758 cases. As shown later, this data quality control (QC) step helps improve the performance of MCMC. The exponential function in Figure 5 is a simple but good approximation to the supposedly continuous cumulative number of infections in the early phase of an infectious disease, which is used here to smooth backwards the abrupt jump on Feb 12 by assuming that such a sudden leap was due to the previous under-reporting.

4.2 Evaluation and prediction under time-varying quarantine

We applied our proposed models, algorithms and R package `eSIR` to analyze the COVID-19 data collected from the public website DXY.cn. The earliest public records for the provincial data are available since Jan 20, 2020. According to an existing R package on GitHub `GuangchuangYu/nCov2019` (Yu, 2020), the total counts of confirmed infections and deaths are dated back on Jan 13, 2020. We assumed that before Jan 17 all the reported cases and deaths were from Hubei. We imputed by the linear interpolation the missing cases on Jan 18-19. Therefore, the data used in our analyses starts from Jan 13. The data used in analyses for the other provinces starts on Jan 23, which is the earliest time with non-zero values in the removed compartment. Note that there exist some minor discrepancies between different data sources, and the under-reporting issue is addressed in Appendix D by a calibration procedure. This section aims to provide a demonstration of our software to analyze the current public COVID-19 data, through which users may understand the proposed methods. We will also elaborate ways to export and interpret the MCMC results. The R package may be applied to analyze infectious data from other countries.

First, we show the analysis of the calibrated Hubei COVID-19 data after introducing in a time-varying transmission rate modifier $\pi(t)$ using our R function `txt.eSIR` in the package `eSIR`. As described in Subsection 4.1, we partially corrected the under-reported proportion of infections in Hubei province prior to Feb 12, when a big jump occurred on one day. The corresponding results are shown in Figure 6, in comparison with the ones without data calibration in Web

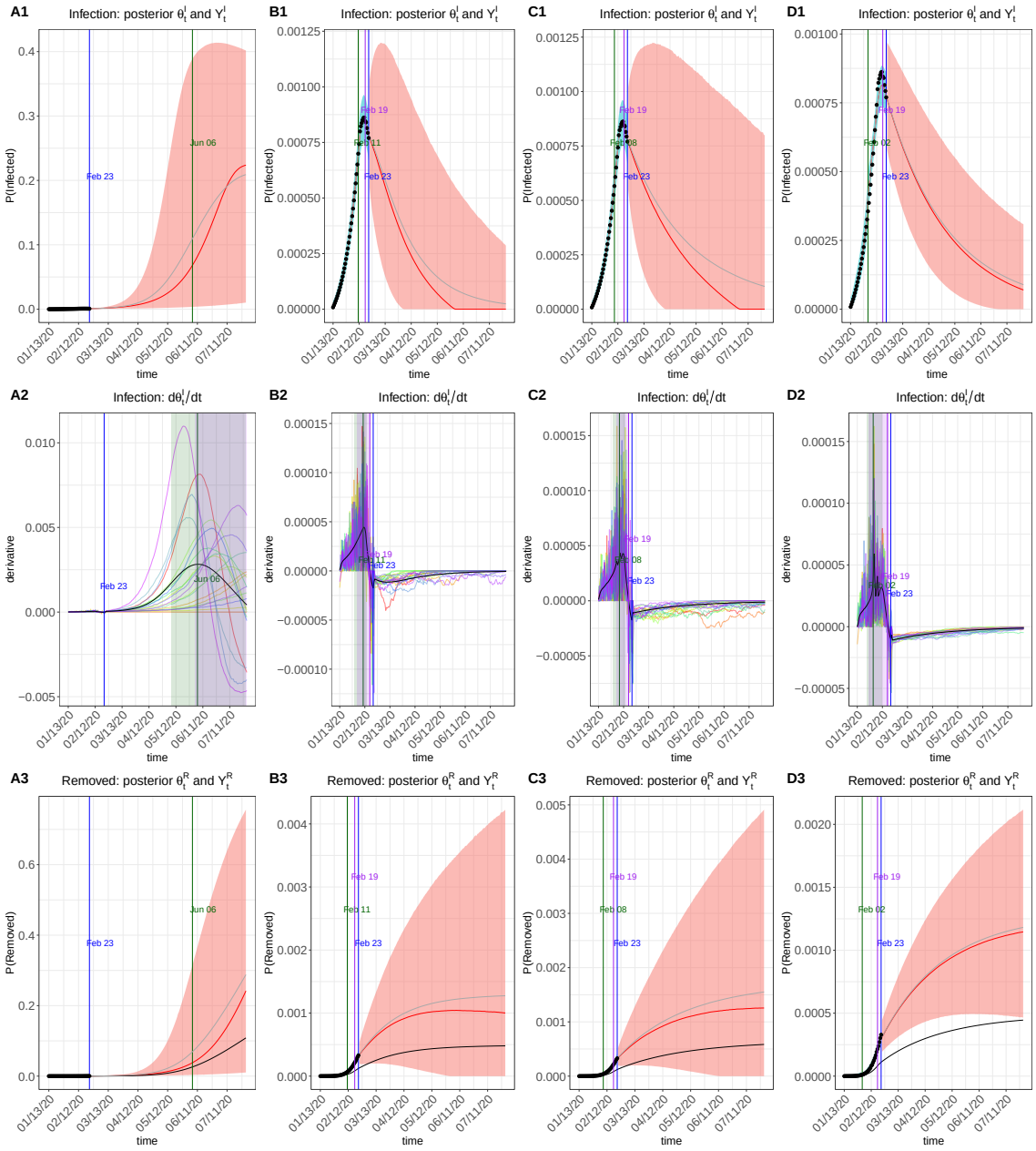

Figure 6: Prediction plots of θ_t^I and Y_t^I (Row 1), $\dot{\theta}_t^I$ (Row 2), θ_t^R and Y_t^R (Row 3) for Hubei after data calibration. Subfigures in Column A display the results of basic SIR model with $\pi(t) \equiv 1$ or $\phi(t) \equiv 0$, Subfigures in Column B display results of a continuous transmission modifier $\pi(t) = \exp(-0.05t)$, subfigures in Column C display results of a step-function transmission rate modifier with $\boldsymbol{\pi}_0 = (1, 0.9, 0.5, 0.1)$ at change points [Jan 23, Feb 4, Feb 8], and subfigures in Column D display results of a Dirac delta function quarantine process with $\boldsymbol{\phi}_0 = [0.1, 0.9, 0.5]$ at change points [Jan 23, Feb 4, Feb 8].

Figure 1. For both figures, Columns B-C denote a transmission rate following a step function with $\boldsymbol{\pi}_0 = c(1, 0.9, 0.5, 0.1)$ at change points [Jan 23, Feb 4, Feb 8] (Panel C of Figure 3) and an

exponential rate modifier with rate $\lambda_0 = 0.05$ (Panel E of Figure 3), as opposed to a basic model of $\pi(t) \equiv 1$ in Column A. Running R codes were given as Examples 1-3 in Appendix C. The forecast plots for infection and removal compartments are presented in Row 1 and Row 3 respectively, with all the black dots left to the blue vertical line denoting observed proportions by the last observational date. That is, the blue vertical marks time t_0 as defined in Section 3. The green and purple vertical lines denote the first and second turning points, respectively. The salmon color area denotes the 95% credible interval of the predicted proportions $[Y^I_{(t_0+1):T}|Y^I_{1:t_0}, Y^R_{1:t_0}]$ and $[Y^R_{(t_0+1):T}|Y^I_{1:t_0}, Y^R_{1:t_0}]$ after t_0, respectively, while the cyan color area represents either the 95% credible intervals of the prevalence $[\theta^I_{1:t_0}|Y^I_{1:t_0}, Y^R_{1:t_0}]$ or that of the probability of removal $[\theta^R_{1:t_0}|Y^I_{1:t_0}, Y^R_{1:t_0}]$ prior to time t_0. The gray and red curves are the posterior mean and median curves. The black curve in the removal compartment plots from Row 3 denotes the estimated proportion of deaths computed based on a pre-specified ratio (`death_in_R`). Row 2 provide a series of important dynamic features of the infection via a spaghetti plot, in which 20 randomly selected MCMC draws of the first-order derivative of the posterior prevalence of infection, namely $\dot{\theta}^I_t$. The black curve is the posterior mean of the derivative, and the vertical lines mark times of turning points corresponding respectively to those shown in Row 1 and Row 3. Moreover, the 95% credible intervals of these turning points are also highlighted by semi-transparent rectangles in Panel B and summarized in Web Table 1. In Subfigures A-C we displayed the results for time-dependent transmission rate modifiers. One can see that $\pi(t)$ plays an important roles in shortening the key turning points of the epidemic, and its effect on both estimation and prediction of the COVID-19 infection dynamics has been clearly demonstrated. It is also interesting to see that after data calibration, the abrupt rise in the infection proportion on Feb 12 in Web Figure 1 disappeared in Figure 6, and the observed data (i.e. the black dots) align better with the credible intervals of both latent processes.

Next, we analyzed the data from the rest of the Chinese population (i.e. the provinces outside Hubei) starting on Jan 23. We included two change points for the step function $\pi(t)$ at [Feb 4, Feb 8] with $\boldsymbol{\pi}_0 = (0.8, 0.1)$. The exponential function remained the same. It is noted that the spread of COVID-19 outside Hubei has been so far much less severe. Possible reasons for such low proportions of infection and deaths include (i) discontinuing the traffic connections between Hubei and the other provinces, (ii) more timely caution and preventative measures taken, and (iii) a comparatively less dense distribution of infection with respect to the huge population size. When Panel A1 in Web Figure 1 is zoomed in, some of the observed proportions (black dots) are deviated from the posterior mean or median of the fitted prevalence albeit they all fall in the 95% credible intervals, as shown by Panels B1 and C1 in Web Figure 2. Since the latent process follows the SIR differential equations, there may be a lack of fit for the SIR model to accommodate a very large and complex population of 1.3 billion people, in which most of the subjects are not at risk. The proposed models should work much better for individual provinces, but we did not perform such analyses.

The other epidemiological model with an added quarantine compartment as an absorbing state was fitted via our R function `qh.eSIR` in the package `eSIR`. We applied the proposed model in analyses of the data within and outside Hubei following Dirac delta functions with jumps of $\boldsymbol{\phi}_0 = [0.1, 0.9, 0.5]$ at change points [Jan 23, Feb 4, Feb 8] and $\boldsymbol{\phi}_0 = [0.9, 0, 5]$ at change points [Feb 4, Feb 8] respectively. Their results were summarized in Column D of Figure 6 and Web Figures 1-2. Their running codes were given as Examples 4-5 in Appendix C. Our analyses once again clearly indicated that stringent quarantine protocols can largely reduce the spread of COVID-19 both within Hubei and outside Hubei. Yet, it is known that too strict

Table 1: The posterior mean and credible intervals of the reproduction number R_0 obtained from different quarantine models and datasets.

	Within Hubei				Outside Hubei	
	Data Calibration		No Data Calibration			
Model	Mean	95%CI	Mean	95%CI	Mean	95%CI
No quarantine	3.02	[1.86, 4.56]	2.98	[1.90, 4.44]	2.56	[1.50, 4.22]
Exponential	4.82	[2.31, 8.38]	6.34	[2.82, 10.80]	3.16	[1.80, 5.06]
Step-function	4.32	[2.32, 6.94]	4.61	[2.12, 8.16]	2.90	[1.65, 4.76]
Quar. Compart.	4.95	[2.26, 9.25]	4.14	[1.96, 8.08]	3.37	[1.77, 5.73]

quarantine can backfire; people may lose their trust and patience in their committed system, and consequently may try to reduce compliance or even avoid quarantine. We also present the posterior mean probability of staying quarantine compartment in Web Figure 4 within Hubei and outside Hubei. Note that Jan 23 was not set as a change point for the cases outside Hubei, leading only to two jumps. It is evident that by Feb 8, more than 90% of the Chinese population have taken in-home isolation or as such, reflective to a very strict quarantine protocol enforced in the entire country.

The reproduction numbers estimated from different models for within and outside Hubei, with and without the data calibration, together with their 95% credible intervals are summarized in Table 1. It is worth pointing out that the estimates of the basic reproduction numbers obtained from the epidemiological models with time-varying transmission or quarantine rates appear larger than those obtained from the basic model with no quarantine. This is not surprising as our new models explicitly incorporate interventions, so that the estimated R_0 is an adjusted number with the influence of interventions be removed. In contrast, the basic model with no use of the quarantine modifier implicitly integrates the effect of interventions into the transmission rate β, and consequently the estimated R_0 is reduced due to the contribution from interventions. Our analyses suggest that reproduction numbers R_0 of COVID-19 without public health interventions would be around 4-5 within Hubei and around 3-3.5 outside Hubei with relatively big credible intervals. These findings are in agreement with findings from (Li et al., 2020a). We also notice that after the data calibration, the estimated reproduction numbers R_0 became less sensitive towards the intervention assumptions. As pointed out above, the size of credible interval may be reduced with more accessible data of COVID-19, which permits users to specify smaller variances in the prior distributions given in Section 3.1.

Since the turning points in China have been observed by Feb 23, there is an increasing concern about whether and when there would be a second outbreak. We conducted another set of analyses on Hubei calibrated data to forecast the epidemic trends when strict intervention may not last long. We focused on different degrees of relaxation on the intervention. In particular, we added Feb 24 to the step function $\pi(t)$ so that it has change points [Jan 23, Feb 4, Feb 8, Feb 24] with $\boldsymbol{\pi}_0 = (1, 0.9, 0.5, 0.1, \pi_{05})$. Note that in our fitted data, Feb 23 is the last observational date. We considered π_{05} equal to 0.1, 0.3 and 0.5 to describe "strictly continuing", "slightly loosening" or "moderately loosening" the control actions that has made the transmission rate 0.1β since Feb 8. Our results in Figure 7 and Web Table 2 indicate that, on average, increasing the transmission rate from 0.1β to 0.5β would end up with a second outbreak with a maximum prevalence 7.5% and totally 16.7% of the population affected by July 20, increasing from 0.1β to 0.3β would end

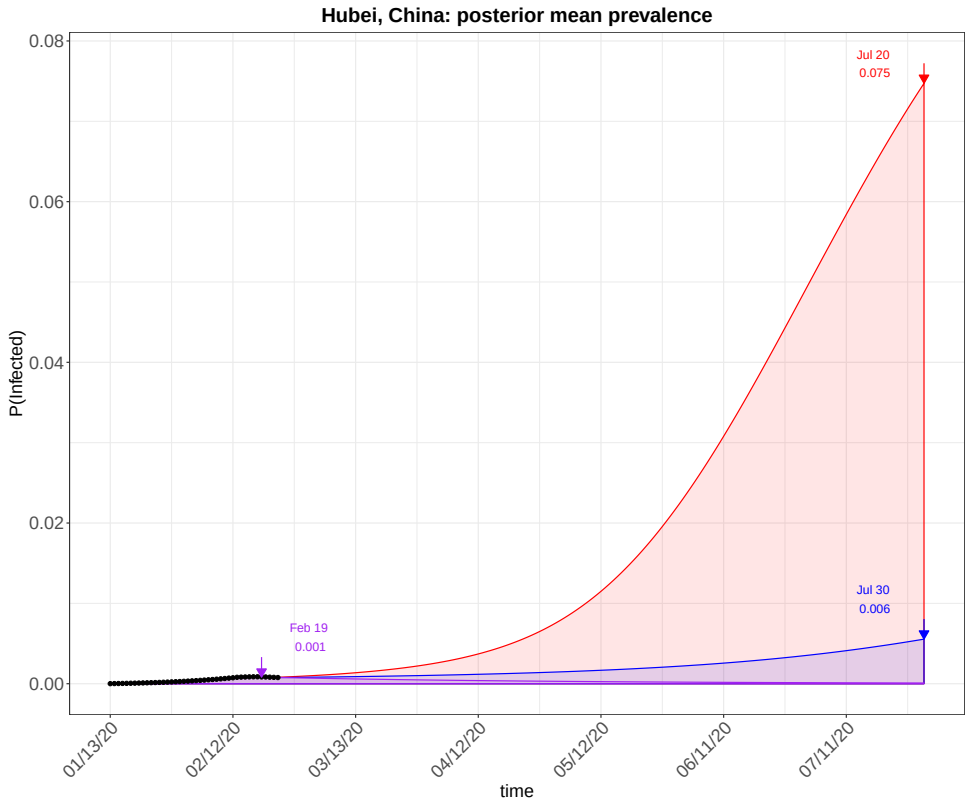

Figure 7: Predicted mean prevalence of infection with or without loosening the strict intervention in Hubei. The red semitransparent area denotes the scenario of moderate relaxation of the strict human intervention ($\pi_{05} = 0.5$), the blue area denotes the slight relaxation of intervention ($\pi_{05} = 0.3$), and the purple area denotes the scenario that stringent control is continued ($\pi_{05} = 0.1$). All their corresponding arrows mark the dates of their maximum mean prevalence.

up with a gradual increase in prevalence to 0.6% and about 1.4% of the population being affected. If we continue keeping the transmission rate to be 0.1β, however, the epidemic will eventually vanish in the population with no second outbreak and in total about 0.1% of the population being affected. All these three scenarios are much better than the one without any intervention (Panel A1 in Figure 6).

5 Concluding Remarks

We develop an epidemiological forecast model with an R software package to assess effects of interventions on the COVID-19 epidemic within Hubei and outside Hubei in China. Since our proposed model utilizes the strength of the SIR's dynamic system to capture the primary mechanism of the COVID-19 infectious disease, we are able to generate potential predictions of future episodes of the disease spread patterns over a prespecified window from the last date of data availability. Some turning points of interest are obtained from these forecasting curves as part of the deliverable information, including the predicted time when daily proportion of infected cases becomes smaller than the previous ones and the predicted time when daily proportion of removed cases (i.e. both recovered and dead) becomes larger than that of infected cases, as well

as the time when the epidemic ends. Our informatics toolbox provides quantification of uncertainty on the prediction, rather than only point prediction values, which are valuable to see the best versus the worst. The key novel contribution is the incorporation of time-varying quarantine protocols to expand the basic epidemiological model to accommodate changing transmission rates over time in the population. The toolbox can be used by practitioners who have better knowledge of quarantine and better quality data to perform their own analyses. Practitioners can use the toolbox to evaluate different types of quarantine strategies in practice. All summary statistics obtained from the toolbox are of great importance for public health workers and government policy makers to take proper actions on stop spreading of emerging epidemics, such as the COVID-19 epidemic examined here.

We choose the MCMC algorithm to implement the statistical estimation and prediction because of the consideration on the prediction uncertainty. Given the considerable complexity in the COVID-19 virus spread dynamics and potentially inaccurate information of quarantine measures as well as likely under-reported proportions of infected and recovered cases and deaths, it is of critical importance to quantify and report uncertainty in the forecast. Note that the publicly reported data of recovery and death of COVID-19 are mostly collected from hospitals where accessibility to such information is warranted. In contrast, it is very difficult, if not impossible, to collect the data of infected individuals with light symptoms who had in-home isolation and recovered, in spite of serious efforts made by the government for a door-to-door inspection to identify suspected cases.

This toolbox is indeed so general that it may be applicable to analyze and evaluate the COVID-19 epidemic in other countries, as well as the future outbreak of other types of infectious diseases. As noted in the paper, our proposed method does need some existing data of similar infectious disease to set up hyper-parameters in the prior distributions of the model parameters to begin the MCMC. For this, we used the epidemic parameters of the SARS outbreak in Hong Kong given some similarity of COVID-19 to SARS. From this perspective, what we learned from this COVID-19 epidemic in this paper is extremely valuable to form initial conditions in the analysis of any future outbreak of similar infectious disease. In addition, understanding forms and strengths of quarantines for the controlling of disease spread is an inevitable path to making effective preventive policies, which is the key analytic capacity that our toolbox offers.

The proposed approach is extremely useful for policy decision makers to conduct interventions forecast. Our analyses have shown that implementing strict intervention can well control the spread of COVID-19 in China. Moreover, continuing relatively strict intervention can help avoid a second outbreak. Though a slight to moderate relaxation on the intervention will lead to increased infection among the population, an interval of stringent control will still largely delay the progression of pandemic and reduce the maximum prevalence, or "flatten" the infection curves. A flattened infection curve means more preparation time and fewer infectious cases at each critical moment, hence more lives can be saved.

The proposed method has several limitations. First, it ignores the compartment of exposure; it is known that incubation period is relevant to disease transmission, which is particularly true for the COVID-19 as asymptomatic individuals are infectious. Second, the number of removed cases may be inaccurate due to the fact that many of deaths occurring outside of hospitals may not be diagnosed for the COVID-19 infection. Third, it assumes that the recovered cases are automatically immune to the coronavirus, which has not been clinically validated yet.

This analysis also has several limitations. Firstly, this analysis used an underlying SIR model structure, which is fairly simple—there are a number of additional processes that are known to be involved in the natural history of COVID-19 and could potentially be incorporated

into the model. For example, the incubation period is known to be approximately a median of 5 days (Lauer et al., 2020), which could be incorporated into the model. Similarly, age structure, potential super-spreading events, asymptomatic infections and variation in transmissibility across individuals, and more complex contact patterns (e.g. accounting for spatial structure when examining larger-scale dynamics such as across the whole country) could all play a potentially important role in the epidemic dynamics, altering the predictions of the model. Further, the model does not explicitly account for the underreporting fraction or how it may change over time, which can affect predictions and forecasts (Gamado et al., 2017, 2014; Eisenberg et al., 2015). Future work to account for more complex dynamics and incorporate these features into the package will be useful, both for model comparison and for extending the model to new contexts and diseases.

A second important future direction for this work is the validation of the predictions made by the model using subsequent data, such as cross-validating the model using data across different countries given that the COVID-19 has become a global pandemic. To fully evaluate the usefulness of this approach, it will be important to compare the model predictions to the actual trajectory of the epidemic—either for COVID-19 or for other epidemics, e.g. as a hindcasting exercise. This is an important next step for this approach to be used as a forecasting tool in public health practice.

Additionally, the proposed epidemiological models can be further extended to accommodate more data reported by the China CDC, which are worth future exploration. Two types of data that may be used in the future extension are the daily number of suspected cases and the daily number of hospitalized cases. We did not use such data due to the concern of data accuracy. For example, the number of suspected cases is largely dependent on the diagnostic protocols, which have been revised a few times since the outbreak of the disease, and the sensitivity of the RNA test. Given such concerns, our strategy in the proposed model was to only use necessary data for analysis, and over the course of improved data quality in the near future, our toolbox may be extended to enjoy greater statistical power and more accurate predictions.

Supplementary Materials

Software website: https://github.com/lilywang1988/eSIR. The online supplementary results can be found on the *Journal of Data Science* website.

A Runga–Kutta Approximation

A.1 Approximation in the Basic SIR model

The forth order Runga–Kutta(RK4) method gives an approximate of $f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)$ in equation (4) as follows:

$$f(\boldsymbol{\theta}_{t-1}, \beta, \gamma) = \begin{pmatrix} \theta^S_{t-1} + 1/6[k^{S_1}_{t-1} + 2k^{S_2}_{t-1} + 2k^{S_3}_{t-1} + k^{S_4}_{t-1}] \\ \theta^I_{t-1} + 1/6[k^{I_1}_{t-1} + 2k^{I_2}_{t-1} + 2k^{I_3}_{t-1} + k^{I_4}_{t-1}] \\ \theta^R_{t-1} + 1/6[k^{R_1}_{t-1} + 2k^{R_2}_{t-1} + 2k^{R_3}_{t-1} + k^{R_4}_{t-1}] \end{pmatrix} := \begin{pmatrix} \alpha_{1(t-1)} \\ \alpha_{2(t-1)} \\ \alpha_{3(t-1)} \end{pmatrix},$$

where

$$
\begin{aligned}
k_t^{S_1} &= -\beta\theta_t^S\theta_t^I, \\
k_t^{S_2} &= -\beta[\theta_t^S + 0.5k_t^{S_1}][\theta_t^I + 0.5k_t^{I_1}], \\
k_t^{S_3} &= -\beta[\theta_t^S + 0.5k_t^{S_2}][\theta_t^I + 0.5k_t^{I_2}], \\
k_t^{S_4} &= -\beta[\theta_t^S + k_t^{S_3}][\theta_t^I + k_t^{I_3}];
\end{aligned}
$$

$$
\begin{aligned}
k_t^{I_1} &= \beta\theta_t^S\theta_t^I - \gamma\theta_t^I, \\
k_t^{I_2} &= \beta[\theta_t^S + 0.5k_t^{S_1}][\theta_t^I + 0.5k_t^{I_1}] - \gamma[\theta_t^I + 0.5k_t^{I_1}], \\
k_t^{I_3} &= \beta[\theta_t^S + 0.5k_t^{S_2}][\theta_t^I + 0.5k_t^{I_2}] - \gamma[\theta_t^I + 0.5k_t^{I_2}], \\
k_t^{I_4} &= \beta[\theta_t^S + k_t^{S_3}][\theta_t^I + k_t^{I_3}] - \gamma[\theta_t^I + k_t^{I_3}];
\end{aligned}
$$

and

$$
\begin{aligned}
k_t^{R_1} &= \gamma\theta_t^I, \\
k_t^{R_2} &= \gamma[\theta_t^I + 0.5k_t^{I_1}], \\
k_t^{R_3} &= \gamma[\theta_t^I + 0.5k_t^{I_2}], \\
k_t^{R_4} &= \gamma[\theta_t^I + k_t^{I_3}].
\end{aligned}
$$

A.2 Approximation in the eSIR model with quarantine compartment

Using the RK4 approximation, $f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)$ in the extended SIR model (6) with a quarantine compartment can be approximated following the two iterative steps:

1. Solve the $f(\boldsymbol{\theta}_{t-1}, \beta, \gamma)$ in Appendix A without considering the quarantine with $f(\cdot)$
$$f(\boldsymbol{\theta}_{t-1}, \beta, \gamma) = [\alpha_{1(t-1)}, \alpha_{2(t-1)}, \alpha_{3(t-1)}]^{\mathrm{T}}.$$
2. Due to the quarantine, we deduct a mass of the susceptible by $\alpha^*_{1(t-1)} = \alpha_{1(t-1)} - \phi(t)\theta^S_{t-1}$, and let $\theta^Q_t = \theta^Q_{t-1} + \phi(t)\theta^S_{t-1}$ with $\theta^Q_0 = 0$.

Let $\boldsymbol{\alpha}^*_{t-1} = [\alpha^*_{1(t-1)}, \alpha_{2(t-1)}, \alpha_{3(t-1)}]^{\mathrm{T}}$, and it is easy to show that the sum $\sum_{k=1}^3 \alpha^*_{k(t-1)} = 1-\theta^Q_t$. Thus we can regenerate the next day's $\boldsymbol{\theta}_t$ following a Dirichlet distribution adjusted by the prevalence of the quarantine compartment $\boldsymbol{\alpha}^*_t \sim \text{Dirichlet}(\kappa\boldsymbol{\alpha}^*_{t-1}/(1-\theta^Q_t))$. The estimated prevalence values become $\boldsymbol{\theta}_t = (1-\theta^Q_t)\boldsymbol{\alpha}^*_t$. We follow above two steps and finish the complete prevalence processes. Note that the deduction of susceptible compartments might cause $\theta^S_t \leq 0$, so we will bound such prevalence value to be consistently 0 or above.

B Moment properties of Beta and Dirichlet distributions

For the sake of being self-contained, we list the moments of both Beta and Dirichlet distributions. The mean and variance of Beta distribution $\text{Beta}(\alpha, \beta)$ are respectively:

$$\text{Mean} = \frac{\alpha}{\alpha+\beta}, \text{Var} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

While to Dirchlet distribution $\text{Dir}(\kappa\boldsymbol{\alpha})$, we have

$$\text{Mean} = \boldsymbol{\alpha}, \text{Var} = \frac{1}{\kappa+1}\begin{pmatrix} \alpha_1(1-\alpha_1) & -\alpha_1\alpha_2 & -\alpha_1\alpha_3 & -\alpha_1\alpha_4 \\ -\alpha_1\alpha_2 & \alpha_2(1-\alpha_2) & -\alpha_2\alpha_3 & -\alpha_2\alpha_4 \\ -\alpha_1\alpha_3 & -\alpha_2\alpha_3 & \alpha_3(1-\alpha_3) & -\alpha_3\alpha_4 \\ -\alpha_1\alpha_4 & -\alpha_2\alpha_4 & -\alpha_3\alpha_4 & \alpha_4(1-\alpha_4) \end{pmatrix},$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^{\text{T}}$ with $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$.

C R Codes

First we conducted analysis of the Hubei COVID-19 data using the transmission rate modifier with function `txt.eSIR` from package `eSIR`. Note that option `dic=TRUE` enables to calculate the deviance information criterion (DIC) for model selection, while options, `save_files=TRUE` and `save_mcmc`, allow the storage of MCMC output tables, plots, summary statistics and even full MCMC draws, which may be saved via the path of `file_add`, or otherwise via the current working directory. The major results returned from the package include predicted cumulative proportions, predicted turning points of interest, two ggplot2 (Wickham, 2016) objects of the summary plots related to both infection and removed compartments, a summary output table containing all the posterior means, median and credible intervals of the model parameters, and DIC if opted. The trace-plots and other useful diagnostic plots are also provided and automatically saved if `save_files=TRUE` is opted. In the package, function `tvt.eSIR()` works on the epidemiological model with time-varying transmission rate in Section 2.2, and `qh.eSIR()` for the other epidemiological model with a quarantine compartment in Section 2.3. Note that in function `tvt.eSIR()`, with a choice of `exponential=FALSE`, a step function is run in the MCMC when both `change_time` and `pi0` are specified. To fit the model with a continuous transmission rate modifier function, user may set `exponential=TRUE` and specify a value of `lambda0`. The default is to run the basic epidemiological model with no quarantine or $\pi(t) \equiv 1$ in Section 2.1. `death_in_R` is usually set to be the average ratio of death and removed proportions at each observation time point, which is used to estimate the death curve in the forecast plot of the removed compartment. Below are the R scripts used in the analysis.

```
### Example 1: Step function pi(t)
### Y and R are observed proportions of infected and removed compartments
change_time <- c("01/23/2020", "02/04/2020", "02/08/2020")
pi0 <- c(1.0, 0.9, 0.5, 0.1)
res.step <- tvt.eSIR (Y, R, begin_str = "01/13/2020", death_in_R = 0.4,
T_fin = 200, pi0 = pi0, change_time = change_time, dic = TRUE,
casename = "Hubei_step", save_files = TRUE,
save_mcmc = FALSE, M = 5e2, nburnin = 2e2)
res.step$plot_infection
res.step$plot_removed
res.step$dic_val

### Example 2: continuous exponential function pi(t)
res.exp <- tvt.eSIR (Y, R, begin_str = "01/13/2020", death_in_R = 0.4,
T_fin = 200, exponential = TRUE, dic = FALSE, lambda0 = 0.05,
```

```
casename = "Hubei_exp", save_files = FALSE, save_mcmc = FALSE,
M = 5e2, nburnin = 2e2)
res.exp$plot_infection
# res.exp$plot_removed

### Example 3: the basic state-space SIR model, pi(t)=1
res.nopi <- tvt.eSIR (Y, R, begin_str = "01/13/2020", death_in_R = 0.4,
T_fin = 200, casename = "Hubei_nopi", save_files = FALSE,
M=5e2, nburnin = 2e2)
res.nopi$plot_infection
# res.nopi$plot_removed
```

The other epidemiological model with an added quarantine compartment as an absorbing state was fitted via our R function `qh.eSIR` in the package `eSIR`. The arguments used in `qh.eSIR()` are almost identical to those in `tvt.eSIR()`. Note that if the quarantine rate function is set at constant 0, this model will be reduced to a basic epidemiological SIR model.

```
### Example 4: Dirac delta function of the quarantine process
change_time <- c("01/23/2020", "02/04/2020", "02/08/2020")
phi <- c(0.1, 0.4, 0.4)
res.q <- qh.eSIR (Y, R, begin_str = "01/13/2020",death_in_R = 0.4,
phi0 = phi0, change_time = change_time, casename = "Hubei_q",
save_files = TRUE, save_mcmc = FALSE, M = 5e2, nburnin = 2e2)
res.q$plot_infection
# res.q$plot_removed

### Example 5: basic state-space SIR model
res.noq <- qh.eSIR (Y, R, begin_str = "01/13/2020", death_in_R = 0.4,
T_fin = 200, casename = "Hubei_noq", M = 5e2, nburnin = 2e2)
res.noq$plot_infection
```

In the above R coding scripts, only very short MCMC chains are specified for the consideration of running time. In practice, we set `M=5e5` and `nburnin=2e5` to achieve desirable burn-ins and yield stable MCMC draws.

D Under-reporting Calibration

As is mentioned in the Introduction, the issue of under-reporting may cause bias in prediction. In order to adjust the under-reported number of infected cases, we apply the following algorithm to calibrate the number of infections before Feb 12, during which time the Chinese government only relies on the RNA test for diagnosis, which was realized later with low sensitivity leading to many false negatives.

We assume that the cumulative number of infected cases between Jan 13 and Feb 12 when a sudden big jump occurs follows an exponential function,

$$y(t) = ae^{\lambda t} + b,$$

where $t \in \{1, 2, \dots\}$ and a, b, λ are parameters to be estimated. Here, $t = 1$ stands for Jan 13 and $t = 31$ stands for Feb 12. Under the condition of $y(0) = 0$, we can easily get that

$$y(t) = ae^{\lambda t} - a.$$

To estimate parameter λ and a, we want to minimize the least square error of the estimated number $\hat{y}(t)$ of infected cases at $t = 32$ (Feb 13), which is one day after the Chinese government changed the diagnosis protocol. It is assumed that the difference between the predicted and observed number of infections on Feb 13 would not be big if the model were established well, although the long term difference might be large due to other interventions. Therefore, the optimization problem we want to solve is,

$$\begin{aligned} \min_{a,\lambda} \quad & \{y(32) - ae^{32\lambda} + a\}^2 \\ \text{s.t.} \quad & ae^{31\lambda} - a = y(31). \end{aligned}$$

The constraint $ae^{31\lambda} - a = y(31)$ is used to ensure that the cumulative number of infected cases till Feb 12 equals to the observed value $y(31)$. The optimization problem can be solved using the method of Lagrange Multipliers. Obtained $\hat{\lambda} = 0.06605, \hat{a} = 7142.80$. The calibrated number of infected cased between Jan 13 and Feb 12 is shown in Figure 5. The proposed calibration method corrected the under-reporting issue, at least partially.

References

Carlin BP, Polson NG, Stoffer DS (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association*, 87(418): 493–500.

Chen H, Guo J, Wang C, Luo F, Yu X, Zhang W, et al. (2020). Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: A retrospective review of medical records. *The Lancet*, 395(10226): 809–815.

Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH (2019). Complexity of the basic reproduction number (R0). *Emerging Infectious Diseases*, 25(1): 1–4.

Dennis DT, Gage KL, Gratz NG, Poland JD, Tikhomirov E (1999). Plague manual: Epidemiology, distribution, surveillance and control. *Technical report*, Geneva: World Health Organization.

Eisenberg MC, Eisenberg JN, D'Silva JP, Wells EV, Cherng S, Kao YH, et al. (2015). Forecasting and uncertainty in modeling the 2014-2015 Ebola epidemic in West Africa. ArXiv preprint: https://arxiv.org/abs/1501.05555.

Fan Y, Zhao K, Shi ZL, Zhou P (2019). Bat coronaviruses in China. *Viruses*, 11(3): 210.

Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, et al. (2009). Pandemic potential of a strain of Influenza A (H1N1): Early findings. *Science*, 324(5934): 1557–1561.

Gamado K, Streftaris G, Zachary S (2017). Estimation of under-reporting in epidemics using approximations. *Journal of Mathematical Biology*, 74(7): 1683–1707.

Gamado KM, Streftaris G, Zachary S (2014). Modelling under-reporting in epidemics. *Journal of Mathematical Biology*, 69(3): 737–765.

Gralinski LE, Menachery VD (2020). Return of the coronavirus: 2019-nCoV. *Viruses*, 12(2): 135.

Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. (2020). Clinical characteristics of 2019 novel coronavirus infection in China. MedRxiv preprint: https://doi.org/10.1101/2020.02.06.20020974.

Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. (2020). First case of 2019 novel coronavirus in the United States. *New England Journal of Medicine*, 382(10): 929–936.

Hu Z, Ge Q, Jin L, Xiong M (2020). Artificial intelligence forecasting of COVID-19 in China. ArXiv preprint: https://arxiv.org/abs/2002.07112.

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223): 497–506.

Hui DS, I Azhar E, Madani TA, Ntoumi F, Kock R, Dar O, et al. (2020). The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—the latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases*, 91: 264–266.

Imai N, Dorigatti I, Cori A, Riley S, Ferguson NM (2020). Estimating the potential total number of novel coronavirus cases in Wuhan City, China. http://hdl.handle.net/10044/1/77150.

Jørgensen B, Lundbye-Christensen S, Song PK, Sun L (1999). A state space model for multivariate longitudinal count data. *Biometrika*, 86(1): 169–181.

Jøsrgensen B, Song PXK (2007). Stationary state space models for longitudinal data. *Canadian Journal of Statistics*, 35(4): 461–483.

Jung Sm, Akhmetzhanov AR, Hayashi K, Linton NM, Yang Y, Yuan B, et al. (2020). Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: Inference using exported cases. *Journal of Clinical Medicine*, 9(2): 523.

Kao YH, Eisenberg MC (2018). Practical unidentifiability of a simple vector-borne disease model: Implications for parameter estimation and intervention assessment. *Epidemics*, 25: 89–100.

Kermack WO, McKendrick AG (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Series A*, 115(772): 700–721.

Kleinsasser M, Barker D, Wang L (2020). Explore analysis and forecast results for China. https://umich-biostatistics.shinyapps.io/eSIR/.

Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9): 577–582.

Li J, Wang Y, Gilmour S, Wang M, Yoneoka D, Wang Y, et al. (2020a). Estimation of the epidemic properties of the 2019 novel coronavirus: A mathematical modeling study. *Preprints with the Lancet*.

Li Q, Feng W, Quan YH (2020b). Trend and forecasting of the COVID-19 outbreak in China. *Journal of Infection*, 80(4): 469–496.

Liu Q, Liu Z, Li D, Gao Z, Zhu J, Yang J, et al. (2020). Assessing the tendency of 2019-nCoV (COVID-19) outbreak in China. MedRxiv preprint: https://doi.org/10.1101/2020.02.09.20021444.

Luk HK, Li X, Fung J, Lau SK, Woo PC (2019). Molecular epidemiology, evolution and phylogeny of SARS coronavirus. *Infection, Genetics and Evolution*, 71: 21 – 30.

Mkhatshwa T, Mummert A (2010). Modeling super-spreading events for infectious diseases: Case study SARS. ArXiv preprint: https://arxiv.org/abs/1007.0908.

Nishiura H, Tsuzuki S, Yuan B, Yamaguchi T, Asai Y (2017). Transmission dynamics of cholera in Yemen, 2017: A real time forecasting. *Theoretical Biology and Medical Modelling*, 14: 14.

Osthus D, Hickmann KS, Caragea PC, Higdon D, Del Valle SY (2017). Forecasting seasonal influenza with a state-space SIR model. *The Annals of Applied Statistics*, 11(1): 202–224.

Peng L, Yang W, Zhang D, Zhuge C, Hong L (2020). Epidemic analysis of COVID-19 in China by dynamical modeling. ArXiv preprint: https://arxiv.org/abs/2002.06563.

Plummer M (2019). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-10.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabajante JF (2020). Insights from early mathematical models of 2019-nCoV acute respiratory disease (COVID-19) dynamics. ArXiv preprint: https://arxiv.org/abs/2002.05296.

Rothe C, Schunk M, Sothmann P, Bretzel G, Froeschl G, Wallrauch C, et al. (2020). Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *New England Journal of Medicine*, 382(10): 970–971.

Smith RD (2006). Responding to global infectious disease outbreaks: lessons from SARS on the role of risk perception, communication and management. *Social Science & Medicine*, 63(12): 3113–3123.

Song PXK (2000). Monte Carlo Kalman filter and smoothing for multivariate discrete state space models. *Canadian Journal of Statistics*, 28(3): 641–652.

Song PXK (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer.

Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE, Decroly E, et al. (2014). One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. *Proceedings of the National Academy of Sciences*, 111(37): E3900–E3909.

Sun H, Qiu Y, Yan H, Huang Y, Zhu Y, Chen SX (2020). Tracking and predicting COVID-19 epidemic in China Mainland. BioRxiv preprint: https://doi.org/10.1101/2020.02.17.20024257.

Wang C, Horby PW, Hayden FG, Gao GF (2020a). A novel coronavirus outbreak of global health concern. *The Lancet*, 395(10223): 470–473.

Wang L, Wang F, Tang L, Zhu B, Zhou Y, He J, et al. (2020b). *eSIR: Extended state-space SIR models*. R package version 0.2.5, https://github.com/lilywang1988/eSIR.

Weitz JS, Dushoff J (2015). Modeling post-death transmission of Ebola: Challenges for inference and opportunities for control. *Scientific Reports*, 5: 8751.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

World Health Organization (2003). Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. http://www.who.int/csr/sars/country/table2004_04_21/en/index.html.

World Health Organization (2020). Emergencies preparedness, response: Pneumonia of unknown origin — China; Disease outbreak news. https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/.

Xiang YT, Li W, Zhang Q, Jin Y, Rao WW, Zeng LN, et al. (2020). Timely research papers about COVID-19 in China. *The Lancet*, 395(10225): 684–685.

Xu XW, Wu XX, Jiang XG, Xu KJ, Ying LJ, Ma CL, et al. (2020). Clinical findings in a group of patients infected with the 2019 novel coronavirus (SARS-Cov-2) outside of Wuhan, China: Retrospective case series. *BMJ*, 368. https://doi.org/10.1136/bmj.m606.

Yu G (2020). *nCov2019: Stats of the '2019-nCoV' Cases*. R package version 0.0.8, https://github.com/GuangchuangYu/nCov2019.

Zhang S, Diao M, Yu W, Pei L, Lin Z, Chen D (2020). Estimation of the reproductive number of

novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. *International Journal of Infectious Diseases*, 93: 201–204.

Zhu B, Taylor JM, Song PXK (2012). Signal extraction and breakpoint identification for array CGH data using robust state space model. ArXiv preprint: https://arxiv.org/abs/1201.5169.

Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*, 382(8): 727–733.