Parameter Estimation for SIR Model Using MCMC and Its Meaning in Epidemic Control

Xinhao Liu
School of Arts and Science
New York University Shanghai
Shanghai, China
Email: xinhao.liu@nyu.edu

Kuntian Chen
School of Arts and Science
New York University Shanghai
Shanghai, China
Email: kc4159@nyu.edu

Hongquan Liu
School of Arts and Science
New York University Shanghai
Shanghai, China
Email: hl3798@nyu.edu

***Abstract*—In this project, by applying the susceptible-infected-removed (SIR) model for epidemic statistics modeling and Monte Carlo Markov Chain (MCMC) method for parameter estimation, researchers discussed about the possibility to estimate real-world epidemic data. More insights beyond the existing models and methods are also provided in this report**

I. Introduction

The Novel Coronavirus Disease 2019 (COVID-19), breaking out in the late January, 2020, has evolved into a global pandemic. Eight months after its first confirmed case, it is still infecting everyone's daily life all over the world. We've seen many articles discussing the spread of the epidemic since the very beginning of its breakout, but very few of them were able to give a rigours analysis. Therefore, in this project, we aim to use mathematical models and inferences, together with integrated programs to estimate and measure the evolution of the epidemic. In this report, we will first introduce our methodology and models, document the process of our experiment, and finally provide some discussion and insights about this whole project.

II. Methodology

A. SIR Model

1) Concept: The susceptible-infected-removed (SIR) model is "developed by Ronald Ross1, William Hamer, and others in the early twentieth century" as introduced by Howard Weiss [1]. The core concept behind it is to view the subjects of the research - individuals into different groups with features. In other words, it classifies the subjects and use a special feature shared by all individuals in the group to simplify our process of research. Actually, this approach is very similar to the pattern of our brains when they receive messages. Since our brains cannot process the huge amount information, especially nowadays with explosive information, classification is a basic approach for them to make our understanding of the outside world easier.

Likewise, when we have a huge amount of population - hundreds of millions or even billions - it's impossible for us to keep track of every individual and look for trends with the evolution of time. Therefore, compartments are needed to classify the huge amount of population. The way of modeling with compartments is generally called compartmental models. Researchers set different compartments with traits that are relative and meaningful to their researches, then put all subjects into this compartments. With time being the independent variable and the population of each compartment being the dependent variable, circulations (unidirectional or bidirectional) are allowed among the compartments. The circulation itself reflects the evolution and trends of the subjects and the changing populations are the statistics that are directly used by researchers for study. Technically speaking, the SIR model is a special type of compartmental model applied in the field of epidemiology. Zhang Fa et al. list several features of SIR model in [2]: individuals are homogeneous, groups are homogeneously distributed, contact is instantaneous, infection and remove rates are both a constant.

2) Application: Considering the special traits of an epidemic, the SIR model have three compartments for the subjects of the research - the susceptible compartment, the infected compartment, and the removed compartment. We can tell from the name that they correspondingly hold populations who are susceptible to the epidemic, who have been infected with the epidemic and who are "removed" from the epidemic. The circulation (transaction) among these compartments are all unidirectional. When a susceptible individual is infected with the epidemic, it will be moved to the infected compartments. Similarly, when an infected individual is recovered from or dead of the epidemic, it will be moved to the removed compartment. No reverse transactions are allowed in both circulations above.

When we use mathematical equations to reflect the population and circulation of these compartments, it looks like this [1]:

$$\frac{dS}{dt} = -\beta SI \tag{1}$$

$$\frac{dI}{dt} = \beta SI - \nu I \tag{2}$$

$$\frac{dR}{dt} = \nu I. \tag{3}$$

Here, t means time unit (in most epidemic statistics, time unit is day); S means the populations of susceptible compartment; I means the populations of infected compartment, R means the populations of infected compartment, β is the rate of

the transaction from susceptible to infected with respect to the population of susceptible and infected, and v is the rate of the transaction from susceptible to removed with respect to the population of infected. The three ordinary differential equations (ODE) reflect all information we need from the three compartments.

Before we move on to talk more about the equations, there are two key points that are worth noting in this setting of compartments. First, although some papers tend to call the removed compartment as "recovered", we prefer the former name. As mentioned above, the removed compartment consists of both recovered and dead individuals. This is because in this setting, there is no difference between recovering or death. Whether an infected individual is end up recovered or dead, due to the unidirectional transaction, it cannot become susceptible or infected (here we only considering the cases where recovering is followed by lifetime immunity, more complicated situation will be covered in the following sections)

The second key point is that the system of the compartments is closed. The whole population - the sum of the population of three compartments - doesn't change. It means the model doesn't take natural born and death into consideration, nor does it consider immigration and emigration. In other words, the model presumes that the total population in an area in a certain period is absolute. Therefore, the final stages of the model are all the same. We can also infer from the equations that at the end, the population of infected will goes to 0 and the population of susceptible and removed will not change consequently (The reason why it's impossible for the susceptible population to go to 0 needs more rigorous proof). Hence, the ratio between susceptible and removed at the final stage is only one indicator that can show the severity of the epidemic and success of actions made by administrations.

Also, because of the second key point, if we know the total population (denoted as P), we can simplify the three equations into two [1]:

$$\frac{dS}{dt} = -\beta S(P - S - R) \tag{4}$$

$$\frac{dR}{dt} = \nu(P - S - R) \tag{5}$$

3) Comprehension: Now that we have understood the mechanism of the SIR model, we still want to talk about the meaning of the two rate variables, β and v. Although we can simplify the rate as a single variable, it is in fact determine by many factors in real world. For example, β, the infection rate, can be determined by how many contact a susceptible individual has with infected individuals and how likely the susceptible is infected during each contact. The former can be lowered by quarantine and the latter can be lowered by wearing masks. However, the majority of epidemic statistics only have the population of infected, recovered, and dead. When we estimate the parameter, there will be only one result, which is the combination of the factors, or β and v.

4) Extension: Since the SIR model only has three compartments, it is easy to imagine that it cannot reflect an epidemic very precisely, nor can it fit more complicated situations. Therefore, more derivative models of SIR with more compartments are created by researchers. For example, the most proper model for COVID-19 might be the SEQIRD (susceptible-exposed-quarantined-infected-recovered-dead) model [2]. Of course, a more comprehensive model means more complicated equations and, in turn, leads to more work during processing.

B. Markov Chain

1) Notations: The probability is written as a column vector while the stochastic matrices is applied to the vector from its left side instead of from the right side.

$$Pv = \begin{bmatrix} P_{1,1} & P_{1,2} & \dots & P_{1,j} & \dots & P_{1,n} \\ P_{2,1} & P_{2,2} & \dots & P_{2,j} & \dots & P_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{i,1} & P_{i,2} & \dots & P_{i,j} & \dots & P_{i,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{n,1} & P_{n,2} & \dots & P_{n,j} & \dots & P_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix}.$$

Then, in order to make the matrices and the vectors be suitable for probability, we need to impose some restrictions on them.

Definition 1. A matrix is a stochastic matrix if it satisfies the following:
1) All its entries are non-negative.
2) For all $j \in [1, n]$ (all column vectors),

$$\sum_{i=1}^{n} P_{i,j} = 1.$$

3) It must be a square matrix.

Definition 2. A vector $\vec{x}$ is a *valid vector* if it satisfies the following:
1) All its entries are non-negative.
2)

$$\vec{x} \cdot \vec{1} = \sum_{i=1}^{n} x_i = 1, \ \vec{1} = [1, 1, ..., 1]^T$$

Comment 1. The set of all valid vectors is NOT a linear space! It in not closed to addition and scaler multiplication.

2) Elementary Properties of the Markov Chain:

Theorem 1. *Let P be a stochastic matrix and $\vec{x}$ be a valid vector. For any $m \in \mathbb{N}$, $P^m\vec{x}$ is also a valid vector.*

Proof. We prove by induction on m. Assume $\vec{x}$ is a *valid vector*,

$$\sum_{i=1}^{n}(P\vec{x})_i = \sum_{i=1}^{n}(\sum_{j=1}^{n} P_{ij}x_j) = \sum_{j=1}^{n}(\sum_{i=1}^{n} P_{ij}x_j) = \sum_{j=1}^{n} x_j = 1$$

□

Theorem 2. *Any stochastic matrix has an eigenvalue* 1.

Proof. We only need to prove that $\det(P-I)=0$.

$$(P-I)=\begin{bmatrix} P_{1,1}-1 & P_{1,2} & \dots & P_{1,j} & \dots & P_{1,n} \\ P_{2,1} & P_{2,2}-1 & \dots & P_{2,j} & \dots & P_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{i,1} & P_{i,2} & \dots & P_{i,j} & \dots & P_{i,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{n,1} & P_{n,2} & \dots & P_{n,j} & \dots & P_{n,n}-1 \end{bmatrix}$$

$$\text{Since } \sum_{i=1}^{n} P_{i,j}=1,$$

$$\begin{bmatrix} P_{1,1}-1 \\ P_{1,2} \\ \vdots \\ P_{1,n} \end{bmatrix}^T + \begin{bmatrix} P_{2,1} \\ P_{2,2}-1 \\ \vdots \\ P_{2,n} \end{bmatrix}^T + \cdots + \begin{bmatrix} P_{n,1} \\ P_{n,2} \\ \vdots \\ P_{n,n}-1 \end{bmatrix}^T = 0$$

We can see that the row vectors of $(P-I)$ are linearly dependent, since we have found an non-trivial linear combination yielding zero.
Therefore

$$\det(P-I)=0$$

□

Theorem 3. *For an eigenvector $\vec{x}$ corresponding to an eigenvalue other than 1, $\vec{x}$ is not a valid vector and $\vec{x}\cdot\vec{1}=0$.*

Proof. Let $\vec{1}=[1,1,...,1]^T$. Let λ be an eigenvalue of P and $\vec{x}$ be the eigenvector corresponding to λ, then

$$P\vec{x}=\lambda\vec{x}$$

where $\vec{x}=[x_1,x_2,...,x_n]^T$.
Since P is a stochastic matrix,

$$\lambda\vec{x}\cdot\vec{1}=P\vec{x}\cdot\vec{1}=\vec{x}\cdot\vec{1}=\sum_{i=1}^{n}x_1=1$$

For an eigenvector v corresponding to an eigenvalue other than 1, $\vec{x}\cdot\vec{1}=0$. □

Corollary 1. A *valid eigenvector* $\vec{x}$ corresponds to the eigenvalue 1.

Theorem 4. *If λ is an eigenvalue of a stochastic matrix P, $|\lambda|\leq 1$.*

Proof. We are going to prove by contradiction.
Assume that λ is an eigenvalue of P such that $|\lambda|>1$. Let $\vec{x}$ be an eigenvector corresponding to λ. Let u be an eigenvector corresponding to eigenvalue 1.
Let $w=u+\alpha v$ for some $\alpha\in\mathbb{R}$ such that all coordinates of w are positive. By previous Theorem, we know that $v\cdot\vec{1}=0$ and $u\cdot\vec{1}=1$. Hence $w\cdot\vec{1}=1$, w is a *valid vector*. Moreover, v is not a zero vector, or it corresponds to eigenvalue 0. Therefore there exists an entry $v_i<0$.
$\forall m\in\mathbb{N}$, we have

$$P^m w=\lambda^m\alpha v+u$$

By **Theorem 1**, $P^m w$ is a *valid vector*. Therefore $(P^m w)_i\geq 0$. Let m be an even number and

$$m\geq\log_\lambda[\frac{(P^m w)_i+u_i}{-v_i}]$$

We would have the LHS to be non-negative and the RHS to be negative, yielding the contradiction. □

3) the Necessary Condition of Convergence: Stochastic matrices are not necessarily diagonalizable.

Proof. Counterexample :

$$P=\frac{1}{12}\begin{bmatrix}5&5&2\\3&3&6\\4&4&4\end{bmatrix}$$

□

However, in reality, almost all stochastic matrices that we will face are diagonalizable and only have one linearly independent eigenvector corresponding to the eigenvalue 1. In fact, if we form the entries of a stochastic matrix by randomly picking up non-negative real numbers, the probability of getting a non-diagonalizable matrix is zero. Now, I will prove the convergence under these normal circumstances.

Theorem 5. *Let P be a diagonalizable stochastic matrix. If u is the only linearly independent eigenvector corresponding to the eigenvalue 1, then for any valid input vector v,*

$$\lim_{m\to\infty}A^m v=\alpha u,\quad \textit{for some } \alpha\in\mathbb{R}.$$

Proof. Because P is diagonalizable, we can find a basis consists of eigenvectors for P. Denote it as $u,v_1,v_2,\cdots,v_{n-1}$. So $v_1,v_2,\cdots,v_{n-1}$ are eigenvectors corresponding to the eigenvalues $\lambda_1,\lambda_2,\cdots,\lambda_{n-1}$, counting multiplicity.
Then, for any v, it can be written as

$$v=\alpha u+\alpha_1 v_1+\alpha_2 v_2+\cdots+\alpha_{n-1}v_{n-1}$$

for some $\alpha,\alpha_1,\alpha_2,\cdots,\alpha_{n-1}\in\mathbb{R}$. Hence,

$$\begin{aligned}P^m v&=\alpha P^m u+\alpha_1 P^m v_1+\alpha_2 P^m v_2+\cdots+\alpha_{n-1}P^m v_{n-1}\\&=\alpha u+\alpha_1\lambda_1^m v_1+\alpha_2\lambda_2^m v_2+\cdots+\alpha_{n-1}\lambda_{n-1}^m v_{n-1}\end{aligned}$$

By Section 3.1, we know that $\lambda_1,\lambda_2,\cdots,\lambda_{n-1}\in(-1,1)$. Hence, for $i\in\{1,2,\cdots,n-1\}$,

$$\lim_{m\to\infty}\alpha_i\lambda_i^m v_i=0$$

Finally, we have

$$\lim_{m\to\infty}P^m v=\alpha u$$

□

4) Other Conditions: In this section, we will try to loose the previous condition and see whether the result will converge or not.

Basically, as long as P is diagonalizable, the situation won't be too bad. By following a similar proof in the previous section, we can get this corollary:

Corollary 2. Let A be a diagonalizable stochastic matrix. Denote the eigenspace of 1 as U, or formally $U = E(P, 1)$. Then for any valid input vector v,

$$\lim_{m\to\infty} P^m v = P_U v,$$

where $P_U v$ denotes the orthogonal projection of v onto U.

When P is not diagonalizable, the whole situation gets much more complicated.

Comment 2. For a stochastic matrix P and a valid input vector v, $P^m v$ does NOT necessarily converge when m gets large.

Proof. I will present an counterexample.

Let $A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $v = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$. As you can see, the vector is trapped in an everlasting desperate iteration between $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$. □

III. Modeling

In this section, we will show the methods we used to estimate the parameters of the SIR model based on real data.

A. Likelihood Maximization

As the Bayesian Inference shows, the parameters we need are the parameters that can maximize the likelihood function of the real-world data. Hence, the work is divided into 2 parts: finding a proper likelihood function and trying to find the parameter that maximizes it.

From the definition of the likelihood function, we know that it should attain its maximum when the predicted data perfectly matches with the real-world data. On the other hand, the data generated by the model and the real world can almost be fully described by two parts, which are the number of the infected people and the removed people at certain time. Therefore, we chose two-dimensional normal distribution as our likelihood function. If we denote the predicted infected and removed population at time t by I'_t, R'_t and the real infected and removed population at time t by I_t, R_t, the exact formula is given by this:

$$\prod_t \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{(I_t - I'_t)^2}{\sigma_1^2} + \frac{(R_t - R'_t)^2}{\sigma_2^2}\right)\right]$$

where σ_1, σ_2 are chosen according to the data.

Now we can denote the observed real-world data as x and the likelihood function as $L(\beta, \nu|x)$, since the SIR model is only determined by two parameters β and ν. Then, our parameter estimation problem is converted into a problem of finding the maximum of a function with 2 variables. We used a quite primitive way to solve it by Python. We found out that β and ν is likely to be in $(0, 10]$. Then, we divide this interval into 100 pieces of equal length. Then, using an embedded loop, we let β and ν individually take the value of $\{0.1, 0.2, \cdots, 9.9, 10\}$. Finally, we record all the results of the likelihood function and find its maximum.

B. MCMC with Hand-Written Codes

From our experiment, the likelihood maximization method doesn't seem to work well. Then, we wrote a code of using MCMC to estimate the parameter.

In this part, I will first go through my understanding of the sampling procedure and then the explanation.

Let π be the distribution that we want to get the sampling. It's expression is known by us. Assume that after n_1 times, the original input distribution eventually converges to $\pi(a, b)$. a, b are the parameters that we are going to estimate. Assume that we need n_2 samples.

Find a σ as you want. It is played as a constant in the whole procedure.

The algorithm begins.

$n = 0$
randomly choose two positive numbers to be a_0, b_0
while $n \le n_1 + n_2$:
  randomly choose a'_{n+1} from $N(a_n, \sigma^2)$
  randomly choose b'_{n+1} from $N(b_n, \sigma^2)$
  $\alpha = \pi(a'_{n+1}, b'_{n+1})/\pi(a_n, b_n)$
  randomly choose u from the uniform distribution in $(0, 1)$
  if $u < \alpha$:
    $a_{n+1} = a'_{n+1}$
    $b_{n+1} = b'_{n+1}$
    $n = n + 1$

$a_{n_1+1}, a_{n_1+2}, \cdots, a_{n_1+n_2}$ and $b_{n_1+1}, b_{n_1+2}, \cdots, b_{n_1+n_2}$ are the samples we want

let a be the average of $a_{n_1+1}, a_{n_1+2}, \cdots, a_{n_1+n_2}$ and b be the average of $b_{n_1+1}, b_{n_1+2}, \cdots, b_{n_1+n_2}$

Note 1: $N(\mu, \sigma^2)$ is the normal distribution with expectation μ and variance σ^2. We know how to get samples from this distribution.

Note 2: if $u \ge \alpha$, then nothing will happen. The a'_{n+1}, b'_{n+1} will be abandoned and n will remain unchanged.

As we have previously proved, if a stochastic matrix π eventually make any distribution converges to π, we have

$$P\pi = \pi.$$

But the question is: how to get this P? As we have seen in the notes, by arbitrarily choosing a stochastic matrix Q, we can find a matrix α such that

$$\alpha Q = P \text{ and } \alpha Q\pi = \pi.$$

One thing to notice is that the "matrices" here are not well-defined matrices, because π is a continuous distribution.

Intuitively, we let the i-th "column" of Q to be $N(i, \sigma^2)$. Why do we do it like this?

The first reason is that we know how to get the samples from a normal distribution. When we know a_n and we want to get a'_{n+1} from it, we are actually going to sample from $Q(x|a_n)$, which is the a_n-th "row" of Q. In this case, it will be $N(a_n, \sigma^2)$, which is easy to compute. This also applies to b.

The second reason that Q is symmetric now. So we can take $\alpha = \pi(a'_{n+1}, b'_{n+1})/\pi(a_n, b_n)$ without considering Q.

C. MCMC with PyMC3

For implementation part of our model. We chose the PyMC3 package as our tool to process artificial and real-world data. This package is introduced by John Salvatier et al. in [3]. Thanks to John Salvatier and his team, this package is integrated with all the basic tools to implement MCMC and other applications of Bayesian models. More specifically, we used the default No-U-Turn-Sampler (NUTS) for our modeling. Compared with our hand-written programs, the program with PyMC3 runs much faster. We believe it is because the package encapsulates lots of optimization for the models. The encapsulation leads to the package being very straightforward in terms of its API. However, every coin has two sides. It leaves our team very little space to alter the way the program runs, neither are we able to determine what was the problem when the program raised errors. This feature directly leads to our outcome of the experiment as will be mentioned in the next section.

IV. Experimenting

In this section, we will briefly introduce our process of doing experiment. We divided it into two parts. First, we tried several sets of artificial data. The second part introduced real-world data of different areas.

A. Artificial Data

To begin with, we need to verify the capability of the modeling tools in the PyMC3 package. Therefore, we set a series of initial values for the parameters and population of each compartment, and solved the ODE to produce a set of artificial data (ODE solving was done by the ode_int function in the SciPy package). In order to simulate the real-world data, we also add random noise to out data. After this, we had a set of SIR data that fits a theoretically perfect SIR curve (as shown in figure 1).

Our next step was to give a prior distribution for β and ν (here we both take a normal distribution with μ=0.5 and σ=2). Then, take iterations of sampling to estimate the two parameters.

We run the same experiment for several times with different initial values and prior distributions. The results of the estimation are all very close to the initial value. Therefore, we reached to the conclusion that the modeling process in PyMC3 is capable of similar parameter estimation tasks.

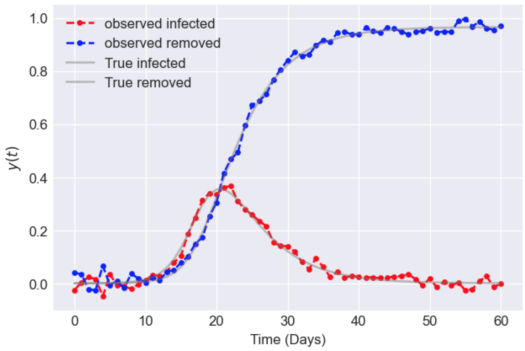

Fig. 1. Artificial data with the original SIR curve (Initial values: S=0.9999, I=0.0001, R=0, β=0.7, ν=0.2).

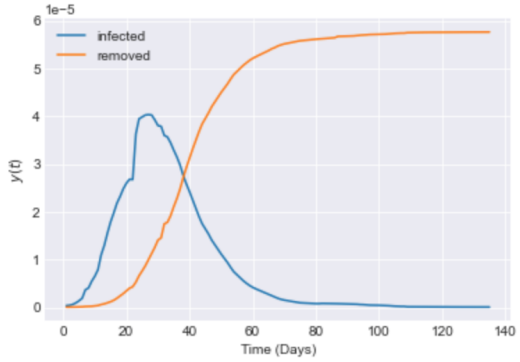

Fig. 2. Real data in China

B. Real-world Data

Our first trial is the data set of all the cases in China. Considering the epidemic broke out in China in the late January and back to normal in the late March, we collected data points from Jan 22, 2020 to Jun 4, 2020 from National Health Commission of the People's Republic of China [1] and China CDC [2]. We can tell from figure 2 that the data is similar to a theoretical SIR model although there is a significant sharp increase around the twentieth day. This increase is due to the change of the definition of infection. In our later processing steps, we were able to curve the data as much as possible.

However, the result of the sampling didn't turn out to be satisfactory. We had a hypothesis that one significant difference between our artificial and real-world data is that they were not in the same magnitude. Artificial data tends to reach a infection peak of 40% the total population and nearly 100% of removed ratio. This is unimaginable in a real world when the data only reaches a peak of 10^{-5} of the total population. Based on this hypothesis, we tried to find an area where the ration of infection and removing was very high. We tried Hubei Province at the beginning, but it turned out to be very similar with the scope of the entire China. After very hard work of searching and comparing, we thought Italy

[1] http://www.nhc.gov.cn/xcs/yqtb/list_gzbd_10.shtml
[2] http://weekly.chinacdc.cn/news/TrackingtheEpidemic.htm

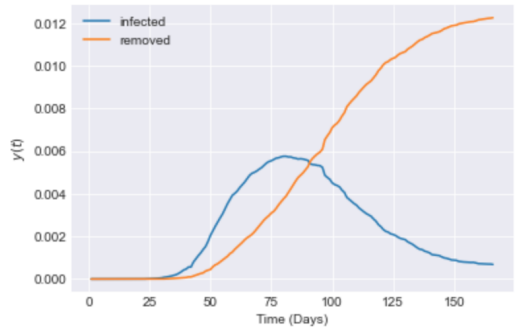

Fig. 3. Real data in Italy

might be the highest area with these two ratios. Therefore, we tried again with the data of Italy (see figure 3).

Nevertheless, we tried many different likelihood functions and sampling methods but didn't get a convergent and reasonable result of the posterior distribution. Figure 4 was one of the estimation results after 5000 times of sampling for each parameter. Hence, it is nearly impossible for us to carry out further analysis based on the estimation of parameters. Due to the time limit, this is so far of our experiment. In the next section, we will write about our discussion and reflection on our experiments.

V. Discussion and Limitations

From the experiments, we found that our methods can correctly estimate the parameters based on the data generated by SIR models, but all of them produce strange results when simulating real-world data. This section is our analysis.

A. the Limitation of the SIR Model

There is one very important assumption of the SIR model: the mass-action mixing of individuals. Even if we ignore the fact that infected people are separated from normal people as soon as diagnosed, people aren't quite "mixed together". For example, when we are estimating the epidemic in China, we took the whole Chinese population as the population of our model, which means that all individuals in China should be move freely and randomly every day. Even looking at this problem at the scope of Wuhan city, the epidemic was also concentrated in several areas. Most people in Wuhan even don't have a chance to meet any infected individuals in their daily life, which contradicts the assumption of the SIR model. (Thank God they don't have this chance, or otherwise We wouldn't have the chance of writing this report safely). Therefore, the susceptible is in fact, lower, instead of being all the population in the city, or the country.

Another important factor is that β, ν and R_0 are viewed as constants in a standard SIR model. However, in reality, β and ν are always changing. At least, the β at the beginning of the epidemic and the β at the middle of it should be different, because uninfected people started to wear masks and infected people were quarantined, which would reduce the infection

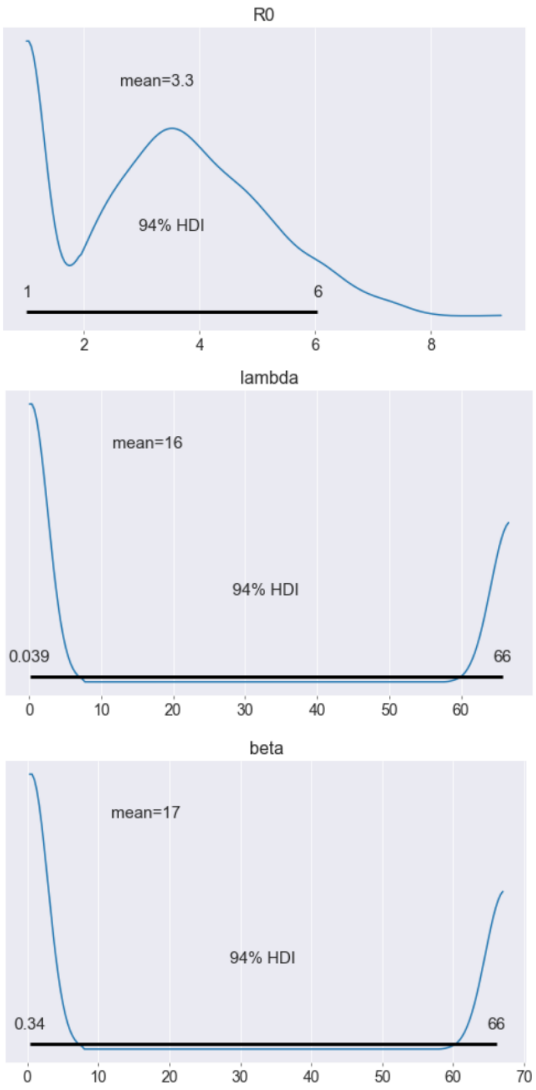

Fig. 4. One of the estimation results

rate. Therefore, we may not find a single pair of β and ν suitable for the whole epidemic.

Mathematically, we can see that the SIR model is strongly deviated. We focus on the cusp of the infected in Italy, the model gives $dI/dt = 0$ and therefore

$$\beta SI - \nu I = 0, \quad S = \frac{\nu}{\beta} \equiv \frac{1}{R_0}$$

is satisfied at the cusp of the infected. Checking the data and we can find that $S \approx 0.99$, making $R_0 \approx 1.01$ and this is absurd. The result would be that this epidemic happens so slowly that on a scale of 100days, the curves would seem linear.(We therefore would like to question those who have totally trusted the data and implemented the simulation using the standard SIR and having acquired some "perfect" results of the parameters.)

B. the Limitation of the Data

We got our data from WHO and all these data are official. However, we believe that the Chinese data is not so accurate. On February 13, there was a sharp growth on the infected population due to the allowance to diagnose based on symptoms, when before, they were only allowed to make diagnosis based on nucleic acid tests. There were 13332 people diagnosed based on their symptoms, but there were only 1820 people diagnosed based on nucleic acid tests. This creates a sudden change in the data and it has become difficult to identify the real case number, because the ratio before Feb 13 is quite different from that after Feb 13.

Moreover, usually the initial data are not convincing since the statistics are not mature enough. For example, the initial infected people in Italy is 2. This can be misleading due to lack of hospital inspections. Hence the error can be high. It is best if we only use later statistics since the number is higher and hence the error is reduced.

C. the Flaws of the Likelihood Function

As we have stated in the previous part, our likelihood function isn't strictly proven to be valid. Hence, it's reasonable to question the reliability of our likelihood function. However, we predicted the parameters of the generated data quite accurately using this likelihood function. If it is completely wrong, then we won't achieve any results in the first place. What's more, even we tried the likelihood function from the documentation of Pymc3, it didn't produce a reasonable result, either. However this is only a possibility of it being the bug, since it worked well with the artificial data.

VI. Breaking the Limitations

In the section we will push a little bit forward from the classical SIR model, adjusting it from the perspectives provided in section V. We will use the data of Italy throughout the section.

A. Data$^+$

We solidify our guesses and present a manual simulation of Italy's data. The smooth curve in Fig. 5. is the mathematical SIR model with manually modified parameters. We may believe this to be a quite good simulation **and now we analyze the reasons why our trials before were imperfect through a reversed logic**, by using this model. In this model the parameter R_0 is around 4, which seems just fine. Now, we need to compute the sample of the REAL susceptible. The data in Italy shows that the cusp of the infected has the value 0.57%, which seems a little bit too low.(Even now in the USA the ratio is above $1/7$.) Instead our manually simulation shows that over 20% population are infected at the cusp.(Here we use the data at the cusp because the data at early stages are unreliable.) Given the population of Italy is 60 million, the susceptible with a reduced factor, becomes less than 1.71 million.

Apart from the change in the size of the susceptible, the model also imported a correction called the incubation period and I set this to be 1 under the unit of $\beta = 1.3, \nu = 0.3$.

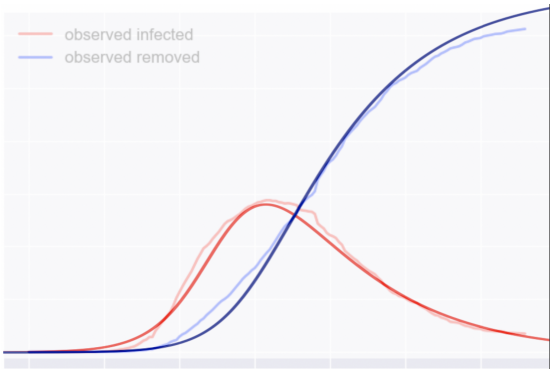

Fig. 5. Manual Simulation for Italy

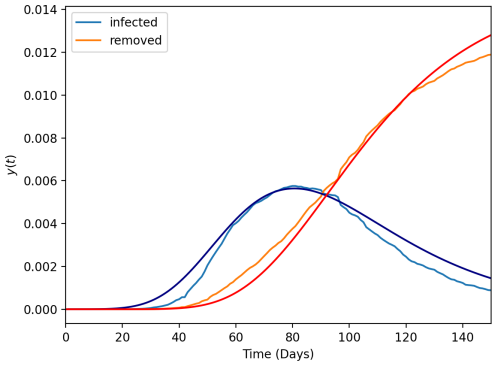

Fig. 6. SIR$^+$ simulation for Italy

This resulted in the blue curve having been shifted towards the right and it seems fit (see figure 5).

B. SIR$^+$

The last Data$^+$ model is still an SIR model, but with funny results. For example, the over-20% population being infected. This is because although the shapes are similar, their values are quite different. It is not convincing enough if we construct this curve without a proper reasoning. But now we may choose to believe Italy's data and change the SIR model as follows to acquire a set of curves similar to the real data in both shapes and values:

1). As discussed before, the parameters change, especially for β. Because people begin to attach importance to the epidemic, β decreases and we now set $\beta = \beta_0 e^{-\alpha t}$, where α is undetermined.

2). The average incubation period is around 14 days, and therefore is added. Now we need to find the value of α. Since the SIR model is modified to be:

$$\frac{dI}{dt} = \beta SI - \nu I = \beta_0 e^{-\alpha t} SI - \nu I,$$

we have $dI/dt = 0$ at the cusp of the infected. This gives

$$\beta_0 e^{-\alpha t} = \frac{\nu}{\beta} \equiv \frac{1}{R_0}, \quad \alpha = \frac{\ln R_0}{t}$$

Since we already know that the cusp of the infected of Italy happens on the 81-th day, $t = 81$. Therefore α is known if β, ν are given, which means the degree of freedom does not increase. We have a rough simulation (see figure 6) with the parameters $R_0 = 15.25, \nu = 0.032$. The reason why R_0 is higher than other result of SIR model is that, R_0 is a decreasing function of time but in SIR model, it is presented as the mean value of $R_0(t)$. In SIR$^+$, this is the maximum of $R_0(t)$, and the decreasing factor $e^{-\alpha t}$ represents the change in the importance attached by the people.

The benefit of this model is that the outcome matches the data of Italy so close and the units are unified. Hopefully and expectantly, this is a promising model describing the COVID-19.

ACKNOWLEDGMENT

The authors would like to thank Prof. Xianbin Gu, who has helped us all through this project. He provided us with many insights and guided us to our first academic research, without whom this project would by no means have these progress.

We also would like to appreciate the Dean's Undergraduate Research Fund (DURF) of NYUSH, which provided us with the initiative and funding for us to carry out this research.

REFERENCES

[1] H. Weiss, "The SIR model and the Foundations of Public Health," *Materials matemàtics*, vol. 2013, pp. 17, 2013.

[2] F. Zhang, L. Li, and H. Xuan, "Survey of transmission models of infectious diseases," (In Chinese) *Systems Engineering —Theory & Practice*, vol. 31, no. 9, pp.1736–1744, Jan. 2011.

[3] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in Python using PyMC3," *PeerJ Computer Science* (April): e55. doi:10.7717/peerj-cs.55.