Supplement: Predicting Long-term Evolution of COVID-19 by On-going Data using Bayesian Susceptible-Infected-Removed Model

Shohei Hidaka
Japan Advanced Institute of Science and Technology (JAIST)

May 9, 2020

1 Probabilistic reformulation of Susceptible-Infected-Removed model

The SIR model (1) is a deterministic model based on the set of ordinary differential equations. The deterministic SIR, *DSIR* in short, is quite useful to grasp the essential quantitative nature of an epidemic dynamics. It is, however, not ready to use for analysis of an empirical data of an epidemic dynamics with noise, especially when a large body of the characteristics of the epidemic nature for the following reasons.

First, DSIR does not offer a systematic treatment for statistical noise which any empirical data inevitably contains.

Second, DSIR has three essential parameters (N, β, γ) and the variables $(S(t), I(t), R(t))$, but only effectively observeable variable is the cumulative number of infectious individuals $F(t) := N - S(t)$. The effective number of susceptive individuals $S(t)$ and the effective whole population N are neither observeable directly, but its difference $F(t)$ can be (partially) observeable. The whole population size N as well as $S(0)$ does not necessarily match with a whole in a country nor region. As the size of an effective social network of interest, in which an infectious individual can move over, can depend on various factors such as individual behaviors, govermental policy, custom, and culture, it is not easy to estimate alone. The infectious individuals $I(t)$ at a certain time point t can be partially observeable as a report on new increase in the cumulative number $F(t)$, but it also depends on unknown parameters γ which reflects the average recovery (removal) rate.

Third, empirical dynamics of an epidemic cannot be necessarily explaineable by a epidemic model alone, due to social and political means taken by a community under the threat of epidemics. For example, a lockdown, which shut down a major body of social activity in a community or a whole country, may have a large impact in a epidemic dynamics – the latent parameters (N, β, γ) may be changed by the policy. Thus, it is preferable to constructe a model, which allows a dynamic change or mixture in the system parameters.

Given these problems of DSIR, we propose a novel modeling framework based on Bayes statistical model. We call this *Bayesian SIR (BSIR)* model as a remedy for it. BSIR is a probabilistic reformulation of DSIR, in which the three state variables (S_t, I_t, R_t) are defined as discrete variables over discrete time $t = 0, 1, \ldots$. BSIR is designed to offer a statistical estimator of the unknown parameters (N, β, γ) as well as unobserveable variables $(S(t), I(t), R(t))$ only from the observation of a short time series of cumulative number of infectious individuals $F = (F_1, F_2, \ldots, F_T)$. Given a set of estimated parameters, BSIR also offers a generative model to predict a timeseries of $F_{T+1}, F_{T+2}, \ldots$ in future. If we identify the difference equations of the expectation of the state variables in BSIR as continuous state variables over continuous time in DSIR, BSIR is a faithful statistical replication of the original DSIR. Thus, an estimate of the parameters (N, β, γ) in BSIR is interpretable the parameters in DSIR as well, up to a time constant.

2 Deterministic SIR model

The deterministic SIR model is the set of the following differential equations, with a $(S(0), I(0), R(0))$ constrained by $S(0) + I(0) + R(0) = N$.

$$\begin{cases} \frac{\mathrm{d}S}{\mathrm{d}t} = & -\frac{\beta}{N} IS \\ \frac{\mathrm{d}I}{\mathrm{d}t} = & \frac{\beta}{N} IS - \gamma I \ , \\ \frac{\mathrm{d}R}{\mathrm{d}t} = & -\gamma I \end{cases} \tag{1}$$

where

- $S(t)$ is a quantity indicating susceptive population size,
- $I(t)$ is a quantity indicating infectious population size,
- $S(t)$ is a quantity indicating removed population size,
- N is the system parameter indicating the whole population size $S(t) + I(t) + R(t) = N$ at any time t,

- β is the system parameter indicating the rate of infection that each infectious individual produces a new infectious individual,
- γ is the system parameter indicating the rate of recovery or removal of infected individuals from the infectious group.

$R_0 := \frac{\beta}{\gamma}$ is called *basic reproduction number*, which is the critical parameter characterizing the SIR dynamics. The SIR dynamics is a "outbreak", which means an broad spread of infection over the given population at the initial state, if

$$R_0 = \frac{\beta}{\gamma} > \frac{N}{S(0)},$$

which implies an increase of the infectious population size with $\frac{\mathrm{d}I}{\mathrm{d}t} > 0$ at $t = 0$. Otherwise, the infectious population decrease to zero, and it shows non-outbreak dynamics. Thus, the basic reproduction number R_0 as well as the thresholding parameters $\frac{N}{S(t)}$ are the critical variables, with respect to the evolution of the epidemic. Although $\frac{N}{S(0)} \approx 1$ is often assumed at initial state as $S(0)$ is quite small at begining, both N and $S(t)$ need to be estimated on the middle of on-going epidemic in order to judge whether the epidemic size is increasing or decreasing.

3 Baysian SIR model

Comparable with the differential equation (1) of the state variables $(S(t), I(t), R(t)) \in \mathbb{R}^3$, Baysian SIR model has the three discrete system variables $(S_t, I_t, R_t) \in \mathbb{Z}^3$ over discrete time $t = 0, 1, \ldots$.

The Baysian SIR model is designed to capture the essential nature of the differential equation (1) by the difference equations for any $t = 1, 2, \ldots$

$$\begin{cases} \overline{S_t} - \overline{S_{t-1}} &= -\frac{\overline{\beta}}{N}\overline{S_{t-1}}\ \overline{I_{t-1}} \\ \overline{I_t} - \overline{I_{t-1}} &= \frac{\overline{\beta}}{N}\overline{S_{t-1}}\ \overline{I_{t-1}} - \gamma\overline{I_{t-1}}\ , \\ \overline{R_t} - \overline{R_{t-1}} &= \overline{\gamma}\overline{I_{t-1}} \end{cases} \tag{2}$$

where $\overline{X}$ denotes the expectation of random variable X. BSIR consists of these three random variables as well as auxilary random variables

$$F_t = N - S_t \tag{3}$$
$$D_t = F_t - F_{t-1}, \tag{4}$$

where F_t and D_t are only observeable variables indicating the cumulative number of infected individuals and the difference between its two consequtive time points, available as a data respectively. As a series $F = (F_0, F_1, \ldots, F_T)$ is supposed given as a observeable data, a dependent variable such as $S_t = N - F_t$ is replaced with $N - F_t$. As the number of removed individuals R_t is a passive variable which does not affect the other variables, we do not explicitly model this variable in this formulaiton. Then the remaining variables explicitly treated includes N, F_t, D_t, I_t.

BSIR is a generative model, meaning that it probabilisticly generates a data of timeseries of F_t with a set of parameters and initial variables. BSIR has the following initial variables:

- the whole population size $N \in \mathbb{N}$,
- the rate of person-to-person contact $\beta \in \mathbb{R}$,
- the rate of removal $\gamma \in [0, 1]$,
- the initial cumulative number of infectious cases $F_0 \in \mathbb{Z}$,
- the initial increase in number of infectious cases $D_0 \in \mathbb{Z}$,
- and the initial unobserved number of infectious cases $I_0 \in \mathbb{Z}$.

As the notational convention in this formulation of BSIR, we denote every integer-valued random variable by a capital letter N, F_t, D_t, I_t and others, and every real-valued random variable by a greek letter β, γ and others, and every set of variables by a capital greek letter. As an additional set of auxiliary random variables, which eases the statistical computation involving integral, we introduce the following variables.

- the number of non-removed infected individuals $H_t \in \mathbb{Z}$ from I_t.
- the rate of infection depending on the infected population I_t at time t: $\lambda_t \in \mathbb{N}$,
- the number of contacts $C_{t,i} \in \mathbb{Z}$, indicating i^{th} susceptible individuals contact other individuals at t.

Given an initial state variables (S_0, I_0, R_0) and the system parameter $\Theta = (N, \beta, \gamma)$ (as well as the auxilary variables F_0, D_0), a series of random variables $F_1, F_2, \ldots$ is generated in the following scheme. For any integer $t > 0$, the random variables $I_t, \lambda_t, C_{t,i}, D_t, F_t$ in this order are drawn as follows.

1. The latent cumulative number of infectious people $I_t \in \mathbb{Z}$ is decided by the random variable $H_{t-1} := I_t - D_{t-1} \in \mathbb{Z}$, which follows the binomial distribution $\mathrm{Bin}(H_{t-1} \mid I_{t-1}, (1-\gamma))$.

2. The latent rate of new infection $\lambda_t \in [0,1]$ follows the beta distribution $\mathrm{Beta}\left(\lambda_t \mid I_t, N - I_t\right)$.

3. The non-negative number of i^{th} individual's contacts at time t $C_{t,i} \in \mathbb{Z}$ follows the Poisson distribution $\mathrm{Po}\left(C_{t,i} \mid \beta\right)$.

4. The new reported number of infectious cases $D_t \in \mathbb{Z}$ follows the Poisson-binomial distribution $\mathrm{PB}(D_t \mid \theta_t)$, where

$$\theta_t = \left(1 - (1-\lambda_t)^{C_{t,i}}\right)_{i=1}^{N-F_{t-1}} \in [0,1]^{N-F_{t-1}}$$

   .

5. The cumulative number of infectious cases adds up the new reported number: $F_t = F_{t-1} + D_t$.

In the above, the probability mass/ density distribution function is defined by

$$\mathrm{Bin}(K \mid N, \theta) := \binom{N}{K} \theta^K (1-\theta)^{N-K},$$

$$\mathrm{Beta}(\theta \mid K_0, K_1) := \frac{\theta^{K_0 - 1}(1-\theta)^{K_1 - 1}}{\mathrm{B}\left(K_0, K_1\right)},$$

$$\mathrm{Po}(K \mid \theta) := \frac{\alpha^K}{K!} e^{-\alpha},$$

and

$$\mathrm{PB}(K \mid (\theta_1, \theta_2, \ldots, \theta_M)) := \sum_{\sum_{i=1}^M C_i = K} \prod_{i=1}^M \theta_i^{C_i} (1-\theta_i)^{1-C_i},$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the number of combinations to draw k from n, $\mathrm{B}\left(n_0, n_1\right) = \frac{\Gamma(n_0)\Gamma(n_1)}{\Gamma(n_0+n_1)}$ is the beta function, and $\Gamma\left(n\right)$ is the gamma function. In a special case, it holds $\Gamma\left(n\right) = (n-1)!$ for any non-negative integer n. By repeating this sampling scheme for $t = 1, 2, \ldots, T$, a time series of an arbitrary length T, $F_1, F_2, \ldots, F_T$ is generated with the initial parameters described above. In BSIR , the parameter N, β, γ are also random variables, and each of them has the prior distribution

$$P(N) \propto N^{-1} \delta(N \in \mathcal{N}), \tag{5}$$

$$P(\beta) = \text{Gamma}(\beta \mid n_0, \beta_0), \tag{6}$$

and

$$P(\gamma) = \text{Beta}(\gamma \mid \gamma_0, \gamma_0), \tag{7}$$

where

$$\text{Gamma}(\alpha \mid K, \theta) := \frac{\alpha^{K-1} e^{\frac{\alpha}{\theta}}}{\theta^K \Gamma(K)},$$

$$\delta(q) = \begin{cases} 1 & \text{if the proposition} q \text{ is true,} \\ 0 & \text{otherwise} \end{cases},$$

and $\mathcal{N} = \{N_{\min}, N_{\min}+1, \ldots, N_{\max}\}$ is the set of integers with $N_{\min} = F_T$ and $N_{\max}$ is the reported population in a given country or domestic region. We set $n_0 = \beta_0 = \gamma_0 = 1$. These prior distributions are designed to give little prior information on N, β and γ. For even a modestly large dataset, this choice of prior has little impact on an estimate of the parameters in BSIR .

In summary, Figure 1 shows the graphical model of BSIR illustrating the dependency among random variables. BSIR has three major components *removal and addition of infected individuals*, *infection rate*, *contact frequency* and *infection sample*, which gives a statistical implmentation of the spirit of the original DSIR model. In the removal and addition of infected individuals (corresponding the sampling process 1. above), a $\gamma \in [0, 1]$ portion of the number of infected individuals I_{t-1} is removed from the infectious pool, and the newly reported number of infected individuals is added to it. The expected number of infected individuals is

$$\overline{I_t} = (1-\gamma)\overline{I_{t-1}} + D_{t-1} = \overline{H_{t-1}} + D_{t-1}. \tag{8}$$

In infection rate (corresponding to the sampling process 2. above), the latent infection rate $\lambda_t \in [0, 1]$ is decided according to the number of infected individuals I_t and the whole population size N. The expected infection rate is

$$\overline{\lambda_t} = \frac{\overline{I_t}}{\overline{N}}. \tag{9}$$

In the contact frequency (corresponding to the sampling process 3. above), the contact frequencies $C_{t,i}$ for the i^{th} susceptible individual at time t is decided depending on the contact rate $\beta > 0$. This contact frequency is not explicit in DSIR. BSIR considers that the rate of contact β characterizes the susceptible individuals' behavioral tendency to contact other individuals, which is a factor deciding infection, which is only possible with at least

one contact between infected and susceptive individuals. The number of contacts $C_{t,i}$ is decided by the contact rate:

$$\overline{C_{t,i}} = \overline{\beta}. \tag{10}$$

In the infection sample (corresponding to the sampling process 4. above), the number of newly reported infected individuals D_t is decided according to the infection rate λ_t, the number of susceptible individuals $N - F_{t-1}$, and the contact frequency $C_{t,i}$ of each susceptible individuals. As the i^{th} susceptible individual has $C_{t,i}$ times of contacts other individuals (without knowing who is infected), the probability to contact at least one infected individual among $C_{t,i}$ contacts is $1-(1-\lambda_t)^{C_{t,i}}$. As this contact is mutually independent, the expected number of new infection is

$$\overline{D_t} = \overline{N} - \overline{F_{t-1}} - \sum_{i=1}^{\overline{N}-\overline{F_{t-1}}} \overline{(1-\lambda_t)^{C_{t,i}}}. \tag{11}$$

Lastly, the cumulative number of infected individuals adds up

$$\overline{F_t} = \overline{F_{t-1}} + \overline{D_t}. \tag{12}$$

In total, the expectation of these random variables (8), (9), (10), (11) and (12) reproduce a part of the difference equation (2), with

$$\overline{D_t} \approx (\overline{N} - \overline{F_{t-1}})\overline{\lambda_t}\ \overline{\beta} = \overline{S_{t-1}}\frac{\overline{I_t}}{\overline{N}}\overline{\beta}.$$

for a sufficiently small $\lambda_t \ll 1$.

3.1 Posterior distribution

We estimate the parameters in BSIR using Gibbs sampler of the posterior distribution $Q(\Theta_T \mid F, D)$ of

$$\Omega = (N, \beta, \gamma), \Theta_T = (I_1, \ldots, I_T, \lambda_1, \ldots, \lambda_T, C_{1,1}, \ldots, C_{i,T})$$

given time series $(F, D) = (F_0, \ldots, F_T, D_0, \ldots, D_T)$. Specifically the posterior probability is

$$\begin{aligned} Q(\Omega, \Theta_T | F, D) \quad &\propto \quad P(\Omega) \prod_{t=1}^{T} \mathrm{Bin}(H_t = I_{t+1} - D_t | I_t = H_{t-1} + D_{t-1}, (1-\gamma)) \\ &\times \quad \mathrm{Beta}(\lambda_t \mid I_t = H_{t-1} + D_{t-1}, N - I_t) \prod_{i=1}^{N-F_{t-1}} \mathrm{Po}(C_{t,i} \mid \beta) \\ &\times \quad \mathrm{PB}(D_t \mid \theta_t) \end{aligned} \tag{13}$$

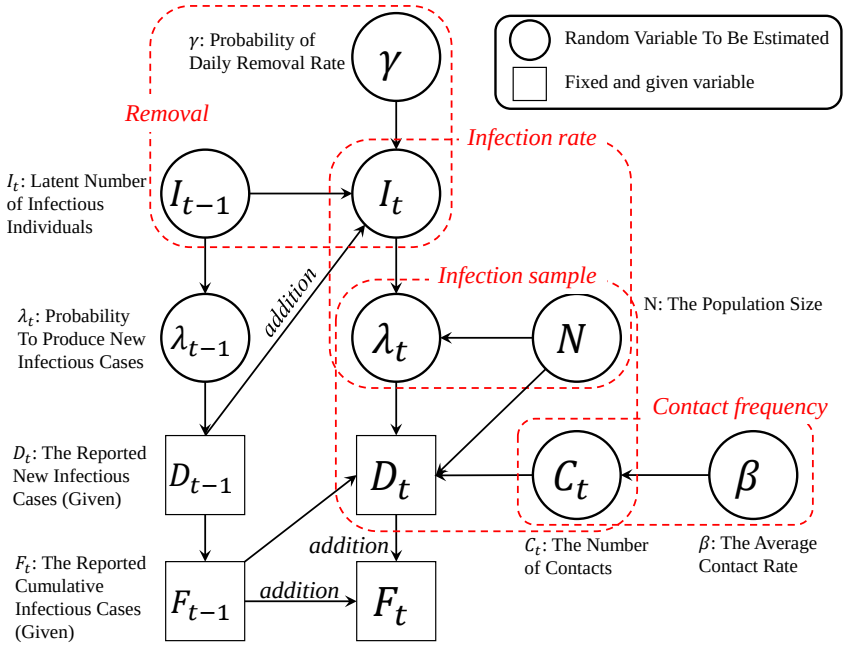

Figure 1: The hierarchical Bayesian generative model of the cumulative infected population size F_t and daily new reported number of infected individuals D_t, based on the whole population size N, daily removal probability γ, the latent cumulative number of infected individuals I_t, the rate of infection λ_t at each time step t in the form of the mutual dependence depicted by this graphical model. The random variables depicted by the circles are random variables to be estimated, and those depicted by the squares are given and fixed in estimation.

where $P(\Omega) = P(N)P(\beta)P(\gamma)$ is the product of the prior probability defined by (5), (6), (7), and

$$\theta_t = (1 - (1 - \lambda_t)^{C_{1,t}}, \ldots, 1 - (1 - \lambda_t)^{C_{N-F_{t-1},t}}) \in [0, 1]^{N-F_{t-1}}.$$

Explicitly writing down, the posterior probability is

$$\begin{aligned} Q(\Omega,\Theta_T|F,D) &\propto P(\Omega)\prod_{t=1}^{T}\binom{H_{t-1}+D_{t-1}}{H_t}(1-\gamma)^{H_t}\gamma^{D_{t-1}+H_{t-1}-H_t} \\ &\times \frac{\lambda_t^{H_{t-1}+D_{t-1}-1}(1-\lambda_t)^{N-(H_{t-1}+D_{t-1})-1}}{\mathrm{B}\left(H_{t-1}+D_{t-1},N-(H_{t-1}+D_{t-1})\right)}\prod_{i=1}^{N-F_{t-1}}\frac{\beta^{C_{t,i}}}{(C_{t,i})!}e^{-\beta} \\ &\times \sum_{D_t=\sum_{i=1}^{N-F_{t-1}}Z_{t,i}}\prod_{i=1}^{N-F_{t-1}}\theta_{t,i}^{Z_{t,i}}(1-\theta_{t,i})^{1-Z_i}, \end{aligned} \tag{14}$$

where the variable $Z_{t,i}\in\{0,1\}$ being 0/ 1 indicates non-infection/ infection of the i^{th} individual at t, and $\theta_{t,i}:=1-(1-\gamma)^{C_{t,i}}\in[0,1]$ is the probability for the i^{th} individual to be infected at t.

Denote the sum of the non-infected/ infected individuals by contacting $C_{t,i}$ times by $K_{t,k,j}:=\sum_{i=1}^{N-F_{t-1}}\delta(C_{t,i}=k\wedge Z_{t,i}=j)$, and the probability of non-infection and infection by

$$\mu_{t,k,0}=\frac{\beta^k}{k!}e^{-\beta}(1-\lambda_t)^k\in[0,1] \text{ and } \mu_{t,k,1}=\frac{\beta^k}{k!}e^{-\beta}\left(1-(1-\lambda_t)^k\right)\in[0,1].$$

Then we have identity

$$\sum_{D_t=\sum_{i=1}^{N-F_{t-1}}Z_i}\prod_{i=1}^{N-F_{t-1}}\frac{\beta^{C_{t,i}}}{(C_{t,i})!}e^{-\beta}\theta_{t,i}^{Z_i}(1-\theta_{t,i})^{1-Z_i}=(N-F_{t-1})!\prod_{k=0}^{\infty}\frac{\mu_{t,k,0}^{K_{t,k,0}}\mu_{t,k,1}^{K_{t,k,1}}}{(K_{t,k,0})!\,(K_{t,k,1})!} \tag{15}$$

$$=\binom{N-F_{t-1}}{D_t}\mu_{t,0}^{N-F_{t-1}-D_t}\mu_{t,1}^{D_t}\frac{(N-F_{t-1}-D_t)!}{\prod_{k=0}^{\infty}(K_{t,k,0})!}\prod_{k=0}^{\infty}\left(\frac{\mu_{t,k,0}}{\mu_{t,0}}\right)^{K_{t,k,0}}\frac{(D_t)!}{\prod_{k=0}^{\infty}(K_{t,k,1})!}\prod_{k=1}^{\infty}\left(\frac{\mu_{t,k,1}}{\mu_{t,1}}\right)^{K_{t,k,1}} \tag{16}$$

where $\mu_{t,0}:=\sum_{k=0}^{\infty}\mu_{t,k,0}^{K_{t,k,0}}=e^{-\beta\lambda_t}$ $\mu_{t,1}:=\sum_{k=0}^{\infty}\mu_{t,k,1}^{K_{t,k,1}}=1-e^{-\beta\lambda_t}$ Thus, summing out the random variables $K_{t,k,j}$, we have

$$\begin{aligned} Q(\Omega,\Theta_T|F,D) &\propto P(\Omega)\prod_{t=1}^{T}\binom{H_{t-1}+D_{t-1}}{H_t}(1-\gamma)^{H_t}\gamma^{D_{t-1}+H_{t-1}-H_t} \\ &\times \frac{\lambda_t^{H_{t-1}+D_{t-1}-1}(1-\lambda_t)^{N-(H_{t-1}+D_{t-1})-1}}{\mathrm{B}\left(H_{t-1}+D_{t-1},N-(H_{t-1}+D_{t-1})\right)} \\ &\times \binom{N-F_{t-1}}{D_t}e^{-\beta\lambda_t(N-F_{t-1}-D_t)}(1-e^{-\beta\lambda_t})^{D_t} \end{aligned} \tag{17}$$

3.2 Estimation: Gibbs sampler

We employed Gibbs sampler to draw a sample from the posterior distribution $Q(\Theta, \Omega_T \mid F, T)$ by conditioning with each variable in Θ, Ω. Denote the set of all variables but X by $\Theta_{-X} : \{X \in \Theta \cup \Omega_T \cup \{F, D\} \setminus \{X\}\}$. The conditional posterior distribution of γ is

$$Q(\gamma \mid \Theta_{-\gamma}) = \text{Beta}\left(\gamma \mid \gamma_0 + \sum_{t=1}^{T} \left(D_{t-1} + H_{t-1} - H_t\right), \gamma_0 + \sum_{t=1}^{T} H_t\right). \quad (18)$$

The conditional posterior distribution of β is

$$Q(\beta \mid \Theta_{-\beta}) \propto \text{Gamma}\left(\beta \mid n_0 + \sum_{t=1}^{T}\sum_{k=1}^{\infty}\sum_{j=0,1} kK_{t,j,k}, \left(\beta_0^{-1} + \sum_{t=1}^{T}(N - F_{t-1})\right)^{-1}\right). \quad (19)$$

The conditional posterior distribution of λ_t is

$$Q(\lambda_t \mid \Theta_{-\lambda_t}) \propto \lambda_t^{I_t - 1}(1-\lambda_t)^{N - I_t - 1 + \sum_{k=0}^{\infty} kK_{t,k,0}} \prod_{k=1}^{\infty}\left(1 - (1-\lambda_t)^k\right)^{K_{t,k,1}}.$$

The conditional posterior distribution of N is

$$Q(N) \propto P(N)\prod_{t=1}^{T} \frac{\lambda_t^{H_t + D_t - 1}(1-\lambda_t)^{N - (H_t + D_t) - 1}}{\text{B}\left(H_t + D_t, N - (H_t + D_t)\right)} \binom{N - F_{t-1}}{D_t} e^{-\beta\lambda_t(N - F_{t-1} - D_t)}\left(1 - e^{-\beta\lambda_t}\right)^{D_t} \quad (20)$$

$$\propto P(N)\prod_{t=1}^{T} \frac{\left\{(1-\lambda_t)e^{-\beta\lambda_t}\right\}^N \Gamma(N)\,\Gamma(N - F_{t-1} + 1)}{\Gamma\left(N - (H_t + D_t)\right)\Gamma(N - F_{t-1} - D_t + 1)}. \quad (21)$$

For $\max(0, H_{t+1} - D_t) \le H_t \le H_{t-1} + D_{t-1}$, The conditional posterior distribution of H_t is

$$Q(H_t | F, D) \propto \frac{\left(\frac{\lambda_{t+1}(1-\gamma)}{(1-\lambda_{t+1})\gamma^{\delta(t=T)}}\right)^{H_t}\left(\frac{\lambda_{t+1}}{(1-\lambda_{t+1})}\right)^{H_t\delta(t<T)}\binom{H_t + D_t}{H_{t+1}}^{\delta(t<T)}}{\Gamma(H_t + D_t)\,\Gamma\left(N - (H_t + D_t)\right)\Gamma(H_t + 1)\,\Gamma(H_{t-1} + D_{t-1} - H_t + 1)}.$$

3.3 Special case $\beta = 1$

Although the full model described in the previous section is attractive to estimate all the parameters of BSIR, it is quite technically dense to sample from its posterior distribution. As an approximation to the full model, we also offer a computationally-light special case of the full model by fixing $\beta =$

1. With this additional assumption, we can marginalize out the continuous random variables γ and λ_t, which both follows the beta distributions, and that make the sampling drastically simple. In this case, the conditional posterior distribution for N and H_t are as follows.

$$\begin{aligned} Q(H_t|N, H_{T,-t}F, D) &\propto B\left(1 + \sum_{t=1}^{T} H_t, 1 + F_{T-1} + H_0 - H_T\right) \\ &\times \delta(H_t \in \mathcal{H}_t)\binom{I_{t-1} + H_{t-1}}{H_t}\binom{I_t + H_t}{H_{t+1}}^{\delta(t<T)} \\ &\times \frac{B\left(I_t + H_t, 2(N - F_{t-1}) - H_t - I_t\right)}{B\left(H_t, N - F_{t-1} - H_t\right)}, \end{aligned} \tag{22}$$

and

$$\begin{aligned} Q(N|H_T, F, D) &\propto P(N)\prod_{t=1}^{T}\binom{N - F_{t-1}}{I_t} \\ &\times \frac{B\left(I_t + H_t, 2(N - F_{t-1}) - H_t - I_t\right)}{B\left(H_t, N - F_{t-1} - H_t\right)} \end{aligned}$$

With the conditional posterior probabilities above, we estimate them specifically by the following algorithm. For initial parameter Ω'_T, we resorted a grid search over the parameter space $(N, \alpha) \in \mathcal{N} \times [0, 1]$ to generate H_T and tentatively maximize the posterior distribution $Q(\Omega_T|F, T)$. Then, we repeated two sampling scheme alternatively: (1) Given N and $H_{T,-t} := (H_0, \ldots, H_{t-1}, H_{t+1}, \ldots, H_T)$, sample H_t from the marginalized posterior distribution $Q(H_t|N, H_{T,-1})$ for every $t = 1, \ldots, T$. (2) Given N, sample N from the marginalized posterior distribution $Q(N|H_T)$. We found this sampling scheme typically mixed quickly, and we discarded the first 100 sample of Ω'_T as a burn-in period, and used the rest of 1000 samples of it for statistical estimation of the posterior distribution.

4 Numerical validation: Bayesian estimator vs linear regression

To test our model, we simulated by generating 100 synthetic timeseries datasets following the BSIR model, and applied the Bayesian estimator proposed in the previous section. In this test, we fixed the parameters to $N = 10^6, \beta = 1, \gamma = 0.4$, and changed the sample size (observed cumulative number of infected individuals). The estimator is supposed to capture

the true parameters within its credible interval (the counterpart concept for the confident interval in Bayesian statistics). Figure 2 showed the median, mean, mode (MAP) estimate of N as well as the averaged 95% credible interval based on the 1000 samples drawn from the posterior distribution of N.

As expected, even with a small sample size such as 1% observed data out of of N, the median estimate of N is close to the true N. Although the credible interval is quite large for a small sample size, the bias of the estimator is reasonably small. None of the estimates of the credible intervals missed the true parameter. Since the posterior distribution of N is quite skewed and has a long tail in general, the maximum a posteriori (MAP) estimate tended to underestimate the parameter N. Similarly, the mean estimate tended to overestimate it. Therefore, we choosed the median estimate of the parameters, for a point estimate of the BSIR model.

For comparison, we also estimated the whole population size N by a linear regression of the cumulative-to-difference (CD) relationship in data $(F_t, \frac{F_{t+1}-F_t}{F_{t-1}})$ (see also Section 6). Since the CD ratio $\frac{F_{t+1}-F_t}{F_{t-1}}$ is supposed to approach a first order polynomial function of F_t, this linear regression is one of the simplest ways to estimate the final outbreak size, which can be a proxy of the whole population N. The estimated N by the linear regression was showed as the broken line in Figure 2. The estimated N by the linear regression is quite poor – at best, it is just correlated to the observed sample size. Thus, this result suggests that linear-regression estimate cannot be effective with noisy data, even though the series of the CD ratio $\left(F_t, \frac{F_{t+1}-F_t}{F_{t-1}}\right)$ approachs to an asymptotic "line" in theory.

Moreover, this naive linear-regression estimator did not work for most of datasets. Since the linear regression is not constrained to give a set of regression coefficients, which is "valid" to esimate N, most of estimates based on the linear regression was "invalid". A valid estimate of N is neither infinite (in case the slope of the line is non-nengative) nor smaller than the observed cumulative number of infected individuals $\max(F)$. The probability of "valid" estimate provided by the linear regression is shown at each point in the broken line in Figure 2.

This osbservation is quite suggestive – it is important to develop a reliable statistical estimator in order to utilize the information underlying the cumulative number of infected individuals.

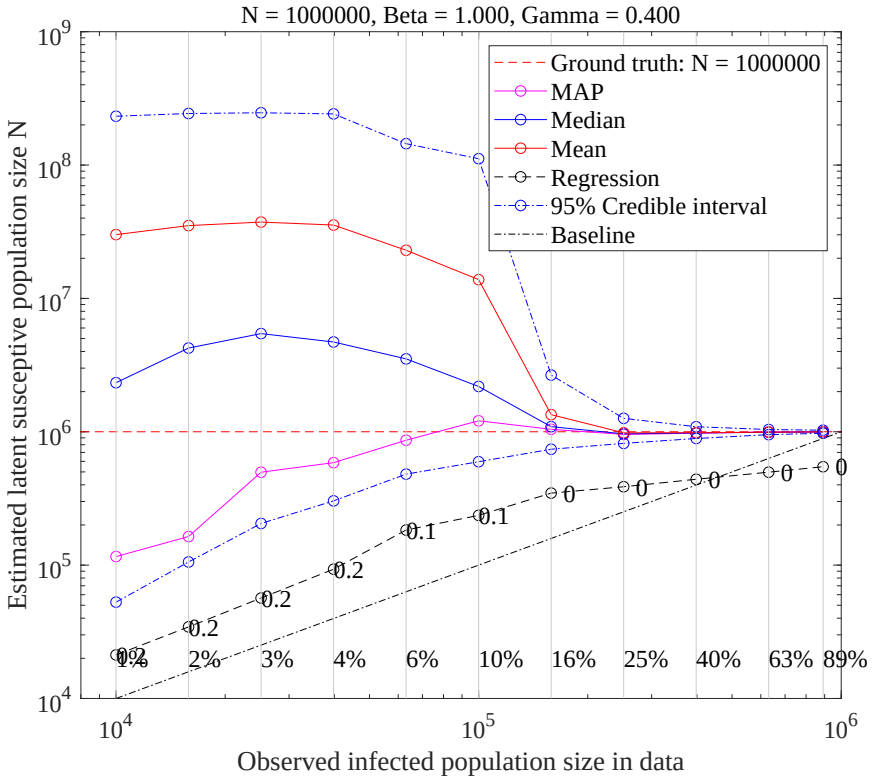

Figure 2: The validation of the proposed estimator for the BSIR. We fixed the parameters to $N = 10^6, \beta = 1, \gamma = 0.4$, and changed the sample size (observed cumulative number of infected individuals). The estimated population size N is converged into the true value after the sample size $\max(F)$ is 16 of N.

5 Simulated analysis with empirical data

Next, the predictive accuracy was analyzed more systematically for US and Italy data using the day with more than 100 infected cases for each country. For each country, we performed three different types of analyses: (1) Prediction of the cumulative number of infected cases on each date up to April 22nd, using each of datasets before the date of prediction target. If the target date is March 31st, for example, prediction using the data up to March 12th, that using the data up to March 13th, and so forth until March 30th, were calculated and take their geometric mean to give the grand prediction on March 31st. (2) Prediction of the newly reported number of infected

cases on each date up to April 22nd. We did the same procedure for this as well. (3) Prediction of the particular date April 22nd was targetted, and each dataset up to each date before then was used to make prediction. The first two analyses tested the bias and variance of prediction averaged across multiple days before the targetted day. The third analysis tested how far in future the prediction is successful. For both US and Italy, their general trend in these three types of analyses are similar. The first two analyses on prediction of cumulative and new infectious cases in both countries showed predictions showed small bias up to three weeks to a month away from the observed time point. The third analysis gives consistent results for both countries – prediction was reasonably accurate using the dataset of three weeks before the targetted date. In summary, these analyses suggest that the proposed method is likely to give a reliable prediction up to a month long future.

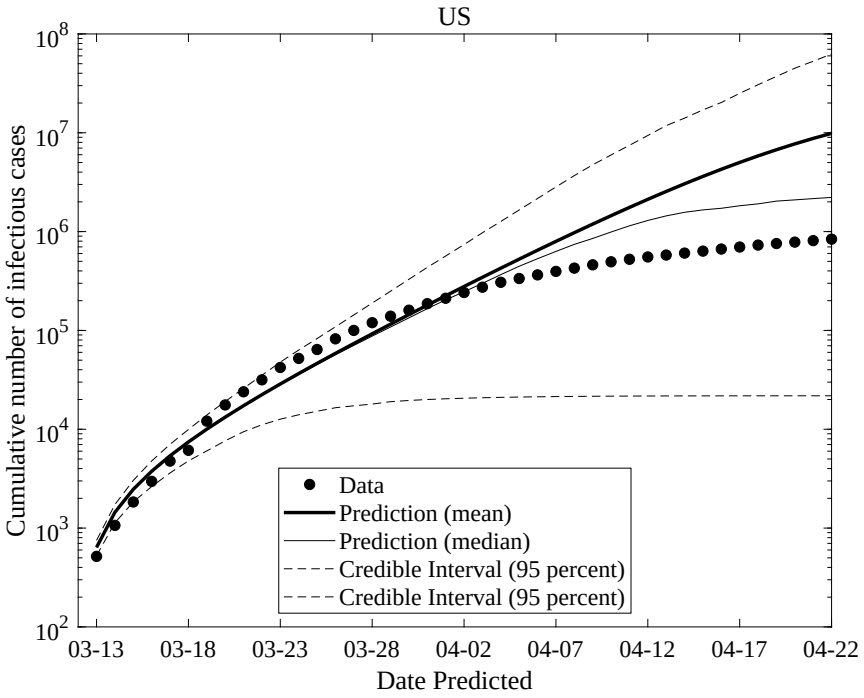

Figure 3: Prediction of the cumulative number of infected individuals in the US (from March 13th to April 22nd). The prediction of date X is the geometric mean of all prediction from March 12th to the date $X - 1$.

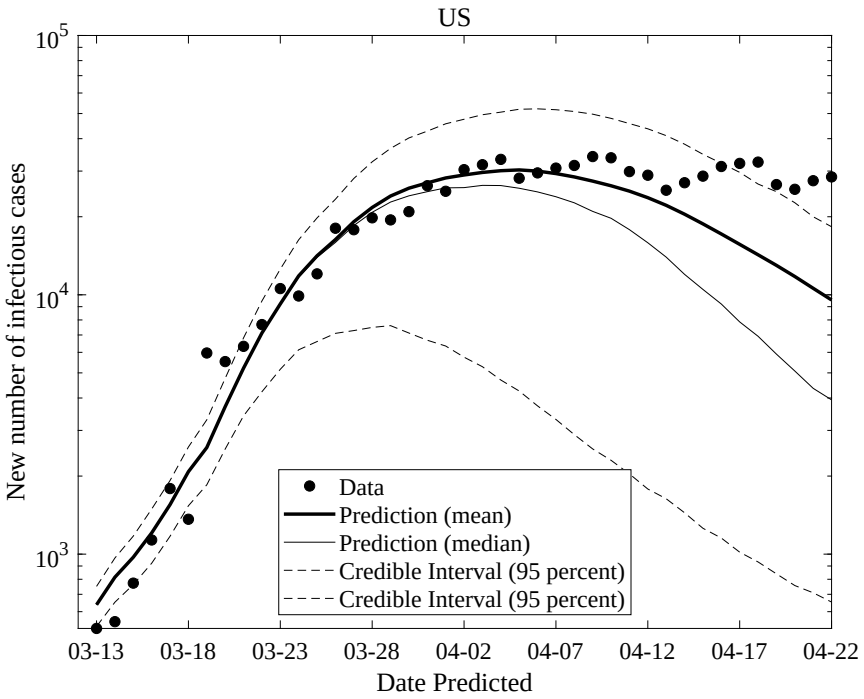

Figure 4: Prediction of the daily new reported number of infected individuals in the US (from March 13th to April 22nd). The prediction of date X is the geometric mean of all prediction from March 12th to the date $X-1$.

6 Cumulative-to-difference identity for the expected values in BSIR

In this section, all random variables X are treated by its expected value $\overline{X}$, and thus we simplify the notation of expected value of random variable X by just writing it X. In our paper, we provided the identity for the cumulative number of individuals F_t and the system parameters N, β, γ:

$$\overline{F_{t+1}} - \overline{F_t} = \frac{\overline{\beta}}{N}\left(\overline{N} - \overline{F_t}\right) g_{t-1}(\gamma, F).$$

This identity can be derived by (2) as follows.

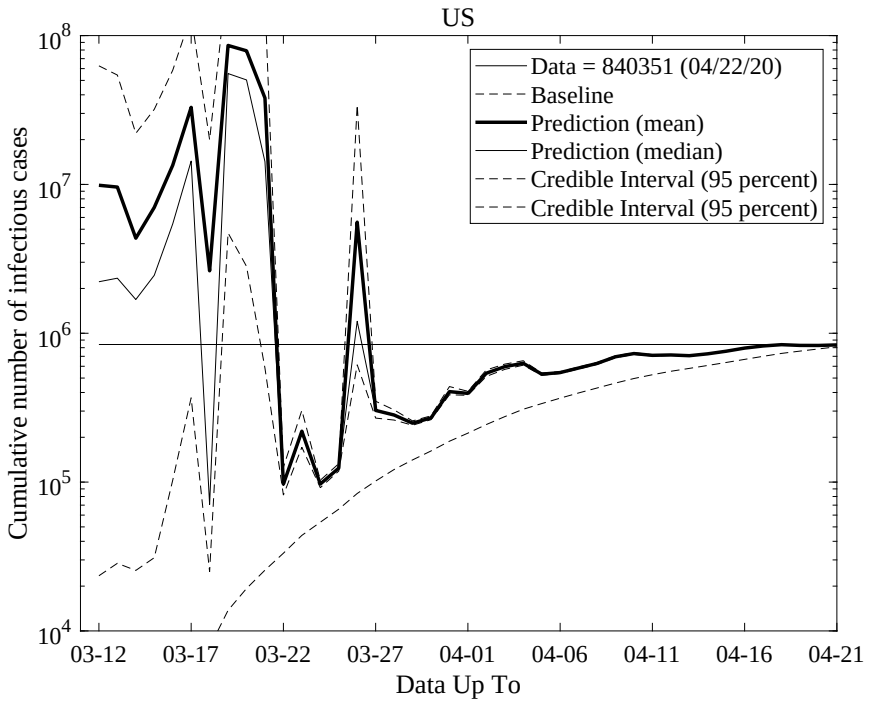

Figure 5: Prediction of the cumulative number of infected individuals in US on April 22nd. Each prediction was based on the data up to March 12th, March 13th, .., April 21st.

Lemma 1. *If the difference equation* (2) *of the expected values holds*

$$\begin{cases} S_t - S_{t-1} & = -\frac{\beta}{N} S_{t-1}\ I_{t-1} \\ I_t - I_{t-1} & = \frac{\beta}{N} S_{t-1}\ I_{t-1} - \gamma I_{t-1} \\ R_t - R_{t-1} & = \gamma I_{t-1}, \end{cases}$$

for any $t = 1, 2, \ldots$, *then we have*

$$F_{t+1} - F_t = \frac{\beta}{N} (N - F_t)\, g_{t-1}(\gamma, F), \tag{23}$$

where

$$g_{t-1}(\gamma, F) = F_{t-1} - \gamma \sum_{s=0}^{t-2} (1-\gamma)^{t-2-s} F_s$$

for any $t = 2, 3, \ldots$.

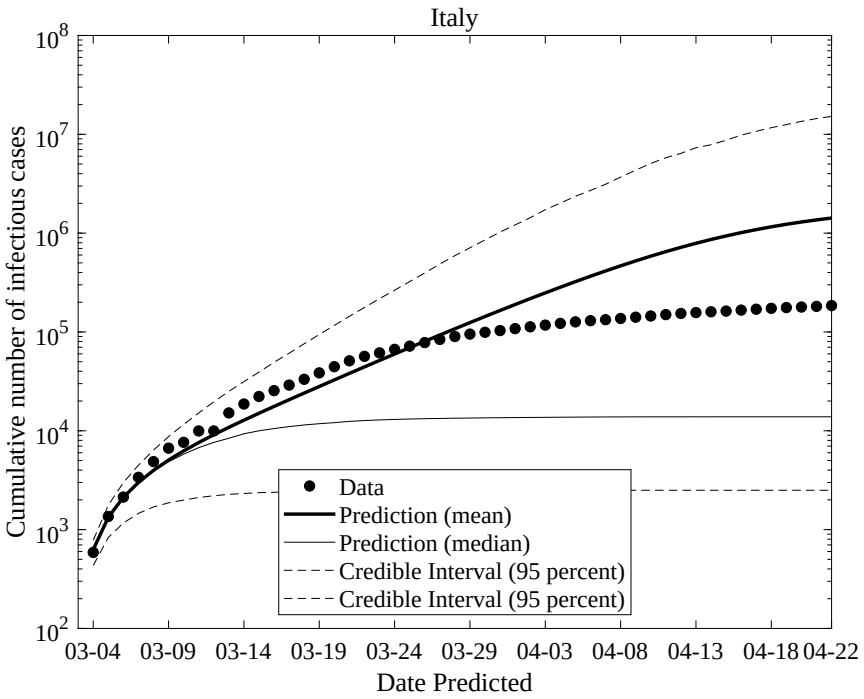

Figure 6: Prediction of the cumulative number of infected individuals in the Italy (from March 4th to April 22nd). The prediction of date X is the geometric mean of all prediction from March 3rd to the date $X-1$.

Proof. Let us define $\hat{\gamma} := 1-\gamma$, the vectors

$$I_{t'}^t := \begin{pmatrix} I_t \\ I_{t-1} \\ \vdots \\ I_{t'} \end{pmatrix},\ D_{t'}^t := \begin{pmatrix} D_{t-1} \\ D_{t-2} \\ \vdots \\ D_{t'} \end{pmatrix},\ S_{t'}^t := \begin{pmatrix} S_{t-1} \\ S_{t-2} \\ \vdots \\ S_{t'} \end{pmatrix},\ F_{t'}^t := \begin{pmatrix} F_{t-1} \\ F_{t-2} \\ \vdots \\ F_{t'} \end{pmatrix}, \quad (24)$$

and the matrices

$$A_t(a) := \begin{pmatrix} 1 & a & \dots & a^{t-1} \\ 0 & 1 & \dots & a^{t-2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{t\times t},\ B_t\left(\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_t \end{pmatrix}\right) := \begin{pmatrix} x_1 & x_2 & \dots & x_t \\ 0 & x_2 & \dots & x_t \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_t \end{pmatrix} \in \mathbb{R}^{t\times t} \quad (25)$$

for any $a, x_1, \dots, x_t \in \mathbb{R}$.

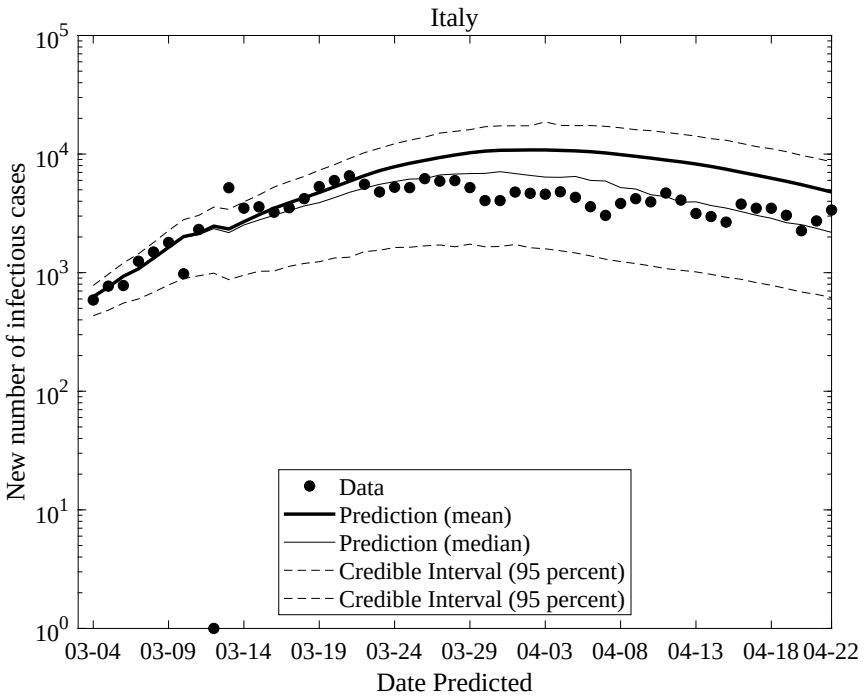

Figure 7: Prediction of the daily new reported number of infected individuals in the Italy (from March 4th to April 22nd). The prediction of date X is the geometric mean of all prediction from March 3rd to the date $X-1$.

By the difference equation (2), we have

$$D_t = S_{t-1} - S_t = \frac{\beta}{N} S_{t-1} I_{t-1}, \tag{26}$$

and

$$F_t = \sum_{s=0}^{t} D_s = \frac{\beta}{N} \left(S_0^{t-1}\right)^\top I_0^{t-1}, \tag{27}$$

where $x^\top$ denotes the transpose of the vector x.

By applying $I_t = D_{t-1} + \hat{\gamma} I_{t-1}$ recursively, we have

$$I_t = D_{t-1} + \hat{\gamma}(D_{t-2} + \hat{\gamma} I_{t-2}) = D_{t-1} + \hat{\gamma} D_{t-2} + \ldots + \hat{\gamma}^{t-1} D_0 + \hat{\gamma}^{t-1} I_0. \tag{28}$$

and thus, for any $t > 0$ we have

$$I_1^t = A_t\left(\hat{\gamma}\right) D_0^{t-1}. \tag{29}$$

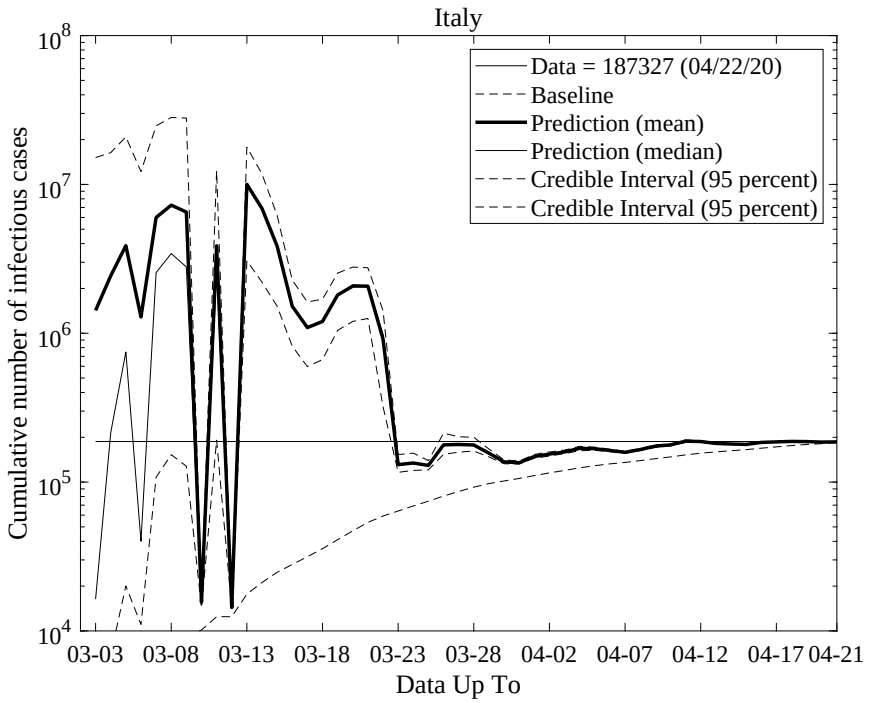

Figure 8: Prediction of the cumulative number of infected individuals in Italy on April 22nd. Each prediction was based on the data up to March 3th, March 13th, .., April 21st.

With (27) and (29), we have

$$F_{T+1} = \frac{\beta}{N}(S_1^T)^\top A_T(\hat{\gamma}) D_0^{T-1}. \tag{30}$$

C_t is the inverse matrix of $A_t(1)$, by defining

$$C_t := \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{t \times t},$$

we have

$$F_{T+1} = \frac{\beta}{N}(S_1^T)^\top A_T(\hat{\gamma}) C_T A_T(1) D_0^{T-1}. \tag{31}$$

As $A_T(1)D_0^{T-1} = F_0^{T-1}$,

$$F_{T+1} = \frac{\beta}{N}(S_1^T)^\top \begin{pmatrix} 1 & -\gamma & -\gamma\hat{\gamma} & \dots & -\gamma\hat{\gamma}^{T-2} \\ 0 & 1 & -\gamma & \dots & -\gamma\hat{\gamma}^{T-3} \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} F_0^{T-1}. \tag{32}$$

Take the differene $F_{T+1} - F_T$, we have

$$F_{T+1} - F_T = \frac{\beta}{N} S_T \begin{pmatrix} 1 & -\gamma & -\gamma\hat{\gamma} & \dots & -\gamma\hat{\gamma}^{T-3} \end{pmatrix} F_0^{T-1}. \tag{33}$$

By calculating the inter product of the vectors explicitly and replace $S_T = N - F_T$, we have (23):

$$F_{T+1} - F_T = \frac{\beta}{N}(N - F_T)\left(F_{T-1} - \gamma \sum_{t=0}^{T-2} (1-\gamma)^{T-2-t} F_t\right). \tag{34}$$

□