Stochastic Processes Report

Modelling and predicting Covid19 infections using the SIR model and MCMC sampling

By: Sourodeep Datta, 21CS10064
Tanishq Prasad, 21CS30054
Anit Mangal, 21CS10005
Anwesha Das, 21CS30007
Mihir Mallick, 21CS30031
Soumik Mandal, 21CS10063
Link to Repository: GitHub

Data Scraping:

The data was scraped from WorldoMeters, which keeps a copy of the data published by the Ministry of Health and Family Welfare. The data is a time series, starting from Feb 15, 2020, to March 22, 2023. The data consists of the Total Daily Cases, New Cases, Active Cases, Total Deaths, Daily Deaths, and New Recoveries. From there, the data used is the Active Cases (which is I), and the cumulative sum of the New Recoveries (which is R).

The SIR Model:

The SIR model is a compartmental model consisting of:

- S: The number of susceptible individuals. When a susceptible and an infectious individual come into "infectious contact", the susceptible individual contracts the disease and transitions to the infectious compartment.
- I: The number of infectious individuals. These are individuals who have been infected and are capable of infecting susceptible individuals.
- R: The number of removed (and immune) or deceased individuals. These are individuals who have been infected and have either recovered from the disease and entered the removed compartment or died. It is assumed that the number of deaths is negligible with respect to the total population.

The SIR system without vital dynamics (birth and death) can be described above can be expressed by the following system of ordinary differential equations:

$$\frac{\mathrm{dS}}{\mathrm{d}t} = -\frac{\beta I S}{N} \tag{1}$$

$$\frac{\mathrm{dI}}{\mathrm{d}t} = \frac{\beta I S}{N} - \gamma I \tag{2}$$

$$\frac{\mathrm{dR}}{\mathrm{d}t} = \gamma I \tag{3}$$

$$S(t) + I(t) + R(t) = N \tag{4}$$

$$R_0 = \frac{\beta}{\gamma} \tag{5}$$

Here, R_0 is the basic reproduction number, the expected number of total cases generated by one active case. In our model, the N = 1366417754, which was the population of India in 2019. We aim to determine the values of β and γ, given the data and some initial states, using a

Metropolis Hastings sampler.

As the value of N and S tends to be high, when considering the data for India, in order to prevent numerical instability, we made the following change:

$$s = \frac{S}{N}$$

This changes equations (1) and (2) to:

$$\frac{\mathrm{ds}}{\mathrm{d}t} = -\frac{\beta I s}{N} \tag{6}$$

$$\frac{\mathrm{dI}}{\mathrm{d}t} = \beta I s - \gamma I \tag{7}$$

Metropolis Hastings Algorithm:

The Metropolis Hastings algorithm is a Markov chain Monte Carlo (MCMC) method for obtaining a sequence of random samples from a probability distribution from which direct sampling is difficult. This sequence can be used to approximate the distribution, in this case, being used to find the posterior distributions of β and γ.

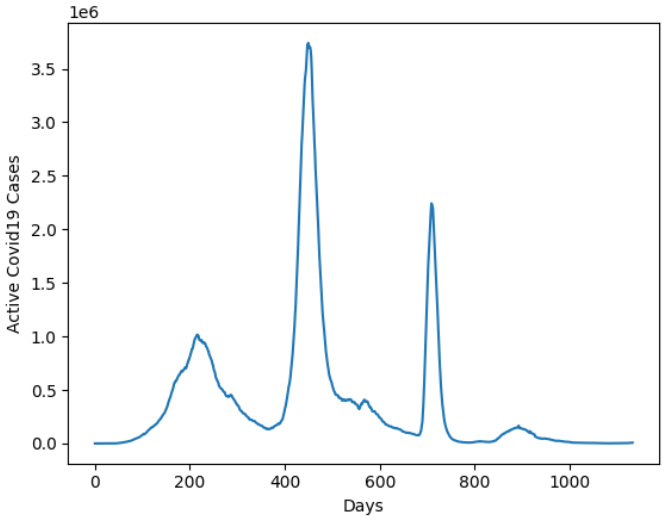
The algorithm for a symmetric proposal distribution (in our case a Normal Distribution) is:

1. Choose an initial starting point x_0 form prior distributions and a proposal distribution $g(x|y)$. The proposal distribution suggests the next candidate x given the previous sample value y. As $g(x|y)$ is assumed to be symmetrical, we can say $g(x|y) = g(y|x)$.
2. For each iteration t:
   (a) Sample a candidate x' from $g(x|y)$.
   (b) Calculate the acceptance ration $\alpha = \frac{f(x')}{f(x_t)}$.
   (c) Accept or reject the candidate by sampling u from a $U(0,1)$ distribution. If $u \leq \alpha$, accept it, otherwise reject it.

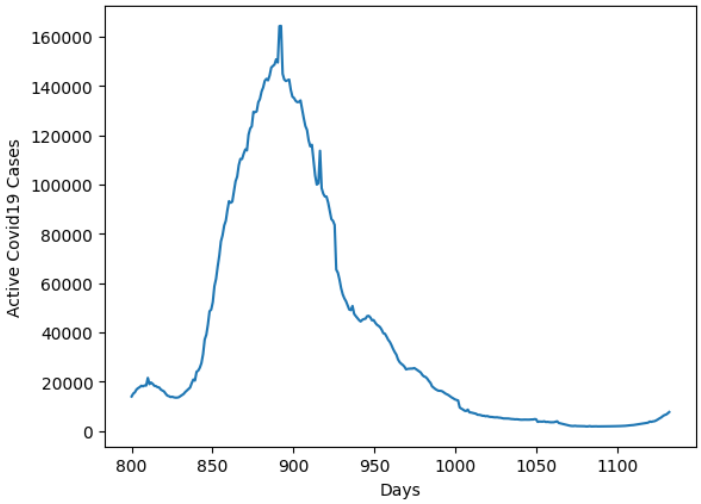
In our model, we use the DEMetropolisZ Algorithm, which is an extension of the Metropolis Hastings Algorithm.

Calculating β and γ:

The data collected is for a large period of time, and the three major covid waves are visible in it. The SIR model is meant for only a single wave, so we clip the first and second waves and only consider data gathered from day number 800 (Considering the start day to be 0).



(a) Active Cases v/s Days for all waves



(b) Active Cases v/s Days from day 800

Similarly, the data for S and R is also truncated. This is then used to make a hierarchical bayesian model, with the number of active cases being assumed to be a Poisson process, with means being dependent on the ODE, which in turn is dependent on beta and gamma, which are sampled from LogNormal prior distributions, with hyperparameters being based on empirical data.

$$\beta \sim LogNormal\,(log(0.4), 0.5)$$
$$\gamma \sim LogNormal\,(log(0.125), 0.2)$$

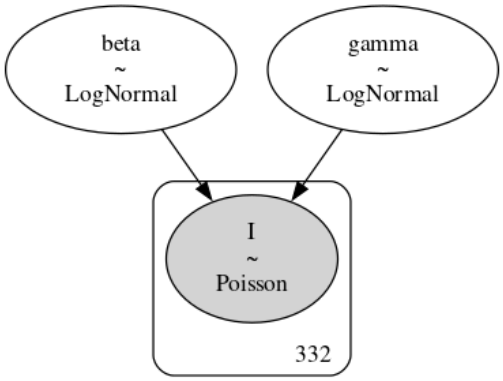


Figure 2: Bayesian Model

After creating the model, we ran the sampler for 4 chains, each with 21000 samples, out of which 1000 are for tuning. This leads to a total of 84000 samples. The trace plots for the samples are:

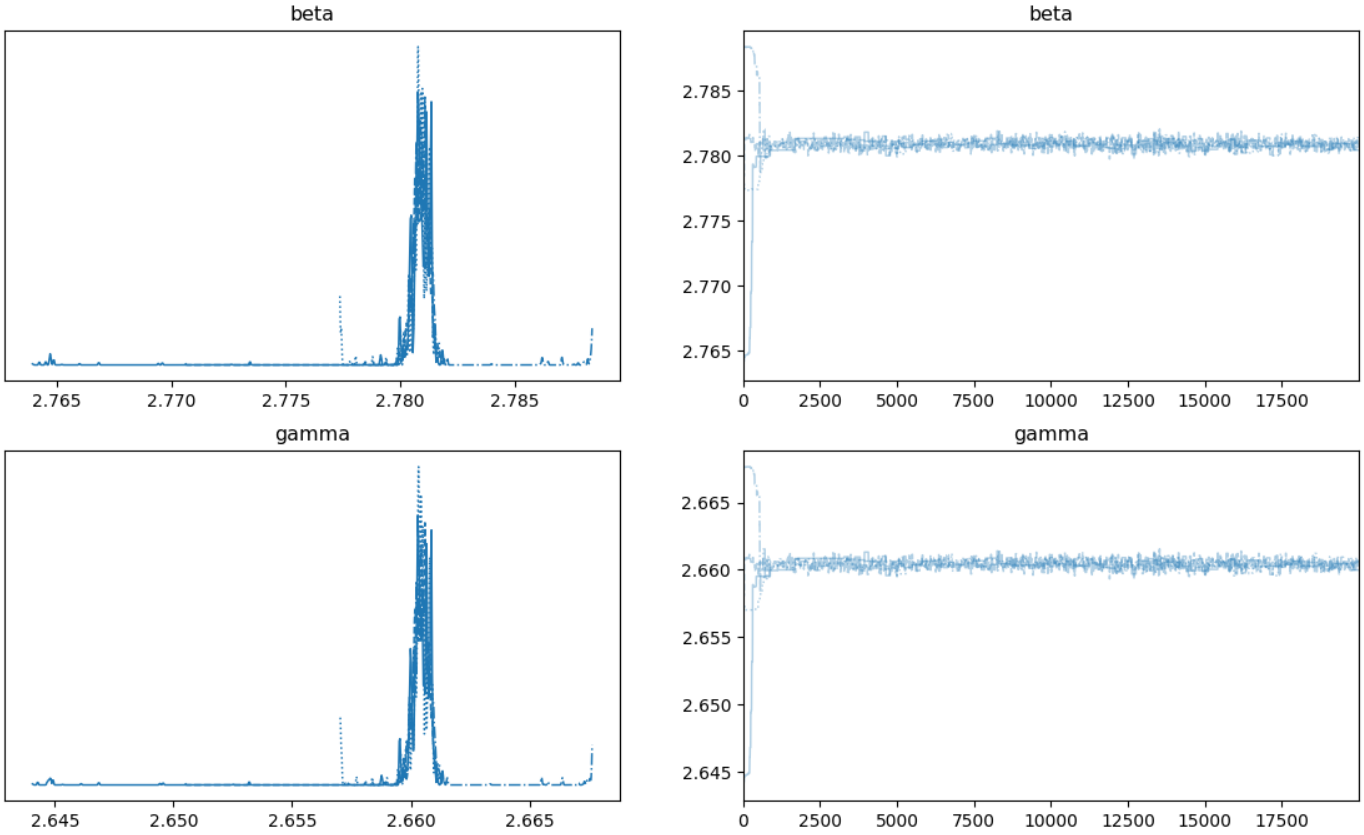


Figure 3: Trace

We take the first 1000 iterations to be burn-in, and so slice them off. The remaining trace is the burned trace.

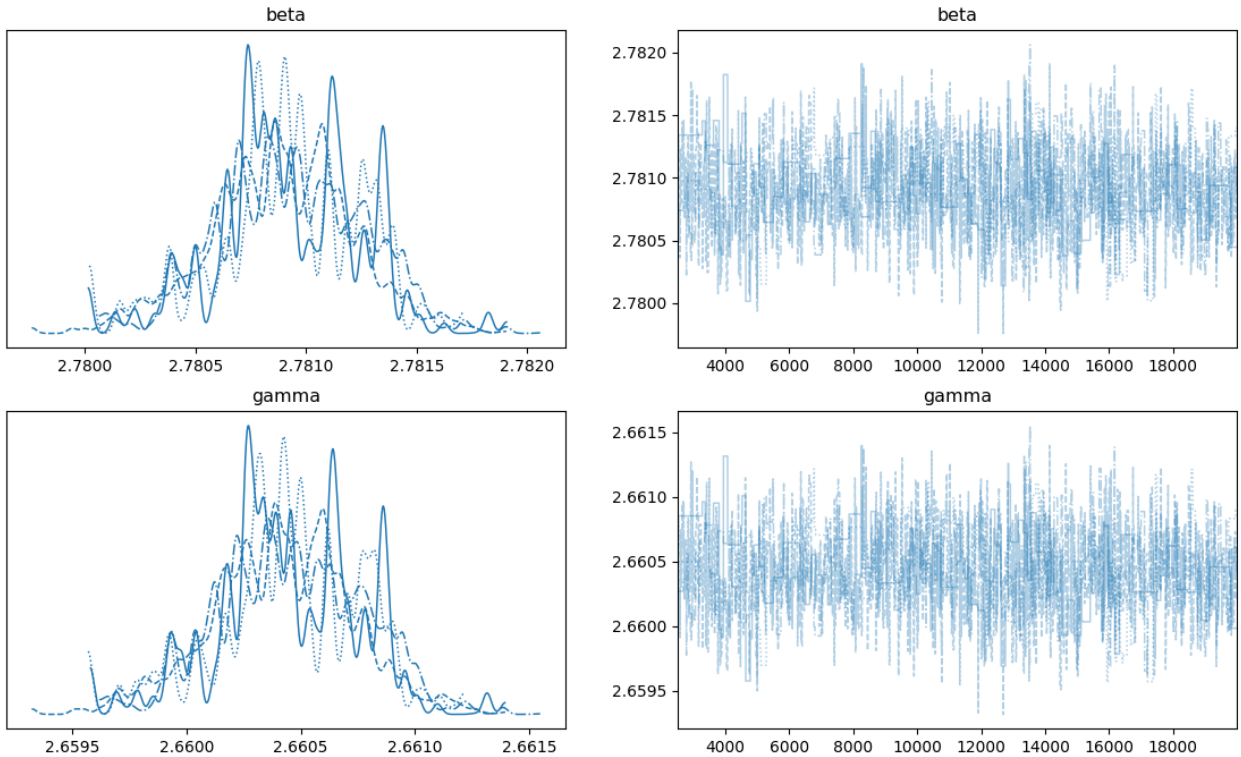


Figure 4: Burned Trace

The posterior distributions of β and γ have been found. Using them, we can find their means to be 2.781 and 2.660 respectively.

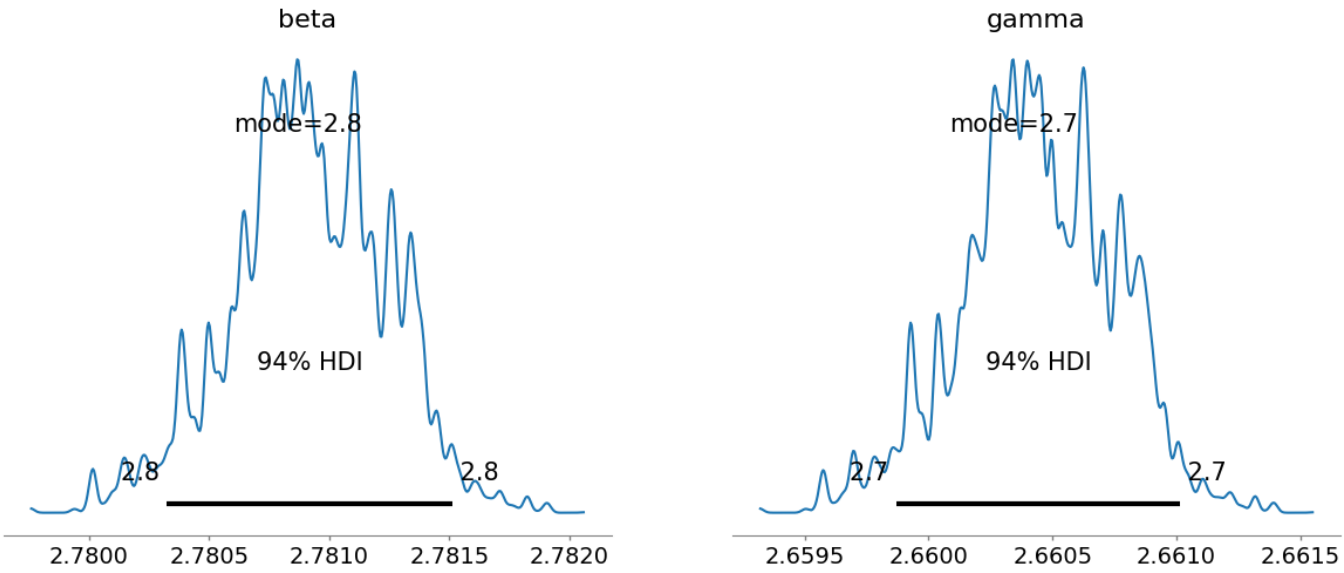


Figure 5: Posterior of β and γ

Using these values, we can calculate R_0, which is 1.045. Using this, we predict the number of active cases on March 23, 2023 (our data used to make the model was till March 22, 2023) to be 7950. Based on the actual data, it is found to be 7927. Thus, our estimate is very close to the actual data.

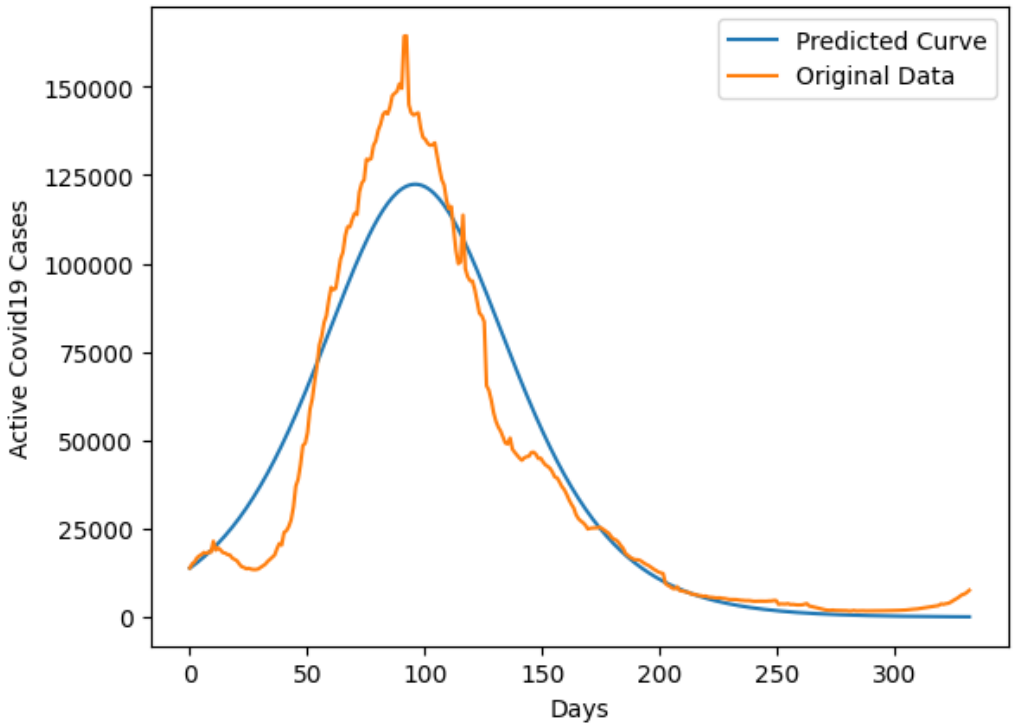


Figure 6: Predicted Active Cases Curve

Conclusion:

The SIR model is one of the simplest compartmental models, and is reasonably effective for making predictions on diseases where recovery confers lifelong immunity. Covid19 is not one such disease, as reinfection is possible, and so we can not expect it to be able to model the disease well. Nevertheless, this model serves as the base for many other compartmental models, such as the SEIS and SIRV models, which are better suited for modelling Covid19. This project is an exercise on the application of various mathematical tools and methods which can be used for the epidemiological modelling of infectious diseases. By applying mathematical methods and tools to real-world data, we can gain valuable insights into how diseases like Covid19 are spreading, and what measures might be most effective in slowing or preventing their spread. Overall, this project has highlighted the importance of mathematical modelling in epidemiology, and the potential it holds for helping us better understand and combat infectious diseases.