# Advanced Techniques in Bayesian Parameter Estimation for Disease Outbreak Prediction
# Course Project Report

## 1. Introduction

Predicting the spread of infectious diseases is one of the most significant applications of mathematical modelling.
The Susceptible–Infected–Recovered (SIR) model ,a compartmental model that tells us about how the population changes and how disease spread happens in the population.
This model categorises the population in three categories Susceptible , Infected , Recovered then studies the infection spread.
This project used one of the  Bayesian estimation techniques, namely MCMC to find the posterior parameters such as infection rate  and recovery rate from the model to predict the outbreak and the future prediction of the cases.
This approach allows for better uncertainty quantification compared to classical curve fitting.
Disease studied in the project is covid-19 which was prevalent in the past and widely studied .
The model used here is basic and the first step towards prediction of outbreak however either by introducing more compartments and branches or using the time varying parameter changes can help in better prediction and dealing with uncertainty.

## 2. Data Preparation

The data set used in the project included cumulative and the individual data about the number of cases registered , cases recovered and people who died daily for 8 months .Data we used has been divided into two sets to use for training and testing sets. We have taken a data set from the website kaggle and used it here.
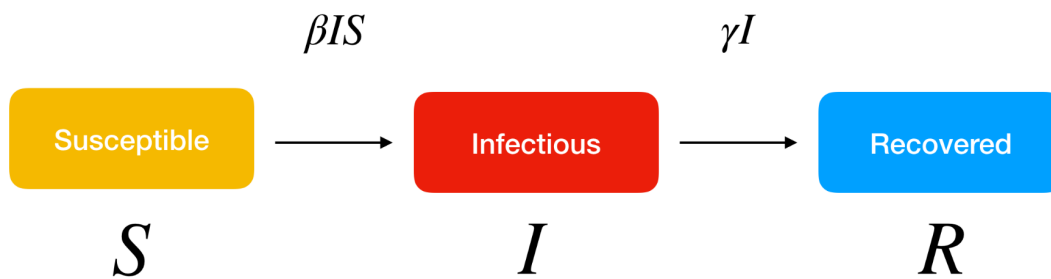
## 3. Mathematical Model

SIR model used here refers:
S: Susceptible , the population prone to the infection . This section reduces as the infection spreads.
I: Infected, the population which has been infected . This population increases as the injection spreads and decreases as the recovery fastens.
R: Removed, the population which is either dead or recovered and also not prone to infection again is in this category.

$$\beta IS \qquad \gamma I$$

Susceptible → Infectious → Recovered

$$S \qquad I \qquad R$$

dS/dt = -βSI/N
dI/dt = βSI/N - γI
dR/dt = γI

Parameters β and γ represent the transmission and recovery rates. $R_0$ represents the reproduction number which tells the average number of the infections caused by a fully infected person in a susceptible population. As soon as the reproduction number is greater than 1 the outbreak happens. The use of vaccines actually reduces the reproduction number and peak happens earlier so disease dies early.

## 4. Parameter Estimation Techniques

To obtain the posteriors we used the MCMC (Markov chain and Monte Carlo) techniques. The curve-fitting estimates for β and γ were used as informative priors in the Bayesian model to guide the sampling process and improve convergence.
MCMC uses both the Monte Carlo and Markov Chain .
Monte Carlo method uses random sampling to estimate the values of parameters and the Markov chain is moving to find values or distribution where your next step depends on the present value. So MCMC uses the randomness that moves and takes more time in the area where the probability density is high .

Algorithm:
1.Pick a random spot to begin — that's your **starting guess** for the parameter (let's call it $\theta_0$).
2.From your current position, take a small random step in any direction to a **new possible spot (θ*)**.
3. Compare its likelihood  if it is more then move next otherwise stay where you are.
4. Repeat 2 and 3 for a number of times and stop after a certain number of epochs or when sampling becomes stable.
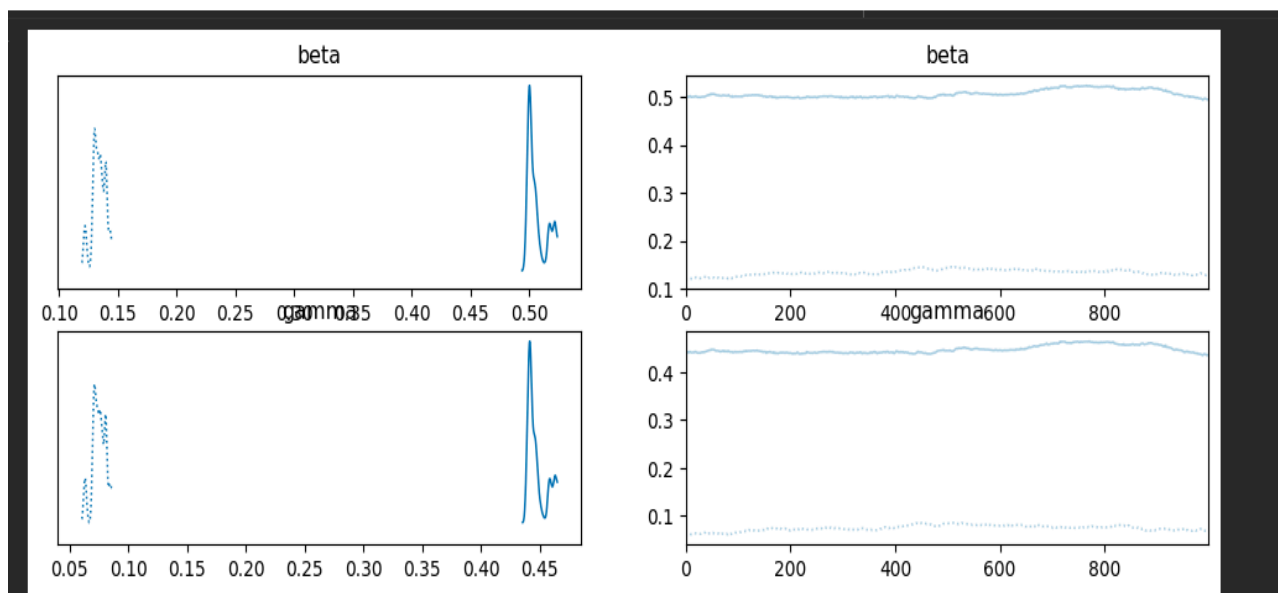
# 5. Results and Discussion

```
Posterior mean β: 0.3201
Posterior mean γ: 0.2610
Posterior R₀ = 1.23
```

These are the result after the MCMC algorithm to get the reproduction number is 1.23 suggests that each infected person could potentially infect about 1 other,
aligning with typical early COVID-19 estimates suggesting when the outbreak happens.
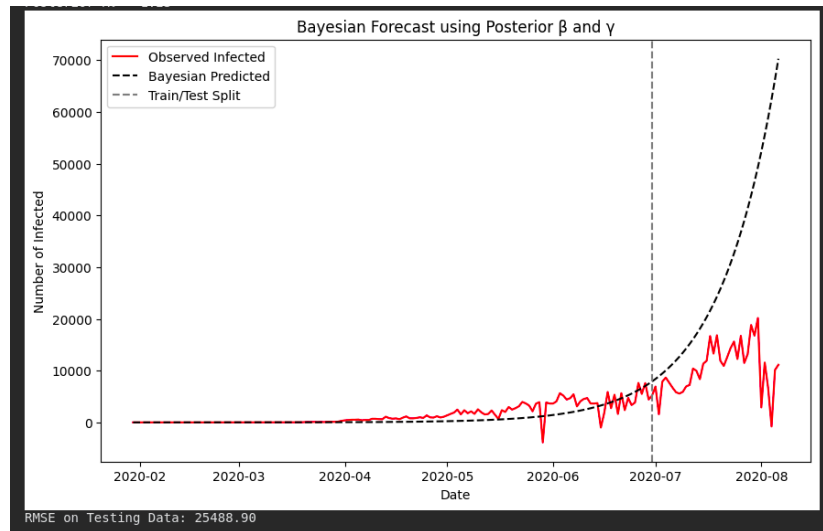


The trace plots show how the sampled values of β and γ changed during training.
At first, the algorithm explores different possibilities (tuning), and once the lines stabilize, it means the model has learned a consistent range of values.
This indicates the Bayesian sampler converged well and produced reliable estimates.

The forecast plot shows the observed and predicted number of infected individuals over time.The Bayesian SIR model follows the general trend of the real data, capturing the rise and stabilization of infections.This indicates that the estimated parameters (β and γ) provide a good fit for the observed outbreak dynamics.

Bayesian Forecast using Posterior β and γ

RMSE on Testing Data: 25488.90

## 6. Improvements and Future Work

The model can be improved by incorporating more compartments such as Extinct , Recovered but Susceptible (Rs)or Vaccinated (V), leading to SEIR or SEIRV extensions. In future work, prior distributions could be informed by epidemiological literature, and advanced sampling algorithms such as No-U-Turn Sampler (NUTS) can be further optimized to improve convergence speed and reliability.

Improvements needed:

The forecasted curve diverges . This suggests no damping or changing contact rate.

There are some dips in the active cases plot caused due to how daily infections  have been computed.

## 7. Conclusion

This project explored how Bayesian parameter estimation can improve traditional models used to predict the spread of diseases.

By comparing the usual curve-fitting (deterministic) method with a Bayesian approach, it became clear that Bayesian analysis provides more flexibility and a better understanding of uncertainty in the results.While the deterministic model gives a single best estimate of parameters like infection and recovery rates, the Bayesian model shows how these parameters can vary.This kind of probabilistic modelling is very useful in public health, where decisions often need to be made even when the situation is uncertain.