Topics learned

Learning a new programming language (R), and all the numerical approaches such as numerical integration calculation and root-finding methods aside, what I found extremely useful and helpful for me are:

1- **Data structure:** In computer science, a data structure is a data organization, management, and storage format that enables efficient access and modification. Working with big structures how to classify and clean them how to build such data structures that I have never learned before. I think the first step of doing any kind of process on your experimental data is preparing and cleaning them and having a specific structure for them so any other subsequent process can be done on that easily.

2- **Random variables:** random variable is a function that maps a sample space to real numbers, in a way that a number is assigned to any event that is a subset of that sample space. then we can have a probability for each event and finally a probability distribution for different possible outcomes. Frequentist and Bayesian perspective of probability. And finally, the likelihood function which is the joint probability of observing a specific outcome given a specific condition (a specific parameter that we have chosen), helps us (in the Bayesian perspective) to update our probability distribution (posterior) from the initial one (prior). I found these concepts extremely useful. As I took a reinforcement learning course last semester I had many difficulties understanding the meaning behind these concepts which highly limited my problem-solving abilities. Now I have more confidence and ideas in modifying the algorithms because of the insight this course and especially these topics provide me with. I believe in working with any kind of stochastic methods, these are essential to understand

Modeling Covid

A brief explanation of how the metropolis-Hastings Markov Chain Monte Carlo works:

1- We start with an initial guess for our SIR model parameters
2- Calculate the likelihood function for them
3- Propose a new step or in other words propose to update the parameters by adding normally distributed random numbers to the previous ones   xnew<- xold + N(0,1)
4- Calculate the error of simulated data extracted from the SIR model
5- Calculating the likelihood for the new set of parameters based on the error of the previous step
6- Calculating the probability of accepting this point which is the ratio of new parameters likelihood over that of old ones (the ratio of likelihood is equal to the ratio of posterior distributions, in this case, that's why for simplicity we used likelihood)
7- Draw a uniformly distributed random number and compare it with likelihood ratio u(0,1)
8- If ratio > u(0,1) then we accept the step else we use the previous parameters again
9- In case of acceptance we go to the first step with the new parameters as initial ones
10- We do this for the predetermined number of steps

Project goals

As shown in the figure below covid new daily cases number has several peaks. We can extract features of data using models such as SIR, helping us have an idea of the rate of infection so we can what is the probability of us dealing with a new variant of the virus for instance.
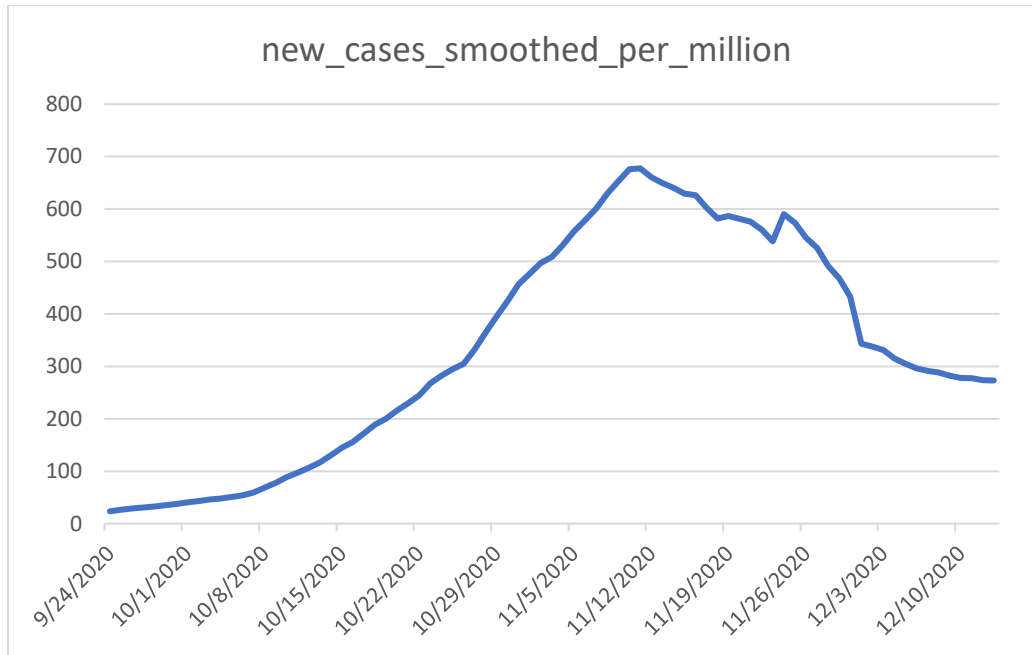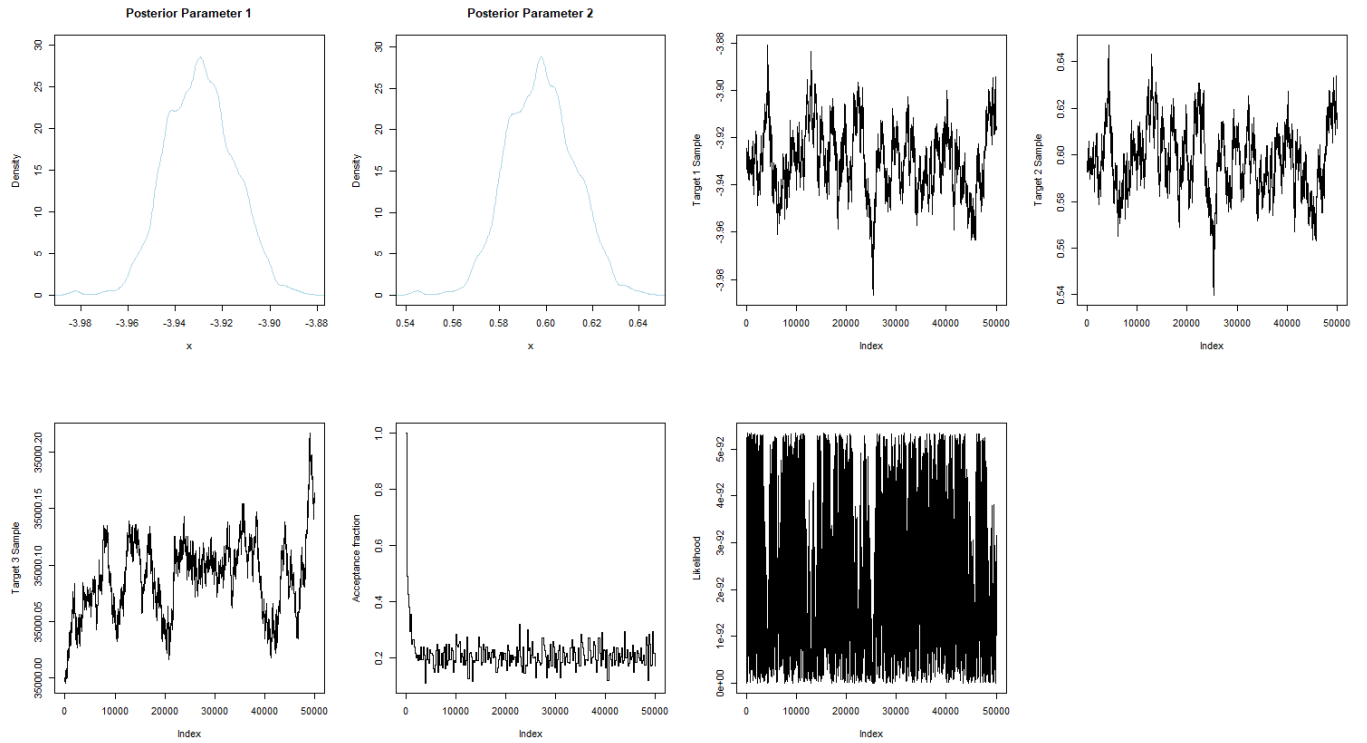

New daily cases of Poland per million for peak one

After convergence of the algorithm, we could extract the parameters of our model and have some discussion about them
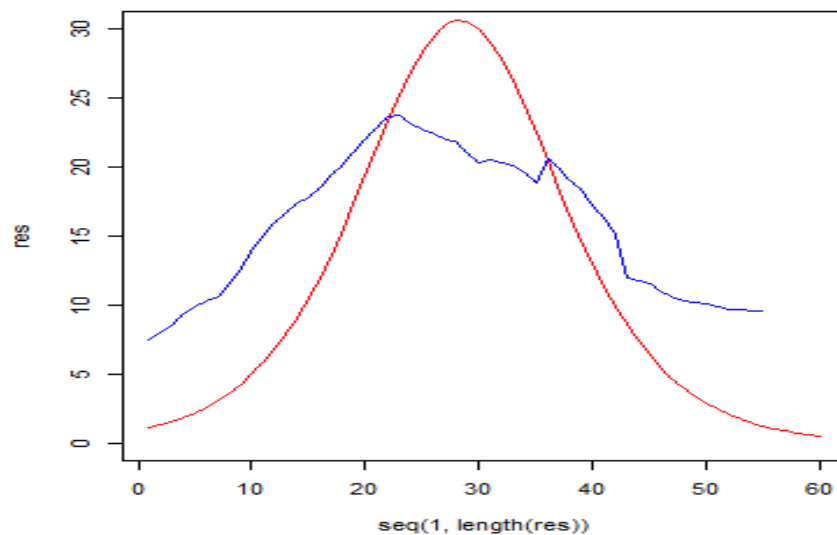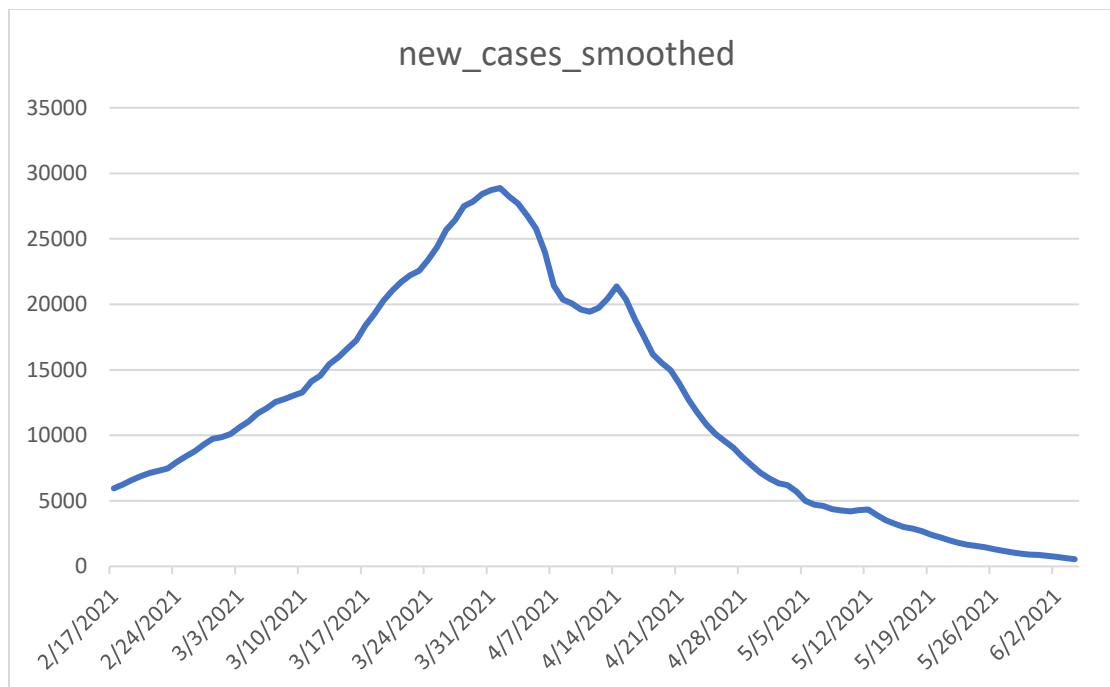
New daily cases of Poland per 35000 bule is collected data and red is simulated data for peak 1
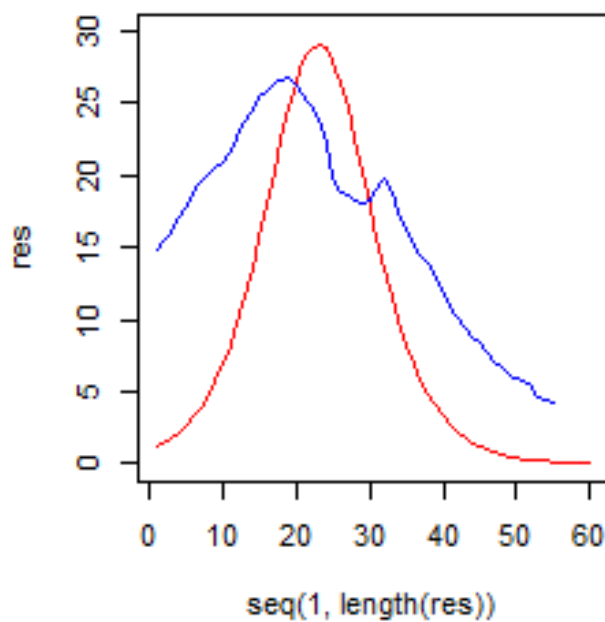

On the top left we have posterior distribution of beta and eta for the first peak, on the top right we have the variation of those parameters during MHmcmc algorithm solving
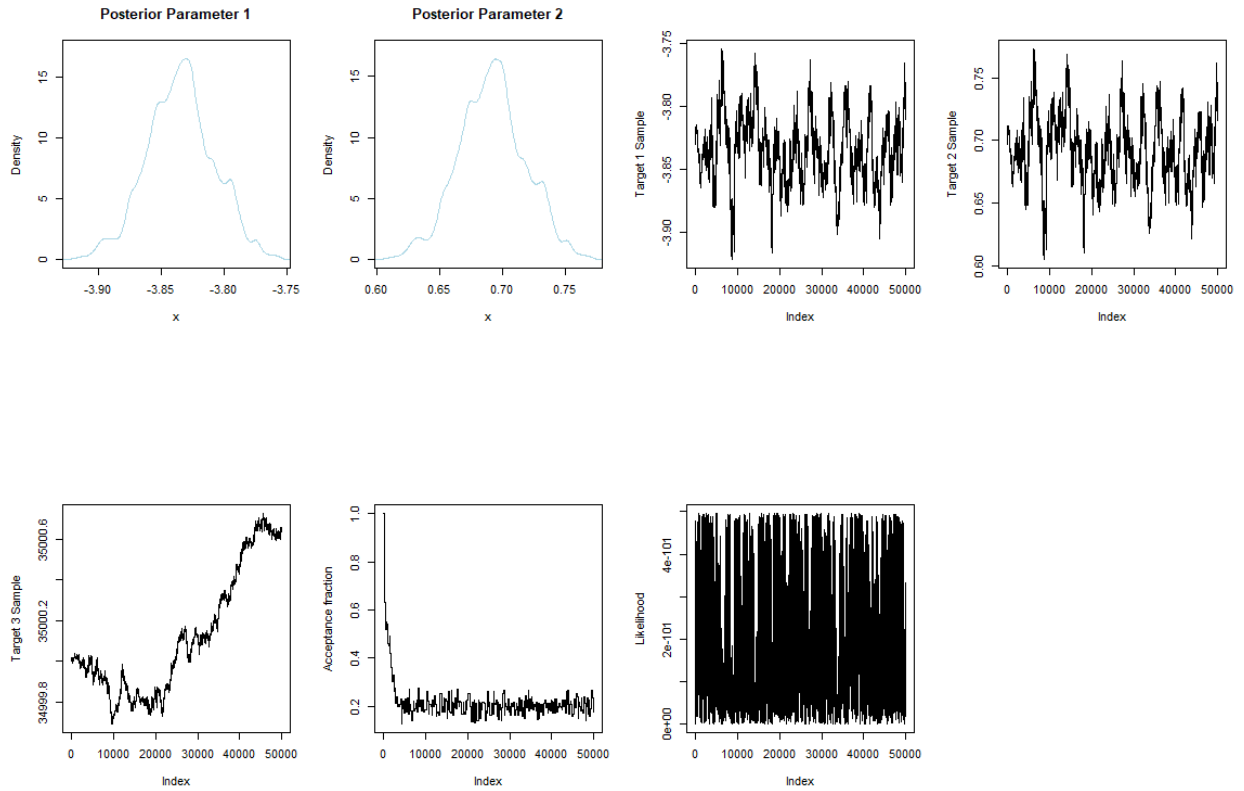
Peak2

New daily cases of Poland per million for peak two

New daily cases of Poland per 35000 bule is collected data and red is simulated data for peak 2


On the top left we have posterior distribution of beta and eta for the second peak, on the top right we have the variation of those parameters during MHmcmc algorithm solving

Finally, in the figure below there is not a high correlation between beta parameters for different peaks so we can draw the conclusion that we are dealing with a new variant.

Beta Posterior peak1 : blue, peak2: red
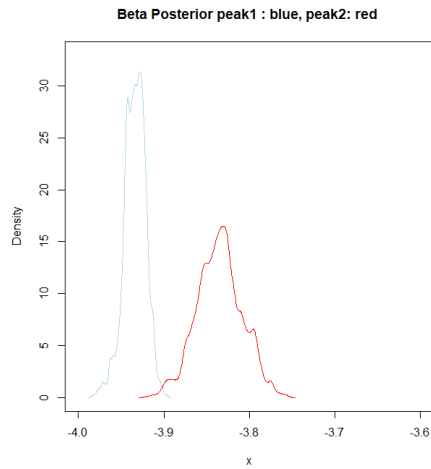
The Beta distribution for peak one and peak 2, As can be seen these two values has small correlation which means there is a high possibility that we are dealing with a new variant

eta Posterior peak1 : blue, peak2: red
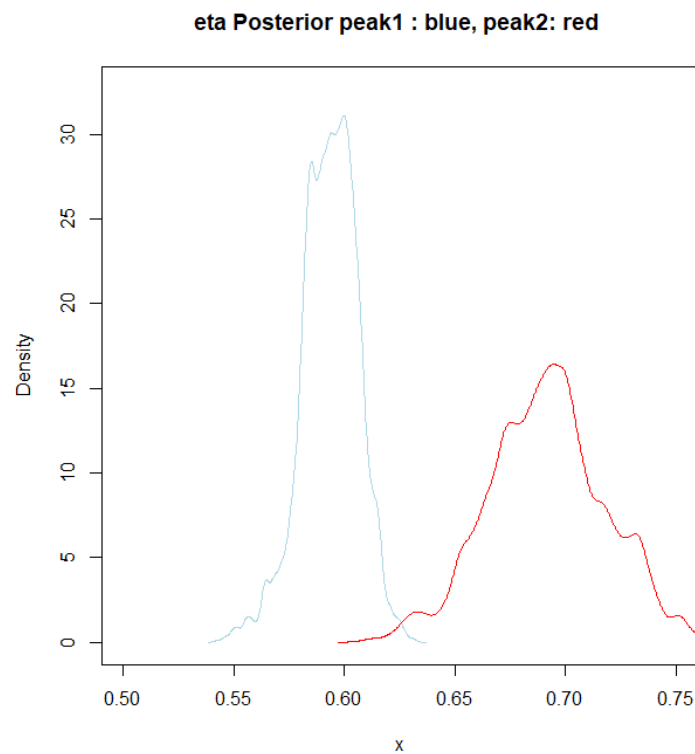
The eta distribution for plotted for the first and second peak similar to distribution for the Beta we have small correlation here as well which suggest the fact that we are dealing with a new variant