SIR Modelling of Twitter Hashtags

Mariam Ali

Computational Physics Lab

10/07/2019

Outline

1. Introduction
2. The deterministic model
3. Model Parameters
4. The Data
5. Reading The Data
6. Infection, Recovery, Susceptibility Count
    - Counting Functions
    - Algorithm
7. Analysis
8. Time Series Analysis
9. MCMC
10. Conclusion
11. Further Work

Introduction

- SIR Stands for Susceptible-Infected-Recovered which are three stages that could reoccur and describe an epidemic.
- Susceptible are the people who never got the disease, but are willing to, (do not have immunity) which is in our case twitter users that might be in contact with hashtag users.
- Infected are the people who used the hashtag, and recovered are those who stopped after two time windows.

The deterministic model

$$\frac{dS}{dt} = -\beta S(t) \times I(t) \tag{1}$$

$$\frac{dI}{dt} = \beta S(t) \times I(t) - \gamma \times I(t) \tag{2}$$

$$\frac{dR}{dt} = \gamma \times I(t) \tag{3}$$

Model Parameters

The model parameters γ and β are the epidemic quantifiers, thus a part of this project would be finding them. Solving the previous model analytically would result in these parameters.

$$\beta = \frac{S(t) - S(0)}{\int_0^t S(\tau)I(\tau)} \tag{4}$$

$$\gamma = \frac{R(t)}{\int_0^t I(t)} \tag{5}$$

The Data

The Data is sampled public tweets from Twitter streaming API. The parts used in the analysis:
”timeline$_t$*ag.anony.dat”whichisaseriesofhashtagwithuserIDsofuserswh*
”timeline$_t$*ag*$_r$*t.anony.dat”whichisverysimilartothepreviousone,butItcoun*
And finally the followers of the network to be the defining parameter of Susceptibility.

Reading The Data

Reading the data is a challenging task since the rows of each file are not of equal size, and the data is kind of large. So the main function defined by the Cython library marks the hahtags with space character and the time stamps and user IDs with commas and these hashtags are apended to a dictionary such that the tag is the key and defined by the user IDs and time stamps. The next part is filtering the hashtags based on number of tweets for each hashtag to be greater tha ten thousand tweets, in both the tweets and retweets datasets combined.
The reason We combined both datasets is that the users will be dealt with equally either being a tweeter OR a retweeter.

Infection, Recovery, Susceptibility Count

- Seriestimeconvert
- InfectionHistory
- SusceptibleHistory

Algorithm

- The time series converter is a function the takes in the raw data for each hashtag and the time window decieded by the user (8 hours) and returns the users who tweeted or retweeted that hashtag in the increment of the time windows, meaning that the real time is divided into time windows resulting in a series that has users at each time step.
- Infectionhistory function basically counts a user to be infected if they are in the series and count them as recovered if they are not in the series after two time windows.
- SusceptibleHistory This function looks at the infected and recovered at a certain time window, and if an ID is in the following list but it Neither Infected Nor Recovered, then it would count as a susceptible.

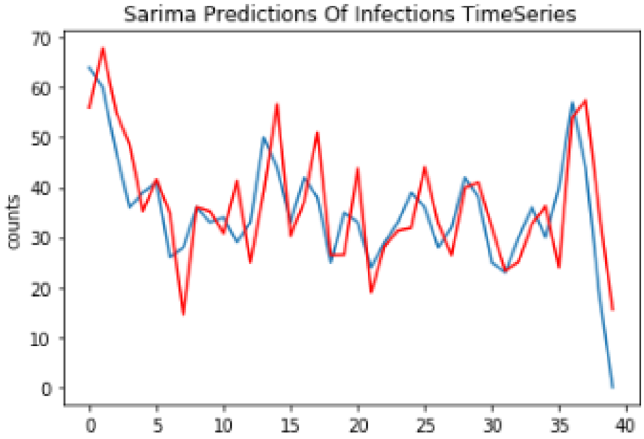
Analysis

- Time Series Analysis
- Marov Chain Monte Carlo

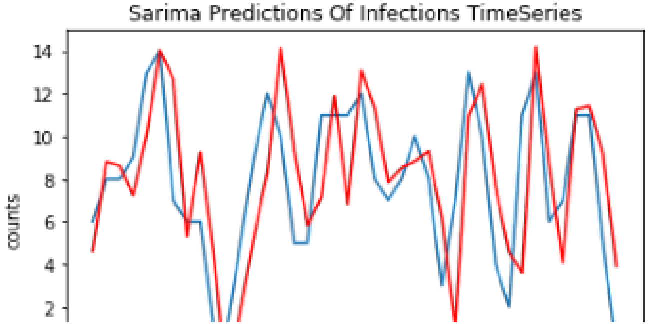
Time Series Analysis

- Autoregressive Model
  $I_t = \beta_0 + \beta_1 I_{t-1} + \epsilon_t$
- Moving Average
  $Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + ... + \phi_q \epsilon_{t-q}$
- ARIMA
- Seasonal ARIMA

```
500aday
Test RMSE: 8.398
```



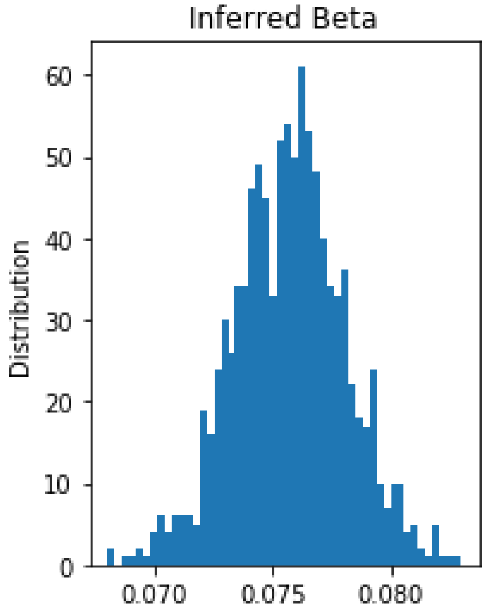
```
gnation
Test RMSE: 3.090
```

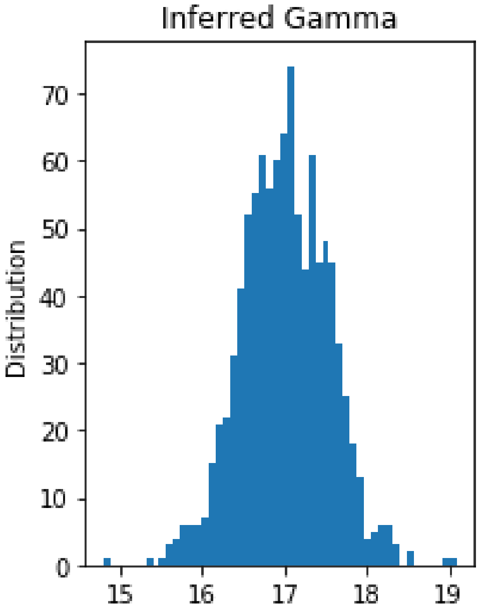


Seasonal Arima

MCMC

A likelihood Function Calculated based on the deterministic model, the value of this likelihood is calculated with three functions: Counts, DoubleSum and LogProduct of Counts.
Initial values of β and γ are assumed to follow a gamma distribution.
The final values of β and γ for each hashtag vary but they seem to follow a normal distribution.





Conclusion

- Arima with grid search of parameters gives the least root mean squared error for predictions of infections and recovery counts, meaning that Arima is the best Prediction method of such data.
- Hashtags with $R_0 > 1$ will go viral, being too little I Also checked for $R_0 > 0.5$.

Further Work

- Try Different time windows because maybe Different Time windows show different behavior for each hashtag in both infection and recovery.
- A specific Likelihood that's custom made to this data with the same assumptions.