AN APPLICATION OF BAYESIAN INFERENCE TO EPIDEMIC MODELING

João Pedro Valeriano Miranda*
Instituto de Física Teórica (IFT)
Universidade Estadual Paulista (Unesp)
São Paulo, Brazil
joaopedromv98@gmail.com

Pedro Henrique Pinheiro Cintra*
Instituto de Física "Gleb Wataghin" (IFGW)
Universidade Estadual de Campinas (Unicamp)
Campinas, Brazil
pedrohpc96@hotmail.com

ABSTRACT

In this work, we use the Metropolis-Hastings algorithm to infer the infection rate from artificial epidemic data during the initial stage of an outbreak. The MAP of the posterior distribution is shown to be in good agreement with the value obtained via least squares estimation. We show that the model can fit the data more precisely if the initial condition is also considered a free parameter to be fit, in which case the initial number of infected individuals and the infection rate are shown to be strongly correlated.

1 Introduction

Epidemics are characterized by a initial phase of exponential growth [2]. During this phase, simple exponential models may be implemented in order to estimate the growth rate of the disease. Consider a population divided into three categories: Susceptible (S), Infected (I) and Removed (R), where removed includes both recovered individuals who acquire immunity against the disease or deceased individuals. The dynamics of the disease, in a mean field regime, may be described by

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI \tag{1}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta SI - \alpha I \tag{2}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \alpha I \tag{3}$$

where β is the infection rate and α is the removal rate, characterized as the sum of the recovery rate from the disease and the death rate.

During the initial stages of the disease, the infected population is very small in comparison to the susceptible population. Therefore $S \approx N$, being N the total number of people in the community. If we define $N = 1$, S, I and R become fractions of the total population that are susceptible, infected and recovered, respectively. Under this consideration, equation (2) is rewritten as

$$\begin{aligned} \frac{\mathrm{d}I}{\mathrm{d}t} &= (\beta - \alpha)I \Rightarrow \\ I_\theta(t) &= I_0 e^{\theta t};\ \theta = \beta - \alpha, \end{aligned} \tag{4}$$

where I_0 stands for the initial number of infected, at $t = 0$, and θ is the growth rate of the disease. If $\theta > 0 \Rightarrow \beta > \alpha$, the disease spreads among the community. The higher θ is, the faster the spread.

*Both authors contributed equally to this work.

In this work, we implement an Metropolis-Hastings (MH) algorithm to provide Bayesian inference of the growth rate of a given disease in a population. To do so, we consider an artificial dataset that imitates the initial stage of a disease outbreak.

2 Methodology

In order to adjust the dataset and estimate the posterior distribution of the growth rate, we implement a Metropolis-Hastings sampling algorithm [1], detailed in Algorithm 1.

Algorithm 1 Metropolis-Hastings algorithm for sampling a given distribution $\pi(\theta)$

```
Set the initial value for a first sample θ_0 and the number of samples to be generated N.
Set the function g(φ|θ) that describe the probability of a sample candidate φ, given the actual candidate θ.
n ← 0
while n < N do
    Generate a candidate sample φ from g(x|θ_n).
    p ← π(φ)/π(θ_n)
    Generate a random variable r from U(0, 1)
    if p > 1 then
        θ_{n+1} ← φ                                  ▷ Accept candidate sample
    else if r < p then
        θ_{n+1} ← φ                                  ▷ Accept candidate sample
    else
        θ_{n+1} ← θ_t                                ▷ Reject candidate sample
    end if
    n ← n + 1
end while
```

According to Bayes theorem, the posterior distribution for a given parameter θ, given an observed data y is

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)}, \tag{5}$$

where $\pi(\theta)$ is the prior distribution for the parameter θ, which can be built with the information one may have about the parameter. $\pi(y|\theta)$ is the likelihood function, which describes the probability of observing the y data given model parameter has value θ. $\pi(y)$ is the probability distribution of the observed data, which is usually inaccessible, but this is not a problem, as it can be seen as a normalization for the expression for the posterior distribution, so that it can be rewritten as

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\displaystyle\int \mathrm{d}y\, \pi(y|\theta)\pi(\theta)}. \tag{6}$$

In our application, we assume a normal likelihood $\pi(y|\theta)$ described by a normal distribution $\mathcal{N}(F_\theta, \sigma)$, where F_θ is the model output for the parameter θ. Therefore we have an additional parameter σ that describes the width of our likelihood. In the epidemic model example, $F_\theta = I_\theta(t)$. As for the prior distributions of θ and σ, we assume uniform priors

$$\pi(\theta) = \mathcal{U}(0.1, 0.2); \tag{7}$$
$$\pi(\sigma) = \mathcal{U}(10, 15). \tag{8}$$

Having this in mind, the posterior distribution we will be sampling via MH algorithm is given by:

$$\pi(\theta, \sigma|y) = \begin{cases} N \exp\left[-\dfrac{(I_\theta(t) - y)^2}{2\sigma^2}\right], & \textbf{if } \theta \in [0.1, 0.2] \textbf{ and } \sigma \in [10, 15]; \\ 0, & \textbf{otherwise}, \end{cases} \tag{9}$$

where N is just a normalization constant, in which is embedded $\pi(y)$ and the normalizations of the normal likelihood and the uniform priors. As we'll see in the next section, our data y consists of a vector describing infected number values over a 40 day span, and so will the model $I_\theta(t)$ generate 40 points of estimation for the number of infected individuals. This means we are dealing with a high-dimensional likelihood distribution, which often leads to very concentrated probability mass in the parameter space. This means that the likelihood probability density function (PDF) can be very small almost everywhere in the parameter space, often so small that computer precision is not enough to discern these PDF values from zero, impeding comparison of the PDF between different points in parameter space. To avoid that, when running the MH algorithm , instead of using the PDF of the posterior distribution $\pi(\theta|y)$, we will be working with its logarithm, $\log(\pi(\theta|y))$, to compare the value of the PDF at the actual candidate sample ϕ to the past one θ_t (refer to Algorithm 1).

3 Results

The mock data we are going to analyze is presented in Figure 1, where we also present the curve generated by the model $I_\theta(t)$ (4) with θ obtained by a least squares fit of the data via Levenberg-Marquadt (LM) algorithm, with the `curve_fit` method implemented by Scipy [3]. The estimated optimal value for θ is $\theta_{opt} = 0.1193198$ with estimated variance of 3×10^{-7}. As we can see from the plot, the model deviates from the data in a considerable range of data. As we will see soon, this can be improved by considering the initial number of infected individuals I_0 as a free parameter to be fit as well.

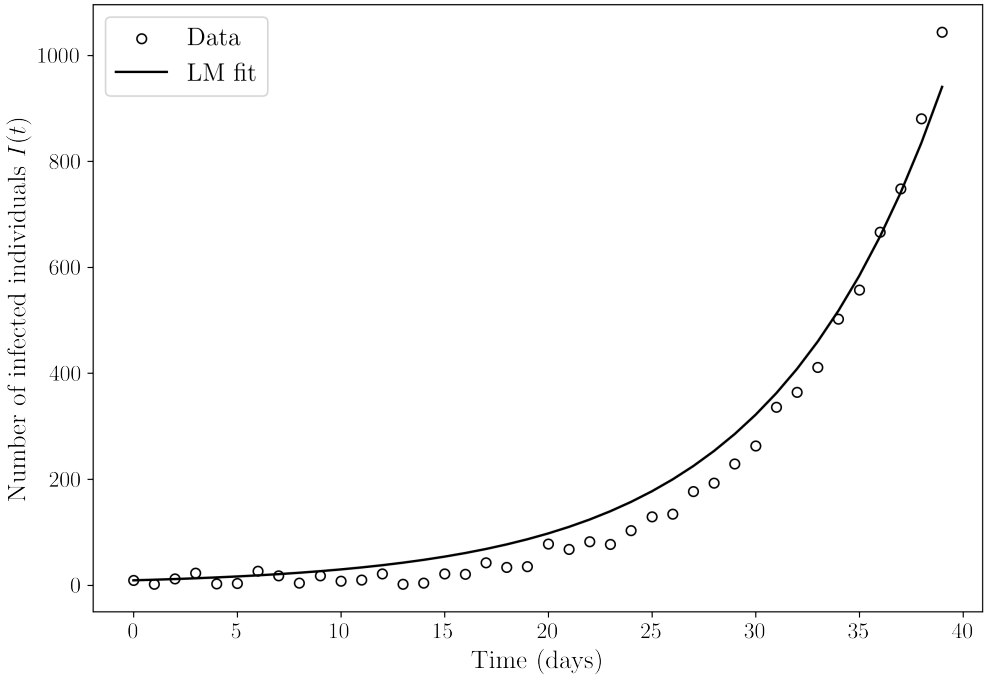

Figure 1: Mock data with fit of $I_\theta(t)$ model.

Proceeding to the Bayesian inference of θ and σ parameters, we need a function $g(x|\theta_t)$ (refer to Algorithm 1) to generate candidate samples from the last one. We choose a normal distribution centered in the last sample, with standard deviations δ_θ and δ_σ in the θ and σ directions of the parameter space, respectively, $x \sim \mathcal{N}(\theta_t, \{\delta_\theta, \delta_\sigma\})$.

To run the diagnostics of the MH algorithm's performance for the problem at hand, we vary the values of δ_θ and δ_σ over the values $\{0.001, 0.01, 0.1\}$. The result is presented in Figure 2, where we can see the evolution of the Monte Carlo

Markov chain (MCMC) over sorting 10^5 samples. It is clear that the pair $(\delta_\theta, \delta_\sigma) = (0.001, 0.01)$ provides the most accepted samples, over 20% of the total sampled, only converging a little slower than cases $(\delta_\theta, \delta_\sigma) = (0.001, 0.1)$ and $(0.01, 0.1)$, but the convergence problem is very easy to deal with by simply adding a burn-in period.

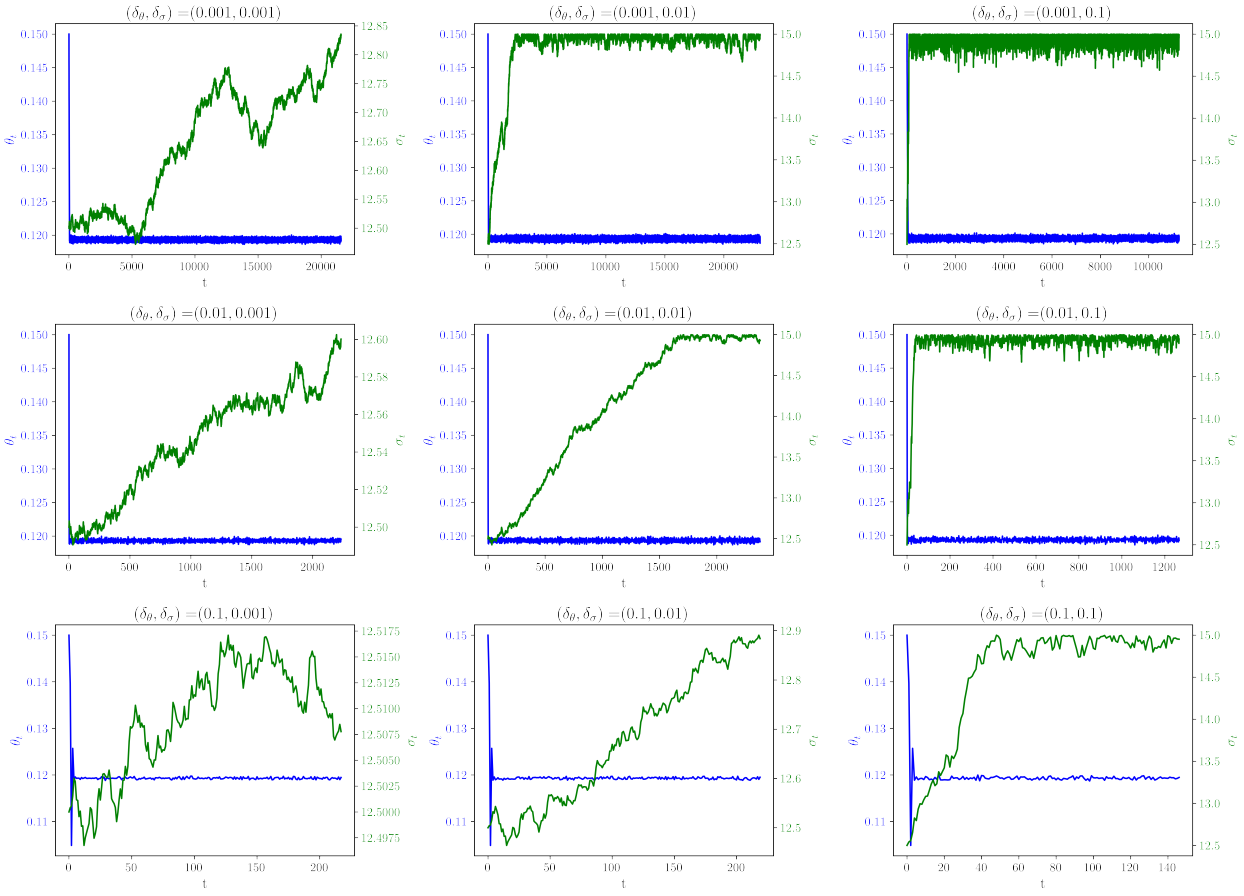

Figure 2: MCMC for different $(\delta_\theta, \delta_\sigma)$ pairs.

Choosing the pair $(\delta_\theta, \delta_\sigma) = (0.001, 0.01)$ from the diagnostic of the MCMC, we can proceed to generating more samples from the posterior distribution, now with a burn-in of 5×10^3 samples and then saving the results for 10^4 samples. The results are displayed in Figure 3, where we can see both the joint and marginal posterior distributions for θ and σ. From the histogram, there seem to be little correlation between both parameters. Indeed, the Spearman correlation is $r = 0.02474$, with p-value $p = 0.01335$. Notice we also display the value of the estimated parameter values for the maximum a posteriori value (MAP), and the MAP for θ is actually very close to the one obtained by the least squares fit, as one would expect. Our maximum a posteriori estimate for θ is 0.1193206, while the estimate by the least squares is 0.1193198. It is interesting to notice that the MAP for σ is found at the upper limit allowed by its prior distribution, which indicates the actual peak of the distribution is to be found at a value higher than 15.

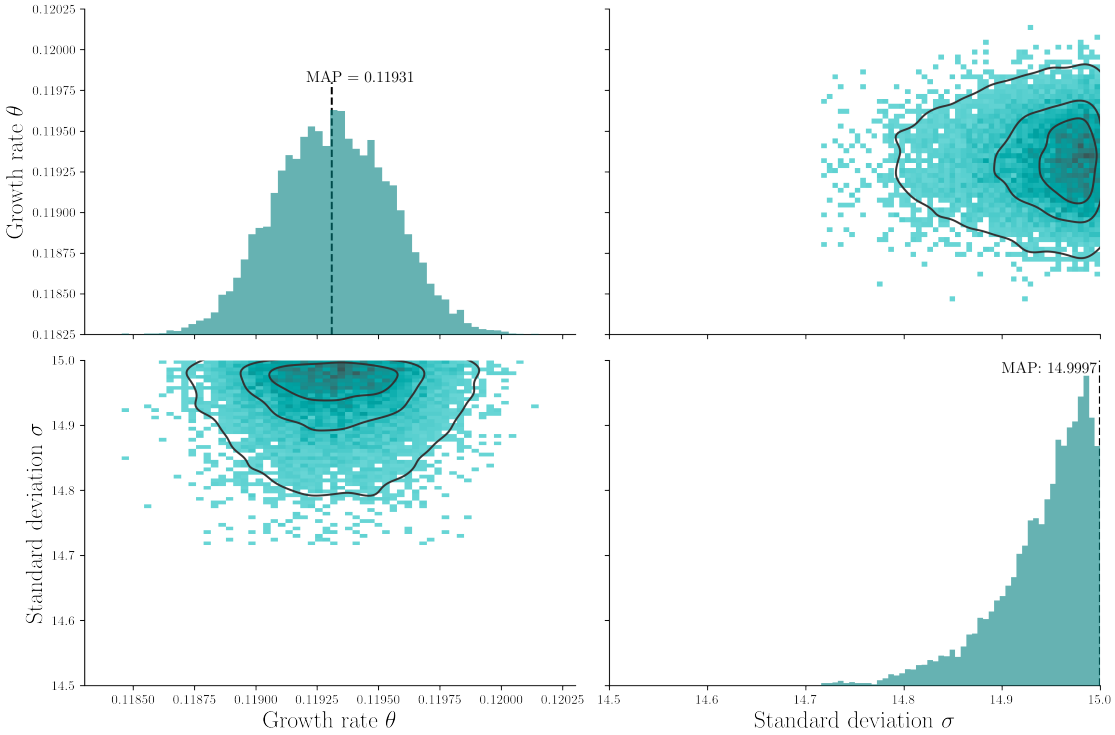

Figure 3: Posterior distributions for the growth rate θ and the standard deviation σ. The dashed lines in each histogram indicates the maximum a posteriori (MAP) estimated during the MH sampling, continuous lines in the 2D histograms indicate confidence intervals of 68, 95 and 99%.

3.1 Setting the initial condition free

As we saw in Figure 1, the model cannot fit very well to a considerable range of the dataset. In an attempt to improve that, we decided to consider that the initial condition I_0 can be a free parameter to be fit as well. Doing so, by a least squares fit we obtain the result shown in Figure 4, where we can see it is possible to achieve a much better fit. It makes sense, as one must be careful that the first data point may be subject to noise in the same way the rest of the data is.

In order to include I_0 in the MH method, we assume a log-normal distribution as a prior. The choice of a log-normal distribution is based on the fact that I_0 can only have positive values, and small values of initial number of infected are more probable than large values, based on the observed data. The parameters of the PDF were chosen in order to make the mode = y_0, that is, the mode is equal to the first data point, and the standard deviation of the logarithm of the random variable is set to 5% of $y_0 = y(t = 0)$.

$$\pi(I_0) = \mathcal{L}_{\mathcal{N}}(\ln(y_0) + y_0^2/400, y_0/20). \tag{10}$$

The posterior distribution therefore assumes the expression

$$\pi(\theta, I_0, \sigma|y) = \begin{cases} \frac{N}{I_0} \exp\left[-\frac{(I_\theta(t) - y)^2}{2\sigma^2} - \frac{(\ln(I_0) - \ln(y) - y^2/400)^2}{y^2/200}\right], & \textbf{if } \theta \in [0.1, 0.2] \textbf{ and } \sigma \in [10, 15]; \\ 0, & \textbf{otherwise}, \end{cases} \tag{11}$$

Figure 5 shows the joint and marginal posterior distributions for θ, I_0 and σ. One may notice in this plot that θ and I_0 have a strong negative correlation (Spearman $r = -0.994$, p-value < 0.001), that means that larger values of growth rate must imply in lower infected at $t = 0$, in order to correctly reproduce the observed data.

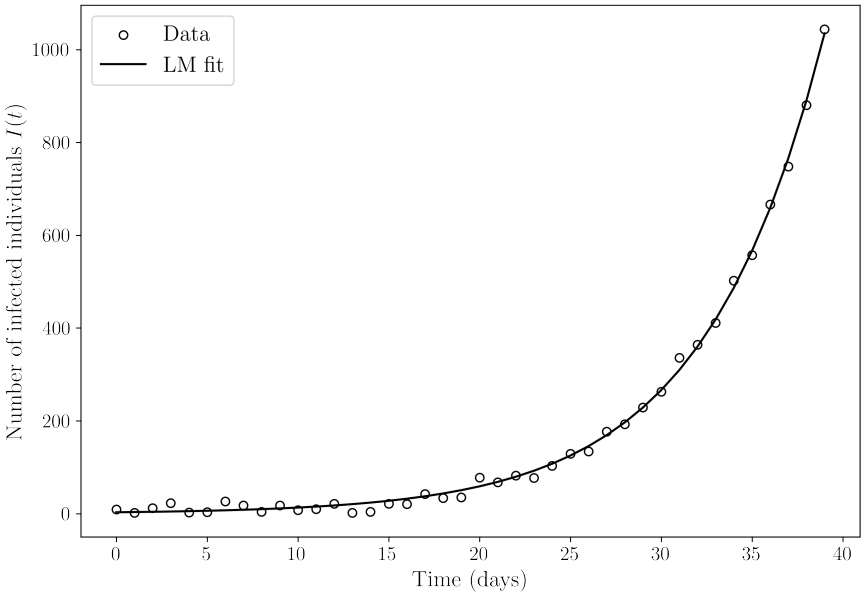

Figure 4: Mock data with model fit, now considering I_0 as a free parameter as well.

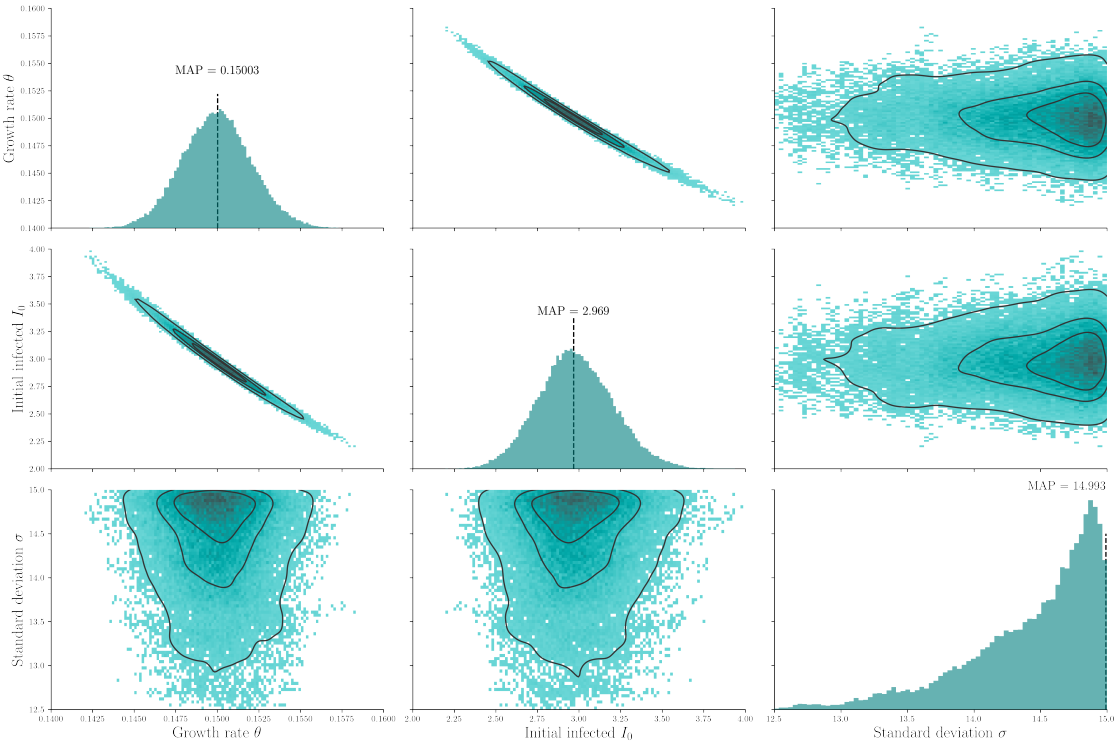

Figure 5: Posterior distributions for the growth rate θ, initial number of infected individuals I_0 and the standard deviation σ. The dashed lines in each histogram indicates the maximum a posteriori (MAP) estimated during the MH sampling, continuous lines in the 2D histograms indicate confidence intervals of 68, 95 and 99%.

Having the posterior distribution, we can see how the inference predicts the data should fluctuate around the average. Figure 6 shows the best fit to the data in the continuous line, using the MAP for θ and I_0. For the last 10^3 samples of the posterior distribution, we generate a realization of the model with the sampled values of I_0 and θ, and add to it

gaussian noise with standard deviation equal to the σ value of the same sample, and this is represented in white blue with a high transparency, so that it gets darker with the overlap of data generated from multiple samples.

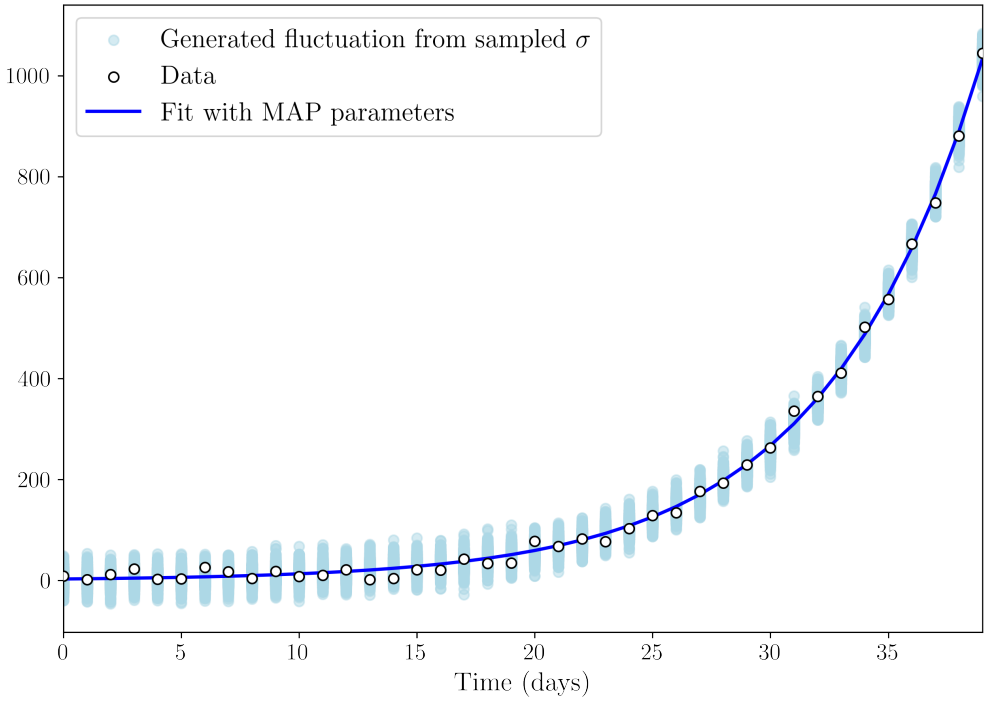

Figure 6: Model realization with parameters selected by the MAP values, described by the blue line. In light blue, we have data generated by the model with different samples of the posterior.

4 Conclusion

The Metropolis-Hastings provides good results for the inference of θ and σ parameters, and the MAP for θ is in good agreement with the value obtained by least squares estimation. It is also clear that allowing the initial value I_0 to also be fit as a free parameter leads to an improvement in the capacity of the model to describe the observed data.

Further investigation could be done on the inference of the σ parameter as the peak of its marginal posterior distribution is very close to the boundary imposed by the prior distribution.

Tests could also be made to consider how the effect of changing the prior distribution can affect the inference results but, for doing so, there must be a good reason to consider a different prior. One possibility would be to change the prior of θ to be a gaussian centered of the value obtained by a least squares fit, as we would expect it to give a good first approximation, and we actually checked that the least squares estimation results in a value close to the MAP obtained by the MH algorithm. The least squares estimation could be used also to change the prior of I_0, in the case we consider it to be a free parameter of the model as well.

References

[1] Link, W.A., B.R., *Bayesian Inference: with ecological applications*. Academic Press (2010), ISBN 0123748542.

[2] Ma, J., Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling* (2020). 5, 129–141, doi:10.1016/j.idm.2019.12.009.

[3] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* (2020). 17, 261–272, doi:10.1038/s41592-019-0686-2.