Bayesian Epidemic Modelling

Hannah Craddock
*S*upervisors: Dr. Simon Spencer, Prof Xavier Didelot

CDT Department of Statistics
University of Warwick

August 28, 2021

Abstract

Infectious diseases continue to pose a major threat to human health on a global scale. The current COVID-19 pandemic and other recent outbreaks such as SARS (2002) and Ebola (2013) highlight the immense burden they can impose on society. Statistical models have been established as an essential tool for understanding the outbreak of epidemics and their transmission dynamics. Here, a discrete time model with branching process characteristics was developed to model the spreading of an epidemic through a susceptible population with the aim of producing a model as realistic to true events as possible. A key characteristic of the model is the time varying infectivity allowed for each infected. This was incorporated within a framework of Bayesian inference and an adaptive Markov chain Monte Carlo (MCMC) scheme was used to infer the model parameters, specifically the reproduction number R. The model framework is adaptive and generalizable to a wide range of diseases with a flexible generation time and the hope would be that the tool would prove useful for epidemiologists investigating a variety of diseases. The transmission dynamics of an epidemic are rarely homogeneous and both a super-spreading events model and super-spreader model were also developed to capture the uneven transmission patterns of real-life epidemics. Simulations of all models were carried out to aid understanding of the model dynamics involved and the inferred parameters were plotted to ensure correctness. The long-term, overall aim of this work is to get a better understanding of the underlying dynamics of epidemic outbreaks. Such an understanding could be used to inform public policy and control strategies with the goal of mitigating the worst effects of infectious diseases and epidemics.

Contents

1 Introduction

Statistical models are an essential tool for understanding the outbreak of epidemics and their transmission dynamics (Keeling and Rohani, 2011). They enable predictions to be made on the future course of an outbreak, the evaluation of potential control strategies to be carried out and estimates on the modes of transmission to be determined (Ferguson et al., 2006). Stochastic models are often used as they account for the randomness which is inherent to the manifestation of a disease in a large population. One aim of this project is to develop and apply a stochastic model of infectious disease outbreaks that accounts for the complexity of the real epidemiological process while remaining statistically tractable.

The quantification of transmissibility during an epidemic is vital to planning and adjusting public health responses (Cori et al., 2013). Transmissibility can be measured by the reproduction number R which is the average number of secondary cases caused by a typical infectious individual (Zhang et al., 2020). This statistic, which is time and situation specific, can be monitored over time to provide feedback on the efficacy of interventions and on the necessity to intensify control efforts (Ferguson et al., 2006). A super-critical outbreak, such as the the current Covid-19 pandemic, corresponds to estimates of $R > 1$. This is indicative of a significant risk of the infectious pathogen to cause a major outbreak (Zhang et al., 2020). A sub-critical outbreak would correspond to estimates of $R < 1$, as was the case in the Middle East respiratory syndrome coronavirus (MERS-CoV) in South Korea 2018 (Kucharski and Althaus, 2015). Such estimates indicate a trend of declining incidence and that the outbreak is slowing down. In general the objective of control efforts is to reduce R below 1, the threshold value, and as close to 0 as possible, bringing the epidemic under control as a result (Fraser et al., 2004). The R number can refer to either the basic reproduction number, known as the R nought or zero (R_0), or the effective reproduction number (R_e). R_0 describes how many people each infected person will infect on average, assuming that there is no pre-existing immunity in the community (Delamater et al., 2019). R_e is the average

number of people that can be infected by an individual at any specific time, and it changes as the population becomes increasingly immunised (Nishiura and Chowell, 2009). This could occur through individuals gaining immunity after being infected, through vaccination. Both are often referred to as the R number. In this work the number is referred to as R_0.

A discrete time model with branching process characteristics was developed to model the outbreak of an epidemic. A key characteristic of the model is the time varying infectivity allowed for each infected. This model framework is adaptive and generalizable to a wide range of diseases given the flexible generation time and the hope would be that the tool would prove useful for epidemiologists investigating a variety of diseases. The model was incorporated within a framework of Bayesian inference and an adaptive Markov chain Monte Carlo (MCMC) scheme was used to infer the model parameters, specifically the reproduction number R_0. MCMC is an increasingly popular method for sampling from a distribution (Van Ravenzwaaij et al., 2018). Unlike other methods like point estimation or maximum likelihood estimation which only output a single value, with MCMC we are able to quantify the uncertainty around the estimate from the MCMC sample or chain. The ability to quantify the uncertainty about an estimate is highly advantageous and one of the reasons why it was chosen for the purposes of this research. A simple MCMC scheme was firstly implemented, specifically the Random Walk Metropolis Algorithm (Metropolis and Ulam, 1949). An Adaptive Scaling Metropolis algorithm (Vihola, 2011) was then adopted to improve the algorithm performance, particularly that of the acceptance rate.

The transmission dynamics of an epidemic however are rarely homogeneous and generally display highly uneven transmission patterns and fluctuating trends. Super-spreading events and super-spreaders are two such examples of heterogeneous transmissibility. A super-spreading event is an event in which an infectious disease is spread to a much greater extent than usual, while an unusually contagious organism infected with a disease is known as a superspreader (Shen et al., 2004). Super-spreading, which can be brought about by any such biological, social environmental or random factors can give rise to significant individual variation around the R_0 average (Vespignani et al., 2020). This pertains to all pathogens to varying degrees, however there is evidence to suggest that the emerging coronaviruses causing severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS) and

COVID-19 are systematically prone to superspreading (Lloyd-Smith et al., 2005, Kucharski and Althaus, 2015). According to Vespignani et al., 2020, this variation around the R_0 statistic within an infectious population matters most when case numbers are small as countries attempt to prevent the establishment of community transmission. For example early in the pandemic, or after successful suppression of the outbreak where the population remains susceptible. Public health officials must remain vigilant, acknowledging that many occurrences will die out but a portion will instigate outbreaks that spread at shocking rates (Vespignani et al., 2020). Given the significance of such super-spreading events and individuals, a significant focus of this work involved the development, simulation and inspection of such models. The overall objectives of this research are outlined below.

Objectives

1. Develop a stochastic model of epidemic outbreaks, applicable and generalizable to many diseases, using a branching process and a time-varying infectivity curve, a unique and defining property of the model

2. Implement simulations under this model in order to validate it's correctness and applicability to epidemic outbreaks

3. Derive a mathematical formulation for the likelihood of this model

4. Establish a Bayesian framework to infer the parameters of the model, specifically the posterior distribution of R_0, based on case report time series

5. Implement a Markov Chain Monte Carlo (MCMC) algorithm to sample from the model parameters, i.e R_0, implementing a standard Metropolis algorithm approach and subsequently a tuned adaptive scaling variation of the Metropolis algorithm to improve the results

6. Inspect the results of the MCMC algorithms, including trace plots, histograms, acceptance rates etc.

7. Develop a stochastic model of superspreading events (SSE) during an epidemic, again using a branching process and a time-varying infectivity curve

8. Implement simulations under this model in order to validate it's correctness and applicability to super-spreading events (SSE)

9. Develop a stochastic model of super-spreaders events again using a branching process and a time-varying infectivity curve

10. Implement simulations under this model in order to validate it's correctness

11. Formulate a report of the work completed to date, including a literature review of relevant, comparative studies

2 Research Methodology

The models developed include a baseline stochastic epidemic model, a super-spreading events model and a super-spreader model. All models were subsequently implemented in the R programming language, the code for which can be found on the github page **here** (Craddock, 2021). The model frameworks are adaptive and generalizable to a wide range of diseases with a flexible generation time and the hope would be that the tool would prove useful for epidemiologists investigating a variety of diseases. MCMC was then used to sample from R_0. A simple Metropolis scheme was firstly implemented (Metropolis and Ulam, 1949) and then an Adaptive Scaling Metropolis algorithm (Vihola, 2011) was adopted to improve the algorithm performance, particularly that of the acceptance rate.

2.1 Baseline Model

A discrete-time model with branching process characteristics was developed to model the spreading of an epidemic through a susceptible population. A key characteristic of the model is the time varying infectivity allowed for each infected. This was incorporated within a framework of Bayesian inference and an adaptive MCMC scheme was used to infer the model parameters, specifically R_0.

Model

To model the spreading of the epidemic let x_t be the number of infecteds at time t. The number of infecteds x_t was assumed to be Poisson distributed with rate $R_0 \cdot \lambda_t$. Thus the probability of x_t is given by;

$$P(x_t) = exp(-R_0 \cdot \lambda_t) \cdot \frac{1}{x_t!} \cdot (R_0 \cdot \lambda_t)^{x_t},$$

whereby λ_t represents the infectious pressure of the entire population. This is the sum of the time-varying infectivity of each individual and defined as

$$\lambda_t = \sum_{i=1}^{t-1} x_i \cdot \left(\Gamma\Big((t-i); k, \theta\Big) - \Gamma\Big((t-i-1); k, \theta\Big) \right)$$

The gamma distribution was chosen to represent the time-varying infectivity of each individual. This distribution was chosen as it has positive support to infinity and allows for the temporal variability. For the purposes of simulation a $\Gamma(6, 1)$ was chosen, that is with shape 6 and scale 1, as it accounts for an initial buildup in the infectivity before peaking and subsequently dropping off (See Figure 2.1). This was deemed quite a realistic pattern of events for the infectivity of an individual. Simulations under this model were carried out as presented in the results section. The simulated data was used to emulate real epidemic data, in this instance the number of reported cases per day. This data was subsequently used for inference and as part of the MCMC sampling scheme.

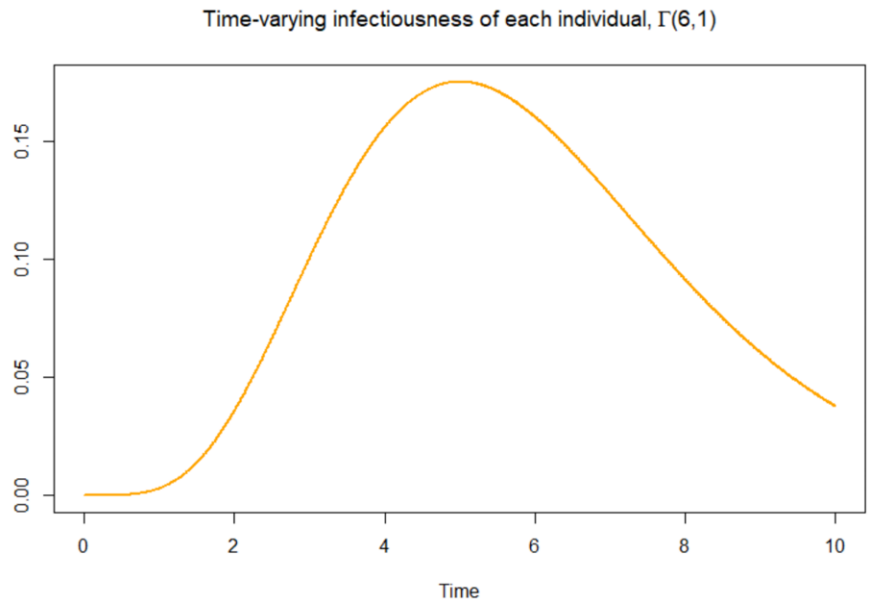


Figure 2.1: *Simulation of time-varying infectivity of each individual assumed to be a* $\Gamma(6, 1)$*, i.e with shape 6 and scale 1.*

2.1.1 Maximum Likelihood Estimation

Likelihood function

The likelihood function $L(R_0, k, \theta|\ \mathbf{x})$ is the joint probability density function viewed as a function of the model parameters with x fixed to the observed data. The likelihood principle says that all information about the parameters is contained within the likelihood function. The likelihood of this model is

$$L(R_0, k, \theta|\ \mathbf{x}) = \prod_{t=1}^{Ndays} exp(-R_0 \cdot \lambda_t) \cdot \frac{1}{x_t!} \cdot (R_0 \cdot \lambda_t)^{x_t}$$

It is often more convenient to work with the log-likelihood

$$l(R_0, k, \theta|\ \mathbf{x}) = ln\left(\prod_{t=1}^{Ndays} exp(-R_0\lambda_t) \cdot \frac{1}{x_t!} \cdot (R_0\lambda_t)^{x_t} \right)$$

$$l(R_0, k, \theta|\ \mathbf{x}) = -\sum_{t=1}^{Ndays} R_0 \cdot \lambda_t \ + \sum_{t=1}^{Ndays} ln(\frac{1}{x_t!}) \ + \sum_{t=1}^{Ndays} x_t \cdot ln(R_0 \cdot \lambda_t)$$

The maximum likelihood estimator MLE is the value of a model parameter, say $\widehat{R_0}$, which maximises $L(R_0, k, \theta|\ \mathbf{x})$. The MLE also maximises $l(R_0, k, \theta|\ \mathbf{x})$ because $log(\cdot)$ is a monotonic function and it is usually easier to maximise. Finding the MLE of R_0 involves a simple optimization problem, which, given that it is a simple case, has a closed-form expression for R_0 as a function of x;

$$\frac{dl}{dR_0} = -\sum_{t=1}^{Ndays} \lambda_t \sum_{t=1}^{Ndays} \frac{x_t}{R_0}$$

$$\widehat{R_0} = \frac{\sum_{t=1}^{Ndays} x_t}{\sum_{t=1}^{Ndays} \lambda_t}$$

2.1.2 Bayesian Inference

In the Bayesian approach to inference, parameters are treated as random variables and so have a probability distribution. Uncertainty is handled through probability, and thus assigns probability distributions to anything that is unknown, typically parameters in a model. Prior information about the model parameter, say R_0, is combined with information from sample data to estimate the distribution of the parameter. The prior information about the parameter is given by the prior distribution $p(R_0)$ and information from sample data, simulated data in this case, is given by the likelihood $L(\mathbf{x}|R_0, k, \theta) = f(x, R_0, k, \theta)$. For fixed prior parameters k and θ the posterior distribution of R_0 is given by

$$f(R_0|x) = \frac{f(x|R_0) \cdot f(R_0)}{f(x)}$$

where $f(x)$ is the marginal probability of the data x. For a continuous sample space, this marginal probability is computed as:

$$f(x) = \int f(x|R_0) \cdot f(R_0) \ dR_0$$

The marginal distribution acts as a normalizing constant to make the posterior density proper. As this denominator scales the posterior density, the posterior can be simplified as;

$$f(R_0|x) = \frac{L(R_0|x) \cdot p(R_0)}{f(x)} \propto L(R_0|x) \cdot p(R_0)$$

The likelihood is as before and a fixed gamma prior with shape k and scale θ is assigned to R_0 as follows;

$$p(R_0) = \Gamma(k, \theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot R_0^{k-1} \cdot exp(-\frac{R_0}{\theta})$$

Given $\Gamma(k, \theta)$ prior on R_0, specifically $\Gamma(1, 1)$ the posterior distribution of R_0 can be written

as follows;

$$
\begin{aligned}
f(R_0|x) &\propto \prod_{t=1}^{Ndays} exp(-R_0 \cdot \lambda_t) \cdot \frac{1}{x_t!} \cdot (R_0 \cdot \lambda_t)^{x_t} \cdot {R_0}^{k-1} \cdot exp(-\frac{R_0}{\theta}) \\
&\propto \prod_{t=1}^{Ndays} exp(-(R_0 \cdot \lambda_t + \frac{R_0}{\theta})) \cdot {R_0}^{x_t+k-1} \cdot {\lambda_t}^{x_t} \cdot \frac{1}{x_t!} \\
&\propto \prod_{t=1}^{Ndays} exp\big(\frac{R_0 \cdot \theta \cdot \lambda_t + R_0}{\theta} \big) \cdot {R_0}^{x_t+k-1} \cdot {\lambda_t}^{x_t} \cdot \frac{1}{x_t!} \\
&\propto \prod_{t=1}^{Ndays} exp\big(\frac{R_0(\theta \cdot \lambda_t + 1)}{\theta} \big) \cdot {R_0}^{x_t+k-1} \cdot {\lambda_t}^{x_t} \cdot \frac{1}{x_t!} \\
&\propto exp\big(\frac{R_0(\theta \sum_{t=1}^{Ndays} \lambda_t + 1)}{\theta} \big) \cdot {R_0}^{\sum_{t=1}^{Ndays} x_t+k-1}
\end{aligned}
$$

$$
\therefore\ f(R_0|x)\ \propto\ Gamma\Big(\sum_{t=1}^{Ndays} x_t + k,\quad \frac{\theta}{\theta \sum_{t=1}^{Ndays} \lambda_t + 1} \Big)
$$

Note the uninformative prior $\Gamma(1,1)$ which is simply an $exp(1)$ distribution. Such a prior was chosen so that the data, rather than any prior assumptions, would have the biggest effect on the posterior distribution.

2.1.3 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) was also used to infer R_0. A simple MCMC scheme was firstly implemented, specifically the Random Walk Metropolis Algorithm (Metropolis and Ulam, 1949). Although the exact analytical solution was available as above, this was carried out as an exercise in MCMC and as a good test-bed for future, multi-parameter inference. The algorithm is as below;

Algorithm 1: Random Walk Metropolis Algorithm to infer R_0

Input : Data; daily number of infecteds X, $R_0^{(0)}$, σ

Starting with $R_0^{(0)} \leftarrow 0$;

for $t \leftarrow 2$ **to** N *days* **do**

1. Draw $\epsilon \sim N(0, \sigma^2)$ and set $R_0^{'} = {R_0}^{t-1} + \epsilon$;
2. Compute; $\alpha(R_0^{'} \mid R_0^{t-1}) = \min\{\ 1,\ \dfrac{f(R_0^{'})}{f({R_0}^{t-1})}\ \}$;
$$= \min\{\ 1,\ \frac{L({R_0}^{t-1} \mid x^{1:t}, k, \theta) \cdot p(R_0)}{L(R_0 \mid x^{1:t}, k, \theta) \cdot p({R_0}^{t-1})}\ \}\ ;$$
3. With probability $\alpha(R_0^{'} \mid R_0^{t-1})$ set ${R_0}^t$ to ${R_0}^{'}$;
Otherwise set ${R_0}^t$ to ${R_0}^{t-1}$

end

A variation to this algorithm involving adaptive scaling was then implemented to improve the algorithm performance, for example the acceptance rate. The general form of the algorithm is as below (Vihola, 2011) followed by the adaption for R_0.

Adaptive Scaling

The scaling adaptation involves a function ϕ which maps real-valued parameter values S_n to a scaling in $(0, \infty)$. The scaling function $\phi : \mathbb{R} \to (0, \infty)$ is increasing and surjective, piecewise differentiable and there are constants $h, c > 0$ and $\kappa \geq 1$ such that;
$\phi^{'}(s+h) <= c \cdot max\{1,\ \phi^{\kappa}(s)\ \}$ for all $s \in \mathbb{R}$ and all $0 \leq \bar{h} \leq h$. (Vihola, 2011). The adaptive scaling algorithm is as below;

Algorithm 2: Adaptive Scaling Metropolis Algorithm (Vihola, 2011)

Input : Let $\{U_n,\ W_n\}_{n\geq 2}$ be a set of independent random variables where each U_n is uniformly distributed in the unit interval $[0, 1]$ and each W_n has the distribution q for all $n \geq 2$. Let $X_1 = x_1 \in \mathbb{R}^d$ with $\pi(x_1) > 0$, $S_1 \equiv s_1 \in \mathbb{R}$ and α^*, the optimal acceptance rate

for $n \leftarrow 2$ **to** N **do**

1. $Y_n = X_{n-1} + \phi(S_{n-1}) \cdot \Sigma^{1/2} \cdot W_n$;
2. **if** $U_n \leq \alpha_n$ **then**
   $X_n \leftarrow Y_n$;
   else
   if $U_n \leq \alpha_n$ **then**
   $X_n \leftarrow X_{n-1}$;
   end
   end
3. $S_n = S_{n-1} + \eta_n \cdot (\alpha_n - \alpha^*)$,
   where $\alpha_n = \min\{1, \frac{\pi(R_0^{'})}{\pi(R_0^{t-1})}\}$ stands for the acceptance probability

end

The adaptive scaling algorithm was applied to infer R_0 in the Bayesian epidemic model as follows;

Algorithm 3: Adaptive Scaling Metropolis Algorithm to infer R_0

Input : Data - daily number of infecteds X, $R_0^{(0)}$, σ, α^*; the optimal acceptance rate

Starting with $R_0^{(0)} \leftarrow 0$ and $S^{\,0} \leftarrow 0$;

for $t \leftarrow 2$ **to** N *days* **do**

1. Draw $Z \sim N(0,1)$ and set $R_0^{'} = R_0^{\,t-1} + exp(S^{\,t-1}) \cdot Z$;
2. Compute; $\alpha(R_0^{'} \mid R_0^{t-1}) = \min\{\ 1,\ \dfrac{f(R_0^{'})}{f(R_0^{\,t-1})}\ \}$;

$$= \min\{\ 1,\ \frac{L(R_0^{\,t-1} \mid x^{1:t}, k, \theta) \cdot p(R_0)}{L(R_0 \mid x^{1:t}, k, \theta) \cdot p(R_0^{\,t-1})}\ \}\ ;$$

3. With probability $\alpha(R_0^{'} \mid R_0^{\,t-1})$ set $R_0^{\,t}$ to $R_0^{\,'}$;

   Otherwise set $R_0^{\,t}$ to $R_0^{\,t-1}$;

4. Compute $S^{\,n} = S^{\,n-1} + \dfrac{1}{t} \cdot (\alpha_t - \alpha^*)$,

end

2.2 Super-spreading Model

The transmission dynamics of an epidemic are rarely homogeneous and generally display highly uneven transmission patterns and fluctuating trends. Super-spreading events and super-spreaders are two such examples of heterogeneous transmissibility. A super-spreading event is an event in which an infectious disease is spread to a much greater extent than usual, while a super-spreader refers to an unusually contagious organism infected with a disease (Shen et al., 2004). Both a super-spreading events model and a super-spreader model were developed, in addition to the baseline model, to capture such heterogeneous transmission dynamics. Simulations of all models were carried out to aid understanding of the model dynamics involved.

According to Shen et al., 2004, one of the most intriguing aspects of such coronaviruses as SARs has been the events under which the virus is transmitted to large groups of people. Such a superspreading event occurred in a hotel in Hong Kong in which the transmission from an ill individual from Guangdong led to export of the virus to several other countries (Shen et al., 2004). Another such super-spreading event occured onboard China Air's flight 112 from Hong Kong to Beijing in March 2003 (Liang et al., 2004). Such super-spreading events are responsible for a significant portion of resulting infections. Thus it was deemed an important aspect of epidemic transmission dynamics to consider for this piece of research. The super-spreading events model is outlined below.

2.2.1 Model

Let $x_t = y_t + z_t$

where

y_t = Number of infecteds not by a super-spreading event (NSSE)

$z_t = x_t - y_t$ = Number of infecteds infected by a super-spreading event (SSE)

Probability

$$P(x_t) = \sum_{y_t=0}^{x_t} P(y_t) \cdot P(z_t)$$

i. y_t - Non super-spreading events

$y_t \sim Poisson(\alpha \cdot \lambda_t)$,

$$p(y_t) = exp(-\alpha \cdot \lambda_t) \cdot \frac{1}{y_t!} \cdot (\alpha \cdot \lambda_t)^{y_t},$$

with

$$\lambda_t = \sum_{i=1}^{t-1} x_i \; (\; Gamma((t-i); k, \theta) - Gamma((t-i-1); k, \theta)$$

ii. z_t - super-spreading events

For each SSE event we get Poisson(γ) infections

$z_t|n_t \sim Poisson(\gamma \cdot n_t)$ = Number of infecteds by all Super-Spreading Events at time t

$n_t \sim Poisson(\beta \cdot \lambda_t)$ = Number of super-spreading events

$$p(z_t) = \sum_{n_t=0}^{\infty} p(n_t = n_t) \cdot p(z_t|n_t = n_t)$$

$\therefore p(z_t)$ is a Poisson-Poisson Compound

Determining the Mean and Variance of z_t

$(z_t) = ((z_t|n_t)\,)$

$= (\gamma \cdot n_t)$

$= \gamma(n_t)$

$$= \gamma \cdot \beta \cdot \lambda_t$$

$Var(z_t)$= Var$((z_t|n_t))$ + (Var$(z_t|n_t)$)

As $z_t|n_t$ and n_t are poisson rvs the mean and variance are equivalent giving;

$$= Var(\gamma \cdot n_t) + (\gamma \cdot n_t)$$

$$= \gamma^2 \cdot \beta \cdot \lambda_t + \gamma \cdot \beta \cdot \lambda_t$$

Poisson-Poisson Compound z_t in the form of a Negative Binomial RV

The mean and variance of the Poisson-Poisson Compound z_t is equated to that of a Negative Binomial distribution of size $\boldsymbol{n}$ with probability of success $\boldsymbol{p}$ and density given by;

$$\frac{\Gamma(x+n)}{\Gamma(n) \cdot x!} \cdot p^n (1-p)^x$$

for $x = 0, 1, 2, ..., n > 0$ and $0 < p <= 1$.

This represents the number of failures which occur in a sequence of Bernoulli trials before a target number of successes is reached. The mean and variance are respectfully;

$$\mu = \frac{n(1-p)}{p}, \qquad Var = \frac{n(1-p)}{p^2}$$

Equating the poisson-poisson compound to that of the negative binomial distribution gives;

I. $\mu = \dfrac{n(1-p)}{p} = \gamma \cdot \beta \cdot \lambda_t$

II. $Var = \dfrac{n(1-p)}{p^2} = \gamma^2 \cdot \beta \cdot \lambda_t + \gamma \cdot \beta \cdot \lambda_t$

From **I** we can say that the size

$$n = \frac{p \cdot \gamma \cdot \beta \cdot \lambda_t}{1-p} \qquad \textbf{III.}$$

Substituting **III** into **II** gives;

$$\frac{(\frac{p \cdot \gamma \cdot \beta \cdot \lambda_t}{1-p}) \cdot (1-p)}{p^2} = \gamma^2 \cdot \beta \cdot \lambda_t + \gamma \cdot \beta \cdot \lambda_t$$

Cancelling like terms gives;

$$\frac{(\frac{p \cdot \gamma \cdot \beta \cdot \lambda_t}{1-p}) \cdot (1-p)}{p^2} = \gamma^2 \cdot \beta \cdot \lambda_t + \gamma \cdot \beta \cdot \lambda_t$$

$$\frac{1}{p} = \gamma + 1$$

$$1 = p \cdot \gamma + p$$

$$1 = p \cdot \gamma + p$$

$$\therefore p = \frac{1}{\gamma + 1} \qquad \textbf{IV.}$$

Substituting **IV** into **III** gives;

$$n = \frac{\frac{1}{\gamma+1} \cdot \gamma \cdot \beta \cdot \lambda_t}{1 - \frac{1}{\gamma+1}}$$

$$n = \frac{\frac{\gamma \cdot \beta \cdot \lambda_t}{\gamma+1}}{\frac{\gamma+1-1}{\gamma+1}}$$

$$n = \frac{\frac{\gamma \cdot \beta \cdot \lambda_t}{\gamma + 1}}{\frac{\gamma}{\gamma + 1}}$$

$$\therefore n = \beta \cdot \lambda_t$$

z_t can thus be written as a Negative Binomial of size n with probability of success p, i.e

$$z_t \sim \text{NB}(\beta \cdot \lambda_t, \ \frac{1}{\gamma + 1})$$

$$p(z_t) = \frac{\Gamma(z_t + \beta \cdot \lambda_t)}{\Gamma(\beta \cdot \lambda_t) \cdot z_t!} \cdot \frac{1}{\gamma + 1}^{\beta \cdot \lambda_t} \cdot (1 - \frac{1}{\gamma + 1})^{z_t} \qquad \text{for } z_t = 0, 1, 2, ..., n > 0 \text{ and } 0 < p <= 1$$

2.2.2 Likelihood

$$P(x_t) = \sum_{y_t=0}^{x_t} P(y_t) \cdot P(z_t)$$

$$L(\alpha, \beta, \gamma, \lambda | \ \mathbf{x_t}) = \prod_{t=1}^{Ndays} \sum_{y_t=0}^{x_t} P(y_t) \cdot P(z_t)$$

$$= \prod_{t=1}^{Ndays} \sum_{y_t=0}^{x_t} exp(-\alpha \cdot \lambda_t) \cdot \frac{1}{y_t!} \cdot (\alpha \cdot \lambda_t)^{y_t} \times \frac{\Gamma(z_t + \beta \cdot \lambda_t)}{\Gamma(\beta \cdot \lambda_t) \cdot z_t!} \cdot \frac{1}{\gamma + 1}^{\beta \cdot \lambda_t} \cdot (1 - \frac{1}{\gamma + 1})^{z_t}$$

$$= \prod_{t=1}^{Ndays} \sum_{y_t=0}^{x_t} exp(-\alpha \cdot \lambda_t) \cdot \frac{1}{y_t!} \cdot (\alpha \cdot \lambda_t)^{y_t} \times \frac{\Gamma(z_t + \beta \cdot \lambda_t)}{\Gamma(\beta \cdot \lambda_t) \cdot z_t!} \cdot \frac{1}{\gamma + 1}^{\beta \cdot \lambda_t} \cdot (\frac{\gamma}{\gamma + 1})^{z_t}$$

Log Likelihood

$$l(\alpha, \beta, \gamma, \lambda | \ \mathbf{x_t}) = ln\left(\prod_{t=1}^{Ndays} \sum_{y_t=0}^{x_t} P(y_t) \cdot P(z_t)) \right)$$

$$= \sum_{t=1}^{Ndays} ln\left(\sum_{y_t=0}^{x_t} (exp(-\alpha \cdot \lambda_t) \cdot \frac{1}{y_t!} \cdot (\alpha \cdot \lambda_t)^{y_t} \times \frac{\Gamma(z_t + \beta \cdot \lambda_t)}{\Gamma(\beta \cdot \lambda_t) \cdot z_t!} \cdot \frac{1}{\gamma + 1}^{\beta \cdot \lambda_t} \cdot (\frac{\gamma}{\gamma + 1})^{z_t} \right)$$

2.3 Super-spreaders Model

A super-spreader, as stated, refers to an unusually contagious individual. For example, sexually transmitted diseases where promiscuity drives risk (Lloyd-Smith et al., 2005) The super-spreaders model is outlined below

2.3.1 Model

Let $x_t = \psi_t + \zeta_t$

whereby

ψ_t = Number of non-superspreaders

$\zeta_t = x_t - \psi_t$ = Number of superspreaders

Probability

$$P(X_t = x_t) = \sum_{k=0}^{x_t} P(\psi_t = k) \cdot P(\zeta_t = x_k - k)$$

2.3.2 i. ψ_t - Non super-spreading individuals

$\psi_t \sim Poisson(a \cdot \lambda_t)$,

$$p(\psi_t) = exp(-a \cdot \lambda_t) \cdot \frac{1}{\psi_t!} \cdot (a \cdot \lambda_t)^{\psi_t},$$

with

$$\lambda_t = \sum_{i=1}^{t-1} \Big(\psi_i + c \cdot \zeta_i \Big) \cdot \Big(\Gamma(t-i;k,\theta) - \Gamma(t-i-1;k,\theta) \Big)$$

2.3.3 ii. ζ_t - Super-spreading individuals

$\zeta_t \sim Poisson(b \cdot \lambda_t)$,

$$p(\zeta_t) = exp(-b \cdot \lambda_t) \cdot \frac{1}{\zeta_t!} \cdot (b \cdot \lambda_t)^{\zeta_t},$$

2.4 Likelihood

$$P(x_t) = \sum_{\psi=0}^{x_t} P(\psi) \cdot P(z_t)$$

$$L(a, b, p, \lambda | \, \mathbf{x_t}) = \prod_{t=1}^{Ndays} \sum_{\psi=0}^{x_t} P(\psi) \cdot P(\zeta)$$

$$= \prod_{t=1}^{Ndays} \sum_{\psi=0}^{x_t} \left(p \cdot exp(-a \cdot \lambda_t) \cdot \frac{1}{\psi!} \cdot (a \cdot \lambda_t)^{\psi} \quad + \quad (1-p) \cdot exp(-b \cdot \lambda_t) \cdot \frac{1}{\zeta!} \cdot (b \cdot \lambda_t)^{\zeta} \right)$$

Log Likelihood

$$l(\alpha, \beta, \gamma, \lambda | \, \mathbf{x_t}) = ln \left(\prod_{t=1}^{Ndays} \sum_{\psi=0}^{x_t} P(\psi) \cdot P(\zeta)) \right)$$

$$= \sum_{t=1}^{Ndays} ln \left(\sum_{\psi=0}^{x_t} p \cdot exp(-a \cdot \lambda_t) \cdot \frac{1}{\psi!} \cdot (a \cdot \lambda_t)^{\psi} \quad + \quad (1-p) \cdot exp(-b \cdot \lambda_t) \cdot \frac{1}{\zeta!} \cdot (b \cdot \lambda_t)^{\zeta} \right)$$

3 Results

The main results presented include simulations of the baseline model, simulations of the super-spreading events model and simulations of the super-spreaders model. The results of the Adaptive Scaling Metropolis algorithm for generating samples of R_0 as in the baseline model are then presented.

3.1 Simulations

Baseline Model

Simulations from the baseline model are shown below. The model is as detailed in the Methodology section. A $\Gamma(6, 1)$ was again used to model the time-varying infectivity. The simulated data is used to emulate real epidemic data, i.e the number of reported cases per day.

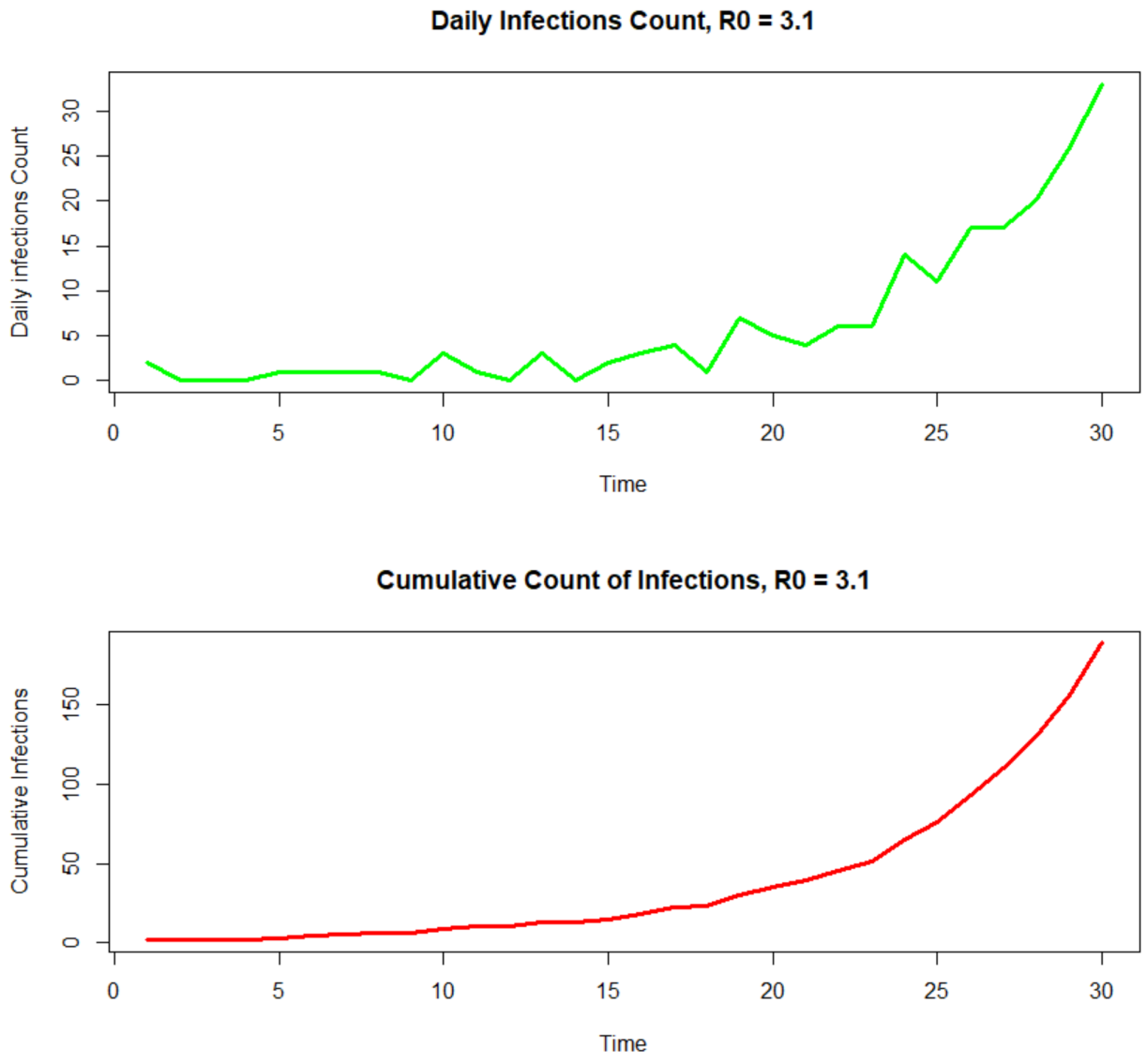


Figure 3.1: *Data generated from simulation of the baseline model (Raw count; top, cumulative count; bottom). R_0 was set to 3.1 and the time-varying infectivity of each individual was assumed to be a $\Gamma(6, 1)$.*

Super-Spreading Events Model

Data generated from simulations of the super-spreading events model are shown below. The number of infecteds from non super-spreading events (*y; NSSE*) and the number of infecteds from super-spreading events (*z; SSE*) along with the total (*NSSE + SSE; x*) are displayed. The increased infection count compared to the baseline model is significant. As expected, the number of infections from the super-spreading event is greater than that of the non super-spreading event.

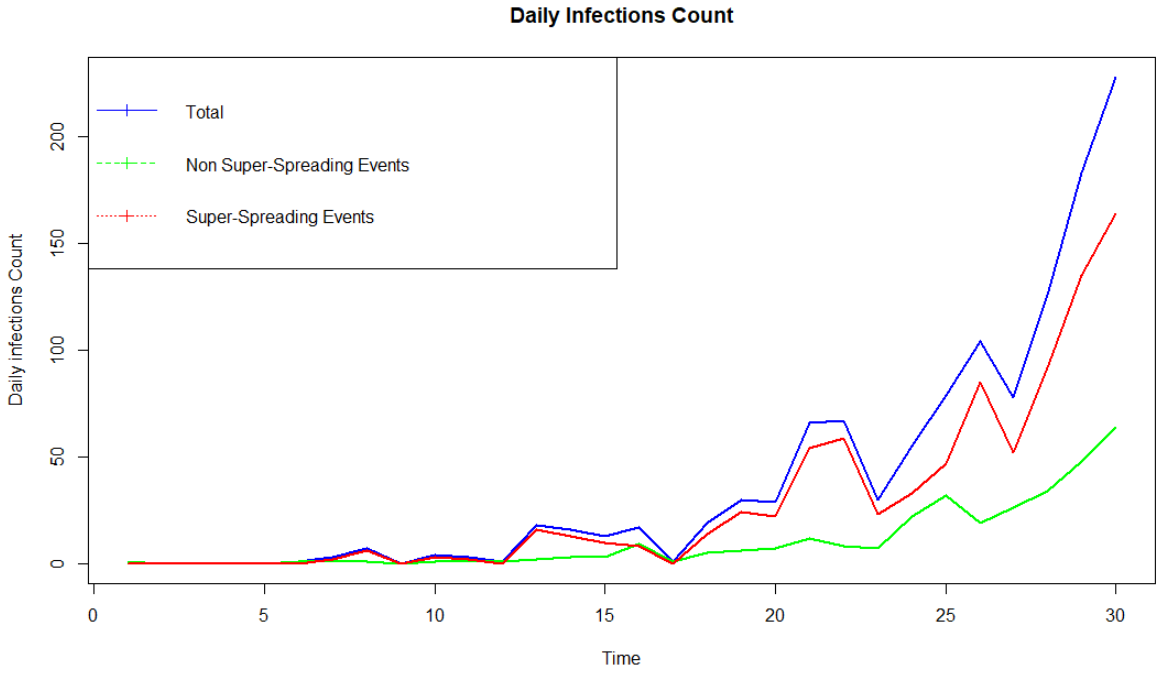


Figure 3.2: *Raw count of the daily infections generated from simulation of the super-spreading events model. The number of infecteds from non super-spreading events (*y, NSSE*) and the number of infecteds from super-spreading events (*z, SSE*) along with the total (*NSSE + SSE, x*) are shown*

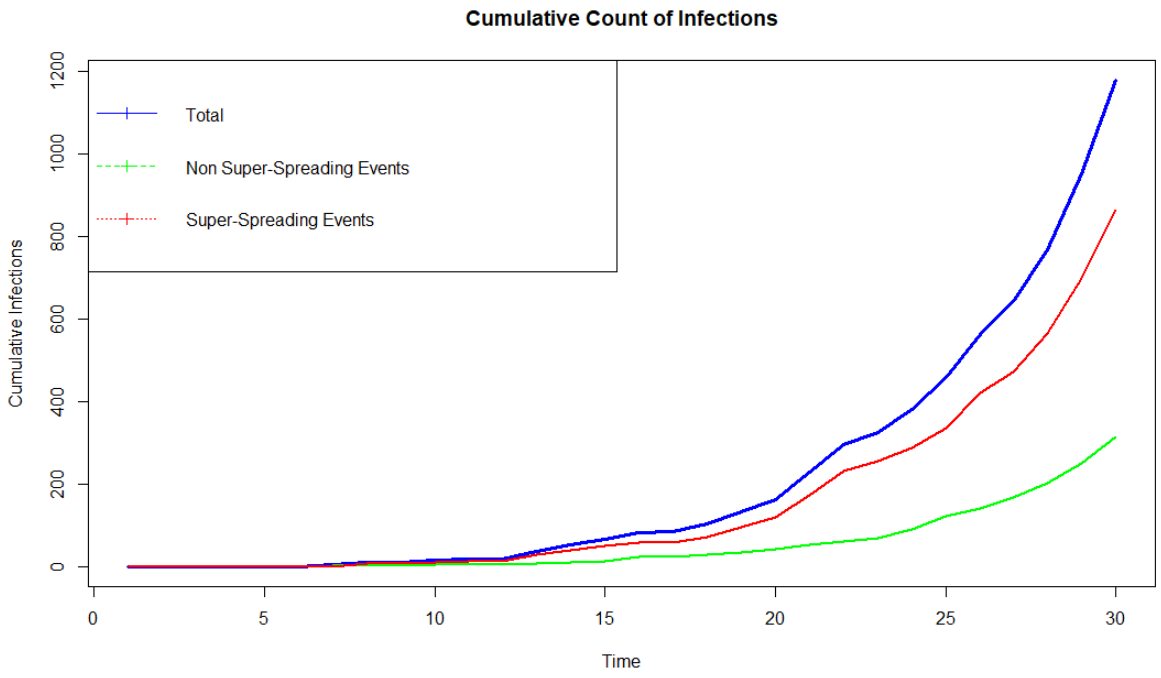


Figure 3.3: *Cumulative count of the daily infections generated from simulation of the super-spreading events model. The number of infecteds from non super-spreading events (*y, NSSE*) and the number of infecteds from super-spreading events (*z, SSE*) along with the total (*NSSE + SSE, x*) are shown*

3.1.1 Super-Spreaders Model

Data generated from Simulation of super-spreaders model. The number of non super-spreader infecteds (ψ_t) and the number of super-spreader infecteds (ζ_t) along with the total ($x_t = \psi_t + \zeta_t$) are shown. The increased infection count compared to the baseline model is significant. As expected, the number of infections due to super-spreaders is greater than that of non super-spreaders.

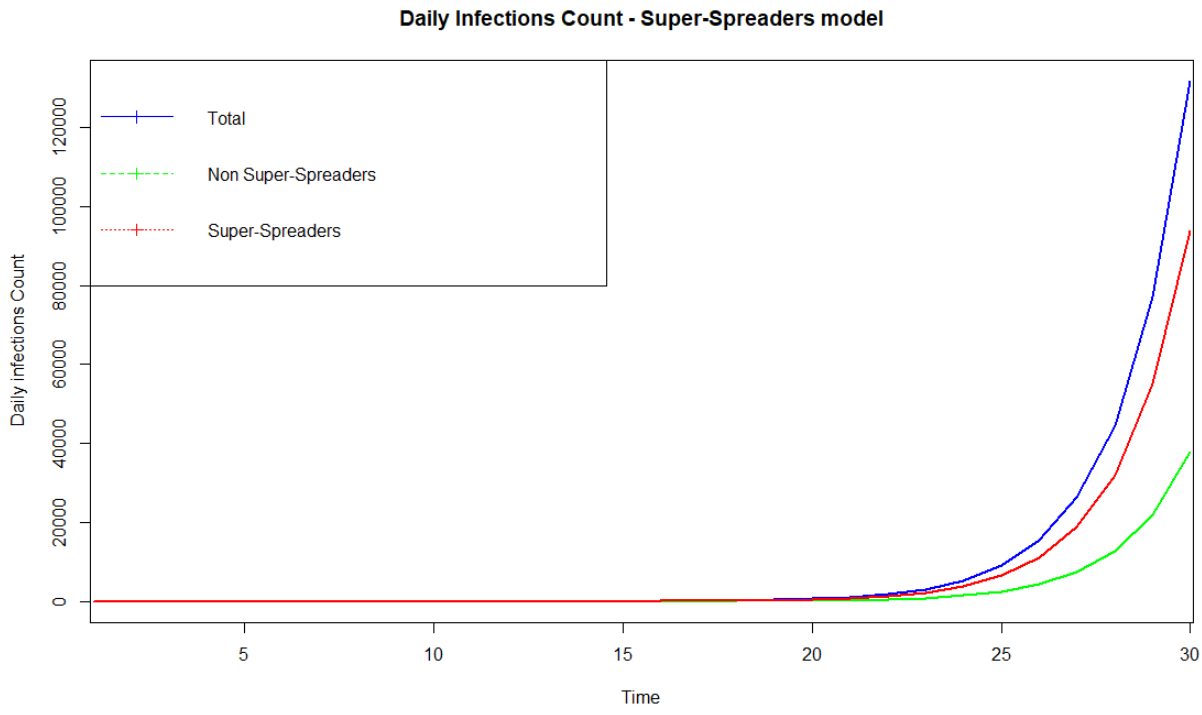


Figure 3.4: *Raw count of the daily infections generated from simulation of the super-spreaders model. The number of non super-spreader infecteds (ψ_t) and the number of super-spreader infecteds (ζ_t) along with the total ($x_t = \psi_t + \zeta_t$) are shown* $x_t = \psi_t + \zeta_t$

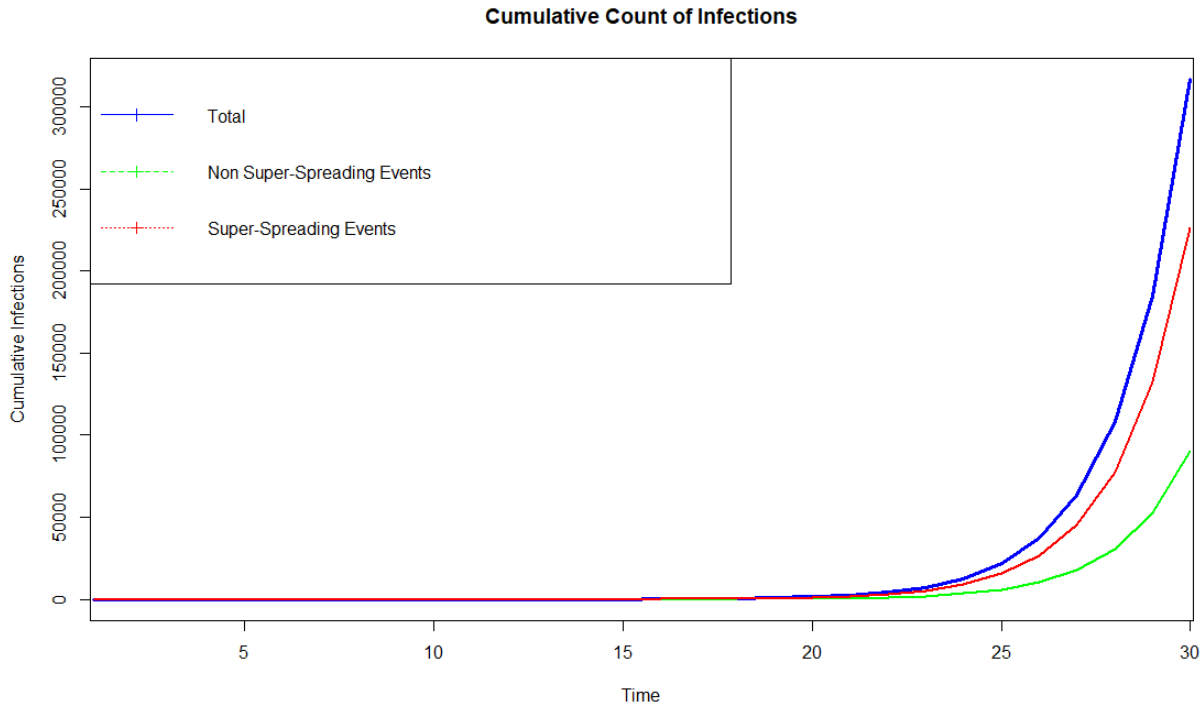


Figure 3.5: *Cumulative count of the daily infections generated from simulation of the super-spreaders model. The number of non super-spreader infecteds (ψ_t) and the number of super-spreader infecteds (ζ_t) along with the total ($x_t = \psi_t + \zeta_t$) are shown* $x_t = \psi_t + \zeta_t$

3.2 MCMC Results

The results of the MCMC scheme, specifically the 'Optimal Scaling Metropolis Algorithm' (Vihola, 2011) are shown below. The data were simulated with a range of R_0 values to test the efficacy of the algorithm. In all cases the MCMC chain displays good mixing and exploration of the state space, as seen in the trace plots below. In most cases the algorithm appears to have reached an adequate value of σ, the standard deviation of the proposal. That is, a value at which it mixes in an optimal manner, finding a good balance between exploring the space sufficiently without proposing candidates that are too extreme. For certain values of R_0 the results are more accurate, with the ultimate mean of the sampled chain closely matching that of the original simulated data. For other values there is a greater discrepancy between the true value and that of the mean of the mcmc sample. To illustrate this a comparison is made between the true R_0 value and the mean of the sampled MCMC chain in the plot below. The 95% confidence intervals of the MCMC samples are given by the blue error bars. The small error bars indicate that the MCMC samples are generally good estimates. The ability to generate such confidence intervals highlights the well-founded capacity of MCMC to quantify the uncertainty of its generated sample. This is unlike other statistical methods such as point estimation or maximum likelihood estimation which are unable to quantify the uncertainty around their estimate. Instead these methods only output a single value. In the case of $R_0 = 3.5$, the MCMC chain displays very good mixing and exploration of the state space as shown in Figure 3.6. The final mean of the sampled MCMC chain is a good estimation of the true R_0 with a mean value of 3.52. The algorithm is effectively able to sample values of R_0.

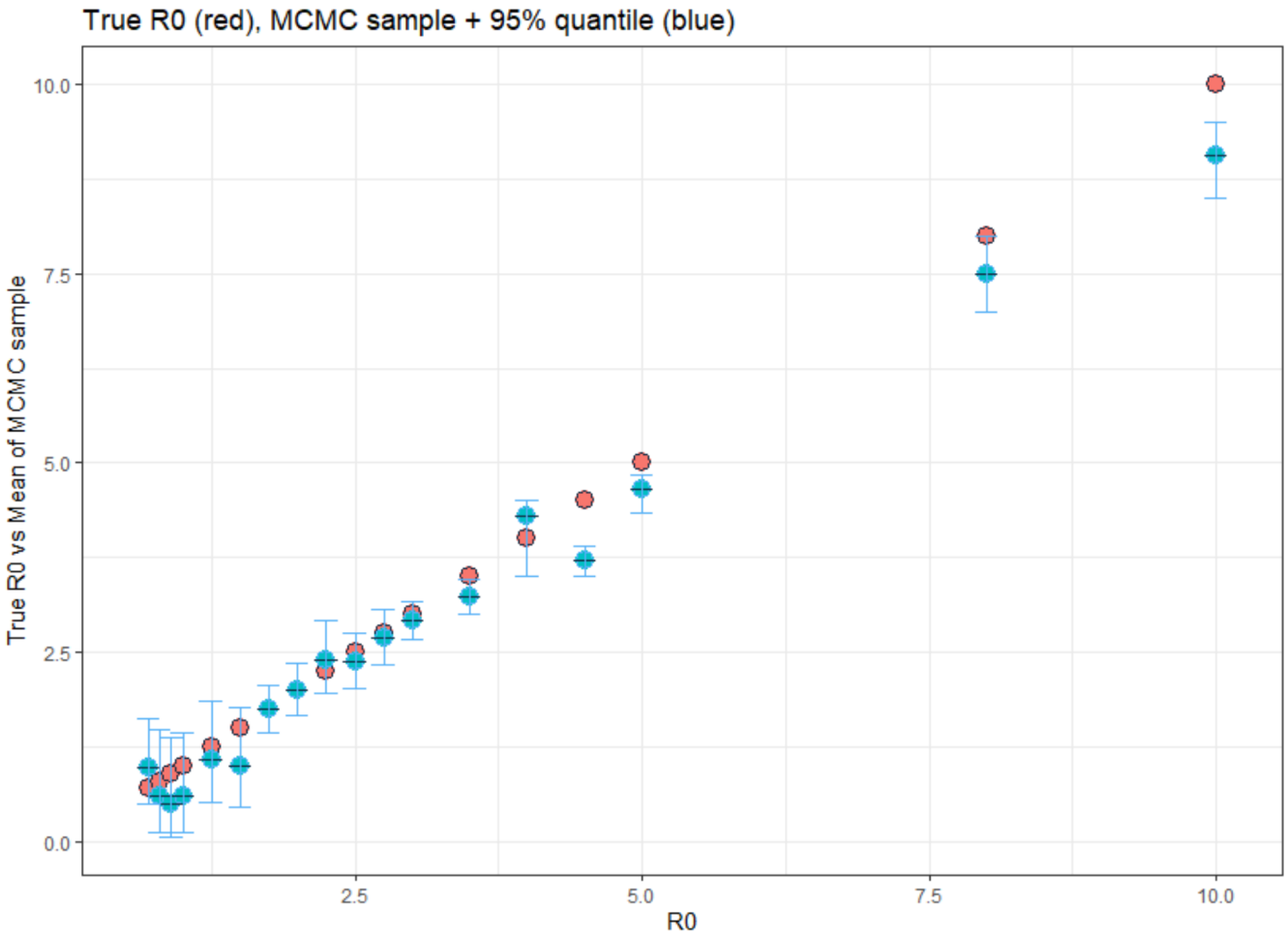
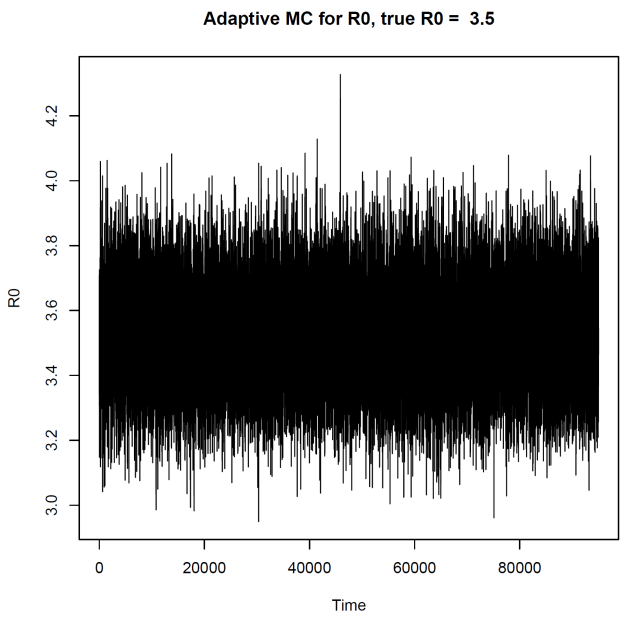
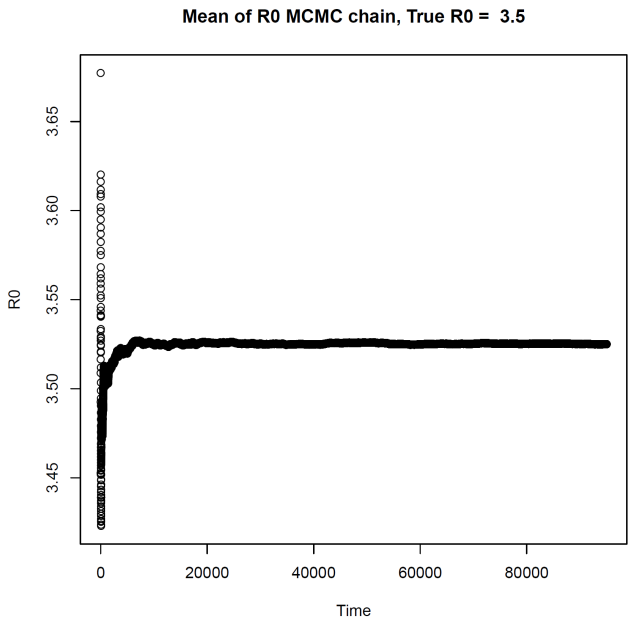


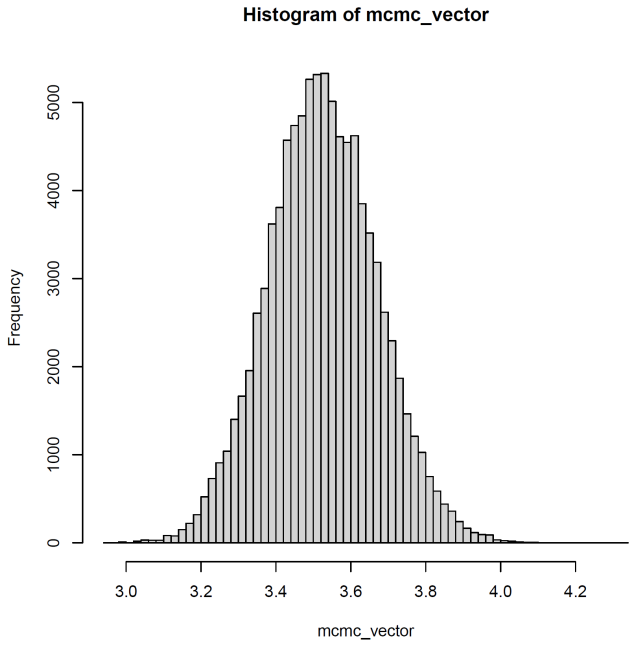
Figure 3.6: *The true R_0 value of the simulated data vs the mean of the sampled MCMC chain. The 95% confidence intervals of the MCMC samples are shown in blue. The ability to generate such confidence intervals highlights the well-founded capacity of MCMC to quantify the uncertainty of its generated sample.*



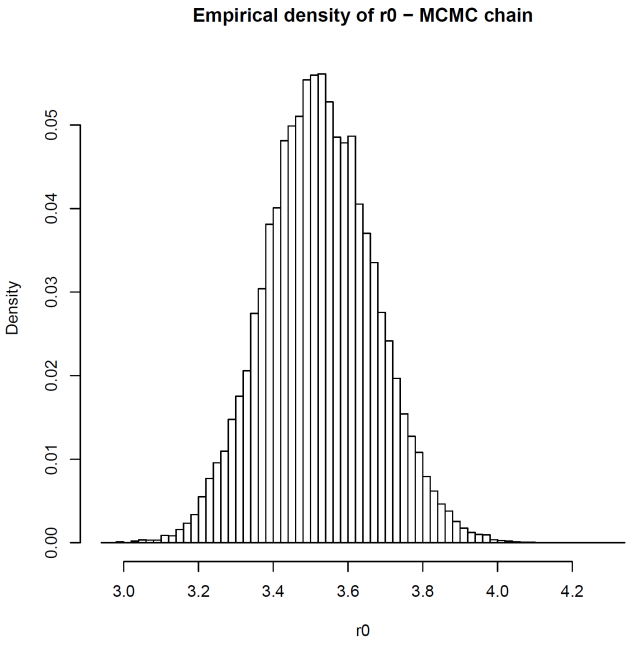
(a) Trace plot



(b) Cumulative Mean



(c) Histogram



(d) Empirical density

Figure 3.7: Trace plot, cumulative mean, histogram and empirical density of the MCMC samples of R_0 obtained from the Adaptive Scaling Metropolis Algorithm.

Acceptance Rate

The acceptance rate of the 'Optimal Scaling Metropolis Algorithm' (Vihola, 2011) across all values of R_0 is relatively similar, as plotted below. This illustrates that the Adaptation in the algorithm is working.

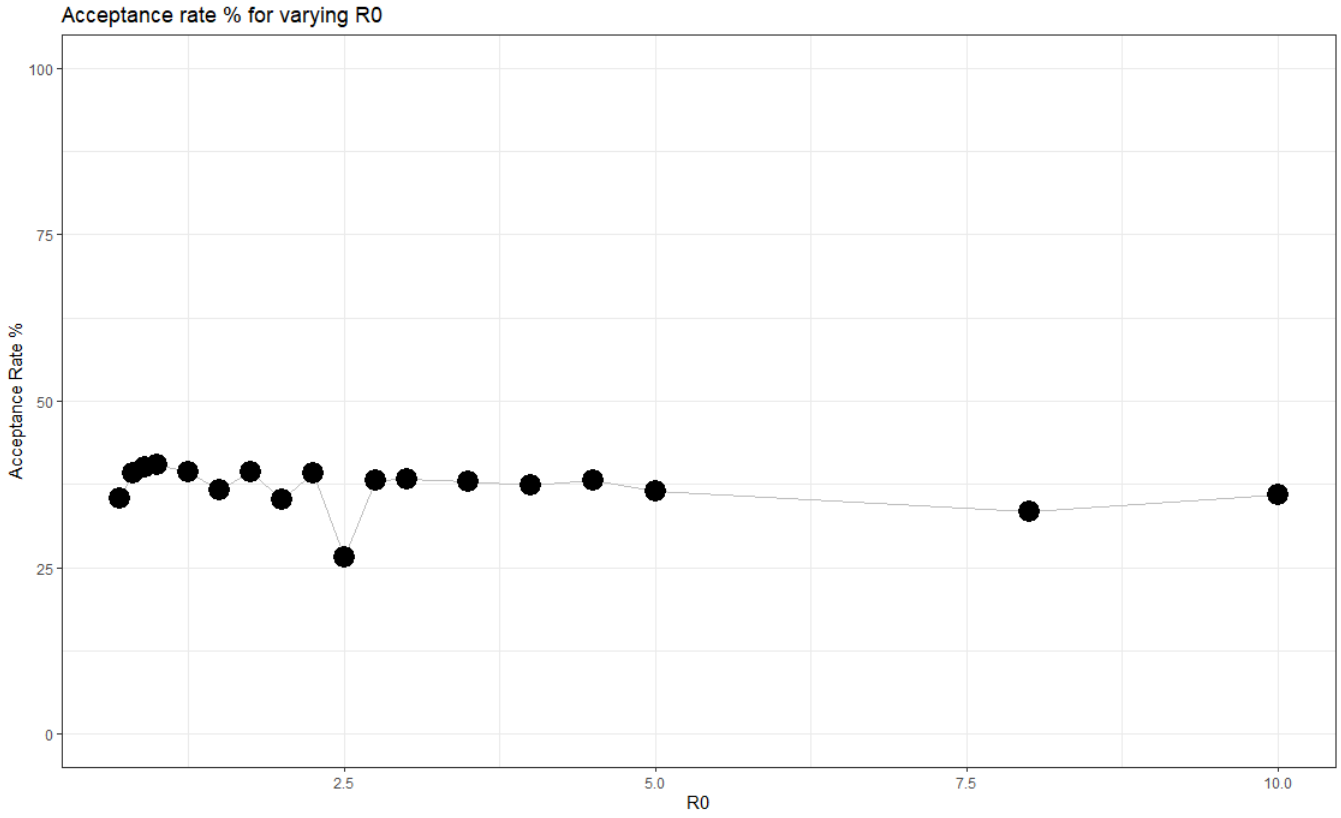


Figure 3.8: *Acceptance rate of the Optimal Scaling Metropolis Algorithm for varying* R_0.

4 Discussion and Conclusion

Infectious diseases continue to pose a major threat to human health on a global scale. The current COVID-19 pandemic and other recent outbreaks such as SARS (2002) and Ebola (2013) highlight the immense burden they can impose on society. Statistical models have been established as an essential tool for understanding the outbreak of epidemics and their transmission dynamics. (Cori et al., 2013, Vespignani et al., 2020)

Here, a discrete time model with branching process characteristics was developed to model the outbreak of an epidemic within a susceptible population with the aim of producing a model as realistic to true events as possible. A key characteristic of the model is the time varying infectivity allowed for each infected, allowing for flexibility and applicability to a wide range of diseases with different infectivity curves. The model was incorporated within a framework of Bayesian inference and an adaptive MCMC scheme was used to infer the reproduction number R_0. MCMC has become one of the foremost methods for sampling from a distribution (Van Ravenzwaaij et al., 2018). Unlike other methods like point estimation or maximum likelihood estimation which only output a single value, with MCMC the uncertainty around the estimate can be easily quantified from the MCMC sample or chain. The ability to quantify the uncertainty about an estimate is one of method's key advantages and one of the reasons why it was chosen for the purposes of this research. A simple Metropolis scheme was firstly used to sample from R_0 (Metropolis and Ulam, 1949) and an Adaptive Scaling Metropolis algorithm (Vihola, 2011) was then adopted to improve the algorithm performance, particularly that of the acceptance rate.

The transmission dynamics of an epidemic are rarely homogeneous and both a super-spreading events model and a superspreader model were also developed to capture such heterogeneous transmission dynamics. Simulations of all models were carried out and the increased infection

count generated from the super-spreading models compared to that of the baseline model was significant.

Immediate future plans envisaged for this work include a detailed review of both the baseline model and super-spreading models. The correctness of the specific mathematical models derived, as detailed in the Research Methodology section, will be further examined. A variety of prior distributions will also be considered. A review of the model assumptions will be carried out and whether these assumptions are justified by the data will be considered. Further development of the super-spreading events and super-spreader frameworks in order to infer the model parameters via the Adaptive scaling Metropolis algorithm will also be carried out.

A further vision for the project includes model validation and exploring the concept of influential statistics or influence. Validation involves a formal verification of the model and the task of confirming that the outputs of a statistical model are acceptable with respect to the true data-generating process (Poolla et al., 1994). This will include making use of Bayes factors and validation of the inferred posterior predictive distributions of the models. The concept of influence involves quantifying the effect of individual data points or groups of data points on a statistical analysis (Chatterjee and Hadi, 1986). An MCMC tool will be developed to investigate the influence of certain data points. Specifically, importance sampling will be used to re-weight the MCMC samples, with certain, potentially influencing data points removed, to investigate their influence on the inferred model parameters and results. Importance sampling will prevent having to run the MCMC algorithm repeated times with each data points removed separately.

Another key, future objective is the deployment of the models to open-source, software packages that can be freely accessible and used by researchers in the field. Given that our model framework is adaptive and generalizable to a wide range of diseases with a flexible generation time, the hope would be that the tool would prove useful for epidemiologists investigating a variety of diseases. The R code developed for this research is already open source on github **here** (Craddock, 2021), so we would envisage this objective as a natural next step. The in-

corporation with data and application of the model framework to real data sources is another aim of the project ("Public Health Engalnd", 2021, World Health Organisation, 2021). The number of reported cases per day for a specific epidemic is one of the key data sources of an epidemic. Here we emulate such data by means of simulation, with the longer term aim of applying the Bayesian model framework to real epidemic data. As stated the framework is generalizable to any human-to-human infectious disease so we are not restricted to a specific dataset. Identifying potential sources of accessible and relevant data will be necessary. Data pertaining to super-spreading events and super-spreaders will also be of keen interest.

Of course we would anticipate several challenges in this future endeavor. Inferring the dynamics of a rapidly changing infectious disease and performing inference of imperfect data in complex environments poses many technical challenges (Vespignani et al., 2020). In the early stages of an outbreak, and sometimes even later on, the only data available is the number of reported cases per day. Such case reports are typically the result of infection that happened days before, may exclude some mild or even asymptomatic cases, and may also include unrelated cases that are not part of the outbreak. Thus data issues such as missing or uncertain counts at certain times or spaces will have to be addressed.

In general, statistical models remain an essential component in the handling of infectious diseases and the outbreak of epidemics. This has been emphasised to a great extent in recent times in the context of the COVID-19 pandemic. Epidemic models have been fundamental to understanding the course of the disease and essential to health care systems and government at the highest level worldwide, enabling effective control strategies to be implemented (Giordano et al., 2020, World Health Organisation, 2007). By understanding the underlying dynamics of the disease and it's transmission, we can better target interventions to decelerate it's spread. Epidemic models will continue to be of major importance in combating COVID-19 and in the context of future epidemic outbreaks. We feel that the generalisability of our framework and the relevance to the prevailing epidemic nicely positions this piece of research. The hope is that, through collaborative work between epidemiologists, field workers, public health officials, statisticians and society at large, we can strive to mitigate the worst effects

of infectious diseases and epidemics.

A Appendix

Super-Spreading Model - log sum exp implementation

Simplest case x_1, x_2

Let $L_i = log(x_i)$

$$log(x_1 + x_2) = log\Big(exp(L_1) + exp(L_2)\Big)$$

$$= log\Bigg(\big(exp(L_1) + exp(L_2)\big) * \frac{exp(M)}{exp(M)}\Bigg)$$

$$= log\big(exp(M)\big) \;\; + \;\; log\Bigg(\frac{exp(L_2)}{exp(M)} + \frac{exp(L_2)}{exp(M)}\Bigg)$$

$$= M + log\Big(exp(L_1 - M) + exp(L_2 - M)\Big)$$

In general;

$$LSE(x_1, ..., x_n) = x^* + log\Big(exp(x_1 - x^*) + ... + exp(x_n - x^*)\Big)$$

where

$$x^* = max(x_1, ..., x_n)$$

For this super-spreading events model x_i is underlined in the likelihood;

$$l(\alpha,\beta,\gamma,\lambda|\ \mathbf{x_t}) = \sum_{t=1}^{Ndays} ln\left(\sum_{y_t=0}^{x_t}(exp(-\alpha\cdot\lambda_t)\cdot\frac{1}{y_t!}\cdot(\alpha\cdot\lambda_t)^{y_t}\times\frac{\Gamma(z_t+\beta\cdot\lambda_t)}{\Gamma(\beta\cdot\lambda_t)\cdot z_t!}\cdot\frac{1}{\gamma+1}^{\beta\cdot\lambda_t}\cdot(\frac{\gamma}{\gamma+1})^{z_t}\right)$$

$$x_i = exp(-\alpha\cdot\lambda_t)\cdot\frac{1}{y_t!}\cdot(\alpha\cdot\lambda_t)^{y_t}\times\frac{\Gamma(z_t+\beta\cdot\lambda_t)}{\Gamma(\beta\cdot\lambda_t)\cdot z_t!}\cdot\frac{1}{\gamma+1}^{\beta\cdot\lambda_t}\cdot(\frac{\gamma}{\gamma+1})^{z_t} \text{ for a given } y_t$$

and

$$L_i = log(x_i)$$

$$= -(\alpha\cdot\lambda_t)\ -\ lfactorial(y_t)\ +\ y_t\cdot log(\alpha\cdot\lambda_t)\ +\ lgamma(z_t+\beta\cdot\lambda_t)\ -\ lgamma(\beta\cdot\lambda_t)$$

$$-\ lfactorial(z_t)\ -\beta\cdot\lambda_t\cdot log(\gamma+1)\ +z_t\cdot log(\gamma)\ -\ z_t\cdot log(\gamma+1)$$

Replacing z_t with $x_t - y_t$ gives;

$$L_i = -(\alpha\cdot\lambda_t)\ -\ lfactorial(y_t)\ +\ y_t\cdot log(\alpha\cdot\lambda_t)\ +lgamma((x_t-y_t)+\beta\cdot\lambda_t)-\ lgamma(\beta\cdot\lambda_t)$$

$$-lfactorial(x_t-y_t)\ -\beta\cdot\lambda_t\cdot log(\gamma+1)\ +(x_t-y_t)\cdot log(\gamma)\ -(x_t-y_t)\cdot log(\gamma+1)$$

A.0.1 Github

Repository link: https://github.com/hanmacrad2/epidemic_modelling

References

Chatterjee, S., & Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical science* (1986), 379–393.

Cori, A., Ferguson, N. M., Fraser, C., & Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology. 178.*(9) (2013), 1505–1512.

Craddock, H. (2021). Epidemic modelling; github.

Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T., & Jacobsen, K. H. (2019). Complexity of the basic reproduction number (r0). *Emerging infectious diseases. 25.*(1) (2019), 1.

Ferguson, N. M., Cummings, D. A., Fraser, C., Cajka, J. C., Cooley, P. C., & Burke, D. S. (2006). Strategies for mitigating an influenza pandemic. *Nature. 442.*(7101) (2006), 448–452.

Fraser, C., Riley, S., Anderson, R. M., & Ferguson, N. M. (2004). Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences. 101.*(16) (2004), 6146–6151.

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., & Colaneri, M. (2020). Modelling the covid-19 epidemic and implementation of population-wide interventions in italy. *Nature Medicine* (2020), 1–6.

Keeling, M. J., & Rohani, P. (2011). *Modeling infectious diseases in humans and animals.* Princeton University Press.

Kucharski, A., & Althaus, C. L. (2015). The role of superspreading in middle east respiratory syndrome coronavirus (mers-cov) transmission. *Eurosurveillance. 20.*(25) (2015), 21167.

Liang, W., Zhu, Z., Guo, J., Liu, Z., He, X., Zhou, W., Chin, D. P., Schuchat, A., Group, B. J. S. E., et al. (2004). Severe acute respiratory syndrome, beijing, 2003. *Emerging infectious diseases. 10.*(1) (2004), 25.

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature. 438.*(7066) (2005), 355–359.

Metropolis, N., & Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association. 44.*(247) (1949), 335–341.

Nishiura, H., & Chowell, G. (2009). The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. *Mathematical and statistical estimation approaches in epidemiology* (pp. 103–121). Springer.

Poolla, K., Khargonekar, P., Tikku, A., Krause, J., & Nagpal, K. (1994). A time-domain approach to model validation. *IEEE Transactions on automatic control. 39.*(5) (1994), 951–959.

Public health engalnd [Accessed: 2021-08-27]. (2021).

Shen, Z., Ning, F., Zhou, W., He, X., Lin, C., Chin, D. P., Zhu, Z., & Schuchat, A. (2004). Superspreading sars events, beijing, 2003. *Emerging infectious diseases. 10.*(2) (2004), 256.

Van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to markov chain monte–carlo sampling. *Psychonomic bulletin & review. 25.*(1) (2018), 143–154.

Vespignani, A., Tian, H., Dye, C., Lloyd-Smith, J. O., Eggo, R. M., Shrestha, M., Scarpino, S. V., Gutierrez, B., Kraemer, M. U., Wu, J., et al. (2020). Modelling covid-19. *Nature Reviews Physics. 2.*(6) (2020), 279–281.

Vihola, M. (2011). On the stability and ergodicity of adaptive scaling metropolis algorithms. *Stochastic processes and their applications. 121.*(12) (2011), 2839–2860.

World Health Organisation. (2007). Global public health threats in the 21st century.

World Health Organisation. (2021). World health organisation.

Zhang, Y., Li, Y., Wang, L., Li, M., & Zhou, X. (2020). Evaluating transmission heterogeneity and super-spreading event of covid-19 in a metropolis of china. *International journal of environmental research and public health. 17.*(10) (2020), 3705.