Using Probabilistic Models to Infer Infection Rates in Viral Outbreaks

Candidate Number: 34904

Hilary Term 2007

Abstract

Mathematical modelling of epidemics is very important in gaining an understanding of the spread of infections. Recently, there has been interest in stochastic models for epidemics. We analyse two epidemic models - the Susceptible-Infective-Recovered (SIR) model and the Reed-Frost Epidemic Model. We try to estimate the infection rate for the Reed-Frost Epidemic model using both classical maximum likelihood methods and Gibbs Sampling. We also use Approximate Bayesian Computation to estimate the rate of infection and recovery for the SIR model. We cannot apply maximum likelihood and Gibbs Sampling methods to the SIR model as its likelihood is not readily computed. The results for the maximum likelihood and Gibbs Sampling methods for the Reed-Frost model agree, but the model does not appear to fit the data well. When considering the SIR model Approximate Bayesian Computation accurately returns parameter values when we have complete data. It performs less well when we have incomplete data. Further work needs to be carried out before we can say whether Approximate Bayesian Computation is useful when considering real-life data sets, which often have incomplete data.

Acknowledgements

Many thanks to Jotun Hein for his guidance. It has been much appreciated. Also, special thanks to Dave Dale for his help and for writing the Approximate Bayesian Computation program.

Contents

List of Figures

Chapter 1

Introduction and Motivation

Diseases cause more deaths throughout the world than wars and famines, so their study is vital. Over the years many different models have been discovered and analysed. Using mathematical models to describe epidemics can be useful in giving estimates for the level of vaccination required and, as is the case in this project, to estimate the rates of infection and recovery for the epidemic.

One of the simplest models for infections and epidemics is the Susceptible-Infective-Recovered (SIR) model. This model has been analysed both in its deterministic and probabilistic form. In this project we are primarily concerned with looking at the probabilistic form.

When analysing the stochastic form of a model, Bayesian models can be useful as the parameters of interest are generally given in terms of individuals in the population so lead us to consider the distribution of these parameters over the whole model. Bayesian methods of inference are also preferential as we often have incomplete data when dealing with the general epidemic model (the SIR model). However, it is not possible to readily calculate the likelihood function for the SIR model so Approximate Bayesian Computation is applied to undertake inference on the model.

Generally, we only know the time of recovery of an individual and the time when symptoms of the infection appear, not the precise time of infection - it is possible to be infected before the onset of symptoms. Bayesian methods are better at dealing with incomplete data, they regard it as nuisance parameters, whereas the classical method of inference requires a complete likelihood function for the infection.

Another model for infection is the Reed-Frost Epidemic Model. This is a variation of the SIR model, though not a strict subcase of the model. It has a stochastic rate of infection and a deterministic recovery rate, so assumes every individual has the infection for an identical period of time. Whereas, the SIR model is either completely deterministic or stochastic.

It is possible to calculate the likelihood function for the Reed-Frost Model so we can infer the infection rate by using classical maximum likelihood methods. Also, Markov Chain Monte Carlo Methods, specifically the Gibbs' Sampling Method, can be used to infer the same rate of infection parameter.

In this project we are comparing the classical maximum likelihood and the Markov Chain Monte Carlo methods for finding the rate of infection of the Reed-Frost Epidemic Model. Being able to estimate this rate accurately will enable us to find out more about the progress of infection for those infections which can be modelled by the Reed-Frost epidemic model.

We will undertake Approximate Bayesian Computation on the SIR model to try and estimate the rates of infection and recovery of an epidemic when we have incomplete data. It is useful to know these rates for incomplete data as it is very difficult to find complete data sets for epidemics - generally infections are not diagnosed at the moment of infection. Individuals can be infected for a period of days before they display symptoms.

Chapter 2

Modelling Epidemics

There are many ways to model disease epidemics. Two of the simplest models to carry out analysis on - the Susceptible-Infective-Recovered (SIR) model and the Reed-Frost Model, are analysed in this project.

The aim of this chapter is to set up these models so we can undertake inference on them later in this project.

There are many more complicated models in existence, which are more difficult to analyse and draw inference from, such as the SIRS model and the Multi-Strain Model, which have been described in this Chapter.

2.1 The SIR Model

One of the simplest, most successful and most used ways for modeling dynamics of epidemics is the deterministic Susceptible-Infective-Recovered (SIR) epidemiological model. Recently, there has been interest in transforming the SIR model into a probabilistic context so parameters and quantities of interest can be estimated using various statistical inference techniques.

In our descriptions of the SIR model we have made the following assumptions:

- The final size of the epidemic is the number of individuals who contract the disease less those individuals who were infected at time 0.
- The epidemic continues until there are no infected individuals in the population.
- The population is closed, that is there are no entries to the population (e.g. births and immigration) and no departures (e.g. deaths and emigration).

2.1.1 The Deterministic Model

The deterministic model was first explicitly defined by Kermack & McKendrick (1927).

With a deterministic model we know with certainty the values of the parameters and variables in the model, as there are no random fluctuations in value. The system can be completely defined at any time using the initial conditions we have specified.

Under the deterministic SIR model, a population of N individuals are each categorised according to their infection status - susceptible (S), infected (I) and recovered (R).

- An individual is said to be susceptible (S) if they are healthy and can contract the infection after sufficient contact with an infected individual.
- An individual is infected (I) if they have become infected by having sufficient contact with an infected individual, while still susceptible, and can transmit the infection to susceptible individuals.
- An individual is recovered (R) if they have had the infection and are no longer infected. They cannot become reinfected with contact with an infected individual, so are effectively removed from the epidemic.

The differential equations specifying the movement between the states are:

$$\frac{ds}{dt} = -\alpha' si \tag{2.1}$$
$$\frac{di}{dt} = \alpha' si - \beta' i \tag{2.2}$$
$$\frac{dr}{dt} = \beta' i \tag{2.3}$$

where $s = \frac{S}{N}$ and $i = \frac{I}{N}$ are the proportions of susceptible and infected individuals in the population and α' is the rate of infection and β' is the rate of recovery. We have assumed the population is closed, so the proportion of recovered individuals $\left(r = \frac{R}{N}\right)$ is given by $r(t) = 1 - i(t) - s(t)$.

The dynamics of these equations are shown in a diagrammatic form in below.

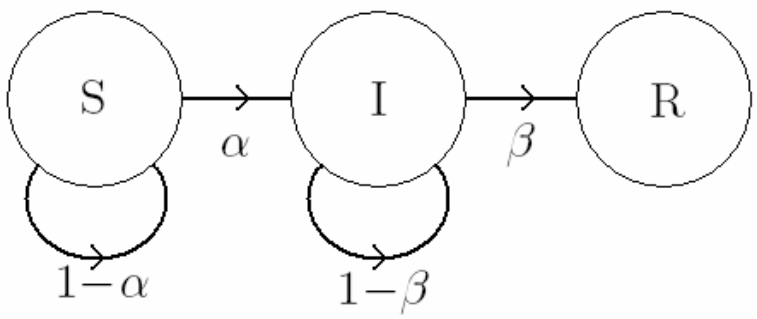

Figure 2.1: The dynamics of the deterministic SIR Model

This system can be reduced to Equations 2.1 and 2.2 as Equation 2.3 can be decoupled due to the population being closed and thus removed from the system.

2.1.2 The Probabilistic Model

McKendrick (1926) first published a paper on the probabilistic SIR Model, however it was not until Bartlett (1949) did the model really establish itself.

A probabilistic model takes into account there may be some element of variability and randomness in at least one of the parameters or variables. So, predictions from the model are probability distributions of possible numbers of susceptibles, infecteds and recovereds. Hence, the transmission and recovery of an infection is a stochastic process which can be described by a probability distribution.

The Markovian Probabilistic Model

A process is Markovian if the conditional probability distribution of the future states of the process depends only on the current state of the process and not on any past states.

The stochastic random infection dynamics are given by:

$$S + I \xrightarrow{\alpha} 2I \tag{2.4}$$

$$I \xrightarrow{\beta} R \tag{2.5}$$

Equation 2.4 shows the encounter of a susceptible and an infective results in two infective individuals (the original infective individual infects the susceptible individual with the virus) with probability α at any particular time instant.

Equation 2.5 states infective individuals recover from the virus, and so are effectively removed from the population, at any particular time instant with probability β.

These stochastic infection dynamic equations are represented graphically below

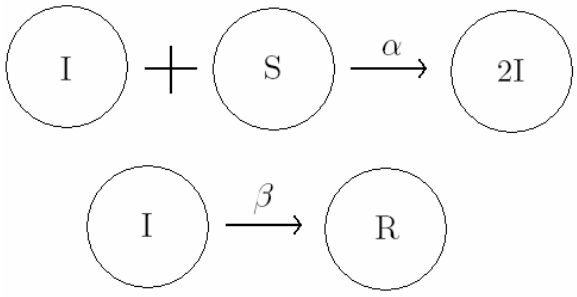

Figure 2.2: Diagram showing the stochastic SIR model dynamics

We have assumed the process is Markovian on our timescale, so the dynamics of this probability are given by:

$$\partial_t p(S, I; t) = \frac{\alpha}{N}(S+1)(I-1)p(S+1, I-1; t) + \beta(I+1)p(S, I+1; t) - \left(\frac{\alpha}{N}SI + \beta I\right)p(S, I; t) \tag{2.6}$$

with $p(S, I; t = t_0)$ having a small but fixed number of infective individuals, I_0, and a small but fixed number of susceptible individuals, S_0, giving the initial conditions

$$p(S_0, I_0; t = t_0) = 1 \text{ and } p(S, I; t = t_0) = 0 \text{ for } I \neq I_0 \text{ and } S \neq_0$$

Equation 2.6 shows the rate of change with respect to time t of the probability of there being S susceptible individuals and I infective individuals at time t can be expressed in terms of the probabilities of there being S+1 susceptible individuals and I-1 infective individuals at time t, S susceptible individuals and I+1 infective individuals at time t and S susceptible individuals and I recovered individuals at time t.

There is also a Non-Markovian Probabilistic model which has the same infection rate as the Markovian model but the recovery rate has a distribution which has a majority of values clustered around a central value, such as a Weibull distribution.

2.1.3 Deterministic versus Probabilistic Models

This section follows the discussion in Andersson and Britton (2000) [4].

In the past deterministic models have been more widely used, as they are simpler to analyse, especially when numerical solutions are adequate. For a probabilistic model to be mathematical manageable, it needs to be quite simple, which can lead to unrealistic models.

However, probabilistic models appear to be a more intuitive way to explain the spread of an infection. It makes more sense to talk about the probability of an infection spreading between two individuals in the population as in the probabilistic model.

The Law of Mass Action states the rate of a reaction is directly proportional to the product of the concentrations of each participating molecule, and is used to describe the spread of infection in the deterministic model, where the reactions are described by Equations 2.1 and 2.2 and the participating molecules being the number of individuals in the classes S, I and R.

Also, the Law of Large Numbers (the average of a randomly selected sample from a large population is likely to be close to the average of the whole population), as used in the deterministic model, cannot explain all phenomena in infection transmission and so models are genuinely stochastic - there are many infections that lead to either a minor outbreak infecting only a few individuals or to a major outbreak that infects a deterministic proportion of the community. The probability of these events occurring can only be calculated when using a deterministic model.

Only a probabilistic model can give knowledge of uncertainty, so are preferred when we need to estimate quantities of interest, such as the infection and recovery rates.

Both models play an important role in understanding the mechanisms of the spread of infectious diseases. However, probabilistic models are preferred when their analysis is possible - the model has not been simplified so much that is distorts their analysis. Deterministic models are best as introductory models when studying new phenomena.

In this project we have concentrated on probabilistic models, as we are wanting to estimate the numbers of susceptible, infective and recovered individuals in the population along with the rates of infection and recovery.

2.2 The Reed-Frost Model

This is a discrete time model for an epidemic spreading in a closed population, defined as in Section 2.1. This model was initially proposed, but never published, by Reed and Frost in the 1920s, and was subsequently published with some extensions to the model by Abbey (1952). See also Bailey (1975) [21] and O'Neill and Roberts (1999) [16] for further explanation of the model.

We define the *latent period* as the period of time between infection and becoming infective and the *incubation period* as the time between infection and the first symptoms.

In this model we make the following assumptions:

- Latent and incubation periods are constant.
- The period of infectiousness is reduced to a single point.
- A single attack of the epidemic gives immunity.
- The latent period is taken to be the unit of time.

If an epidemic started in a group of susceptible individuals with a single infective, the epidemic continues in a series of stages separated by time intervals, of length the latent period, until there are no infected individuals in the population.

Consider an initial population of N susceptible individuals and a infected individuals. For $t = 0, 1, \ldots,$ let S_t and I_t denote the number of susceptible and infected individuals in the population at time t. We assume spatial homogeneity (random mixing within the population with each individual equally likely to mix with all other individuals) so at a given time point each susceptible individual has a probability $1 - q$ of having had adequate contact with an infected individual in the population at that time point. We define $q \in (0, 1)$ as the *avoidance probability* - the probability a susceptible individual avoids contact with an infected individual at a given time point. Assuming infections occur independently of each other, each susceptible individual has a probability q^{I_t} of avoiding infection at time t. Once an individual is infected they can infect others, but only from the next time point.

We can calculate the probability there will be I_{t+1} newly infective persons at $t + 1$ given we know there are S_t and I_t susceptible and infected individuals at time t.

The probability I_{t+1} people will become infected is given by $(1 - q^{I_t})^{I_{t+1}}$, by above. If I_{t+1} individuals become infected, there are S_{t+1} individuals in the population who do not have sufficient contact with an infected individual in (t,t+1) and so remain susceptible. This is given by $(q^{I_t})^{S_{t+1}}$. Noting $I_{t+1} = S_t - S_{t+1}$, in a population of size $S_{t+1} + I_{t+1} = S_t$ there are $\binom{S_t}{S_{t+1}}$ possible combinations of individuals that will form the required numbers of susceptible individuals and individuals as the period of infectiousness is a single time, so giving:

$$P(I_{t+1}|S_t, I_t) = \binom{S_t}{S_{t+1}} (1 - q^{I_t})^{I_{t+1}} q^{I_t S_{t+1}} \tag{2.7}$$

This is a binomial distribution with parameters S_t and $(1 - q^{I_t})$.

The progress of an epidemic can be expressed as a vector of the number of infectives at each time in the epidemic $(I_0, I_1, \ldots, I_m)$ where $m+1 = \min\{t : I_t = 0\}$.

An example of such a vector is (a,b,c) which denotes an epidemic in which a initial infected individuals infect b susceptible individuals, who infect a further c susceptible individuals at the next time point, who fail to infect any of the remaining susceptible individuals in the population. From this we can infer the number of recovered individuals in the population, as can be seen in the example below.

From the vector (a,b,c) we get this population distribution for the progress of the epidemic for a population of size N.

t	S_t	I_t	R_t
0	N-a	a	0
1	N-a-b	b	a
2	N-a-b-c	c	a+b
3	N-a-b-c	0	a+b+c

We can use this model to calculate the probability there is a particular vector $(I_0, I_1, \ldots, I_m)$. This is the product of each I_t, $0 \leq t \leq m$, occurring, as we assume the probability of each individual becoming infected is independent. This gives:

$$\begin{aligned} P(I_0, I_1, \ldots, I_m) &= P(I_m|S_{m-1}, I_{m-1})P(I_{m-1}|S_{m-2}, I_{m-2}) \ldots P(I_1|S_0, I_0)P(I_0) && (2.8) \\ &= \frac{S_0!}{I_1!I_2! \ldots I_m!S_{m+1}!} q^{\sum_{j=0}^{m} I_j S_{j+1}} \prod_{k=0}^{m-1} (1-q^{I_k})^{I_{k+1}} && (2.9) \end{aligned}$$

$P(I_0) = 1$ as we are given I_0, the initial number of infected individuals.

2.3 The SIRS Model

The SIRS (Susceptible-Infective-Recovered-Susceptible) model is similar to the SIR model except after a period of time a recovered individual can become resusceptible to the infection. The equations corresponding to this model are:

$$\begin{aligned} \frac{ds}{dt} &= -\widetilde{\alpha}is + \widetilde{\delta}(1-s) + \widetilde{\beta}r && (2.10) \\ &= -\widetilde{\alpha}is + (\widetilde{\delta} + \widetilde{\beta}) - (\widetilde{\delta} + \widetilde{\beta})s - \widetilde{\beta}i && (2.11) \\ \frac{di}{dt} &= \widetilde{\alpha}si - \widetilde{\alpha}i - \widetilde{\delta}i && (2.12) \end{aligned}$$

where s,i and r are the proportions of susceptible, infected and recovered individuals in a closed population of size N, so r(t)=1-s(t)-i(t), $\widetilde{\alpha}$ the rate of infection, $\widetilde{\beta}$ the rate of recovery and $\widetilde{\delta}$ the rate of becoming resusceptible and $\widetilde{\alpha}$, $\widetilde{\beta} + \widetilde{\delta}$, $\widetilde{\delta} + \widetilde{\alpha} > 0$.

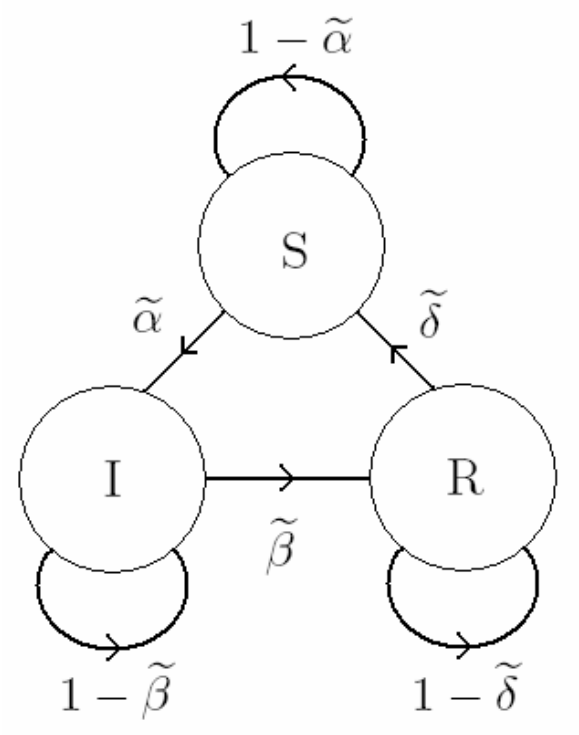

Figure 2.3: Diagram showing the dynamics of the SIRS Model

These equations are illustrated in a diagrammatic form in Figure 2.3:

The stochastic random infection dynamics are given by:

$$S + I \xrightarrow{\alpha} 2I \tag{2.13}$$

$$I \xrightarrow{\beta} R \tag{2.14}$$

$$R \xrightarrow{\delta} S \tag{2.15}$$

These are represented in diagrammatic form in Figure 2.4:

Assuming the process is Markovian, the dynamics of the probability are given to be:

$$\begin{aligned}\partial_t p(S,I,R;t) &= \frac{\alpha}{N}(S+1)(I-1)p(S+1,I-1,R;t) + \beta(I+1)(R-1)p(S,I+1,R-1;t) \\ &\quad +\delta(S-1)(R+1)p(S-1,I,R+1;t) \\ &\quad -(\frac{\alpha}{N}SI + \beta I + \delta R)p(S,I,R;t)\end{aligned}$$

with $p(S,I,R;t=t_0)$ having a small but fixed number of infective individuals, I_0, a small but fixed number of susceptible individuals, S_0 and a small but fixed number of recovered individuals, R_0. This gives the initial condition:

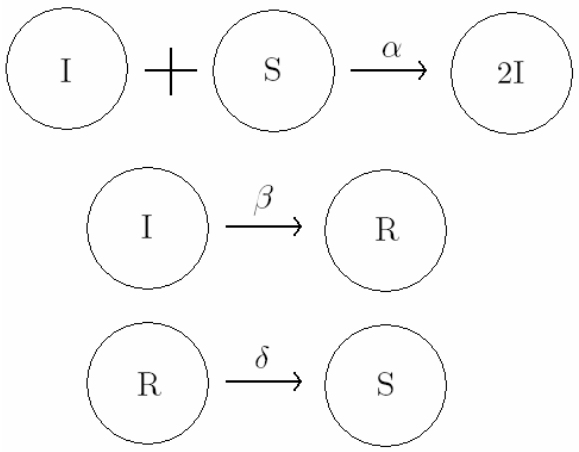

Figure 2.4: Diagram showing the stochastic SIRS model dynamics

$p(S_0, I_0, R_0, ; t = t_0) = 1$ and $p(S, I, R; t = t_0) = 0$ for $S \neq S_0$, $I \neq I_0$ and $R \neq R_0$.

2.4 The Multitype Model

The SIR and the SIRS models rely on the population being homogeneous with respect to the disease (i.e. the population is wholly human and the disease is the influenza A virus) and that the population mixes uniformly, so each member of the population is equally likely to mix and infect any other member of the population. This is very simplistic and is rarely the case.

Explained briefly below are three models which do not rely on these simplifying assumptions.

2.4.1 The Household Model

The main reference for this model is Ball et al (1997). Clearly, small social groups such as households, work places and schools form within a population. Within such groups, the spread of infection is increased as each individual interacts more with individuals in their group. This leads to an increased susceptibility, so we have different rates of infection corresponding to each group in the population.

2.4.2 Varying susceptibility

This model was first suggested independently by Proschan and Sethuraman (1976) and Ball (1985).

In this model, each individual has their own infection rate. We assume all infected are equally infected and only the susceptibility of the susceptible changes. Within a population there are groups of individuals such as the young, the elderly and the infirm who are more susceptible to infection, so making the model more realistic.

2.4.3 The Multi-Strain Model

This model was first defined by Ball (1986), see also Andreasson et al (1997) [6].

If we assume a finite number of strains of an infection an extension of the SIR model can be used to describe the outcome of interacting strains of the infection. Subdividing the S, I and R classes by assuming immunity from a particular virus strain is lifelong and independent of the order in which an individual experienced the various strains of the virus, it is possible to create an "immunity profile" of an individual by giving the set of strains of infections they are now immune from. This can be used to calculate the probability of a particular individual becoming infected and recovered from each strain they have not been infected with.

Chapter 3

Methods of Statistical Inference

The aim of this chapter is to describe various statistical procedures, that are to be used when carrying out inference on the SIR and Reed-Frost Epidemic Models.

3.1 Bayesian Inference

The Bayesian method of inference is particularly suited to epidemic models, as the parameters of interest are usually defined in terms of individuals, so lead us to consider the distribution of these parameters over the whole model.

3.1.1 Bayes' Theorem

Bayes' Theorem will be used in carrying out Bayesian Inference, specifically the Markov Chain Monte Carlo methods, to calculate the posterior distribution of the avoidance probability, q, in the Reed-Frost Model.

Suppose $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is a vector of n observations whose probability distribution $p(\mathbf{y}|\theta)$ depends on a set of k parameters $\theta' = (\theta_1, \theta_2, \ldots, \theta_k)$. θ has a probability distribution $p(\theta)$. Then

$$p(\mathbf{y}|\theta)p(\theta) = p(\mathbf{y}, \theta) = p(\theta|\mathbf{y})p(\mathbf{y}) \tag{3.1}$$

We observe the data $\mathbf{y}$, so the conditional distribution of θ is

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} \tag{3.2}$$

$$p(\theta|\mathbf{y}) = cp(\mathbf{y}|\theta)p(\theta)$$

c is a normalising constant to ensure that as the posterior distribution is a probability distribution it integrates (if θ is continuous) or sums (if θ is discrete) to 1.

Equation 3.2 is Bayes Theorem. $p(\theta)$ tells us what we know about θ when we have no knowledge of the data. This is known as the *prior distribution of* θ.

Similarly $p(\theta|\mathbf{y})$ tells us what we know about θ given we have knowledge of the data. This is the *posterior distribution of θ given* $\mathbf{y}$.

3.1.2 Likelihood Function

Given the data $\mathbf{y}$, we can consider $p(\mathbf{y}|\theta)$ in Equation 3.2 to be a function of θ, known as the *likelihood function of θ for a given* $\mathbf{y}$, denoted by $L(\theta|\mathbf{y})$ and so enables Bayes' Theorem to be rewritten as

$$p(\theta|\mathbf{y}) \propto L(\theta|\mathbf{y})p(\theta) \tag{3.3}$$

so

$$\textit{posterior distribution} \propto \textit{likelihood} \times \textit{prior distribution}$$

The likelihood function plays an important role in Bayes' Theorem, we use it to show how data, $\mathbf{y}$, modifies our prior knowledge of θ. The likelihood function is defined up to a multiplicative constant (multiplication of the likelihood by a constant has no effect on the posterior distribution of θ).

3.1.3 Markov Chain Monte Carlo Methods

This section follows Gilks et al. (1996) [23] and Ross (2002) [18].

Markov Chain Monte Carlo methods can be used to make predictions about the posterior distribution of q, the avoidance probability, in the Reed-Frost Model.

Markov Chain Monte Carlo works by generating a vector with a distribution approximately the same as a random vector $\mathbf{X}$. In $\mathbf{X}$, each of the component random variables are dependent on each other. These methods are useful when we do not know the explicit distribution of $\mathbf{X}$, as we only need the mass or density of $\mathbf{X}$ to be defined up to a multiplicative constant for these methods to be implemented.

Metropolis-Hastings Algorithm

Suppose we have a large number, m, of positive numbers $b(j) = j$, $j = 1, \ldots, m$ and the sum of these, $B = \sum_{j=1}^{m} b(j)$, is hard to calculate. We want to simulate random variables with probability mass function

$$\pi(j) = \frac{b(j)}{B}$$

One way of carrying this out is to find a Markov Chain which is easy to simulate from and has $\pi(j)$ as its limiting probabilities. The Metropolis-Hastings Algorithm constructs a time reversible Markov chain with our desired limiting probabilities.

The Metropolis-Hastings Algorithm is not used in this project, as we do not know the full joint distribution of the parameters but we can find their conditional distributions. We instead use Gibbs' Sampler, a special case of the Metropolis-Hastings Algorithm. We are merely describing the Metropolis-Hastings Algorithm

for completeness. See also Gilks et. al. (1996) [23] for further explanation of the Metropolis-Hastings Algorithm and Gibbs' Sampling.

To get the limiting probabilities $\pi(j)$ we need

$$\pi(i)P_{i,j} = \pi(j)P_{j,i} \tag{3.4}$$

$P_{i,j}$ is the transition probability from state i to state j in an irreducible Markov Chain, Q, with transition probabilities defined as

$$\begin{aligned} P_{i,j} &= q(i,j)\alpha(i,j) \; for \; i \neq j \\ P_{i,i} &= q(i,i) + \sum_{i \neq k} q(i,k)(1-\alpha(i,k)) \; for \; i = j \end{aligned}$$

where $q(i,j) = P(X = j)$ when $X_n = i$ where X is a random variable and $\alpha(i,j) = P(X_{n+1} = j)$ if $X = j$.

By above Equation 3.4 is equivalent to

$$\pi(i)q(i,j)\alpha(i,j) = \pi(j)q(j,i)\alpha(j,i) \tag{3.5}$$

The Metropolis-Hastings Algorithm uses Equation 3.5 to generate a time-reversible Markov Chain, with algorithm:

1. Choose an irreducible Markov transition probability matrix **Q** with transition probabilities q(i,j), $i, j = 1, \ldots, m$. Then choose an integer $k = 1, \ldots, m$.
2. Let n=0 and $X_0 = k$
3. Generate a random variable X such that $P(X = j) = q(X_n, j)$ and generate a random number U.
4. If $U < [b(X)q(X, X_n)]/[b(X_n)q(X_n, X)]$ then $X_{n+1} = X$, else $X_{n+1} = X_n$.
5. Set $n \to n+1$, $X_n \to X_{n+1}$.
6. Go back to step 3

When running this algorithm a number of results from the beginning of the run are discarded. This is so we give the algorithm a chance to "forget" its initial state. This is known as the burn-in period.

Gibbs Sampling

Gibbs Sampling is a commonly used special case of the Metropolis-Hastings Algorithm.

Suppose a sample X is taken from a distribution depending on a parameter $\theta \in \Theta$, with prior distribution $\pi(\theta_1, \theta_2, \ldots, \theta_d)$. d could be very large, so using numerical integration to evaluate the θ_i's would have a long computation time. Instead, Gibbs Sampling creates a Markov chain on Θ. Using distributions $\pi(q|D, r)$ and $\pi(r|D, q)$ we can use the Gibbs sampling scheme to sample from $\pi(q, r|D)$, where D is known data and q,r are the unknown parameters.

Gibbs Sampling can be implemented by using the following recursive algorithm:

1. Set B, the burn-in time
2. Set T, the number of cycles between samples after burn-in
3. Set M, the desired sample size
4. Set R, the run size, equal to MT
5. Define $S(.)$ as the vector for sample output
6. Set initial values for q(-B) and $r(-B)$
    - *Loop*: for i=-B to R:
        - sample q(i+1) according to equation $\pi(q|D, r)$ with $r = r(i)$;
        - sample $r(i+1)$ according to equation $\pi(r|D, q)$ with q=q(i+1);
        - if $i > 0$ and k=i/T is an integer then set S(k)=(q(i+1),$r(i+1)$);
    - *end of loop*

R code has been written for this algorithm and can be seen in Appendix A.1.

This is the algorithm used in O'Neill and Roberts (1999) [16].

If we have a model with unknown parameters q and r with known data, D, we know the marginal posterior distributions for q and r are proportional to the complete posterior distribution as

By Bayes theorem:

$$p(D, q, r) = \prod_{i=1}^{n} p(D_i|q, r)p(q)p(r)$$

$$\Rightarrow p(q, r) = p(q, r|D) = \frac{p(D, q, r)}{\int p(D, q, r)dqdr}$$

$$\begin{aligned} \Rightarrow p(q|r, y) &= \frac{p(q, r|y)}{p(q|y)} \\ &= \frac{p(y, q, r)}{p(y, r)} \propto p(y, q, r) \end{aligned}$$

as $p(q|ry)$ is a distribution for q and so its denominator does not depend on q. Similarly for $p(r|q, y)$.

There are some issues around the convergence of Markov Chain Monte Carlo and the Gibbs Sampling Method.

The Gibbs Sampler method can only approximate the target distribution, there will always be some residual effect from the starting distribution. However, this effect has been minimised by implementing a burn-in time in the algorithm. This burn-in time is set so that samples from the posterior distribution have been taken only once the Gibbs Sampler has converged, yet we cannot safely say whether or not these distributions are independent of our starting distribution.

Burn-In

When carrying out Gibbs Sampling and the Metropolis-Hasting Algorithm we need to specify an initial distribution for the parameters. In both methods we have convergence for (θ_n) as $n \to \infty$, for every starting value (θ_0). However, we do not know the exact rate of convergence. In order to remove the time while the chain is converging we have a burn-in period. This is found by running the algorithms over a large number of time intervals and only considering the chain once convergence has taken place. If this is successful; we observe random scattering over a fixed range. R code has been written for this and can be seen in Appendix A2.

Thinning

We carry out these Algorithms over a very large number of time intervals. After, removing the burn-in period, it is still impractical to plot all values of the posterior distribution. So we use every Tth, say, sample. This has the added benefit of ensuring that all points plotted are independent.

3.2 Maximum Likelihood

We have also used Maximum Likelihood to calculate the avoidance probability q, for the Reed-Frost Model, see also Bailey (1975) [21] and Daly et al (1995) [22].

Consider a discrete data set with random variable X, as is the case in this project. We want to model the variation of the observations in X by a probability distribution with an unknown parameter θ, which can be multidimensional. In general the probability mass function is written:

$$P(X = x) = p(x; \theta) \quad x = 0, 1, 2, \ldots$$

To estimate a value of θ, assume a random sample of size n is collected. Let the probability that X_i takes the value x_i be $p(x_i; \theta)$ for $i = 1, \ldots, n$. As the X_i's are independent the joint probability of the sample is:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = p(x_1; \theta)p(x_2; \theta) \ldots p(x_n; \theta) \tag{3.6}$$

Equation 3.6 gives the probability that the observed sample arose. It is not the true value as we do not the exact value of θ. Thinking of Equation 3.6 as a function of θ, shows how likely we are to obtain our observed samples for each value of θ. We want to find the value of θ which maximises Equation 3.6, known as the *likelihood of* θ. The value of θ for which the likelihood is maximised, the *maximum likelihood estimator of* θ, $\hat{\theta}$- a random variable. The mean and variance of the distribution of $\hat{\theta}$ are useful in determining information about the precision of the sample. The differential of the log-likelihood is known as the *score function*.

In order to find the value of $\hat{\theta}$, we need to differentiate the likelihood, set it equal to zero and solve for θ. However, it is often best to find the *log-likelihood* by taking logs of the likelihood and use this to find the value of $\hat{\theta}$.

The application of this method to the Reed-Frost Model is explained in Section 4.1.

3.3 Testing Goodness of Fit

We test of the goodness of fit of $\hat{q}$, on a data set. See also Bailey (1975) [21] and Daly et al (1995) [22].

We have a random sample of size n where each of the n observations belongs to one of k distinct classes. We let the observed number of observations in class i be O_i.

Suppose we can infer the probabilities θ_i, from a statistical hypothesis, of an observation falling into class i. In this project we have used maximum likelihood estimation to gain estimates for the parameters. These values of θ_i can be used to calculate the expected number of observations in class i for a population of size n. Let the expected number of observations in class i be denoted E_i. For large enough n, the distribution of:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where the right hand side is the sum of the scaled squared differences, provides a measure of the goodness of the fit of the model to the data.

This is approximately a χ^2 random variable with (k-1) degrees of freedom. So we test the value of χ^2 against a $\chi^2(k-1)$ distribution, where k is the number of classes.

If we have estimated some or all of the parameters we test against a different χ^2 distribution. Suppose we estimate p parameters from the data then we test against a $\chi^2(k-p-1)$ distribution.

Using the χ^2 test is only a satisfactory measure of the goodness of fit of data to a model if no expected frequency is less than 5. If this occurs we have to "pool" two or more classes together so that the new class has an expected frequency greater than 5.

3.4 Approximate Bayesian Computation

These methods have also been described in Tavaré et. al. (2003) [24].

If we cannot easily find the likelihood function for a probability model, it is not possible to generate observations from a posterior distribution, as to use Bayes' Theorem we need the likelihood function so we use a different simulation method to generate observations.

We do not know the likelihood function for the SIR Model, so we use Approximate Bayesian Computation to estimate values for the number of initial infectives, the rate of infection and the rate of recovery of the epidemic.

Suppose we have data $\mathfrak{D}$ which has been generated by a model $\mathfrak{M}$ with parameter θ. If we cannot explicitly calculate the likelihood $L(\mathfrak{D}|\theta)$, values of θ could be simulated by:

1. Generate θ from $p(.)$.
2. Simulate $\mathfrak{D}'$ from the model $\mathfrak{M}$ with parameter θ.
3. Accept θ if $\mathfrak{D}' = \mathfrak{D}$, return to step 1.

where $p(.)$ is the prior distribution for θ.

This is repeated numerous times to give a distribution for the value of θ. However, this assumes the model $\mathfrak{M}$ can be easily simulated.

If the acceptance rate from the algorithm is too small we can use the following approximate method.

1. Generate θ from $p(.)$.
2. Simulate $\mathfrak{D}'$ from the model $\mathfrak{M}$ with parameter θ.
3. Calculate the distance $\rho(\mathfrak{D}, \mathfrak{D}')$ between $\mathfrak{D}'$ and $\mathfrak{D}$.
4. Accept the value of θ if $\rho \leq \epsilon$; return to step 1.

This algorithm relies on a suitable function ρ. Commonly used functions for ρ are:

- least squares differences
$$\sum_{i=1}^{n}(O_i - S_i)^2$$
- scaled least square differences
$$\sum_{i=1}^{n}\frac{(O_i - S_i)^2}{S_i}$$
- difference
$$\sum_{i=1}^{n}(O_i - S_i)$$

where O_i is the observed value, S_i is the simulated value and $i = 1, \ldots, n$ is the number of groups/time values observations and simulated data points are taken. We use the least squares difference function.

We also need to select a suitable value for ε. As $\varepsilon \to \infty$ we accept all values of θ and so we are generating observations from the prior distribution of θ, which is not what we require. If $\varepsilon = 0$ then an observation is only accepted if $\mathfrak{D}' = \mathfrak{D}$ and so observations come from the required posterior density of θ. In choosing ε we need to make a compromise between accuracy (sampling from the required posterior distribution) and computability (having a large enough sample of values of θ to carry out inference). However, for all values of ρ and ε we are generating independent identically distributed observations from $P(\theta|\rho(\mathfrak{D}, \mathfrak{D}') \leq \varepsilon)$.

For around 1000000 simulations an ϵ value such that the acceptance probability is around 0.005 is seen as an appropriate compromise. This is one of the key issues around the use of Approximate Bayesian Computation. Choosing the wrong function ρ, or an inappropriate value for ε can lead to incorrect values for the parameters that we are estimating.

Approximate Bayesian Computation methods are usually easy to code and generate independent observations. However, it is unlikely to be a sensible method when the posterior is a long way away from the prior.

Chapter 4

Applying the Methods of Statistical Inference to the Epidemic Models

In this Chapter we aim explain how we used the different methods of Statistical Inference on the Reed-Frost and SIR Model.

In the Reed-Frost Model we are concerned with populations that consist of households, and in this project we are only concerned with households of size 3. A household is defined as a homogeneous group of people between which the epidemic can be transmitted. In each household we assume 1 of the individuals is infected at time 0 and the other 2 members are susceptible. For simplicity, we assume there is no interaction between households and transmission of the epidemic is independent for each household. We also make the assumption each household has the same rate of infection.

4.1 Applying Maximum Likelihood Methods to the Reed-Frost Model

The methods detailed in this section follow the methods in Bailey (1975) [21].

There are two possible ways of applying the maximum likelihood method to the Reed-Frost Model:

- Dealing with individual paths - the actual progress of the epidemic given by the epidemic vector detailed in Section 2.2.
- Dealing with the final size of the epidemic within each household.

When concerned with individual chains we consider all possible paths of the epidemic within the household, calculating the probabilities of each path and using these to calculate the maximum likelihood. This method is useful for small populations, but if we have a large population it becomes more difficult to calculate the probability for each path. Then it becomes advisable for the maximum likelihood to be calculated using the final size of the epidemic. This second method is less precise as we group possible paths together so lose some data.

In the methods described before we find the maximum likelihood estimator for p, $\hat{p}$ described in Section 3.2. As we have take $p = 1 - q$, the maximum likelihood estimator for $\hat{q}$ is given by

$$\hat{q} = 1 - \hat{p}$$

.

4.1.1 Estimating the chance of Adequate Contact with individual paths

If we have a household of size 3 then the only possible progress of epidemic vectors are (1),(1,1),(1,1,1) and (1,2) with the epidemic vector defined as in Section 2.2.

Using Equation 2.9 we get the following probabilities for the vectors, with $p = 1 - q$.

Vector	Probability
(1)	q^2
(1,1)	$2pq^2$
(1,1,1)	$2p^2q$
(1,2)	p^2

These are the only possible epidemic vectors as their probabilities sum to 1.

The method of complete enumerations of all possible paths is appropriate, in this case, as we have a small household. But, if we had a larger population a more general theory would be more appropriate.

Letting

- a=observed number of chains of form (1)
- b=observed number of chains of form (1,1)
- c=observed number of chains for form (1,1,1)
- d=observed number of chains of form (1,2)

We have the expected number of households in a population of n households with each vector, $E(i)$ where $i = (1), (1,1), (1,1,1), (1,2)$ to be:

$$E(i) \;=\; n \times \; \textit{Probability of vector i occuring}$$

giving the Likelihood function to be:

$$\begin{aligned} L &= E[(1)]^a E[(1,1)]^b E[(1,1,1)]^c E[(1,2)]^d && (4.1)\\ &= n^a q^{2a} (2n)^b p^b q^{2b} (2n)^c p^{2c} q^c n^d p^{2d} && (4.2)\\ &= 2^{b+c} n^{a+b+c+d} p^{b+2c+2d} q^{2a+2b+c} && (4.3) \end{aligned}$$

Carrying out maximum likelihood estimation as detailed in Section 3.2 on this likelihood gives:

$$\hat{p} = \frac{b + 2c + 2d}{2a + 3b + 3c + 2d} \tag{4.4}$$

Using this value of $\hat{p}$ we can calculate expected values for each of the chains and carry out a χ^2 test with two degrees of freedom as we have estimated the value of p, so $\chi^2(4-1-1) = \chi^2(2)$, to see how well the model fits the data.

4.1.2 Estimating the Chance of Adequate Contact using the final epidemic size.

Here we group together the observed and expected values for the vector (1,1,1) and the vector (1,2), as they both have a final epidemic size of 2.

In order to carry out the analysis we need to find n_0, n_1 and n_2, where n_i is a final epidemic of size i=0,1,2. These are gained by amalgamating the classes in the table above to give us:

Final Size of Epidemic	Probability
0	q^2
1	$2pq^2$
2	$p^2(1+2q)$

Letting the probability of final size of epidemic being i be $P(n_i)$ and the expected number of households with final epidemic size n_i is denoted by $E(n_i)$. This gives the likelihood to be:

$$\begin{aligned} L &= E(0)^{n_0} E(1)^{n_1} E(2)^{n_2} \\ &= n^{n_0} q^{2n_0} n^{n_1} 2^{n_1} p^{n_1} q^{2n_1} n^{n_2} p^{2n_2} (1+2q)^{n_2} \\ &= n^{n_0+n_1+n_2} 2^{n_1} q^{2n_0+2n_1} p^{n_1+2n_2} (1+2q)^{n_2} \end{aligned}$$

where $p = 1 - q$.

Maximising using the methods in Section 3.2, gives the score function:

$$S_p = \frac{n_1 + 2n_2}{p} - \frac{2n_0 + 2n_1}{q} - \frac{2n_2}{1+2q}$$

Setting $S_p = 0$ and solving for p gives the maximum likelihood estimator to be:

$$\hat{p} = \frac{(6n_0 + 11n_1 + 12n_2) \pm \sqrt{(6n_0 + 11n_1 + 2n_2)^2 - 12(n_1 + 2n_2)(4n_0 + 6n_1 + 6n_2)}}{2(4n_0 + 6n_1 + 6n_2)} \tag{4.5}$$

$0 < \hat{p} < 1$, is a probability, so we discard any value of $\hat{p}$, not in this range.

As when estimating the chance of adequate contact with individual chains we use this value of $\hat{p}$ to find the expected values of each of the final size of epidemics and so again use the χ^2 test with one degree of freedom to find how well the model fits to the data. We only have one degree of freedom as we have estimated the value of the parameter p, so we have a $\chi^2(3-1-1) = \chi^2(1)$ test.

4.2 Applying Bayesian Inference to the Reed-Frost Model

This method of inference has also been carried out on the Reed-Frost Model by O'Neill and Roberts (1999) [16].

Again, we consider the case where the available data consists of the final sizes of epidemics in households of 3 individuals, with only one initial infected.

The objective is to make inferences about the avoidance probability, q and the number of households with the pattern of infection (1,1,1), as we are assuming that this number is unknown.

For $j = 0, 1, 2$, let n_j be the number of households where an epidemic of final size j occurs, and let (1,2) be expressed as n_{21}. Noting that $n_2 = (1,1,1)+(1,2)$, so $n_2 = (1,1,1)+n_{21}$ and so hence $(1,1,1) = n_2 - n_{21}$, this gives the full conditional likelihood function for q to be, using Equation 4.3 and

$$L(q; n_0, n_1, n_2, n_{21}) = 2^{n_1+n_{21}} q^{2n_0+2n_1+n_{21}} (1-q)^{n_1+2n_2}$$

We assume the prior distribution of q is a $beta(\alpha, \beta)$ distribution. Using Bayes' Theorem we can find the posterior distribution for $q|n_0, n_1, n_2, n_{21}$

$$\begin{aligned} p(q|n_0, n_1, n_2, n_{21}) &\propto L(q|n_0, n_1, n_2, n_{21})p(q) \\ &\propto (2^{n_1+n_{21}} q^{2n_0+2n_1+n_{21}} (1-q)^{n_1+2n_2}) \times \frac{1}{B(\alpha,\beta)} q^{\alpha-1}(1-q)^{\beta-1} \\ &\propto \frac{2^{n_1+n_{21}}}{B(\alpha,\beta)} q^{2n_0+2n_1+n_{21}+\alpha-1}(1-q)^{n_1+2n_2+\beta-1} \\ &\sim Beta(2n_0+2n_1+n_2+\alpha,\ n_1+2n_2+\beta) \end{aligned}$$

We assume the behaviour between households is independent. We know n_{21} is a subset of n_2, as $n_2 = (1,1,1)+(1,2)$ and $n_{21} = (1,1,1)$. If a particular household is in n_2 then it is also in n_{21} with probability γ.

$$\begin{aligned}\gamma = P(n_{21}|n_2) &= \frac{P(n_{21} \cap n_2)}{P(n_2)} \\ &= \frac{P(n_{21})}{P(n_2)} \\ &= \frac{2p^2q}{p^2(2q+1)} \\ &= \frac{2q}{1+2q}\end{aligned}$$

We want to find $p(n_{21}|n_0, n_1, n_2, q)$, the posterior distribution of n_{21} given that we know n_0, n_1, n_2 and q. This distribution does not depend on n_0 and n_1, so is equivalent to $p(n_{21}|n_2, q)$. This has a $Binomial\left(n_2, \frac{2q}{1+2q}\right)$ distribution as at most n_2 households can be in n_{21}, as each household either has the vector of the progress of epidemic of the form (1,1,1) in which case it is in both n_{21} and n_2, else it is of the form (1,2) and so is in n_2 only. Above we calculated the probability that a household is in n_{21} given that it is in n_2 as $\frac{2q}{1+2q}$, the success probability of the posterior distribution.

So

$$P(n_{21}|n_0, n_1, n_2, q) \sim Binomial\left(n_2, \frac{2q}{1+2q}\right) \tag{4.6}$$

We use the posterior distributions $p(q|n_0, n_1, n_2, n_{21})$ and $p(n_{21}|q, n_0, n_1, n_2)$ in the Gibbs' Sampling Algorithm in 3.1.3 to find approximate values for q and n_{21}.

The algorithm has been written into R code using these posterior distributions, the code can be found in Appendix A1.

4.3 Inference on the SIR Model

It is not possible to apply the maximum likelihood nor the Bayesian Inference Method that we applied to the Reed Frost Model to the SIR Model as it is not possible to compute the likelihood function quickly and effectively. Several attempts have been made to compute these likelihoods, these solutions are highly recursive and are impractical for large population sizes. These methods are summarised in Appendix B.

4.3.1 Approximate Bayesian Computation

We can apply Approximate Bayesian Computation methods to the SIR model, as it does not require a likelihood. As detailed in section 3.4 Approximate Bayesian Computation generates a data set $\mathfrak{D}'$ by estimating the value of the parameter θ, where θ can be multidimensional. It then measures the "distance" between the simulated data set $\mathfrak{D}'$ and the true data set $\mathfrak{D}$ using a distance function ρ. If this distance is less that ε then the simulated data set is accepted. Carrying out this simulation numerous times enables us to infer the most likely values for the parameters we wish to estimate.

We are using the number of infective individuals in the population at each time t, only. We are not using the number of susceptible or recovered individuals as when real data is collected it is generally not possible to infer the number of susceptible individuals at each time. Also, the number of recovered individuals in the population is not known, as the time someone recovers is not easily found.

When using Approximate Bayesian Computation on the SIR model we assume the prior for the rate of infection, α to be a $Uniform(0,1)$ distribution. We also assume the prior for the rate of recovery, β, to be a $Uniform(0,1)$ distribution. We are also estimating the number of initial infectives in the population for which we take the prior to have a $Uniform(1,n)$ distribution where $1 \leq n \leq N$, where N is the population size. However, as we generally know the value of n calculating an estimate for n can serve as a "check" to see how well the Approximate Bayesian Computation Methods are estimating the parameter values.

We then use the following algorithm

1. Generate α from $p(\alpha)$ and β from $p(\beta)$.
2. Simulate $\mathfrak{D}'$ from the model $\mathfrak{M}$ with parameters α and β.
3. Calculate the distance $\rho(\mathfrak{D}, \mathfrak{D}')$ between $\mathfrak{D}'$ and $\mathfrak{D}$.
4. Accept α and β if $\rho \leq \epsilon$

The choice of ρ used here is the least squares distance between the simulated infected individuals and observed infected individuals:

$$\rho = \sum_{i=1}^{n} (IT(i) - IS(i))^2 \tag{4.7}$$

where IT(i) is the number of true infected individuals at time i, IS(i) is the number of simulated infected individuals at time i and $i = 1, \ldots, n$ is the time period over which we have modelled the infection.

We have chosen this function for ρ as it is the most widely used function and also, as can be seen in section 6.2, gives accurate results when we use Approximate Bayesian Computation to return parameter values.

This algorithm has been written into C^{++} code by Dave Dale, detailed in Appendix A.4.

Chapter 5

Inference Applied to the Reed-Frost Model I

The data used here is from an investigation by Wilson, Bennet, Allen and Worcester (1939) into measles epidemics occurring in Providence, Rhode Island during 1929-1934. We are only using the data on households of size 3, with 1 initial infected and 2 susceptible individuals at time 0. All members of the population are aged between 7 months and 10 years. In section 2.4.3 we stated that groups of individuals such as the young, the elderly and the infirm can have increased susceptibility. This assists us in finding a value for q, the avoidance probability as we can assume that the population has a single infection rate, as they are all of approximately the same age. We also assume there is no interaction between households, so the epidemic in each household is independent.

We have the following data, with the epidemic vectors as described in Section 2.2 and expected number of households calculated as in Section 4.1.1. We have a population of size n=334.

Type Of Chain	Expected Number of Households	Providence Measles Data
(1)	nq^2	34
(1,1)	$2npq^2$	25
(1,1,1)	$2np^2q$	36
(1,2)	np^2	239

where $p = 1 - q$.

5.1 Maximum Likelihood Method

These maximum likelihood methods follow the work detailed in Bailey (1975)[21].

5.1.1 Estimating the chance of adequate contact with individual chains

This uses the maximum likelihood method detailed in section 4.1.1.

When carrying out maximum likelihood estimation on the data we need to note that n_2 is the number of paths of form (1,1,1) and (1,2) and n_{21} is the number of paths of form (1,2), and so number of paths of form (1,1,1) is $n_2 - n_{21}$.

Setting

$$a = n_0,\ b = n_1,\ c = n_2 - n_{21},\ d = n_{21}$$

By using Equation 4.1 we gain a value of $\hat{p} = 0.789$. As q=1-p we get the maximum likelihood estimator for q to be 1-0.789=0.211

Using this value of $\hat{p}$ we get the fitted values of the data to be:

Type of Chain	Expected Number of Households	Providence Measles Data	Fitted Values
(1)	nq^2	34	14.9
(1,1)	$2npq^2$	25	23.5
(1,1,1)	$2np^2q$	36	87.7
(1,2)	np^2	239	207.9

Carrying out a χ^2 test with 2 degrees of freedom gives a value of 59.8. This has a corresponding p-value of less than 0.001.

5.1.2 Estimating the Chance of Adequate contact using the final sizes of the epidemic.

This method of maximum likelihood estimation is detailed in section 4.1.2.

The total number of cases refers to the final size of the epidemics in a household. As we are using households of size three, the final size is either 0, 1 or 2. This gives us the data in terms of n_0, n_1 and n_2.

which when evaluated using Equation 4.5 gives $\hat{p} = 0.728$. As q=1-p this gives the maximum likelihood estimator of q to be 0.272.

This gives the fitted values:

Final Size Epidemic	Expected Number of Households	Providence Measles Data	Fitted Values
0	nq^2	34	24.7
1	$2npq^2$	25	36.0
2	$np^2(1+2q)$	275	273.3

When testing against a χ^2 distribution with 1 degree of freedom we gain a value of 6.85 which is critical at the 1% level.

5.2 Bayesian Inference

This method is detailed in section 4.2. Also O'Neill and Roberts [16] carried out a Gibbs Sampling analysis on the same data set.

We have the data values $n_0 = 34$, $n_1 = 25$, $n_2 = 275$.

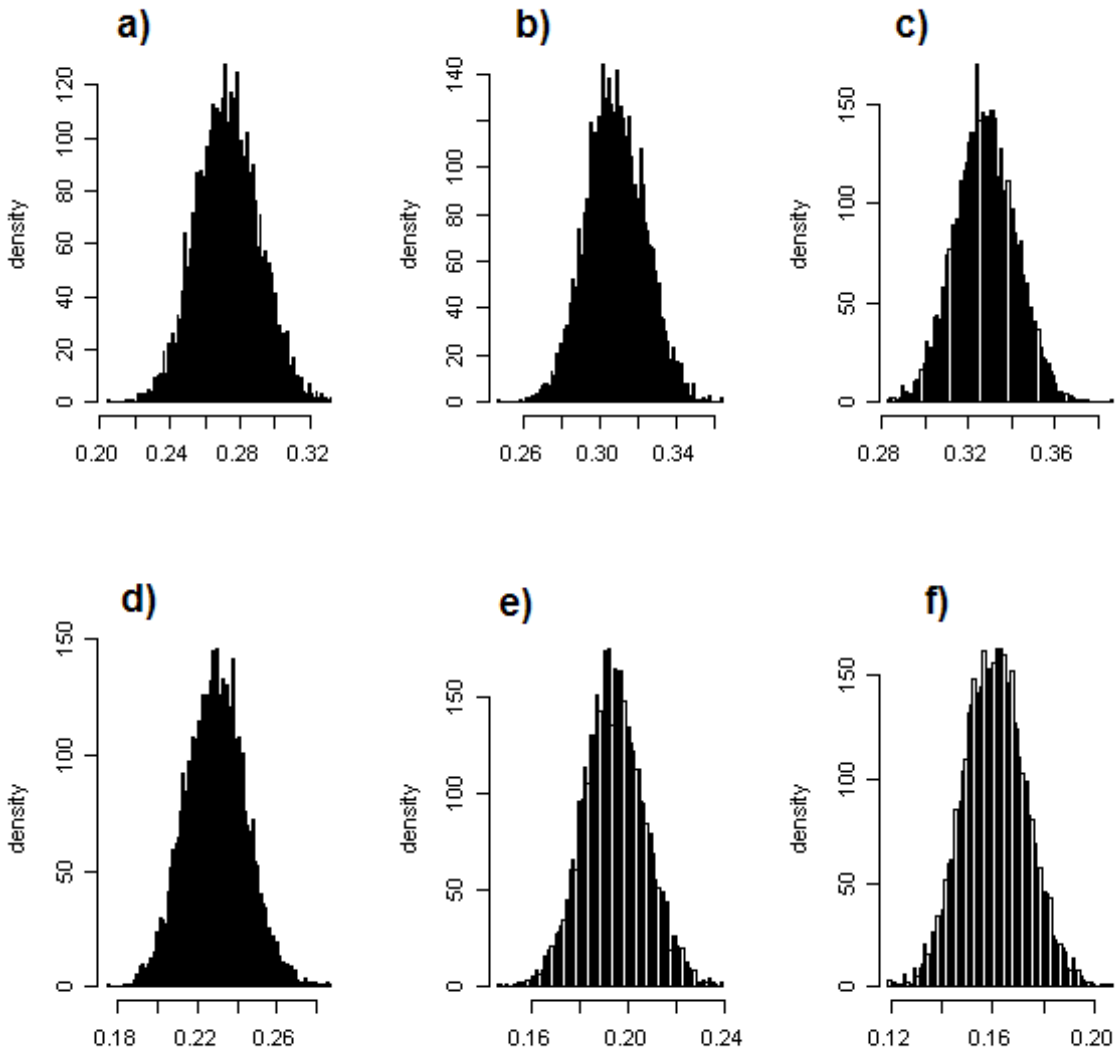

Figure 5.1: The posterior distribution for q with various initial parameter values for α and β where in a) $\alpha = 2$, $\beta = 3$, b) $\alpha = 100$, $\beta = 150$, c)$\alpha = 200$, $\beta = 300$, d)$\alpha = 20$, $\beta = 180$, e)$\alpha = 40$, $\beta = 400$ and f)$\alpha = 2$, $\beta = 400$

Inputting the above values for $n_0,\ n_1,\ n_2$ into the R code in Appendix A1, which we have written, gives the graphs in Figure 5.1 for various initial parameter values for α and β. We set the initial value for n_{21} to be 0, as we know that $0 \leq n_{21} \leq n_2$, by the definition of n_{21}, so choosing this initial value to be 0 seems a sensible thing to do.

The R code inputted into the R code in Appendix A1 was `graphq(alpha,beta,1000,1000,5000)`.

We chose the Burn-In time B to be 1000 as when you look at the graph in Figure 5.2, which were plotted using the R code we have written in A2. we see there is convergence of the value for the posterior distribution for q and also a reasonably random scatter of values by the time the Gibbs Sampler samples its 1000th value. Infact, it reaches convergence very quickly and so we could have take the burn-in time to be much less than 1000. However, taking burn-in to be 1000 enables us to be more certain we have convergence and that the samples are independent. This convergence and random scatter implies that the values for the posterior distribution do not depend on the initial choices of q and n_{21}. This is what we want, as dependence would mean we would not get a true value for the posterior distributions.

We wanted every 1000^{th} value to be sampled and wanted 5000 data points on the plots of the posterior density of q so let T=1000 and M=5000 in the R code. A small value for M may lead to successive values for q that are not independent which would lead to results that are not reliable. We would also like a large number of points on the histograms of posterior density for q as this will enable us to make more reliable and informed inferences from the data, so choosing M=5000.

The graphs in Figure 5.3 show the graphs for the posterior distribution of n_{21} with the same values of α and β as in the posterior distribution for q. These were gained by inputting into the R code in Figure A.2 the command `graphn21(alpha,beta,1000,1000,5000)`. We have chosen the values of B,T and M for the same reasons as when we were finding the posterior distribution of q with the burn-in times being appropriate, as shown in Figure 5.4.

O'Neill and Britton (1999) carried out analysis on this data set using the same Gibbs Sampler methods and gained the same results.

5.3 Discussion

As can be seen from the graphs in Figure 5.1 for all values of α and β we appear to have approximately the same posterior distribution for q. This suggests the posterior distribution is insensitive to the prior distribution. This is further shown by the 95% interval for each of the graphs plotted. These confidence intervals have been calculated using the `meanp` and the `varp` functions in Appendix A1 and then letting calculating the 95% confidence intervals using:

$$\texttt{meanp} \pm 1.96\sqrt{\texttt{varp}}$$

where 1.96 is the 95% 2-tailed value of a standard normal distribution.

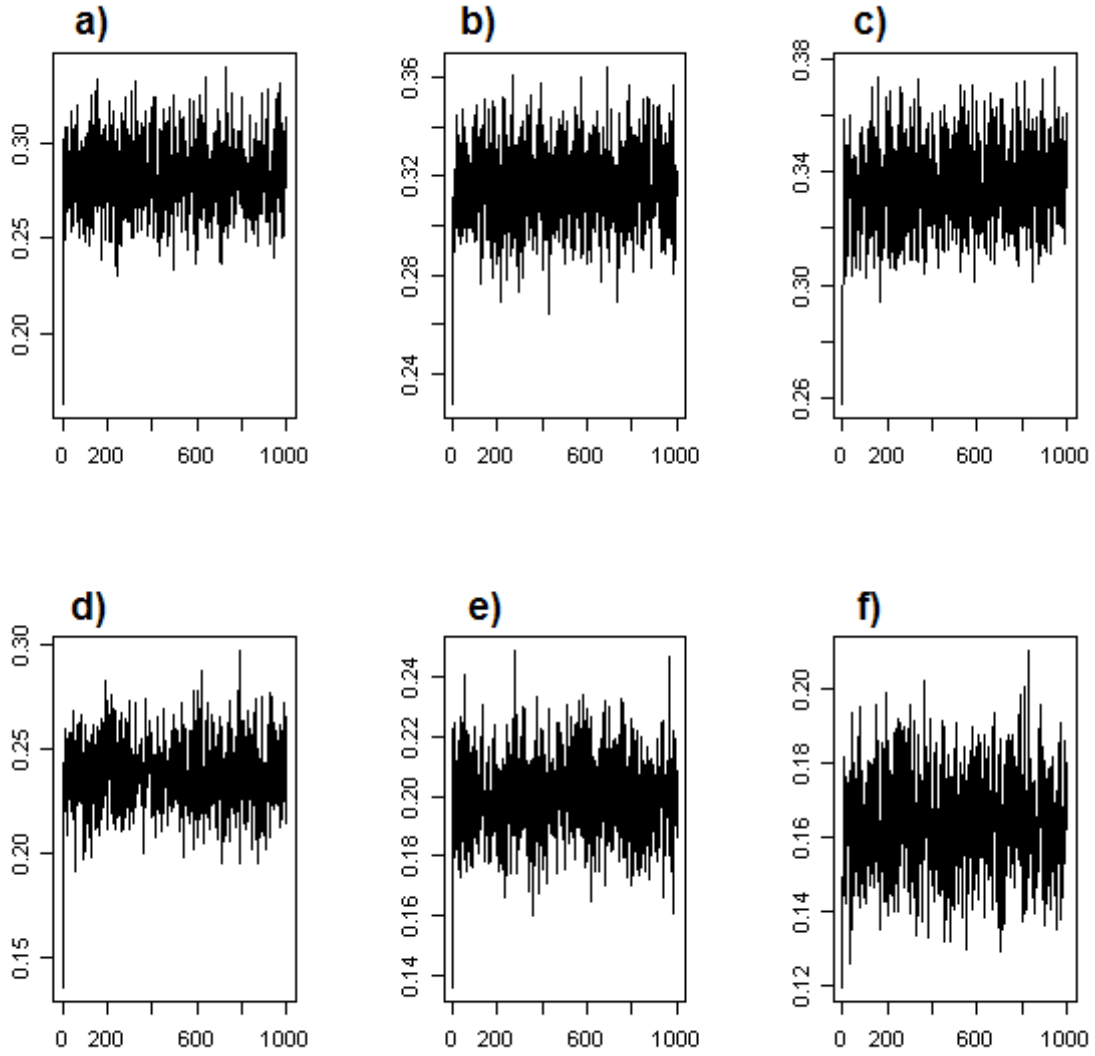

Figure 5.2: Graphs of convergence of the posterior distribution for q for various initial parameter values for α and β where in a) $\alpha = 2$, $\beta = 3$, b) $\alpha = 100$, $\beta = 150$, c)$\alpha = 200$, $\beta = 300$, d)$\alpha = 20$, $\beta = 180$, e)$\alpha = 40$, $\beta = 400$ and f)$\alpha = 2$, $\beta = 400$

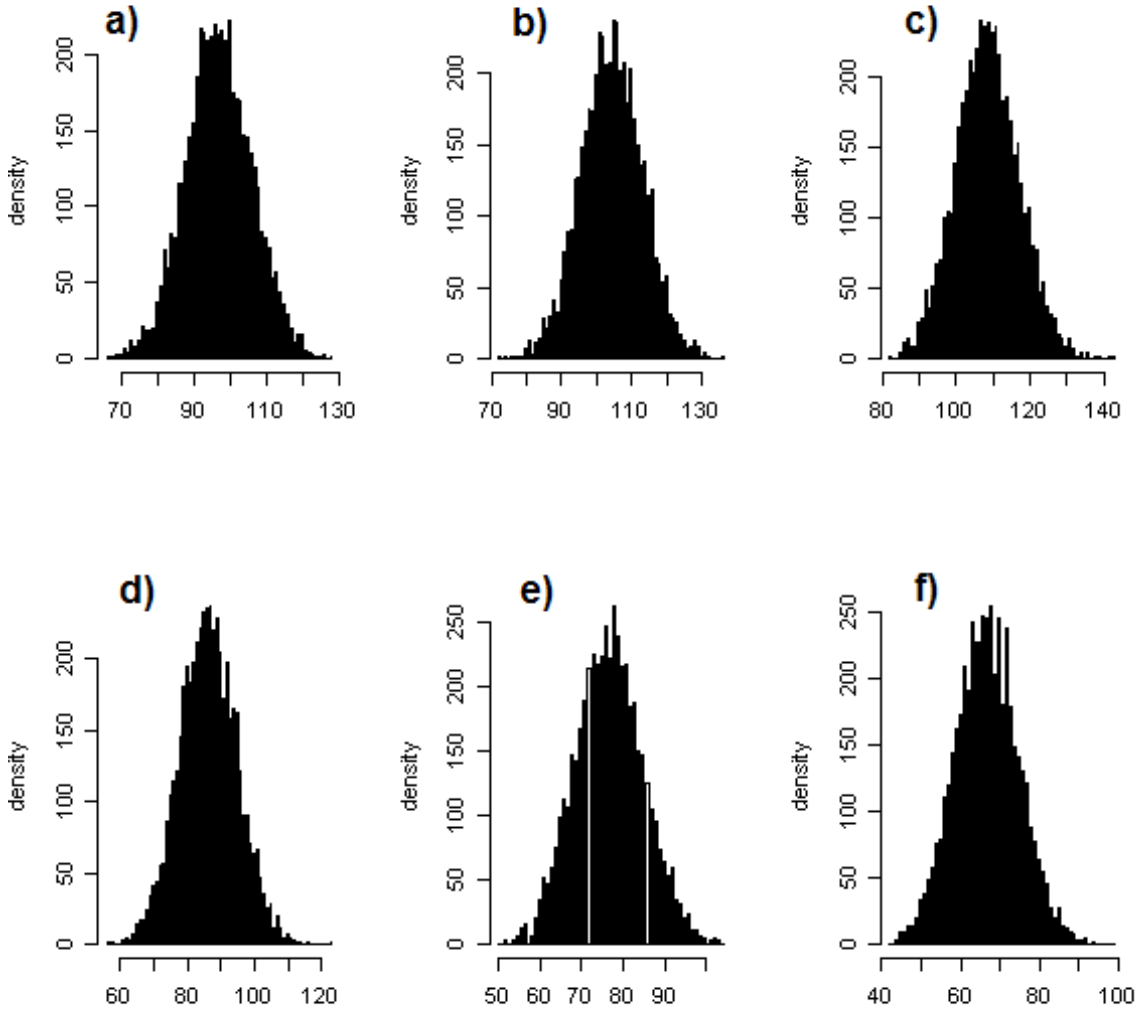

Figure 5.3: The posterior distribution for n_{21} with various initial parameter values for α and β where in a) $\alpha = 2$, $\beta = 3$, b) $\alpha = 100$, $\beta = 150$, c)$\alpha = 200$, $\beta = 300$, d)$\alpha = 20$, $\beta = 180$, e)$\alpha = 40$, $\beta = 400$ and f)$\alpha = 2$ $\beta = 400$, $\beta = 400$

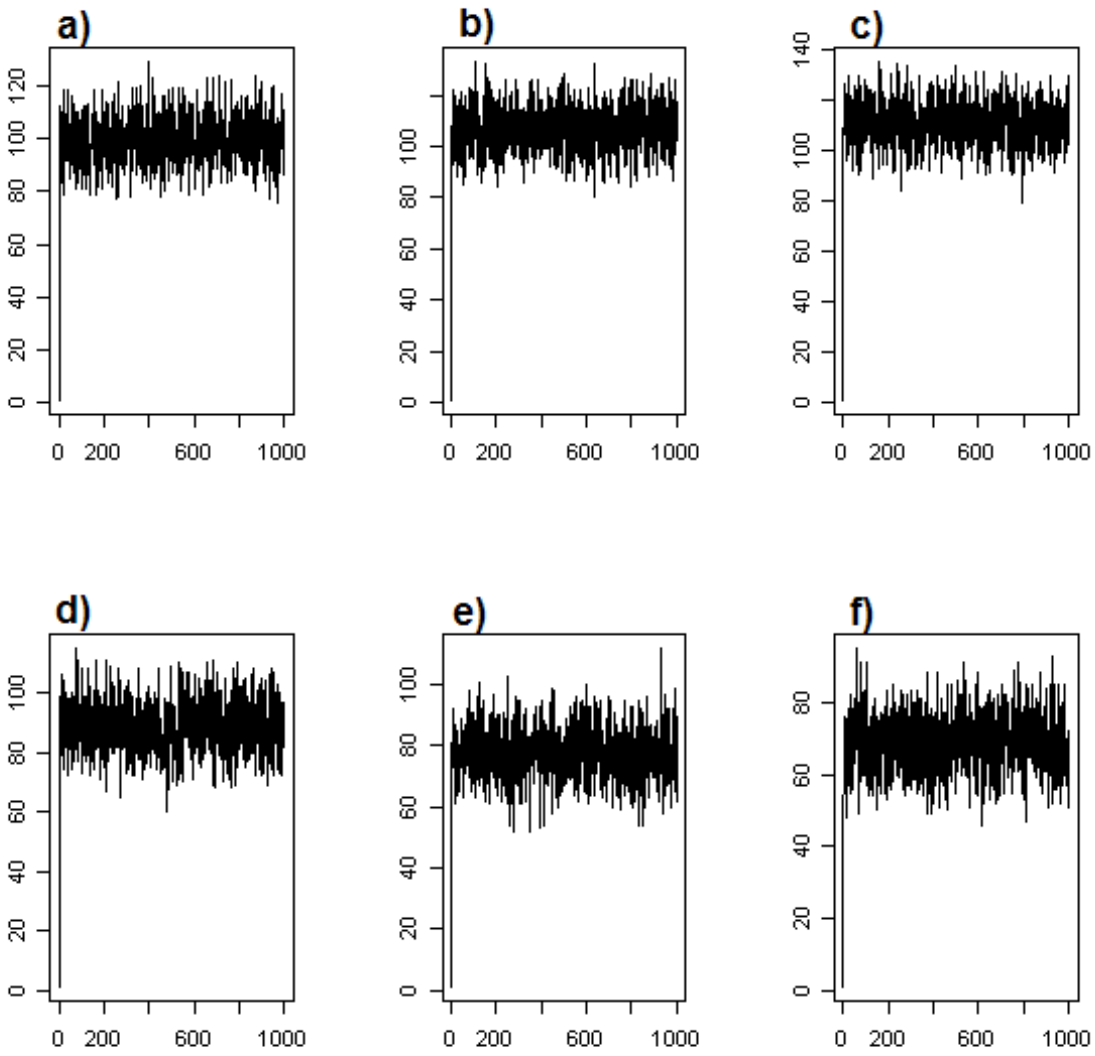

Figure 5.4: Graphs of convergence of the posterior distribution for n_{21} with various initial parameter values for α and β where in a) $\alpha = 2$, $\beta = 3$, b) $\alpha = 100$, $\beta = 150$, c)$\alpha = 200$, $\beta = 300$, d)$\alpha = 20$, $\beta = 180$, e)$\alpha = 40$, $\beta = 400$ and f)$\alpha = 2$, $\beta = 400$

α,β	Mean	Variance	95% CI
2,3	0.27253	0.00032	(0.23759,0.30746)
100,150	0.30802	0.00024	(0.27790,0.33814)
200,300	0.32795	0.00019	(0.30101,0.35488)
20,180	0.22874	0.00022	(0.19950,0.25798)
40,400	0.19393	0.00016	(0.16920,0.21864)
2,180	0.16084	0.00015	(0.13678,0.18490)

An overall estimate for the value of $\hat{q}$ to be 0.24867.

The graphs in Figure 5.3 also suggest that the posterior distribution for n_{21} appears to be insensitive to the prior distribution. This is also supported by the 95% interval, calculated as above, for each of the graphs plotted.

α,β	Mean	Variance	95% CI
2,3	96.7938	78.5743	(79.420,114.168)
100,150	104.7512	78.4431	(87.392,112.111)
200,300	109.0334	72.6794	(92.324,125.743)
20,180	86.3510	71.1943	(69.573,103.129)
40,400	76.9074	71.1943	(60.370,93.445)
2,400	66.8778	65.0813	(51.066,82.690)

In calculating all of these confidence intervals we have assumed the posterior distribution is normally distributed. We actually have an empirical distribution, having a probability of $\frac{1}{n}$ at each of the n points. Making the normal assumption has meant it is easier to calculate 95% confidence intervals. This assumption appears realistic as the graphs in Figures 5.1 and 5.3 have shape approximately that of the normal distribution bell curve.

These posterior distributions give an average estimate for the value of n_{21} to be 90.1191. In this data set we are actually given the value of n_{21}, this value is 36. It is also clear the inference is not particularly effective at gaining the correct values for n_{21} as none of the 95% confidence intervals contain it.

From the three different methods of estimating q, the avoidance probability, we have:

Method	$\hat{q}$
All Possible Paths	0.211
Final Epidemic Size	0.272
Gibbs Sampling	0.249

In Section 5.1 we undertook χ^2 tests for the goodness of fit of the models with the estimated values of $\hat{p}$. The enumerating individual paths maximum likelihood method was significant at the 0.001% level, When carrying out the maximum likelihood method where we considered only the final size of the epidemic we gained a value of $\hat{p}$ which was significant at the 1% level.

So although all three estimates for $\hat{q}$, agree they do not appear to be very good approximations to the data. This suggests that the methods were fine but it was the approximation to the Reed-Frost Model was wrong.

We have found the posterior distribution for q and n_{21} with various parameter values for α and β, giving different values for the prior distribution of $q \sim Beta(\alpha, \beta)$. The posterior distribution for q, shown in Figure 5.1, appears to be invariant to the values of α and β. This along with the graphs of the values after the chosen burn-in period, 1000, in Figure 5.2, showing a reasonably random level of scatter gives us confidence that the residual effect from the starting distribution is negligible when using Gibbs Sampler to approximate from the posterior distribution of q. The same can be said when using Gibbs Sampler to approximate from the posterior distribution of n_{21} by looking at Figures 5.3 and 5.4.

When carrying out both the Maximum Likelihood and the Bayesian Method we made assumptions about the data, such as there being a constant avoidance probability, q. These assumptions could account, in part, for the values calculated for the avoidance probability not being a very good approximation to the data.

The data was collected over a 5 year period. This is long enough for there to be a change in the strain of the measles infection, such a change may lead to a change in avoidance probability at some point as different strains of an infection have different infection rates. The value of q changing during the time in which data was being collected could lead to the data being not fitted by the model effectively. This issue could be rectified by splitting the data into more homogeneous groups - maybe over a smaller time scale or maybe by considering smaller areas/neighbourhoods where strains are likely to be the same.

Overcrowding in a household would lead to an increased probability of infection of measles. This could be seen to be a factor by analysising the number of living rooms per individual. Subdividing the population into groups which had a similar number of living rooms per individual would give a better fit to the data if this was a key factor in the transmission of the disease. This point was raised in Bailey (1975)[21].

Taking into account all of the above possible situations and recalculating the values of q, should give a better fit to the data.

This study agrees with the results from Bailey (1975) [21] and O'Neill and Roberts [16].

Chapter 6

Inference Applied to the Reed-Frost Model II

This data has been taken from studies on outbreaks of influenza A(H3N2) in Tecumseh, Michigan seen in Demiris and O'Neill (2005)[13]. The data is in the required form of the final number of infected individuals in a population that has been divided into households. We are only interested in households of size 3, as we are carrying out the methods of sections 4.1 and 4.2. This gives the data as

Final Epidemic Size	Tecumseh Influenza Data
0	8
1	2
2	3

6.1 Maximum Likelihood Methods

We only have the data above, so cannot carry out the maximum likelihood estimator method where we enumerate all the possible paths as we do not know the number of paths of form (1,2) and (1,1,1). We can, however, carry out the maximum likelihood estimator method using final size epidemics.

Using Equation 4.5 with these values of n_0, n_1 and n_2 gives

$$\hat{p} = 0.268$$

as q=1-p, we get $\hat{q}$ to be 0.7314.

Final Epidemic Size	Expected Number of Households	Tecumseh Influenza Data	Fitted Values
1	nq^2	5	6.95
2	$2npq^2$	2	3.74
3	$np^2(1+2q)$	3	2.30

We cannot carry out a χ^2 test for the goodness of fit as the expected numbers are less than 5 for n_1 and n_2, so they need to be pooled. We have also estimated the value of the parameter p and so we would need to test against a $\chi^2(2-1-1) = \chi^2(0)$ distribution which does not exist.

6.2 Bayesian Inference

Inputting these values into the R code in Figure A.1 and setting the starting value of n_{21} to be 0 gives us the graphs for the posterior value of q in Figure 6.1. These were gained by entering `graphp(alpha,beta,1000,1000,5000)` into R. The values for B, N and T are chosen to be the same as for the Measles Epidemic in Providence, Rhode Island 1929-1934 in Chapter 5 for the same reasons. By looking at Figure 6.2 it is clear that the burn-in time, B=1000, is appropriate.

When carrying the algorithm and plotting the results for the posterior distribution of n_{21} we get Figure 6.3. We have used the same values for α and β as for calculating the posterior distribution for q as well as the values for B, T and M. Figure 6.4 also shows the burn-in time appears appropriate, as we seem to have convergence and a reasonably random scatter.

6.3 Discussion

These show that the posterior value of q does depend on the values of α and β. Due to the mean values for the posterior distributions varying greatly - from 0.2315 when α=2 and β=400 to 0.8459 when α=2 and β=3 This is further supported by the 95% interval for each of the different values of α and β, where the confidence intervals for $\alpha = 2, \beta = 3$, $\alpha = 100, \beta = 150$, $\alpha = 20, \beta = 180$ and $\alpha = 2, \beta = 400$ all being disjoint. These give the average estimate for p to be 0.4633 and thus for q to be 0.5367, which is quite different to the value $\hat{q}$ gained from the maximum likelihood method, 0.7314.

α, β	Mean	Variance	95%CI
2,3	0.84588	0.00094	(0.7859,0.9059)
100,150	0.55838	0.00064	(0.5088,0.6079)
200,300	0.49533	0.00040	(0.4563,0.5343)
20,180	0.40004	0.00072	(0.3480,0.4528)
40,400	0.26648	0.00034	(0.1123,0.3407)
2,400	0.21353	0.00032	(0.1782,0.2488)

We also see the values of α and β have an effect on the posterior distribution for n_{21}. This is further supported by the 95% confidence intervals:

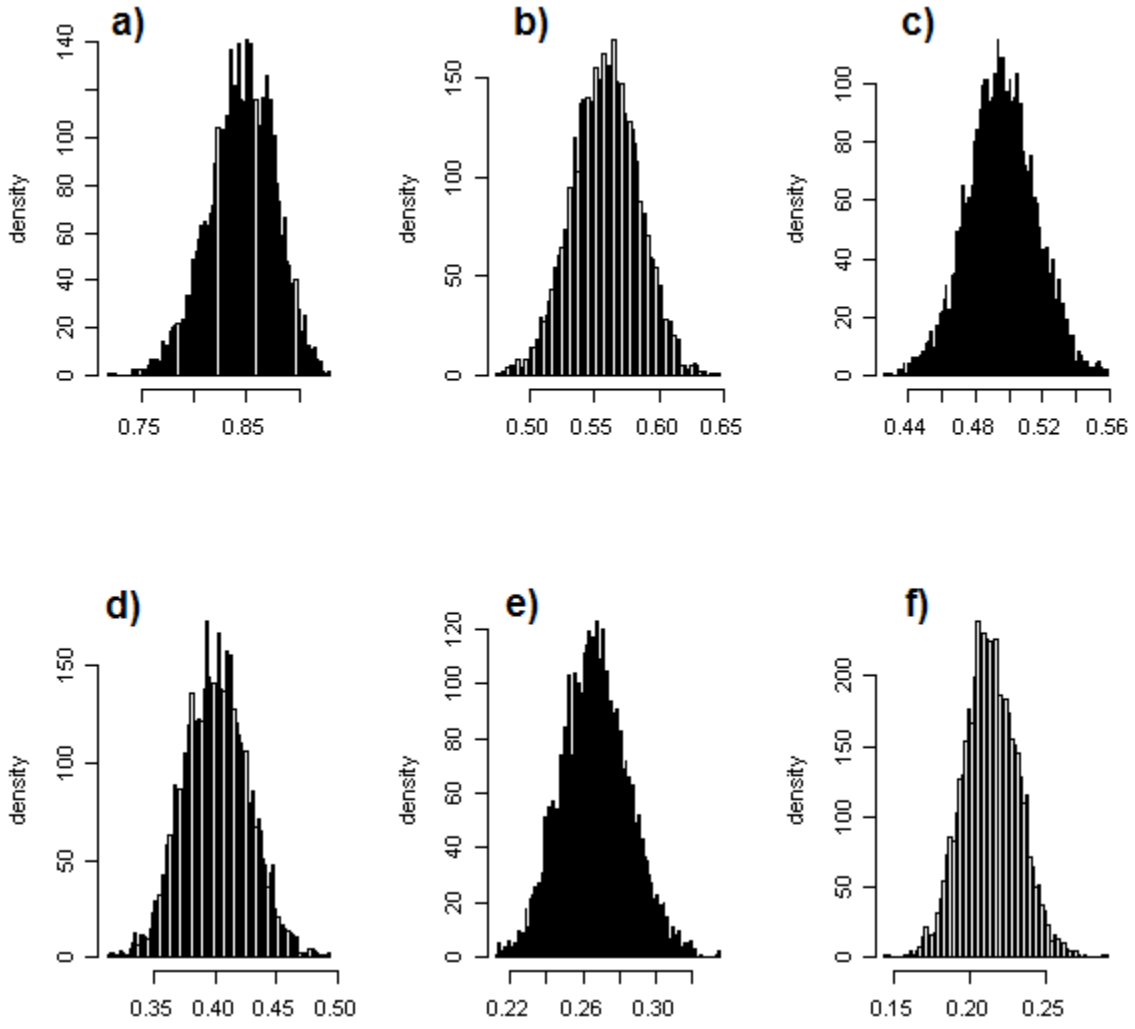

Figure 6.1: posterior distributions for q with various initial parameter values for α and β where in a) $\alpha = 2$, $\beta = 3$, b) $\alpha = 100$, $\beta = 150$, c)$\alpha = 200$, $\beta = 300$, d)$\alpha = 20$, $\beta = 180$, e)$\alpha = 40$, $\beta = 400$ and f)$\alpha = 2$, $\beta = 400$

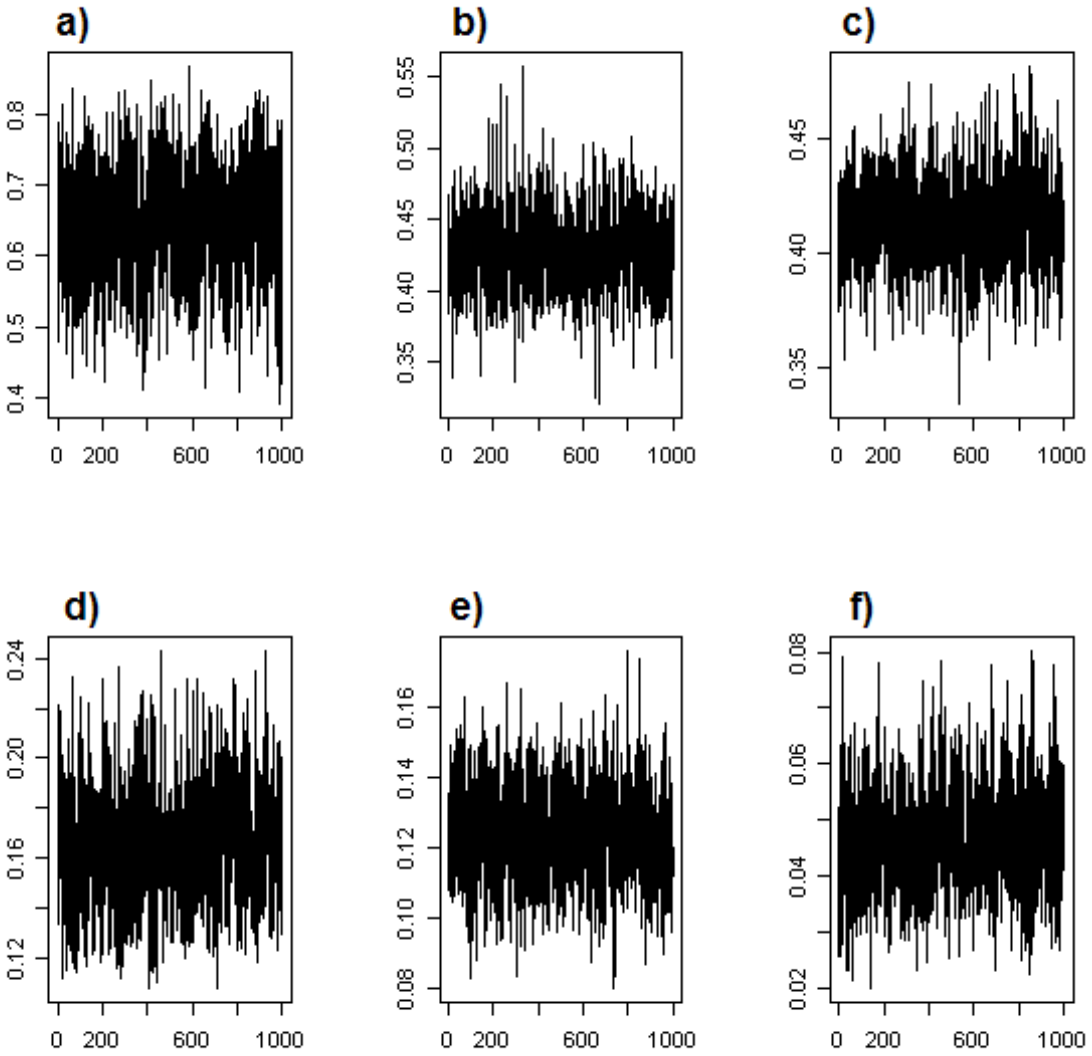

Figure 6.2: Graphs of convergence of the posterior distribution for q with various initial parameter values for α and β where in a) $\alpha = 2$, $\beta = 3$, b) $\alpha = 100$, $\beta = 150$, c)$\alpha = 200$, $\beta = 300$, d)$\alpha = 20$, $\beta = 180$, e)$\alpha = 40$, $\beta = 400$ and f)$\alpha = 2$, $\beta = 400$

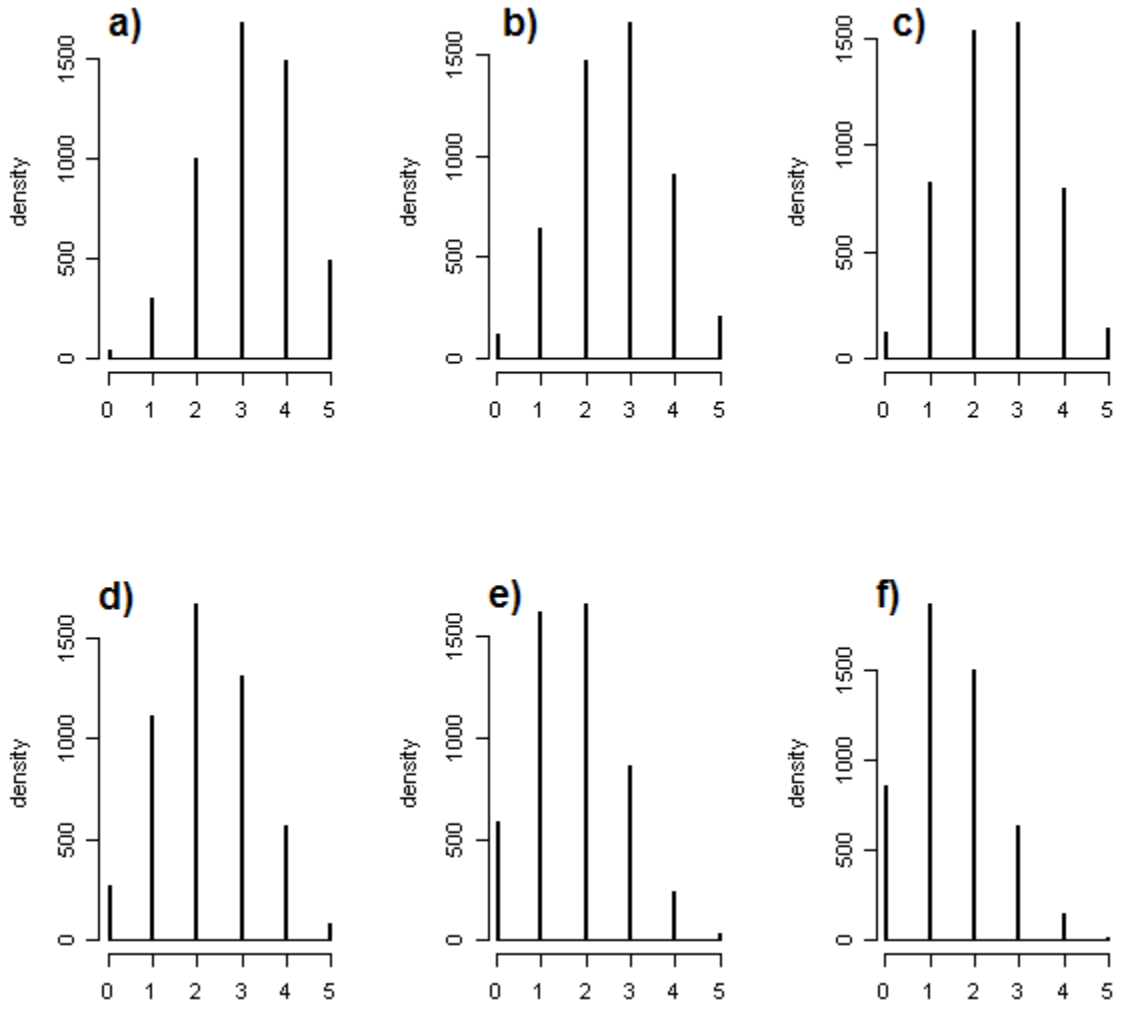

Figure 6.3: posterior distributions for n_{21} with various initial parameter values for α and β where in a) $\alpha = 2$, $\beta = 3$, b) $\alpha = 100$, $\beta = 150$, c)$\alpha = 200$, $\beta = 300$, d)$\alpha = 20$, $\beta = 180$, e)$\alpha = 40$, $\beta = 400$ and f)$\alpha = 2$, $\beta = 400$

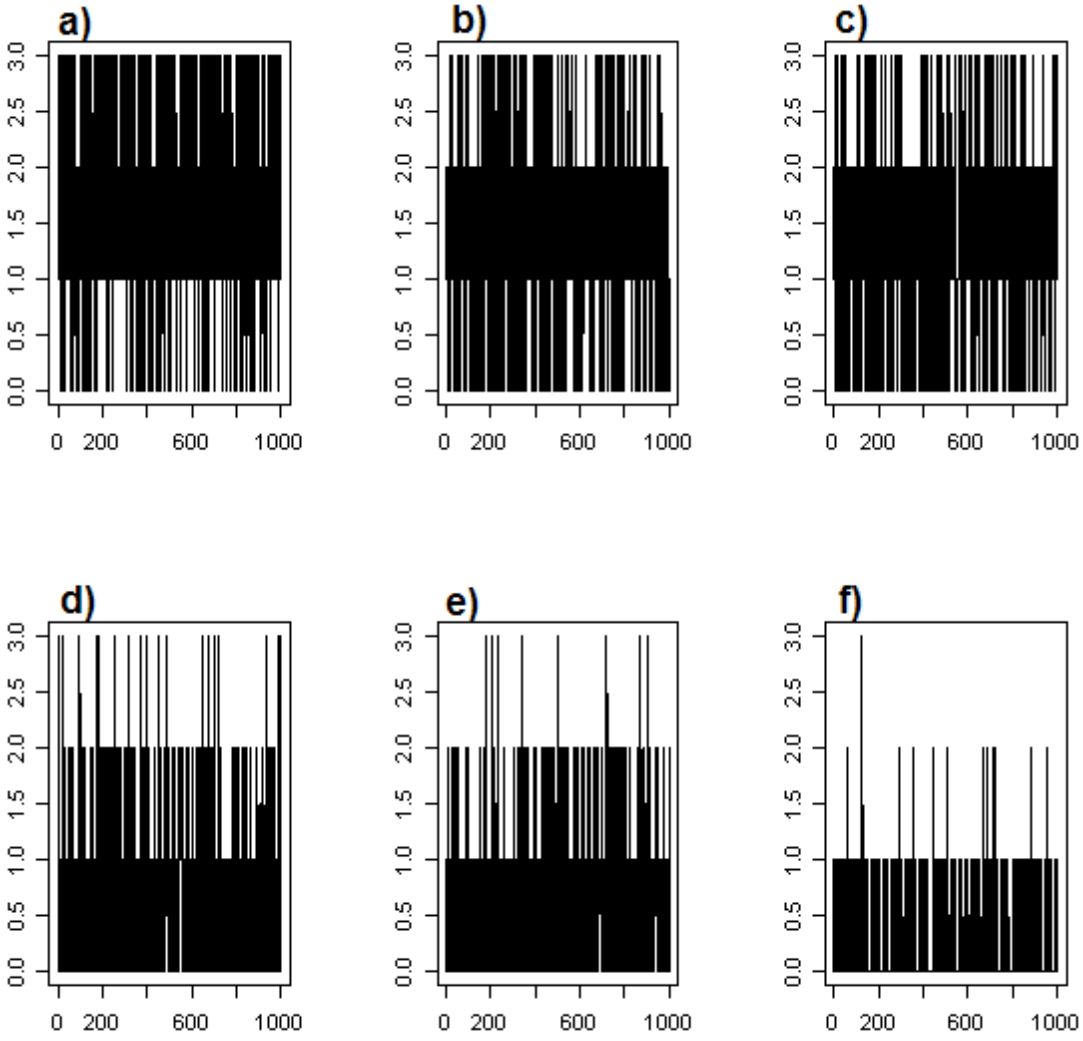

Figure 6.4: Graphs of the convergence of the posterior distribution for n_{21} with various initial parameter values for α and β where in a) $\alpha = 2$, $\beta = 3$, b) $\alpha = 100$, $\beta = 150$, c)$\alpha = 200$, $\beta = 300$, d)$\alpha = 20$, $\beta = 180$, e)$\alpha = 40$, $\beta = 400$ and f)$\alpha = 2$, $\beta = 400$

α, β	Mean	Variance	95%CI
2,3	3.1554	1.2015	(1.0070,5.3038)
100,150	1.6462	1.2100	(0.4813,4.8111)
200,300	2.4870	1.2151	(0.3264,4.6475)
20,180	2.2158	1.2537	(0.0212,4.4104)
40,400	1.7426	1.112871	(-0.3251,3.18103)
2,400	1.5088	1.0215	(-0.4721,3.4897)

However, these do not show as conclusively that the values of α and β depend on the posterior distribution of n_{21} as there are no disjoint confidence intervals. This gives an average value for n_{21} to be 2.1259.

Unlike the results in Chapter 5 the maximum likelihood and Gibbs' Sampling method do not seem to agree on a value for q, the avoidance probability. This could be due to the small sample size we have, here we are dealing with a population of just size 13. This is a very small sample and could lead to incorrect results, and as a result it is difficult to draw any conclusions.

Chapter 7

Applying Real Data to the SIR Model

It is not possible to use the same methods of inference that we used for the Reed-Frost Model as they all depended on having a likelihood function for the model and used this likelihood to carry out the inference. It is not possible to readily calculate the likelihood function for the SIR model and so we need to use a method that does not involve the likelihood function, such as Approximate Bayesian Computation.

7.1 Simulator

Simulation is useful to model the paths of the numbers of susceptible, infectious and recovered individuals in an epidemic with known rates of infection and recovery.

We have written a simulator for the SIR model using R, which can be seen in Appendix A3.

In Figure 7.1 we have carried out a sample simulation on a population which had 25 susceptible individuals, 1 infective and 0 recovered at time 0. The rate of infection was set to be 0.8 and the rate or recovery 0.6. This simulation was carried out for a period of time of length 100. 1000 of these simulations were carried out and the results plotted using the `allgraph` function in Figure A.3.

The command entered into R was `allgraph(1000,100,25,1,0,0.8,0.6)`

The benefits of carrying out a large number of simulations mean you get reasonably smooth lines for the number of susceptible individuals, recovered individuals and infective individuals and so can fairly accurately describe what the dynamics of the population are showing.

In the graph the red line is the number of susceptible individuals, the blue line the number of infective individuals and the green line the number of recovered individuals.

From Figure 7.1 you can see that the number of susceptible individuals decreases as we would expect with time. The number of infective individuals increases to around 13 individuals quite rapidly and then decreases so that there are no infective individuals. As indicated by the initial distribution of the population the number of recovered individuals starts at zero and increases quite rapidly, leveling out at around 24 recovered individuals. The number of susceptible individuals and recovered individuals remains constant once there are no infective individuals in the population.

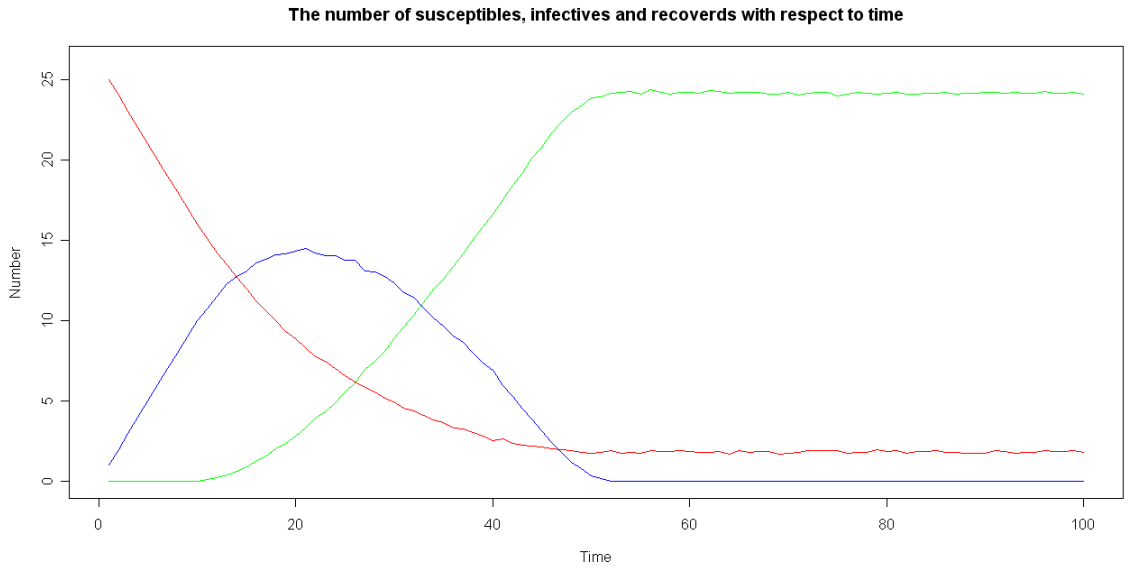

Figure 7.1: A sample simulation with initial susceptibles 25, initial infectives 1, initial recoverds 0 rate of infection 0.8, rate of recovery 0.6, time period 100 and 1000 simulations

This is the form of graph we would expect as if there was no-one infected then it is not possible for any of the susceptibles to become infected and similarly no-one could recover from the infection.

7.2 Using Approximate Bayesian Computation to recover parameter values

We can use Approximate Bayesian Computation to recover parameter values for a simulation. The purpose of this is two-fold. It can help see whether the program is accurate enough to return known values. It also assists in finding an appropriate value of ε.

We undertook a single simulation of a data set where we set the parameter values. This was done using the `mean` computer program written by Dave Dale and detailed in Appendix A4 using the parameter values: infection rate=0.3, recovery rate 0.5, population 250, initial infectives=5 and time period=2 taken over 100 time steps. This gave the output in graphical form in Figure 7.2. This is what we would expect as it agrees with the "shape" of the graph in Figure 7.1. However, it is not the complete graph as the epidemic has not been run to completion as it has been ran over a much shorter period of time.

We can use this output as the true data for the Approximate Bayesian Computation program again written by Dave Dale, the `abc` programme in Appendix A4, to see if we can recover the true values of the number of initially infected individuals, infection rate and recovery rate.

Carrying out this simulation 1000000 times for different values of ε gave the following acceptance rates:

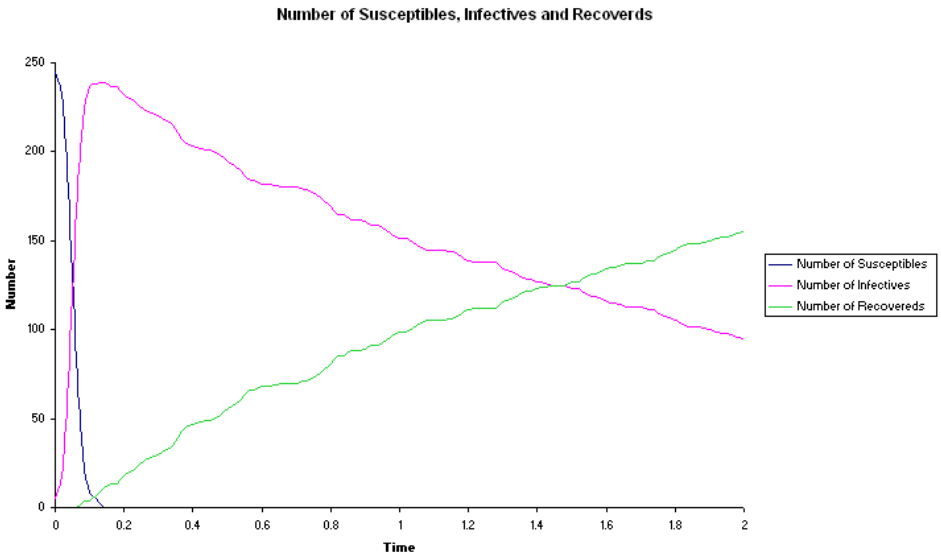

Figure 7.2: Simulation output with population 250, initial infectives 5, time 100, infection rate 0.3 and recovery rate 0.5

Value of ϵ	Acceptance Rate (%)
3000	0.08
4000	0.16
5000	0.32
6000	0.44
7000	0.59
8000	0.75
9000	0.87
10000	1.00

We have chosen to set $\epsilon = 6500$ as it gives an acceptance rate of around 0.5%, which is a good compromise between the amount of data gained and the accuracy of the estimate and then plotted in a histogram.

We use commands to give the prior for the initial number of infectives to be $Uniform(1, 10)$, for the rate of infection $Uniform(0.01, 1)$ and for the rate of recovery $Uniform(0.01, 1)$ for a population of size 250 over 100 time steps over a period of time of length 2.

Using the output from the Approximate Bayesian Computation we can plot a histogram for the Posterior Distribution of the Initial Number of Infective individuals as shown in Figure 7.3. This has a mean value of 5.358.

When plotting the Histogram using the output for the Infection Rate in the Epidemic we get Figure 7.4. This has a mean value of 0.324439.

Figure 7.5 shows the histogram for the output of the Approximate Bayesian Computation for the Rate of Recovery of the infection. This has a mean value of 0.507514.

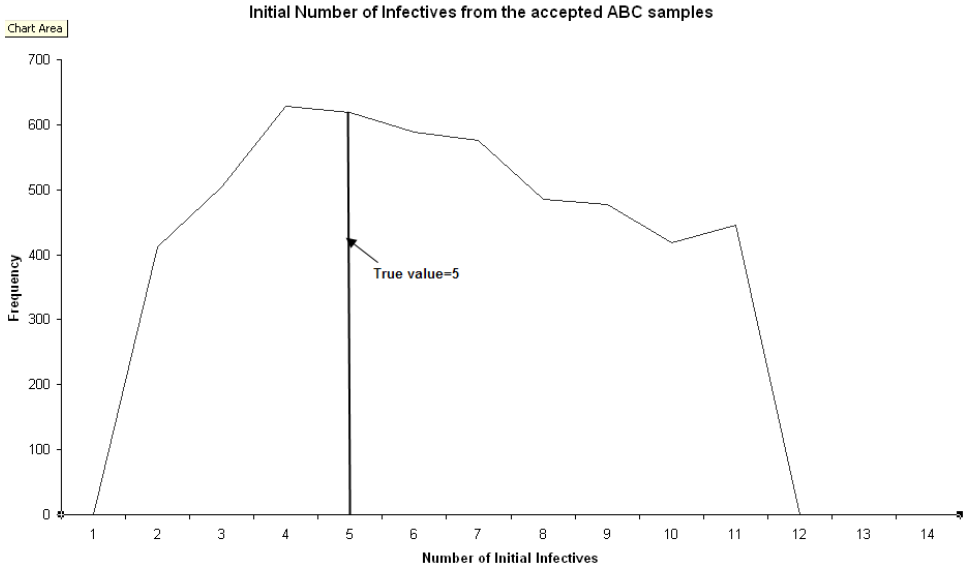

Figure 7.3: Posterior Distribution for the Initial Number of Infectives

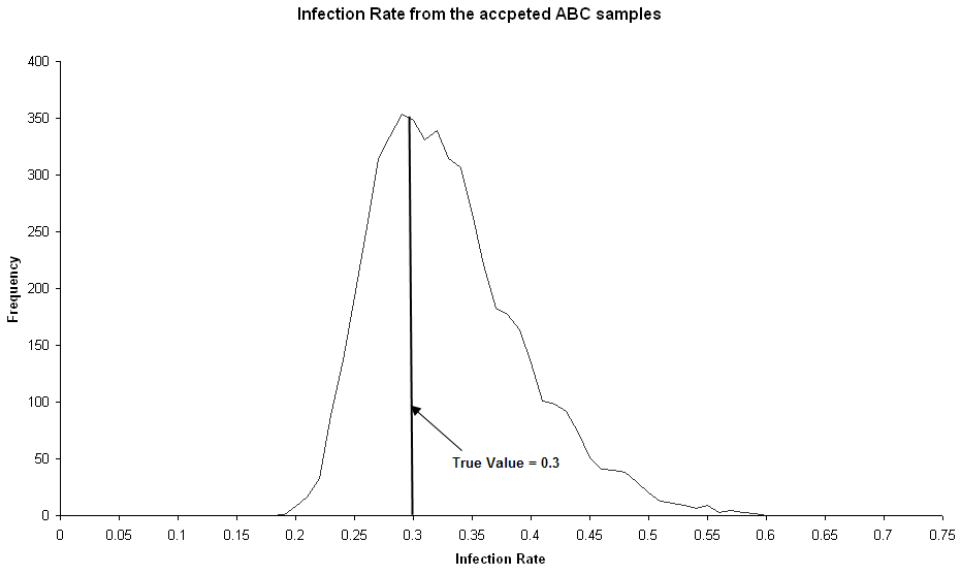

Figure 7.4: Posterior Distribution for the Rate of Infection

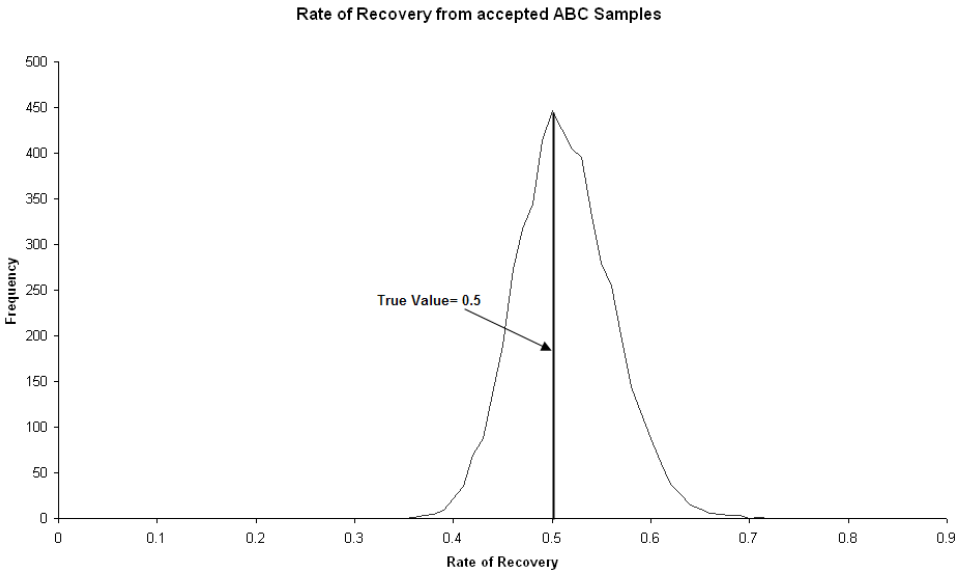

Figure 7.5: Posterior Distribution for the Rate of Recovery

7.3 Using Approximate Bayesian Computation on Real Data

We are using data from an investigation by Chowell, Castillo-Chavez and Diaz-Duenas (2005) [20] into the number of cases of acute hemorrhagic conjunctivitis where the number of cases fully recorded by public health clinics of the Mexican Public Health (IMSS) in the state of Colina, Mexico during the period of September-November 2003.

The data is presented in graphical form of the number of new infective individuals each day in the time period as shown below.

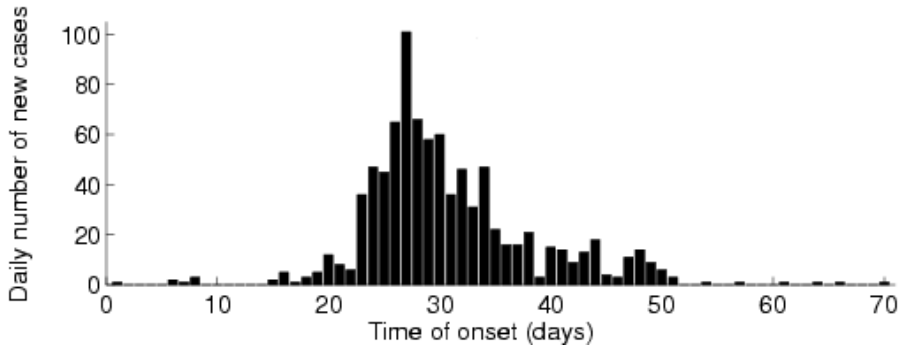

Figure 7.6: Number of New Infectives at each time, taken from Chowell et al (2005)[20]

There is no data on when the individuals recovered. However in the paper it states there is an incubation period of approximately 1-2 days followed by approximately 3-7 days of symptoms. So in calculating the number of infectious individuals at each time we have assumed that each infective stays infective for a length of time which has a geometric$\left(\frac{1}{7}\right)$ distribution. Although this seems high when comparing it to the

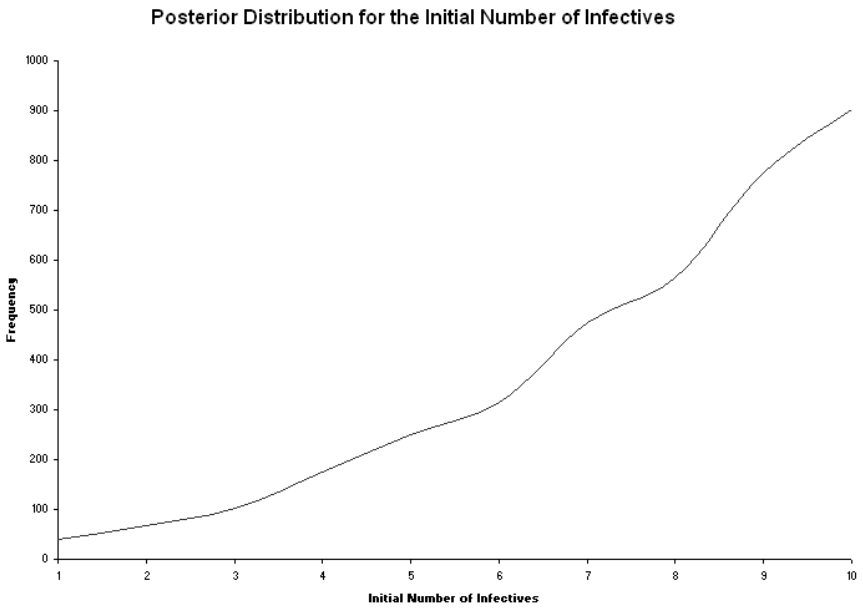

Figure 7.7: Posterior Distribution for the Initial Number of Infectives

average incubation and symptom time it is the smallest value we could take as if it were any smaller the infection would have died out before the second case was diagnosed, which is impossible if the epidemic was to continue.

Inputting this data into the `data` file in Appendix A4, written by Dave Dale, and carrying out Approximate Bayesian Computation on the data with parameters - total time=100, infection rate prior $Uniform(0.01, 1)$, recovery rate prior $Uniform(0.01, 1)$, Initial Infectives Prior $Uniform(1, 10)$, $\varepsilon = 3310000$ and population 886.

This value for ε has an acceptance rate of approximately 0.05% which is a good compromise in the amount of data and the accuracy of the values.

Using the output of this Approximate Bayesian Calculation - the number of initial infective individuals, rate of infection and rate of recovery for the simulations which have been accepted. We can use this data to give us the histogram for the initial number of infective individuals in the population shown in Figure 7.7. This has a mean value of 7.579.

The histogram for the Posterior distribution of the rate of infection is shown in 7.8. This gives the mean value of the rate of infection to be 0.6137.

Figure 7.9 shows the posterior distribution for the rate of recovery. It has the mean value as 0.9807.

7.4 Discussion

Approximate Bayesian Computation appears to work very well in recovering parameters from distribution. All values of the parameters gained from carrying out the Approximate Bayesian Computation were very close to the values we were supposed to be gaining.

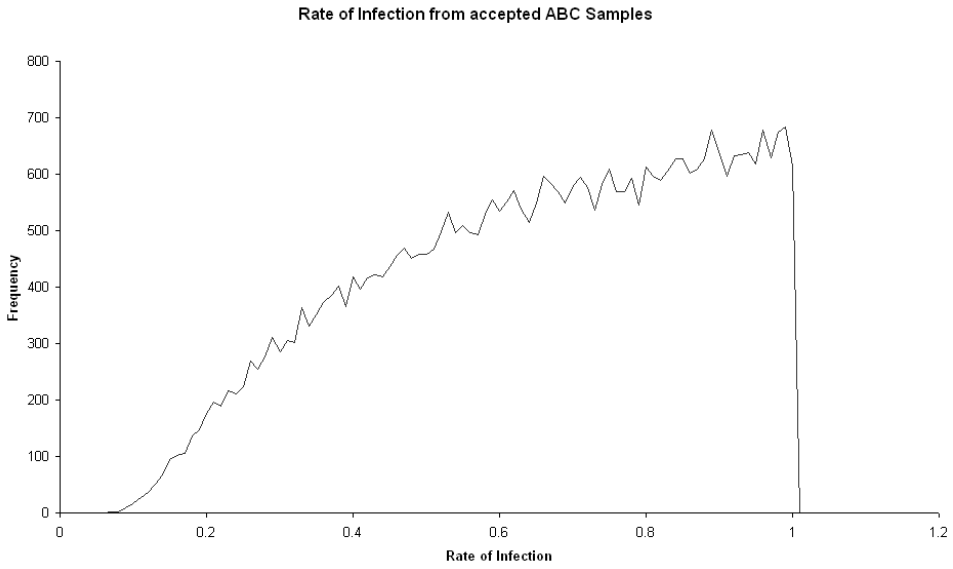

Figure 7.8: Posterior Distribution for the Rate of Infection

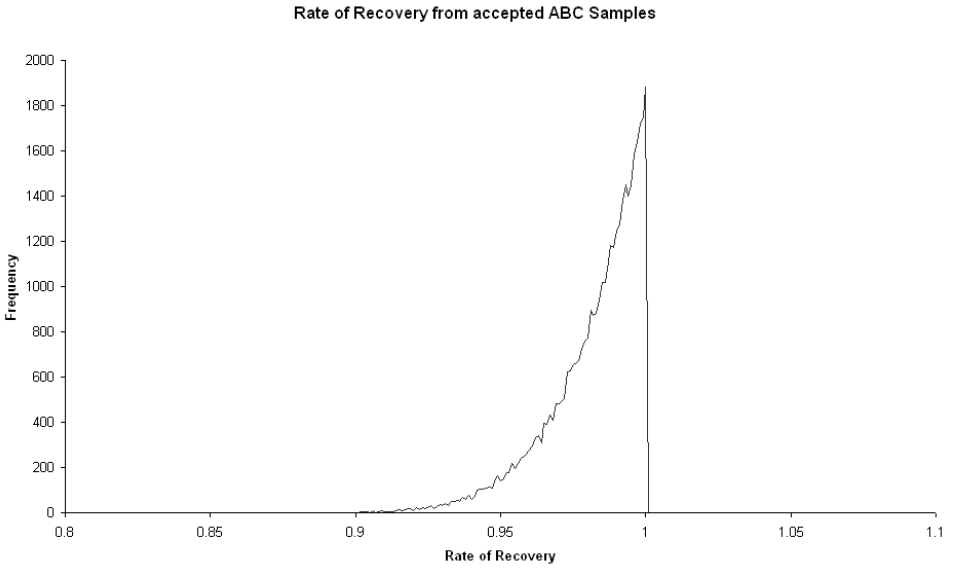

Figure 7.9: Posterior Distribution for the Rate of Recovery

Estimated Parameter	ABC Value	True Value
Initial Number of Infectives	5.358	5
Rate of Infection	0.324	0.3
Rate of Recovery	0.508	0.5

This suggests Approximate Bayesian Computation is an appropriate method to use to estimate the values of the parameters in the SIR model when we have complete data. Also our choice of ε to be 6500 was a good choice as it enabled us to return accurate parameter values without having a large amount of data.

However, when we only know the number of infective individuals at each time point, so having incomplete data, Approximate Bayesian Computation does not appear to be a very effective method for calculating the parameters. As we inputted, the number of infective individuals at each time point into the data file we know that the Approximate Bayesian Computation should give the initial number of infective individuals to be 1. However, it gave us a value of 7.579. When repeating the Approximate Bayesian Computation when we set the initial number of infectives to be 1 in the code it made no difference in the rate of infection and rate of recovery.

Also, in order to work out the number of infective individuals at each time we assumed that each individual was infectious for a geometric time period with average 7 days. This would imply that the recovery rate should be around 0.143. However, the calculated rate of recovery was 0.981.

These suggest that using Approximate Bayesian Computation is only useful when we know the full course of the epidemic - the number of susceptible, infectious and recovered individuals at each time point of the epidemic. Knowing this amount of data for an epidemic would be very rare, as often the actual time of infection and recovery is not known due to individuals being infectious for a period of time before symptoms arise and again not being infectious when they have symptoms of the infection. So questions could be raised on the suitability of Approximate Bayesian Computation on real life data.

Chapter 8

Conclusion

The primary objectives of this project were:

- Compare the Classical and Markov Chain Monte Carlo Methods of Inference to estimate the infection rate for the Reed-Frost Epidemic Model.
- See if Approximate Bayesian Computation can be used to accurately estimate the rates of infection and recovery from the SIR model.

When estimating the rate of infection for the Reed-Frost Epidemic Model, both the maximum likelihood and Gibbs Sampling methods gave approximately the same value. However, when carrying out a χ^2 goodness of fit test to the data we found this estimate was not a particularly good fit to the data, with its value being significant at the 1% level.

Approximate Bayesian Computation was found to be a very good and efficient method of finding parameters for a SIR model when we have complete data. When we simulated some data with specified parameters it accurately returned these values when we carried out Approximate Bayesian Computation on this simulated data as being "real data."

However, when we had incomplete data, as in the data set for the outbreak of Acute Hemorrhagic Conjunctivitis, carrying out Approximate Bayesian Computation did not return the correct values. We know this as the Approximate Bayesian Computation program also gives us an estimate for the initial number of infected individuals in the population at time 0. We know this value as we inputted it into the data set. In this case there was 1 initial infected individual, but the Approximate Bayesian Computation output had an average value of 7.59 initial infected individuals. Also, we assumed each individual was infective for 7 days so the rate of recovery should be approximately $\frac{1}{7} = 0.1429$, but Approximate Bayesian Computation gave a value of 0.9807.

Further work should be done in trying to improve the Markov Chain Monte Carlo Method for finding the infection rate for the Reed-Frost Epidemic model. Also, Approximate Bayesian Computation seems to be a suitable method for estimating parameter values from a SIR Model but further work should be undertaken to improve it so that it can be used to find appropriate parameters for true epidemics where we have incomplete data or even see if it is feasible to use at all. One suggestion for how this could be carried out is by changing the distance function used to measure the distance between the real and simulated data.

Bibliography

[1] Forecast and control of epidemics in a globalised world, L. Hufnagel, D. Brockmann and T. Geisel, PNAS, 101:15124-15129,2004.

[2] On a Partial Differential Equation of Epidemic Theory. I, J. Gani, Biometrika, 52:617-622, December 1965

[3] A Solution of the General Stochastic Epidemic, V. Siskind, Biometrika, 52:613-616, December 1965

[4] Stochastic Epidemic Models and Their Statistical Analysis,Håkan Andersson, Tom Britton, Springer (2000)

[5] Qualitative Analyses of Communicable Disease Models, Herbert W. Hethcote, Mathematical Biosciences 1976 28:335-356

[6] The Dynamics of cocirculating influenza strains conferring partial cross-immunity, Viggo Andreasen, Juan Lin, Simon A. Levin, Journal of Mathematical Biology (1997) 35:825-842

[7] The Stochastic General Epidemic Model Revisited, L. Billard and Zhen Zhao, IMA Journal of Mathematics Applied in Medicine & Biology (1993) 10,67-75

[8] The Transititon Probabilities of the General Stochastic Epidemic Model, Richard J. Kryscio, Journal of Applied Probability, Vol 12, No.3 (Sep.,1975),pp.415-424

[9] Factorial Moments and Proababilities for the General Stochastic Epidemic, L. Billard, Journal of Applied Probability, Vol.10, No2, (Jun.,1973),pp277-288

[10] Two theorems on solutions of differential-difference equations and applications to epidemic theory, Norman C. Severo, J. Appl. Prob. 4,271-280(1967)

[11] A recursion theorem on solving Differential-difference Equations and applications to some stochastic processes, Norman C. Severo, J. Appl. Prob 6,637-681 (1969)

[12] An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations, Desmond J. Higham, SIAM Review, Vol.43, No.3 pp525-546

[13] Bayesian inference of stochastic multitype epidemics in strucured population via random graphs, Nikolaos Demiris and Philip O'Neill, J.R. Statist Soc B(2005) 67 Part 5 pp731-745

[14] Bayesian Inference for Stochastic Epidemics in Populations with Random Social Structure, Tom Britton and Philip O'Neill, Scand. J Statist 29

[15] Bayesian Inference for Stochastic Epidemics in Closed Populations, George Streftaris and Gavin Gibson, Statistical Modelling 2004; 4; 63

[16] Bayesian Inference for Partially Observed Stochastic Epidemics, Philip O'Neill and Gareth Roberts, J.R. Statist. Soc A (1999), 162, Part1, pp121-129

[17] The estimation of parameters from population data on the general stochastic epidemic, Norman Bailey and Anthony Thomas (1971) Advances in Applied Probability.

[18] Simulation, Sheldon M. Ross, Academic Press 3rd Edition (2002)

[19] Bayesian Inference in Statistical Analysis, Box and Tiao, Wiley Classics Library (1992)

[20] Characterisation of an Outbreak of Acute Hemorrhagic Conjunctivitis in Mexico, 2003, G.Chowell, C. Castillo-Chavez, P. Diaz-Duenas, Digital Journal of Opthalmology.

[21] The mathematical Theory of Infectious Diseases and its Application, Bailey, Griffin(1975)

[22] Elements of Statistics, Daly, Hand, Jones, Lunn, McConway, PrenticeHall (1995)

[23] Markov Chain Monte Carlo in Practice, WR Gilks, S Richardson, DJ Spiegelhalter, Chapman & Hall (1996)

[24] Markov Chain Monte Carlo without Likelihoods, P. Marjoram, J. Molitor, V. Plagnol, S. Tavaré, PNAS, December 2003, vol 100, no. 26, pp1534-15328

Appendix A

Computer Code

A.1 R code for Gibbs Sampling of q and n_{21}

In the code in Figure A1 we apply the Gibbs Sampler Algorithm.

The `onecomptrial` command gives the posterior distributions for q and n_{21} if we input values of α, β and the burn-in time (B), the number of samples we require (T) and the frequency at which we want these samples to be taken (M).

`graphq` and `graphn21` plot the graphs of the posterior distributions of q and n_{21} respectively.

Similarly `meanq` and `meann21` give the mean values of the posterior distribution of q and n_{21} respectively and `varq` and `varn21` give the variance of the posterior distribution of q and n_{21}.

A.2 R Code for the Convergence of the Gibb's Sampler

The code in Figure A.2 shows whether or not the Gibb's sampler has converged and so is giving independent values for the posterior distribution of q and n_{21}.

A.3 Code for the Simulation of the SIR Model

I have used R to simulate populations obeying equations 2.4 and 2.5.

In order to carry out such simulations the number of trials that we desire (the number of simulations), the time period over which we are interested in the population, the initial distribution of the population (the initial number of susceptible individuals, infective individuals and recovered individuals) and also the rates at which a susceptible and an infective meet and form two infective individuals, α as in Equation 2.4, and the rate in which a single infective becomes recovered, β, as in 2.5 are all needed. When these are known we can then run the following program which is shown in Figure Equation A.3

In the code `singlesim` gives the number of susceptible individuals, infective individuals and recovered individuals after a single time interval when the number of susceptible individuals is S, number of infective

```
trial<-function(alpha,beta,n21){
a<-rbeta(1,2n0+2n1+n21+alpha,n1+2n2+beta)
b<-rbinom(1,n2,2*a/(2*a+1))
S<-c(a,b)
S}

onecomptrial<-function(alpha,beta,B,T,M){
R<-M*T
S<-array(0,c(length(-B:R),2))
S[1,]<-c(trial(alpha,beta,0)[1],1)
for (i in 2:length(-B:R)){
S[i,]<-trial(alpha,beta,S[i-1,2])}
Q<-array(0,c(length(0:R),2))
for(i in 1:length(1:R)){
Q[i,]<-S[B+i-1,]}
P<-array(0,c(length(1:M),2))
for(i in 1:length(1:M)){
P[i,]<-Q[T*i,]}
P}

par(mfrow=c(2,3))
graphq<-function(alpha,beta,B,T,M){
hist(onecomptrial(alpha,beta,B,T,M)[,1],nclass=100,xlab="",ylab="density",
main="")
}

meanq<-function(alpha,beta,B,T,M){
mean(onecomptrial(alpha,beta,B,T,M)[,1])
}

varq<-function(alpha,beta,B,T,M){
var(onecomptrial(alpha,beta,B,T,M)[,1])
}

par(mfrow=c(2,3))
graphn21<-function(alpha,beta,B,T,M){
hist(onecomptrial(alpha,beta,B,T,M)[,2],nclass=100,xlab="",ylab="density",
main="")
}

meann21<-function(alpha,beta,B,T,M){
mean(onecomptrial(alpha,beta,B,T,M)[,2])
}

varn21<-function(alpha,beta,B,T,M){
var(onecomptrial(alpha,beta,B,T,M)[,2])
}
```

Figure A.1: Gibbs Sampler R code

```
burnin<-function(alpha,beta,B,T,M,a){
R<-M*T
S<-array(0,c(length(-B:R),2))
S[1,]<-c(trial(alpha,beta,0)[1],1)
for (i in 2:length(-B:R)){
S[i,]<-trial(alpha,beta,S[i-1,2])}
Q<-array(0,c(length(0:B),2))
for(i in 1:length(0:B)){
Q[i,]<-S[i,]}
plot(Q[,a],xlab="",ylab="",main="",type="l")
}
```

Figure A.2: Burn-In code

individuals is I and the number of recovered individuals is R. alpha and beta are the α and β respectively that are in 2.4 and 2.5.

`onecompsim` gives the number of susceptible individuals, infective individuals and recovered individuals in a population after an interval of size Nsteps. It gives the numbers of susceptible individuals, infective individuals and recovered individuals after each single interval.

The function `fisim` does `onecompsim` repeatedly `simulations` times.

The function `graph` plots a graph for each of the `onecompsim`'s carried out in `fisim`. The `a` parameter specifies whether we want the graph is of the number of susceptible individuals `(a=1)`, the number of infective individuals `(a=2)` or the number of recovered individuals `(a=3)`.

The `mean` function finds the mean number of susceptible individuals, infective individuals and recovered individuals at each interval using each of the simulations in `fisim`.

The `allgraph` function plots the mean values for each of the number of susceptible individuals, infective individuals and recovered individuals on the same set of axes.

A.4 Approximate Bayesian Computation Code

Dave Dale wrote some code in C^{++} for carrying out Approximate Bayesian Computation on a SIR data set and also a simulator which needs to be run alongside the Approximate Bayesian Computation programme. It was not possible to use my simulator with the Approximate Bayesian Computation programme as it was written using R and so was not compatible.

```
singlesim<-function(alpha,beta,s,i,r){
a<-(alpha/(s+i+r))*s*i
b<-beta*i
d<-runif(1)
ran<-max(a*d,beta*i*d)
z<-c(s,i,r)
j<-ifelse(c(ran>a&ran<=a+b,ran>a&ran<=a+b,ran>a&ran<=a+b),c(s,i-1,r+1),z)
l<-ifelse(c(ran<=a,ran<=a,ran<=a),c(s-1,i+1,r),z)
x<-ifelse(c(ran<=a,ran<=a,ran<=a),l,j)
q<-ifelse(j==z&l==z,z,x)
g<-ifelse(c(ran==0,ran==0,ran==0),z,q)
g}

onecompsim<-function(Nsteps,S,I,R,alpha,beta){
P<-array(0,c(Nsteps,3))
Z<-array(0,c(Nsteps,3))
Y<-array(0,c(Nsteps,3))
P[1,]<-c(S,I,R)
for(i in 2:Nsteps){
P[i,]<-singlesim(alpha,beta,P[i-1,1],P[i-1,2],P[i-1,3])}
P}

fisim<-function(Simulations,Nsteps,S,I,R,alpha,beta){
Q<-array(0,c(Nsteps,3,Simulations))
for(i in 1:Simulations){
Q[,,i]<-onecompsim(Nsteps,S,I,R,alpha,beta)}
Q}

# a is whether we want the graph to be of the number of suceptibles a=1, the number of
infectives a=2 or the number of recoverds a=3
graph<-function(Simulations,Nsteps,S,I,R,alpha,beta,a){
plot(fisim(Simulations,Nsteps,S,I,R,alpha,beta)[,a,1],axes=FALSE,type="l",ylim=c(0,S+I+R))
axis(2,0:(S+I+R))
axis(1)
for(i in 2:Simulations){
lines(fisim(Simulations,Nsteps,S,I,R,alpha,beta)[,a,i])}
}
rm(.Random.seed)

mean<-function(Simulations,Nsteps,S,I,R,alpha,beta,a){
Q<-1:Nsteps*0
for(i in 1:Nsteps){
Q[i]<-sum(fisim(Simulations,Nsteps,S,I,R,alpha,beta)[i,a,])/Simulations}
Q}

meangraph<-function(Simulations,Nsteps,S,I,R,alpha,beta,a){
plot(mean(Simulations,Nsteps,S,I,R,alpha,beta,a),type="l", col="red",ylim=c(0,S+I+R),xlab="Time",
ylab="Number",main="The number of susceptibles, infectives and recoverds with respect to time")}
allgraph<-function(Simulations,Nsteps,S,I,R,alpha,beta){
plot(mean(Simulations,Nsteps,S,I,R,alpha,beta,1),type="l",col="red",ylim=c(0,S+I+R),xlab="Time",
ylab="Number",main="The number of susceptibles, infectives and recoverds with respect to time")
lines(mean(Simulations,Nsteps,S,I,R,alpha,beta,2),col="blue")
lines(mean(Simulations,Nsteps,S,I,R,alpha,beta,3),col="green")}
```

Figure A.3: R code for simulating an SIR model

Appendix B

Solutions of the Probabilistic SIR model

There have been many solutions produced in various forms for the Markovian form of the probabilistic SIR model. The first solutions were in 1965 by Gani [2] and Siskind [3], who independently and simultaneously solved the equations using different methods - Gani used Laplace transforms to find probabilities for populations of sizes N=2,3 and Siskind used a more recursive method to find solutions for populations of any size, however, these solutions are highly recursive and so highly impractical to calculate for large populations.

For example in Siskind's Paper
The individual probabilities $p_t(i,s)$ can be found by identifying the coefficients of

$$p_t(i,s) = (-1)^i(\rho+s)^{i-1}\rho^{n+a-s-i}\frac{n!}{s!}\sum_{m=0}^{n-s}\sum_{h=M_1}^{a+m} e^{-h(\rho+n-m)t}S(m,h)G^*_{n-s}(m,h) \tag{B.1}$$

where

$$G^*_{n-s}(m,h) = \begin{cases} \left[\prod_{\beta=m+1}^{n-s-1}\sum_{k_b=0}^{2}\sum_{j_\beta=M_2}^{h_{\beta-1}-1} h_\beta(\rho+n-1-\beta)^{-h_\beta}A(m,h;j_\beta,k_\beta\right] & \\ \times\sum_{k_{n-s}=0}^{2}\sum_{j_{n-s}=M_3}^{h_{n-s-1}-1} A(m,h;j_{n-s},k_{n-s})(-1)^{h_{n-s}}\binom{h_{n-s}}{i} & m<n-s \\ (-1)^h\binom{h}{i} & m=n-s \end{cases} \tag{B.2}$$

$$M_2 = max(i+s+\beta-n-k_\beta,0)$$

$$M_3 = max(i-k_{n-s},0)$$

$$S(m,h) = \begin{cases} \sum_{j_m+k_m=h}\binom{n}{k_m}(\rho+n-m+1)^{j_m} & \\ \times\left[\prod_{\beta+1}^{m-1}\sum_{k_r=0}^{2}\sum_{(2)j_{m-\beta}}(\rho+n-m+1+\beta)^{j_{m-\beta}}B(m,h,j_{m-\beta},k_{m-\beta})\right] & \\ \times\sum_{j_0=j_1+1}^{a}\binom{a}{j_0}B(m,h;j_0,k_0=0) & m>0 \\ \binom{a}{h} & m=0 \end{cases} \tag{B.3}$$

$$A(m,h;j_\beta,k_\beta)=\binom{2}{k_\beta}\binom{h_{\beta-1}-1}{j_\beta}(\rho+n+1-\beta)^{j_\beta}\frac{h_\beta}{h_\beta(\rho+n-\beta)-h(\rho+n-m)} \tag{B.4}$$

$$B(m,h;j_\beta,k_\beta)=\binom{2}{k_\beta}\binom{h_\beta-1}{j_{\beta+1}}(\rho+n-1-\beta)^{-h_\beta}\frac{h_\beta}{h_\beta(\rho+n-\beta)-h(\rho+n-m)} \tag{B.5}$$

The lower and upper limits of summation of the $\Sigma_{(.)}$ are,

$$\Sigma_{(2)}0\rightarrow h_{r-1}-1$$

In 1973 Billard [9] improved the methods of Gani and Siskind by producing solutions that whilst still recursive are not as recursive as the earlier methods. Unlike the methods of Gani and Siskind, this method can be used to find the asymptotic solutions and properties for the epidemic model for a population of infinite size. In 1975 Kryscio [8] showed how to solve the Kolmogorov forward equations of an epidemic model and applies his technique to the forward equations of the general stochastic model for the special case of there being one initial infective. This gives a solution similar in form to Billard and Siskind, but this formula is explicitly defined and algebraically similar. Severo [10] presented two theorems that provided simple iterative solutions of special systems of differential-difference equations considering the General Stochastic Epidemic.

Then in 1993 Billard and Zhao [7] used a completely different approach in using the number of infective individuals and recovered individuals as opposed to previous solutions which considered the number of susceptible individuals and infective individuals. This gives solutions that are simpler and easier to manage than those previously found. Also, these solutions are not restricted by the size of the population, N - they still work when N is large, a problem which occurs with many of the other methods (Gani, Siskind, Kryscio). It can also help to estimate the unknown model parameters, which before this paper had been an almost untouched area.

Each of these methods have their own advantages and disadvantages. However, they are all computationally very heavy and as a result although it is possible to use some of these for large population sizes it would take a considerable length of time to carry out such calculations, thus making them largely useless for carrying out Maximum Likelihood Methods of Inference.