Computer Intensive Methods for Modelling Household Epidemics.

Author: Zak Varty

Supervisor: Dr. Peter Neal

April 2016

Abstract

Modelling the ways in which infection enters and progresses through a community is key to preventing and intervening in outbreaks of disease. Chain binomial models of infection are considered, focusing on the Reed-Frost model, and Approximate Bayesian Computation methods are considered as a means of parameter estimation. Techniques for handling the posterior samples generated are discussed, including local linear regression and kernel density estimation. Variants on the Reed-Frost model are used to extend the model to a community of households and to allow the infectious period of individuals to vary. Examples of the techniques are given using two example data sets of influenza epidemics in Tecumseh, Michigan and Seattle, Washington.

Contents

1 Introduction 3

2 Chain Binomial Models for Epidemics 4

2.1 SIR Models 5

2.2 The Reed-Frost Model 6

2.3 Properties of an Epidemic 8

2.4 Maximum likelihood estimation and the Reed-Frost model 10

3 Parameter Estimation 12

3.1 Approximate Bayesian Computation 14

3.2 Applying Approximate Bayesian Computation 1 18

3.3 Applying Approximate Bayesian Computation 2 20

3.4 Efficient use of parameter space 21

4 Working with Posterior Samples 23

4.1 Local Linear Regression 23

4.2 Kernel Density Estimation 29

5 Extending the Reed-Frost Model 34

5.1 Introducing Community Structure 34

5.2 Random Infectious Period 38

6 Application of Methods 40

6.1 Community Infection with Uniform Proposal Distribution 41

6.2 Community Infection with Beta Proposal Distribution 42

6.3 Random Infectious Period 45

7 Conclusion 48

8 References 49

1 Introduction

Historically, infectious disease has been the most prominent cause of human mortality. Diseases throughout history have been hugely detrimental to everyday life, from the bubonic plague and dysentery, through to modern cases such as swine-flu and Ebola. While many diseases are now manageable due to medicinal advances, in low-income countries around the world infectious disease still accounts for over one third of all deaths [13]. In higher income countries infectious disease is still an issue, though the predominant cause of mortality is chronic disease there are significant economic costs associated with lost productivity due to illness. For example, in 2013 the United Kingdom lost 27 million working days were lost to minor infectious illness – having major economic impact [8].

Knowledge of infectious diseases is clearly of great importance to the individuals who contract them. In addition, there are many professions who hold a financial, academic or logistical stake in there being a sound theory of infectious disease. In the academic environment, epidemiologists study the relationships between the causes and effects of disease within a population and bio-mathematicians are interested in representing biological processes more generally, which of course includes disease. Taking advice from these academics are health workers, medics, and public health officials making practical treatment plans and policies. The extent of interest goes further still to world leaders, who must draw on sound knowledge to curb global outbreaks of disease and to unite nations in actions against disease.

In addition to medicinal uses, the understanding of infectious disease is compelled by biological warfare. In the development of biological weapons there is the need to develop disease strains both for maximum infection and also for targeted infection. On the side of biological defence, it is necessary to have sufficient modelling techniques to inform the best course of prevention and intervention of attacks. In addition to both of these being theoretical endeavours there is also a practical element. A mathematical background is equally important in informing the development process of these biological agents. The health threat here comes from the risk of accidental exposure during the development process and modelling is used to develop safeguards against failings of the development process.

One of the first considered cases of epidemiological study, the application of statistical theory to disease, was that of John Graunt in 1663 with his text, Natural and Political Observations Made upon the Bills of Mortality. Another major advancement in the field was that of Daniel Bernoulli analysing censored smallpox data in 1766 [4], but major advancements did not come until the adoption of germ theory and bacteriology into standard medical practice during the late nineteenth and early twentieth century. This allowed the development of our modern theory of disease transmission, and progress increased further as data collection, distribution and analysis became easier and possible on far larger scales in recent decades. Work on disease modelling was predominantly deterministic in nature until the introduction of the chain binomial model for disease

transmission was introduced, which will be the model of focus in this text. Further research in the 1940s lead to the development of various stochastic models which are capable of modelling a range of diseases in different community structures, depending on the research question of interest. As mentioned previously, there has been a great increase in the volume of data collected and also in the computing power available to manipulate this quantity of data. This has allowed researchers to observe large communities, which produce fairly homogeneous data when taken in such volume. The increased computational power has also permitted the use of computationally intensive methods, both for modelling epidemics and parameter estimation within those models using the large data sets. These methods are of particular use as direct likelihood methods rapidly become unwieldy as we shall see in Section 2.2, and computational methods will be a theme throughout the text. Worked examples of these computational methods have been included in the programming language R as a part of the text and these were executed using R version 3.2.2 .
Initially chain binomial models of infection are considered, focusing on the Reed-Frost model (Section2). Approximate Bayesian Computation is then considered as a means of parameter estimation for this model (Section 3), as well as techniques for handling the posterior samples generated by the algorithm including local linear regression and kernel density estimation (Section 4). Variants on the Reed-Frost model are introduced in order to model a community of households and to allow the infectious period of individuals to vary (Section 5). Examples of the techniques are given using two example data sets of influenza epidemics in Tecumseh, Michigan and Seattle, Washington (Section 6).

2 Chain Binomial Models for Epidemics

It is clear that not all diseases are infective, or indeed that all infective diseases are equally so. This was brought up in the 1931 paper ‘On the statistical measure of infectiousness’ by Greenwood [10], where an infectious disease was defined as being caused by ‘close association (of some kind) with an immediately pre-existing case’, and comparisons are drawn between infectious diseases such as measles, influenza and gonorrhoea. These diseases are listed in decreasing order of how ‘catching’ they are, to use Greenwood’s terminology, which would be expressed numerically as the ratio of the number of individuals who are infected by the disease and the number of individuals who are exposed to the disease. Looking at a household level, measles would usually infect all individuals in the family and so would be seen as highly catching. Influenza would be seen as less catching, and catching in a different manner to measles in that there is additional association required for transmission to occur, for example being close to the individual when they sneeze. Gonorrhoea would require even further contact for transmission, and so association is once more defined in a different way. We will see later in this section that this concept of quantifying the catching quality of a disease has been brought through into the Reed-Frost model which is the main focus of this dissertation, along with estimation of the ‘catching’ value.

2.1 SIR Models

The Reed-Frost model is an example of an SIR compartmental model, meaning that at any given time the population of interest is partitioned into three sets, defined by the current disease states of the individuals with respect to the disease of interest. These three states are labelled susceptible, infective and recovered. Often the recovered state is replaced by an equivalent group labelled removed, when the disease of interest is fatal. This labelling is not of consequence to the mechanics of the model because it is assumed that once an individual is in the recovered (or removed) state they are no longer capable of contracting or spreading disease. In this way, recovered individuals can be thought of as being immune to the disease of interest.

Within an SIR model, individuals may only move linearly forwards through the three states. A susceptible individual may be infected and move into the infective group, but may not go from susceptible to recovered without first becoming infective. Similarly, an infective individual may recover but can not become susceptible again, and a recovered individual can not move out of this state. The time between infection occurring and an individual becoming infective is the latent period of the disease, and the period between an individual becoming infective and recovering is known as the infectious period. The states and the possible transitions between them are shown in Figure 1.

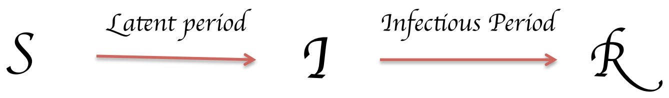


Figure 1: The possible states and transitions in an SIR compartmental model of disease.

For SIR models the following notation is conventional.

S_t The number of individuals who are susceptible to the disease of interest at time t.

I_t The number of individuals who are infectious with the disease of interest at time t.

R_t The number of individuals who are recovered from the disease of interest at time t.

SIR models can be either deterministic or stochastic in nature. Deterministic SIR models assume a fixed rate of transitioning from each state into the next; for example taking β to be the fixed rate at which susceptibles are infected and γ the fixed rate at which infectives recover. This leads to a system of differential equations which can be solved simultaneously to provide information on the types of equilibria a disease could settle into; a pandemic, extinction of the disease or a steady proportion of the population being infected. The differential equations use the principle of mass action, which states that the overall rate of transition is proportional to the size of each of the groups. For the continuous, deterministic model with transitions as shown in Figure 1, the system of equations would be;

$$\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= \frac{-\beta SI}{N}, \\
\frac{\mathrm{d}I}{\mathrm{d}t} &= \frac{\beta SI}{N} - \gamma I, \\
\frac{\mathrm{d}I}{\mathrm{d}t} &= \gamma I,
\end{aligned}$$

where $N = S_t + I_t + R_t$ is the constant total population size [2].
Stochastic SIR models assume that individuals move from one state to another with some given probability. In chain binomial models, time is viewed as discrete and the probability of moving between states is constant. This means that the number of individuals moving from each category is the sum of S_t or I_t Bernoulli random variables, which are binomial random variable describing the transition between states of a discrete time Markov chain. A special case of the chain binomial model, the Reed-Frost model, will be the main focus of study and so these ideas will be elaborated on in Section 2.2.

2.2 The Reed-Frost Model

The Reed-Frost model was developed by Dr. L. Reed and Dr. W. Frost between 1927 and 1928 for a lecture series at John's Hopkins school of hygiene and public health, as an extension of work by G. Sopher. Sopher's model was deterministic, where all members of the population were equally susceptible to the disease and had equal *infective power*, the ability to transmit disease. The model also included an equivalent to the recovered group in SIR models; individuals passed out of observation a short time after they became infective. The model also relied on the law of mass action, so that the rate of infection at a given time was proportional to the number of susceptibles and infectives at that time. The model is effective for moderately large populations but a shortcoming of the model by Sopher is that it does not take into account multiple infectives coming into contact with the same individual. Therefore, in small populations the model greatly overestimates the number of susceptible individuals who become infected. Abbey [1] published a paper in 1952 on the model by Reed and Frost, giving details of the work by Sopher.
The Reed Frost model is a chain binomial model for the spread of disease, and as such is a stochastic SIR model. It is well suited to modelling small, self-contained populations such as a household or the ward of a hospital, as it does not rely on the law of mass action, though the idea of proportional infective power is incorporated into the model. The assumptions of the Reed-Frost model are given below.

Assumptions of the Reed-Frost Model

1. The disease is transferred from one individual to another only by some form of contact, described as *sufficient contact*. When this contact is made disease is always transmitted.

2. The infectious period of the disease is short in comparison to the latent period.

3. The population is closed, homogeneous and well mixed.

4. The probability of sufficient contact occurring between two individuals is proportional to the current number of infective individuals.

5. There are no births or deaths within the population during the outbreak of disease.

The first assumption allows us to switch focus, from individuals infecting one another to coming into contact with one another. This also simplifies the vocabulary and somewhat helps prevent confusion, for example when considering infectives infecting susceptibles during their infectious period. In terms of the model, the most important part of this assumption is that there is no other way in which the disease can be transmitted.
The second assumption allows time to be viewed as discrete in the Reed-Frost model. If the infectious period is short in comparison to the latent period, then in one latent period all infectives become immune and all susceptibles who have come into contact with an infective become infective themselves. It is then possible to view the disease having *generations* and it is logical to scale time to have one unit of time be equivalent to the latent period; one generation of the disease. This can limit the applications of the Reed-Frost model because not all diseases have this property, however the property holds well for diseases such as influenza and chicken-pox.
The third assumption ensures that we are considering all possibilities, and simplifies the contact structure within the population. By having a closed population we eliminate any contact with the world outside of the household, meaning that there is no chance of an individual becoming infected from outside of the population of interest. Having the population be homogeneous means that the latent and infectious period is the same for all individuals within the population and that we do not need to identify *who* is infected at each point, only the number of people that are infected. Finally, the well mixed assumption allows us to assume that the probability of any two individuals coming into sufficient contact with one another is equal. This is what makes the Reed-Frost model particularly well suited to describing small populations, as this assumption is unlikely to hold in reality for large groups.
The fourth assumption adds a similar property to the law of mass action when combined with the third. The assumption allows us to assume that the number of contacts made is greater when the number of susceptibles is greater, and also when then number of infectives is greater. Finally, the fifth assumption means that the total number of susceptible, infective and recovered individuals remains constant over time and that,

because individuals may only move forwards through the states, the number of susceptibles is non-increasing.
The Reed-Frost model can be thought of as a bivariate, discrete time Markov chain with binomial transition probabilities. A discrete time Markov chain is defined as a sequence of random variables $X_1, X_2, \dots$ which display the Markov property; the probability distribution of the next random variable in the chain is completely described by the realisation of the current state and not any state previous. This property is given formally as:

$$P(X_n = x_n | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \dots, X_2 = x_2, X_1 = x_1) = P(X_n = x_n | X_{n-1} = x_{n-1}).$$

The Reed-Frost model can be thought of as a bivariate Markov chain because the probability of having, at time t, S_t susceptible and I_t infectives is only dependent on the numbers of each of these at time $t-1$. The *state space* of the bivariate Markov chain is the set of all possible combinations of numbers of infective and susceptible individuals, $\{0, 1, \dots, s_0\}^2$. The probability of transitioning from state $\{s_t, i_t\} = (s,i)_t$ to $\{s_{t+1}, i_{t+1}\} = (s,i)_{t+1}$ is binomially distributed.

$$S_{t+1} | S_t = s_t, I_t = i_t \ \sim \text{Binomial}(s_t, (1-p)^{i_t}) \quad \text{with} \quad I_{t+1} = S_t - S_{t+1} \quad \text{so that}$$

$$P_{(s,t)_t (s,t)_{t+1}} = \binom{s_t}{s_{t+1}} \cdot ((1-p)^{i_t})^{s_{t+1}} \cdot (1-(1-p)^{i_t})^{(s_t - s_{t+1})}$$

where p is the probability of sufficient contact occurring between any two individuals. The transition probability can be thought of as selecting s_{t+1} of the current s_t susceptibles to avoid sufficient contact with all i_t current infectives. The remaining $s_t - s_{t+1}$ susceptibles fail to avoid sufficient contact and become infectives at time $t+1$. Using this explanation, and letting $q = 1-p$ be the probability of avoiding sufficient contact with between two individuals in the population the transition probability can be expressed more neatly.

$$P_{(s,t)_t (s,t)_{t+1}} = \binom{s_t}{s_{t+1}} \cdot (q^{i_t})^{s_{t+1}} \cdot (1-q^{i_t})^{i_{t+1}}. \quad (1)$$

2.3 Properties of an Epidemic

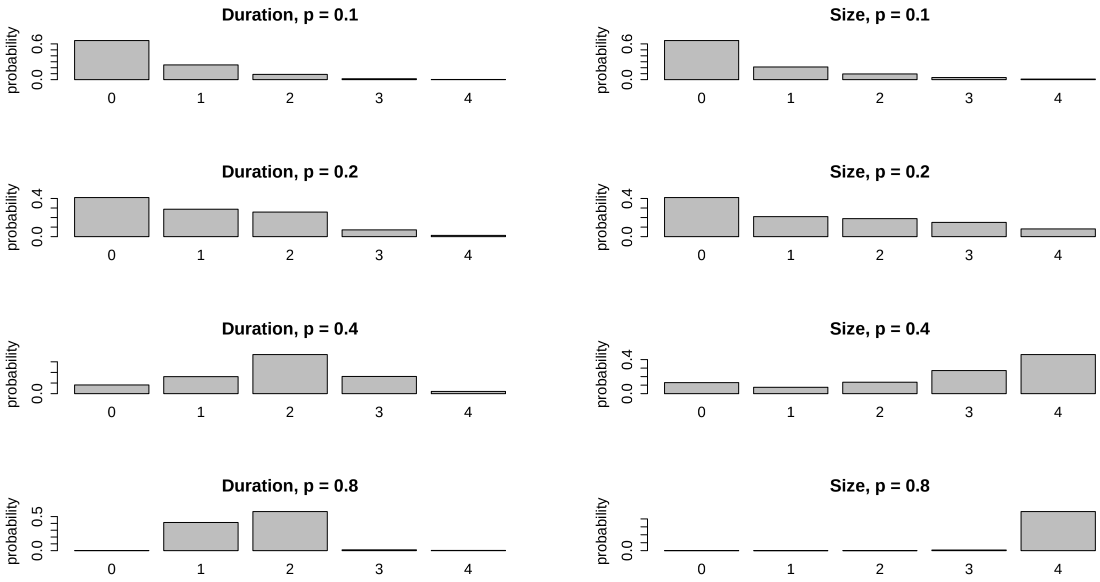
Now that we have the model and notation for a Reed-Frost type epidemic, we introduce a few general properties of epidemics. A *realisation* of an epidemic is an observed sequence of values taken by the Markov chain. This sequence terminates when there are no longer any infectives within the population, as the state is then constant because no more susceptibles can be infected. Since the number of susceptibles at time t is i_t less than previously, $s_t = s_{t-1} - i_t$, and if we know the total number of individuals in the population (which is constant by assumption 5 of the Reed-Frost model), then the

realisation is completely described by the number of infectives at each time. A general realisation of an epidemic would be $\{i_0, i_1, i_2, \ldots, i_T\}$, where T is the *duration* of the outbreak. It should be noted that T is capitalised because it a random variable, like S_t and I_t are at each time $t = 0, 1, \ldots, T$. The *size* of an epidemic W is the total number of individuals who are exposed to the disease over the course of the outbreak, being defined as $W = s_0 - s_T$. Note that the size of an epidemic is also a random variable.
We will now focus briefly on a Reed-Frost outbreak in a household of five people, with one initial infective at $t = 0$ and look at the probability of each realisation of the epidemic using these to examine the distributions of T and W for this set-up. The probability of a given realisation is simple to calculate, if somewhat labour intensive if done by hand, it is the product of the transition probabilities between each of the states observed using Equation 1. In Table 1 we list all possible realisations in our scenario.

Realisation $\{i_0, i_1, \ldots, i_T\}$	Duration T	Size W	Probability
$\{1\}$	0	0	q^4
$\{1,1\}$	1	1	$4pq^6$
$\{1,1,1\}$	2	2	$12p^2q^6$
$\{1,2\}$	1	2	$6p^2q^6$
$\{1,1,1,1\}$	3	3	$24p^3q^7$
$\{1,1,2\}$	2	3	$12p^3q^6$
$\{1,2,1\}$	2	3	$12p^3q^4(1+q)$
$\{1,3\}$	1	3	$4p^3q^4$
$\{1,1,1,1,1\}$	4	4	$24p^4q^6$
$\{1,1,1,2\}$	3	4	$12p^4q^5$
$\{1,1,2,1\}$	3	4	$12p^4q^4(1+q)$
$\{1,2,1,1\}$	3	4	$6p^4q^2(1+q)$
$\{1,2,2\}$	2	4	$6p^4q^2(1+q)^2$
$\{1,1,3\}$	2	4	$4p^4q^3$
$\{1,3,1\}$	2	4	$4p^4q(1+q+q^2)$
$\{1,4\}$	1	4	p^4

Table 1: Enumerating possible realisations of a Reed-Frost type epidemic in a household of size five with one initial infective.

From Table 1 we can sum probabilities over values of T or W to get the probability densities of the summary statistic of our choice. In Figure 2 we see example probability distributions of W and T for a selection of values of p. We can see that as p increases, the distribution of W moves from a positive skew to a negative skew. The distribution of T behaves differently, losing its positive skew until $p = 0.5$ and then regaining it as p increases towards 1. This follows from the fact that the disease is more contagious and so more individuals become infected and these infections happen more rapidly. If the



(a) (b)

Figure 2: Probability densities of T (a) and W (b) for increasing probability of sufficient contact $p = 0.1, 0.2, 0.4, 0.8$ in a household of size 5.

disease is very infectious then many individuals are infected in each period and the epidemic ceases quickly when all individuals have been infected.

2.4 Maximum likelihood estimation and the Reed-Frost model

We will now work fully through a practical example of finding the maximum likelihood estimate for the Reed-Frost model. The data we will use are shown in Table 2, and are from O'Neill and Roberts [14] who summarised it from Wilson et al. [17]. The data look at households of size 3 during outbreaks of measles in Providence, Rhode Island, in the period 1929 to 1934. It is unusual that a paper will contain individual level data, and we deviate from the style of O'Neill and Roberts [14] by not aggregating the data by final size in this example and look at each type of realisation. It is not common that individual level data is available like this because it is often not practical or possible to record the way in which infection spreads between individuals.

Realisation	Final Size, W	Probability of Realisation	General Count	Measles Count
$\{1\}$	0	q^2	n_0	34
$\{1,1\}$	1	$2pq^2$	n_1	25
$\{1,1,1\}$	2	$2p^2q$	n_{21}	36
$\{1,2\}$	2	p^2	$n_2 - n_{21}$	239

Table 2: Household of size 3: possible realisations of the epidemic with probability of occurring, count in a general data set and observed count from O'Neill and Roberts [14].

We will begin by calculating the likelihood function of the general epidemic $L(q)$ in terms of $q = 1 - p$. Either parameter p or q would be equally good to make inference on, the choice is down to ease of interpretation.

$$\begin{aligned} L(q|n_0,n_1,n_2,n_{21}) &= A \cdot (q^2)^{n_0} \cdot (2q^2(1-q))^{n_1} \cdot (2q(1-q)^2)^{n_{21}} \cdot ((1-q)^2)^{n_2} \\ &= A \cdot 2^{n_1+n_{21}}\, q^{2n_0+2n_1+n_{21}}\, (1-q)^{n_1+2n_2}, \\ \text{where } A &= \frac{(n_0+n_1+n_2)!}{n_0!\,n_1!\,n_{21}!\,(n_2n_{21})!}. \end{aligned}$$

The log-likelihood of q, $\ell(q)$, is then calculated to make computation simpler. Letting A be the logarithm of the number of orderings of the realisations:

$$\ell(q|n_0,n_1,n_2,n_{21}) = A + (n_1+n_2)\log(2) + (2n_0+2n_1+n_{21})\log(q) + (n_1+2n_2)\log(1-q).$$

We then differentiate in order to find the greatest point on the likelihood curve:

$$\ell'(q|n_0,n_1,n_2,n_{21}) = \frac{2n_0+2n_1+n_{21}}{q} - \frac{n_1+2n_2}{1-q}.$$

When likelihood is maximised, at $\hat{q}$, this expression is equal to zero. Following some algebraic manipulation we have our candidate value

$$\hat{q} = 1 - \frac{n_1+2n_2}{2n_0+3n_1+2n_2+n_{21}}. \tag{2}$$

We check that this is a local maximum by computing the second derivative of the log-likelihood function.

$$\ell''(q) = \frac{-(2n_0+2n_1+n_{21})}{q^2} - \frac{n_1+2n_2}{(1-q)^2}\ .$$

At the candidate value the second derivative of the log-likelihood is

$$\ell''(\hat{q}) = \frac{-(2n_0+3n_1+2n_2+n_{21})}{(2n_0+2n_1+n_{21})(n_1+2n_2)}\ .$$

Since at least one of our counts will always be greater than zero when we have observations, this value is less than zero and so $\hat{q}$ is indeed a *maximum* likelihood estimate. Substituting the count values of O'Neill and Roberts [14] from Table 2 into Equation 2 we have that $\hat{q} = 0.234$. To construct a confidence interval for our maximum likelihood estimate, we calculate the observed information, $I_O(\hat{q})$, at the maximum likelihood estimate and use this as a good approximation to the expected information $I_E(\hat{q})$. This can be used to construct a 95% confidence interval for $\hat{q}$ based on the result that in the limit as sample size tends to infinity

$$\sqrt{I_E(\hat{q})}(\hat{q} - q) \sim \mathcal{N}(0,1)$$

and so

$$\hat{q} \sim \mathcal{N}(q, [I_E(\hat{q})]^{-1}).$$

In this case

$$I_O(\hat{q}) = -\ell''(\hat{q}) = \frac{2n_0 + 3n_1 + 2n_2 + n_{21}}{(2n_0 + 2n_1 + n_{21})(n_1 + 2n_2)} = 3661.06.$$

$$\sigma^2_{\hat{q}} = I_E(\hat{q})^{-1} \approx I_O(\hat{q})^{-1} = 0.0165,$$

and so a 95% confidence interval for $\hat{q}$ is $0.234 \pm 1.95 \times 0.0165 = (0.202, 0.267)$. This confidence interval can be interpreted as a statement that 95% of samples like the one taken in Providence will have an observed value of q that falls between 0.202 and 0.267, and that we have 95% confidence that the true parameter value lies within this range. In terms of the disease, we interpret this as the probability of avoiding sufficient contact with an individual as falling between 0.202 and 0.267. By the invariance property of the maximum likelihood estimator we have that $\hat{p} = 1 - \hat{q}$ and that the confidence interval bounds also conform to this transformation. That is, we have $\hat{p} = 0.766$ with a 95% confidence interval (0.733,0.798).

From these calculations it should be apparent how quickly direct likelihood maximisation methods would become unwieldy. In practice populations of interest are often of a greater size, for example if the population being considered were wards of a hospital, and usually only summary information available. This is what motivates our methodology in Section 3.1, which allows us to bypass these calculations.

3 Parameter Estimation

In Section 2.4 we saw how the calculations involved in maximum likelihood estimation for the parameters in the Reed-Frost model rapidly became complex as the size of the household being modelled increased. We now approach the problem form a different perspective. Rather than taking a frequentist view that the parameters in the Reed-Frost model have some fixed true value, we take the Bayesian view that the parameters have some unknown probability distribution.

Let the parameter(s) of interest be denoted by θ, where θ is a vector when more than one parameter is to be estimated. The likelihood of the parameter(s) given data $\underline{x}$ is $L(\theta;\underline{x})$. This is the probability (density) of the observed data given θ.

$$L(\theta;\underline{x}) = \pi(\underline{x}|\theta).$$

Under the Bayesian ideology, the parameter θ is a random variable, and so θ has a probability distribution. The *prior* distribution on θ is $\pi(\theta)$ and represents the knowledge of the distribution of θ before observing the data $\underline{x}$. This prior distribution can be either *informative*, providing information about which values of θ are most likely, or assume that all possible values of θ are equally likely. This prior distribution may also be *proper* or *improper*, where a prior distribution (whether or not it is informative) is proper when the distribution integrated over all possible values of θ is equal to one. The *posterior* distribution of θ, $\pi(\theta|\underline{x})$, represents the knowledge of the distribution of θ after we have observed the data.

The joint distribution of the data and parameter(s) can be expressed as either the product of the prior distribution and the likelihood of the observed data or as the product of the marginal distribution of the observed data and the posterior distribution.

$$\pi(\underline{x},\theta) = \pi(\theta)\cdot\pi(\underline{x}|\theta) = \pi(\underline{x})\cdot\pi(\theta|\underline{x})$$

Rearranging these, we find that the posterior distribution is equal to the product of the prior distribution and the likelihood of the parameter(s), divided by the marginal distribution of the data.

$$\pi(\theta|\underline{x}) = \frac{1}{\pi(\underline{x})}\cdot\pi(\theta)\cdot\pi(\underline{x}|\theta).$$

The marginal distribution of the observed data is given by the joint distribution of the data and parameters integrated over the parameter space T, the set of all possible parameter values.

$$\pi(\underline{x}) = \int_T \pi(\underline{x},\theta)\mathrm{d}\theta.$$

Issues can arise here because in various applications this integral is intractable or else impossible to calculate. Markov chain Monte Carlo estimation methods work around this restriction by noticing that for two possible parameter values θ and θ';

$$\frac{\pi(\theta'|\underline{x})}{\pi(\theta|\underline{x})} = \frac{\pi(\theta')\cdot\pi(\underline{x}|\theta')}{\pi(\theta)\cdot\pi(\underline{x}|\theta)}.$$

This states that the relative posterior probability of θ' and θ is equal to the ratio of the product of their respective priors and likelihoods. Using this relationship the optimal choice of parameters can be made, for example if a uniform proir is being used then

$$\pi(\theta') = \pi(\theta) = 1$$

and so the optimum choice of θ, which is approximated by Markov chain Monte Carlo estimation, gives the maximum likelihood estimator $\hat{\theta}$ while evading the intractable calculations explained earlier. The reason that the method works is that when the posterior probability of θ' is greater than that of θ,

$$\frac{\pi(\theta'|\underline{x})}{\pi(\theta|\underline{x})} > 1.$$

However, the disadvantage of this method is that it still relies on the likelihood functions of θ and θ' being tractable. Working around this reliance is what motivates our study of Approximate Bayesian Computation.

3.1 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) originated in the field of population genetics, the study of the frequency and distribution of alleles and how they vary between populations, with Tavaré et al. [16]. While the method began in mathematical biology, it is now used in a wide range of statistical applications as it has the benefit over Markov chain Monte Carlo methods of parameter estimation in that it does not require the likelihood function of the data to be tractable. The ABC algorithm produces posterior samples which are independent and identically distributed without requiring the use of the likelihood function itself.

ABC1 algorithm

1. Propose a value θ simulated from $\pi(\theta)$, the prior distribution on θ.
2. Simulate data $\underline{x}$ using an appropriate model and the proposed value of θ.
3. If $\underline{x} = \underline{x}^*$, the observed data, then accept θ as an observation from $\pi(\theta|\underline{x}^*)$.
4. Otherwise, θ is rejected.
5. Repeat until a sample of the required size has been obtained from $\pi(\theta|\underline{x}^*)$, or until a set number of iterations is exceeded.

We will begin by looking at why the algorithm works. Let $A \subseteq \mathbb{R}^d$ be some region, where d is the dimension of θ. The probability that a proposed value of θ is within A and is accepted is given by the law of total probability as:

$$\begin{aligned} P(\theta \in A \text{ and Accepted}) &= \int_{\theta \in A} P(\text{Accepted}|\theta)\pi(\theta)\ \mathrm{d}\theta \\ &= \int_{\theta \in A} P(\underline{x} = \underline{x}^*|\theta)\pi(\theta)\ \mathrm{d}\theta. \end{aligned}$$

Therefore, using Bayes' Theorem we have

$$\begin{aligned}
P(\theta \in A|\text{ Accepted}) &= \frac{P(\theta \in A \text{ and Accepted})}{P(\text{Accepted})} \\
&= \frac{\int_{\theta \in A} P(\underline{x} = \underline{x}^*|\theta)\pi(\theta)\ \mathrm{d}\theta}{\int_{\mathbb{R}^d} P(\underline{x} = \underline{x}^*|\vartheta)\pi(\vartheta)\ \mathrm{d}\vartheta} \\
&= \int_{\theta \in A} \frac{P(\underline{x} = \underline{x}^*|\theta)\pi(\theta)}{P(\underline{x} = \underline{x}^*)}\ \mathrm{d}\theta \\
&= P(\theta|\underline{x} = \underline{x}^*)
\end{aligned}$$

The evaluation of the integral in the denominator allows us to notice that what we have is in fact the posterior distribution of θ. We therefore have that distribution from which our accepted points are dawn is the posterior distribution of θ as desired.
It is worth noting that in this algorithm the acceptance probabilities are often very small. For continuous data the probability of an exact match occurring is zero, and so the acceptance rate is also zero. We will illustrate the low acceptance probability of the ABC algorithm with a discrete example; 7 observations $\underline{x}^*$ from the Poisson(λ) distribution, where the prior distribution of the parameter λ is $\pi(\lambda) \sim \text{Exp}(2)$.
The acceptance probability is given by

$$P(\text{Accepted}) = \int_{\mathbb{R}} P(\underline{x} = \underline{x}^*|\lambda)\pi(\lambda)\ \mathrm{d}\lambda.$$

The likelihood of λ is given by

$$P(\underline{x} = \underline{x}^*|\lambda) = \prod_{i=1}^{7} \left(\frac{\lambda^{x_i^*} \exp(-\lambda)}{x_i^*!} \right),$$

and the so combining this with the prior distribution of λ we have that

$$\begin{aligned}
P(\text{Accepted}) &= \int_0^\infty 2\exp(-2\lambda) \prod_{i=1}^{7} \left(\frac{\lambda^{x_i^*} \exp(-\lambda)}{x_i^*!} \right)\ \mathrm{d}\lambda \\
&= \frac{2}{\prod_{i-1}^{7} \left(x_i^*!\right)} \int_0^\infty \lambda^{\sum_{i=1}^{7} x_i^*}\ \exp(-9\lambda)\ \mathrm{d}\lambda \\
&= \frac{2}{\prod_{i-1}^{7} \left(x_i^*!\right)} \times \frac{\left\{\sum_{i=1}^{7} x_i^*\right\}!}{9^{\left\{\sum_{i=1}^{7} x_i^*\right\}+1}}.
\end{aligned}$$

This uses the result that for all $N \in \mathbb{N}$, $\int\limits_0^\infty y^N \exp(-\gamma y)\mathrm{d}y = \frac{N!}{\gamma^{N+1}}$. For an example data set from Poisson(3):

$$4, 5, 3, 2, 2, 1, 1$$

the acceptance probability is small enough that generation of a posterior sample would be infeasible under practical time constraints, as $P(\text{Accepted}) = 1.37 \times 10^{-7}$ meaning that just over one in ten million proposed values are accepted as a sample from the posterior distribution.

We can improve on the acceptance probability of the ABC1 algorithm using the concept of a *sufficient summary statistic*. In the previous example, because we have assumed the marginal distribution $\pi(\underline{x})$ and that the likelihood is tractable, we can calculate that the true posterior distribution is $\pi(\lambda|\underline{x}^*) \sim \text{Gamma}(\sum_{i=1}^{7} x_i^* + 1, 7 + 1)$. We notice here that the posterior distribution is dependent on the observations only through their sum. This means that the sum, or alternatively the mean, of our observations contains all of the information required when estimating $\pi(\lambda|\underline{x}^*)$ and therefore are sufficient summary statistics.

As another example of the sufficient summary statistics of a distribution, consider a sample of Normally distributed data $\underline{x}^*$ of size n from a $N(\mu, \sigma^2)$ distribution. We can estimate the mean and variance only knowing $\sum_{i=1}^{n} x_i^*$ and $\sum_{i=1}^{n} (x_i^*)^2$, therefore together together they form a sufficient set of summary statistics. Adapting step 3 of the ABC1 algorithm, we achieve the more efficient ABC2 algorithm when a sufficient (set of) summary statistic(s) $S(\underline{x})$ exist.

ABC2 algorithm

1. Propose a value θ simulated from $\pi(\theta)$, the prior distribution on θ.
2. Simulate data $\underline{x}$ using an appropriate model and the proposed value of θ.
3. If $S(\underline{x}) = S(\underline{x}^*)$, the observed data, then accept θ as an observation from $\pi(\theta|\underline{x}^*)$.
4. Otherwise, θ is rejected.
5. Repeat until a sample of the required size has been obtained from $\pi(\theta|\underline{x}^*)$, or until a set number of iterations is exceeded.

Applying ABC2 to the earlier Poisson example, since each value is an independent sample from a Poisson(λ) distribution, the sum of these variables $Y \sim Poisson(7\lambda)$. Hence, taking m to be the observed sum:

$$
\begin{aligned}
P(\text{Accepted}) &= \int_0^\infty \pi(\lambda)P(Y=m)\mathrm{d}\lambda \\
&= \int_0^\infty 2\exp(-2\lambda)\frac{(7\lambda)^m\exp(-7\lambda)}{m!}\,\mathrm{d}\lambda \\
&= \frac{2\times 7^m}{m!}\int_0^\infty \lambda^m \exp(-9\lambda)\mathrm{d}\lambda \\
&= \frac{2\times 7^m}{m!}\frac{m!}{9^{m+1}} \\
&= \frac{2\times 7^m}{9^{m+1}}.
\end{aligned}
$$

In the case of our data $m = 12$, so we have $P(\text{Accept}) = 0.00241$. The acceptance rate is now more than ten thousand times greater by using the ABC2 algorithm rather than ABC1. While this solves the issues for discrete data, we still have the issue of continuous data having zero probability of an exact match.

So far, we have been taking samples from the exact posterior distribution of θ, there has been no approximation. In order to extend the technique to continuous data, or discrete data where acceptance probabilities are very low, we introduce a distance metric $d(\cdot,\cdot)$. We then define the ε-approximate posterior distribution to be

$$\pi_\varepsilon(\theta|\underline{x}^*) \propto \pi(\theta)\int_{\underline{x}\in d(S(\underline{x}),S(\underline{x}^*))\leq\varepsilon} \pi(S(\underline{x})|\theta)\mathrm{d}\theta.$$

We can sample from the ε approximate posterior distribution using the ABC3 algorithm.

ABC3 Algorithm

1. Propose a value θ simulated from $\pi(\theta)$, the prior on θ.
2. Simulate data $\underline{x}$ using an appropriate model and the proposed value of θ.
3. If $d(S(\underline{x}),S(\underline{x}^*)) \leq \varepsilon$, accept θ as an observation from $\pi_\varepsilon(\theta|\underline{x}^*)$.
4. Otherwise, θ is rejected.
5. Repeat until a sample of the required size has been obtained from $\pi_\varepsilon(\theta|\underline{x}^*)$ or a fixed number of iterations is reached.

When ε is equal to zero the ABC3 algorithm reduces to the ABC2 algorithm - proposed values are accepted only when the summary statistics match exactly, giving a sample from the exact posterior distribution. When ε is greater than zero the proposed value is accepted whenever the summary statistics are within ε of one another, giving a sample from the ε-approximate posterior distribution.

The choice of ε requires compromise; large values of ε increase the acceptance probability (the proportion of points accepted into the posterior sample) as there is a greater window into which these points could fall. A large value of ε will also increase the degree of approximation of the posterior distribution, as points further from the truth are allowed to contribute to the posterior sample. Conversely, choosing a small value of ε reduces the efficiency of the algorithm as we approach the ABC2 acceptance probability. In addition to this the Monte-Carlo error, the uncertainty associated with the results of the simulation, is likely to be large when sampling for a fixed number of iterations. This is because the acceptance probability is small, meaning that the estimate of the posterior distribution is based on only a small number of points, and so will have high variance.

The use of the ABC algorithm is not restricted to models which can be completely described by one or a vector of summary statistics. The model can be extended by adding another layer of approximation, this time to the sufficient summary statistic $S(\cdot)$. Rather than using a summary statistic which is sufficient to completely describe the data, we can use a summary statistic which describes the key features of the data well but does not completely describe it, $T(\cdot)$. Let the altered ABC3 algorithm be equivalent to the ABC3 algorithm with acceptance condition:

ABC3 Algorithm Alteration

3. If $d(T(\underline{x}),T(\underline{x}^*)) \leq \varepsilon$, accept θ as an observation from $\pi_\varepsilon(\theta|\underline{x}^*)$.

3.2 Applying Approximate Bayesian Computation 1

To begin, we will look at the simplest model of a single household with constant infectious period and one initial infective. In order to use ABC, simulations of a Reed-Frost type outbreak are required. The code below simulates the chain binomial outbreak in a single population of size $N = S_0 + I_0$, where the probability of sufficient contact is p.

```
Reed.Frost<- function(p,I_0,S_0){
 #0: Set up.
    #0.1: parameters derivable from initial conditions.
        n<- I_0 + S_0
        q<- 1-p
    #0.2:vectors recording population sizes over time.
        I<- I_0
        S<- S_0
    #0.3: create time counter.
        t<- 0
```

```
  #1: looping to simulate discrete time progression.
    #while there are infectives now
    while(tail(I,1)>0){
    #1.1: generate the number of infectives next and add to I
      I_t<-tail(I,1)
      S_t<-tail(S,1)
      I_n<- rbinom(n = 1,size = S_t,prob = 1-q^(I_t))
      I<- c(I,I_n)
    #1.2: calculate the number of susceptibles next and add to S
      S_n<-S_t - I_n
      S<- c(S,S_n)
    #1.3: progress time
      t<- t+1
    }
  #2: Record Paths
    paths<- list(I=I,S=S,Dur=t,Size=sum(I))
    paths
}
```

We can then use this function in our coding of the ABC2 algorithm for a household epidemic. Since we have no prior knowledge of the distribution of p, and the value of p lies between zero and one, an appropriate proposal distribution is Uniform [0,1].

```
abc.rf<- function(N,I_0,run,Sast){
  output<- c()
  for(i in 1:run){
    p<- runif(1)           #Generate p from Unif(0,1)
    S<-Reed.Frost(p,I_0,N) #Generate data using p
    dist<- S\$Size-Sast    #Calculate distance
    if(dist==0){           #if equal accept and record
      output<-c(output,p)
    }
  }
  output
}
```

To begin the example, we create an example 'observed' data set, to demonstrate the method on a household of size 5 with one initial infective and $p = {}^{1}/_{\pi}$, where π is the 3.1415....

```
set.seed(54321)
xast<-Reed.Frost(1/pi,1,4)
```

The size of the observed outbreak, our sufficient summary statistic is $S(\underline{x}^{*}) = 2$. Performing the ABC we collect a sample from the posterior distribution, with acceptance probability of around 4%, and use this to calculate the empirical posterior mean and variance. Plotting a histogram of our accepted values of p in Figure 3 gives an idea of the shape of the posterior distribution.

```
post.samples<- abc.rf(5,10000,3)
length(post.samples)/10000  #acceptance rate
hist(post.samples,probability = T)
```

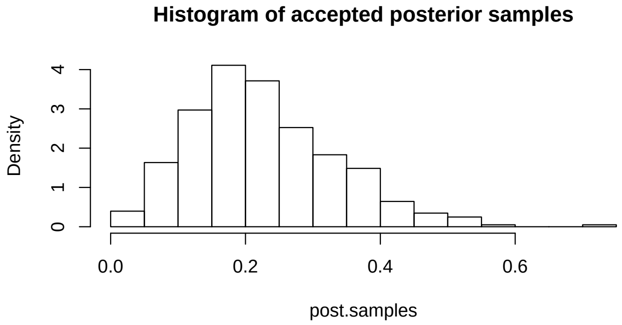


Figure 3: Histogram of points from the exact posterior distribution of p, sampled using ABC.

3.3 Applying Approximate Bayesian Computation 2

We now consider a household where there is a chance of infection both from within and outside of the household, and where the initial number of susceptibles may vary. Once again we must be able to simulate an outbreak in such a household, which is done using the following code.

```
RF.Com<- function(n,ph,pc){
  #0.1: Initial infectives and susceptibles
  I<- rbinom(n = 1,size = n,prob = pc)
  S<- n-I
  #0.2: qh=P(not infected by 1 household infective)
  qh<- 1-ph

  #1: looping to simulate discrete time progression.
  #while there are infectives now
  while(tail(I,1)>0){
    #1.1: Retrieve number of infectives and susceptibles.
    I_t<-tail(I,1)
    S_t<-tail(S,1)
    #1.2: calculate the number of susceptibles next and add to S
    I_n<- rbinom(n = 1,size = S_t,prob = 1- qh^(I_t))
    I<- c(I,I_n)
    S_n<-S_t - I_n
    S<- c(S,S_n)
  }
  #2: Return the size of epidemic
  Size<- sum(I)
  return(Size)
}
```

This model requires two parameters to be estimated, p_c and p_h. The algorithm functions in the same way, but rather than simulating from one proposal distribution we simulate from two. Since we have no prior knowledge of p_c or p_h, both are simulated from the Uniform [0,1] distribution.

```
abc.rf2<- function(N,run,Sast){
  output<- c()
  for(i in 1:run){
    pc<- runif(1)       #Generate p from Unif(0,1)
    ph<- runif(1)
    S<-RF.Com(N,pc,ph) #Generate data using pc and ph
    dist<- S-Sast       #Calculate distance
    if(dist==0){        #If equal accept and record
      output<-rbind(output,c(pc,ph))
    }
  }
  output
}
```

As an example, consider a household of five individuals with total outbreak size of two. Using ABC a posterior sample was collected, as shown in Figure 4. We can interpret this as the chance of infection from the community, p_c, having marginal mean 0.161 and 95% credible interval (0.006,0.4849). The marginal probability distribution of infection from within the household p_h is both more symmetric and more disperse, having mean 0.383 and 95% credible interval (0.074,0.737). When running the algorithm, $100,000$ iterations were used with 3536 proposed values being accepted, giving an acceptance probability of 3.536%.

3.4 Efficient use of parameter space

From the plot of the bivariate samples in Figure 4, we can see that there are few points accepted with both high p_c and high p_h. This means that by proposing values uniformly over the unit square we are not making the most efficient use of our parameter space possible. The posterior distribution in this case is fairly diffuse, and so sampling uniformly is no major disadvantage; the achieved acceptance probability of 3.536% for the above example is perfectly fine. However, it is not difficult to see that in a case where the distribution of parameter values is highly concentrated that this would cause a problem by reducing the acceptance probability of the algorithm and therefore decreasing efficiency and increasing computational time. In addition to this, as more data

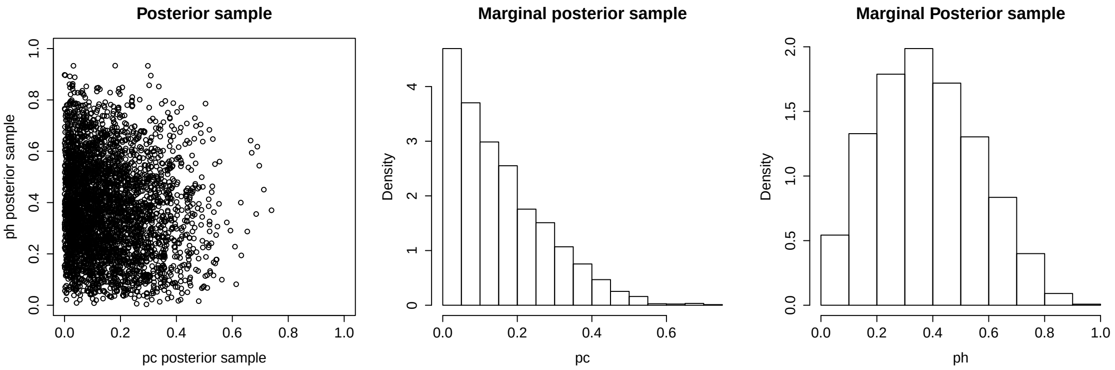


Figure 4: From left to right: Accepted bivariate posterior sample, Histogram of the marginal posterior sample of p_c, histogram of the marginal posterior sample of p_h.

is collected the posterior sample becomes more concentrated in the small part of the parameter space around the maximum of the joint posterior distribution of p_c and p_h, and so by not making use of this additional information we are also making inefficient use of the parameter space.

In order to correct this, we would like to propose values from a distribution closer to the posterior distribution. The proposal distribution being closer to the posterior distribution would mean that simulated samples would be more similar to the observed sample and therefore more likely to be accepted. This can be done by adding an extra stage into the algorithm, to explore the parameter space before running the algorithm in full and to ‘create’ prior knowledge.

The first stage of the algorithm will be as before, using an appropriate choice of ABC for the data and sampling from an sampling distribution that weights all possible values equally. After a low number of iterations, 1000 say, the density of accepted samples is used to define a new proposal density function, taking advantage of the information gained in the first set of iterations. We then begin again in the usual way, but proposing from the informative proposal distribution.

This method can make the ABC algorithm much more efficient, especially when the probability distribution is highly concentrated in a small area [5]. A correction to the accepted values is needed, to compensate for proposing from a distribution other than the prior distribution. ABC works on the principle that

$$\pi(\theta|\underline{x}^*) \propto \pi(\underline{x}^*|\theta)\pi(\theta)$$

and now rather than sampling from the prior distribution $\pi(\theta)$ we are sampling from some other distribution $f(\theta)$. We see that this can be corrected by noticing that:

$$\pi(\theta|\underline{x}^*) \propto \pi(\underline{x}^*|\theta)\pi(\theta)\frac{f(\theta)}{f(\theta)} = \left\{\pi(\underline{x}^*|\theta)\frac{\pi(\theta)}{f(\theta)}\right\} f(\theta).$$

Therefore, our accepted values of θ using the informative prior can be corrected to give true posterior estimates by multiplication by $\frac{\pi(\theta)}{f(\theta)}$. Since we choose the prior distribution, we can choose this in order to make calculation of $\pi(\theta)$ simple. The distribution of $f(\theta)$ can also be selected to be pleasant to work with, as it need not exactly match the observed distribution of points from the exploratory ABC. This means that we can select it to be a function where calculation of $f(\theta)$ is possible.

In the two example cases given above, the parameters being estimated were probabilities and so took values between zero and one. In this case a beta distribution would be a good choice of informative proposal distribution as it covers the parameter space, has a tractable density function and is flexible enough to approximate a range of unimodal distributions. A disadvantage of this choice of distribution would be that it would not approximate well bimodal distributions with modes at either end of the parameter space and would concentrate proposals in the region of lowest acceptance probability. It is therefore important to consider the distribution as well as the summary statistics of the initial running of the algorithm, ensuring that the proposal distribution being used is appropriate.

4 Working with Posterior Samples

Once we have obtained a sample from the exact or approximate posterior distribution of our parameter(s) of interest using Approximate Bayesian Computation, we can move our methods beyond plotting these points or aggregating them into histograms. We will look in this section at how to improve the quality of our estimation and how to avoid aggregation of the data by separating the data into ‘bins’ when producing histograms.

4.1 Local Linear Regression

Let us suppose we have an approximate posterior sample of a single parameter θ which was generated using a single sufficient summary statistic $S(\cdot)$ from an observed data set $\underline{x}$. By the nature of the ABC3 algorithm, not all of the accepted posterior samples will have a summary statistic $S(\underline{x})$ that matches exactly the observed sample summary statistic $S(\underline{x}^*)$. Previously this discrepancy has been acknowledged, but we now consider the method of Beaumont et al. [3] to correct for a discrepancy $|S(\underline{x}) - S(\underline{x}^*)| = \varepsilon > 0$. The assumption that both θ and $S(\cdot)$ are scalars is not necessary here but is taken for simplicity of notation and ease of understanding.

The technique of Beaumont et al. [3] looks at the relationship between the proposed parameter value and the simulated summary statistic; fitting a linear regression model at localised level to describe this relationship. This linear model can then be used to correct the accepted parameter estimates that overestimate or underestimate the summary statistic accordingly. In order to define the ‘local’ region we introduce the idea of a kernel. A kernel is a weighting function $K(t)$ with mean zero and finite variance; formally this is $K(t)$ such that:

$$\int K(t)\,\mathrm{d}t = 1, \quad \int tK(t)\,\mathrm{d}t = 0, \quad \int t^2K(t)\,\mathrm{d}t < \infty \qquad [18].$$

Therefore, any probability density function which is symmetric about zero is a kernel, but kernels are not restricted to such functions. Two example kernels will be introduced, the Gaussian and Epanechnikov kernels. The Gaussian kernel is given by

$$K_G(t) = \frac{1}{\sqrt{2\pi}}\exp\{\frac{-t^2}{2}\} \quad \text{for } t \in \mathbb{R},$$

and is shown in Figure 5, the kernel weights most heavily the values close to zero and assigns a positive weighting to all values which decreases as t increases. This is not well suited to our purposes because we wish to consider only local values because the Gaussian kernel assigns positive weight to all values. This is a disadvantage because it is computationally inefficient to assign near-zero weights to most values in the data set, as they contribute very little. This motivates the introduction of the Epanechnikov kernel:

$$K_\delta(t) = \begin{cases} \frac{3}{4\delta}(1-(\frac{t}{\delta})^2), & t \le \delta \\ 0, & t > \delta. \end{cases}$$

A graph of the kernel for $\delta = \sqrt{2}$ is shown in Figure 5. Once again we see that the weighting is assigned most highly around zero and then decreases with t, but has a finite support, allocating positive weighting only over the interval $[-\delta, \delta]$.

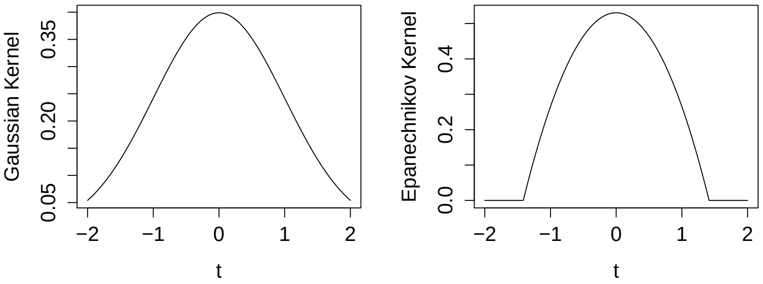


Figure 5: Gaussian kernel (left) and Epanechnikov kernel (right) with $\delta = \sqrt{2}$ so that both have variance of 1.

Using the Epanechnikov kernel we can weight our estimate of (α, β), the regression coefficients in our linear regression, most heavily in the region of the observed summary statistic and only consider posterior samples with summary statistics in the region of the observed summary statistic. These values of α and β are the ones that minimise the weighted residual sum of squares:

$$\sum_{i=1}^{n}\{\theta_i - \alpha - (S_i - S^*)^T\beta\}^2 K_\varepsilon(d(S_i, S^*)). \tag{3}$$

Here we have; $S_i = S(\underline{x})_i$ the summary statistic associated with the i^{th} accepted sample from the ε-approximate posterior distribution, $K_\varepsilon(\cdot)$ is the Epanechnikov kernel with $\delta = \varepsilon$, and $d(\cdot)$ is a scalar distance metric. Had we instead used a uniform weighting by taking $K(t) = 1$, this formula would reduce to standard linear regression. From a computational perspective, we are only interested in values S_i close to the observed value S^* and the local linear regression therefore focuses attention about S^*.
In the following derivation of the the least squares estimates $(\hat{\alpha}, \hat{\beta})$ we generalise to the case of p summary statistics, $S_1, \ldots, S_p$, and z parameters, $\theta_1, \ldots, \theta_z$.
Let W be an $n \times n$ diagonal matrix of weights, with diagonal entries corresponding to the weight of the i^{th} point:

$$W_{ij} = \begin{cases} K_\varepsilon(d(S_i, S^*)) & \text{for } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Also take

$$X = \begin{pmatrix} 1 & S_{11} - S_1^* & \ldots & S_{1p} - S_p^* \\ \vdots & \vdots & \ddots & \vdots \\ 1 & S_{n1} - S_1^* & \ldots & S_{np} - S_p^* \end{pmatrix} \text{ and } \Theta = \begin{pmatrix} \theta_{11} & \cdots & \theta_{z1} \\ \vdots & \ddots & \vdots \\ \theta_{1n} & \ldots & \theta_{zn} \end{pmatrix},$$

so that X is the design matrix of the regression and Θ is a matrix with rows corresponding to the accepted points from the z-dimensional joint posterior distribution of $\theta_1, \ldots, \theta_z$. We will show that the least squares estimator of α and β is given by:

$$\begin{pmatrix} \hat{\alpha}_1 & \ldots & \hat{\alpha}_z \\ \hat{\beta}_{11} & \ldots & \hat{\beta}_{1z} \\ \vdots & \ddots & \vdots \\ \hat{\beta}_{p1} & \ldots & \hat{\beta}_{pz} \end{pmatrix} = (\hat{\alpha}, \hat{\beta}) = \hat{\gamma} = (X^T W X)^{-1} X^T W \Theta.$$

Derivation

Using the notation defined above, the residual sum of squares from Expression 3 can be rewritten as:

$$\begin{aligned} &(\Theta - X\gamma)^T W(\Theta - X\gamma) \\ =&(\Theta^T - \gamma^T X^T)(W\Theta - WX\gamma) \\ =&\Theta^T W\Theta - \Theta^T WX\gamma - \gamma^T X^T W\Theta + \gamma^T X^T WX\gamma. \end{aligned}$$

We wish to find the value of γ that minimises this value, and so we differentiate with respect to γ.

$$\begin{aligned}
&\frac{\partial}{\partial\gamma}\{\Theta^T W\Theta - \Theta^T WX\gamma - \gamma^T X^T W\Theta + \gamma^T X^T WX\gamma\} \\
&\quad = 0 - \Theta^T WX - \Theta^T (X^T W)^T + 2(X^T WX)\gamma \\
&\quad = 2(X^T WX)\gamma - 2\Theta^T WX.
\end{aligned}$$

At $\hat{\gamma}$ this derivative is equal to zero, meaning that

$$X^T WX\hat{\gamma} = \Theta^T WX,$$

and so when $X^T WX$ is non-singular we have shown that the least squares estimator is

$$\hat{\gamma} = (\hat{\alpha}, \hat{\beta}) = (X^T WX)^{-1}\Theta^T W^T X = (X^T WX)^{-1} X^T W\Theta.$$

Now that we have estimates of α and β, we can create a vector of accepted parameter values which have been corrected for having summary statistics that do not match the observed summary statistic.

$$\Theta^*_{i\cdot} = \Theta_{i\cdot} - (S - S^*)^T \hat{\beta} \quad \text{for } i = 1, \ldots, z$$

where S is the vector of simulated summary statistics and $\Theta_{i\cdot}$ is the column of Θ corresponding to parameter i. The posterior density of θ_i can then be estimated at any point ϑ of the parameter space of θ_i, using a kernel estimate (Section 4.2) formed with the adjusted points Θ^*_i;

$$\hat{\pi}(\vartheta_i|S^*) = \frac{\sum_{j=1}^{n} K_\delta(\vartheta_i - \Theta^*_{ij}) K_\varepsilon(d(S_j, S^*))}{\sum_{j=1}^{n} K_\varepsilon(d(S_j, S^*))}.$$

In addition,

$$\hat{\alpha} = \frac{\sum_{j=1}^{n} \Theta^*_{ij} K_\varepsilon(d(S_j, S^*))}{\sum_{j=1}^{n} K_\varepsilon(d(S_j, S^*))},$$

and is therefore an estimator of the posterior mean of θ_i, $E[\pi(\theta_i|\underline{(x)}^*]$.

Local Linear Regression Example

To illustrate the local linear regression technique in practice, we will consider an example with only one summary statistic and one parameter to estimate. Suppose we have 100 days of count data detailing the number of computers which fail each day, which follows a Poisson(λ) distribution, and that the sum of these is 528. We saw in Section 3.1 that the sum of observed counts is a sufficient statistic for a Poisson random variable. Suppose that our prior distribution of λ is given by $\pi(\lambda) \sim Exp(0.2)$; we can then use ABC to obtain a sample from an ε-approximate posterior distribution.

```
Poisson.ABC<-function(runs,Sast){
  output<-c()                                   #Create Storage
  for(i in 1:runs){
    lambda<- rexp(1,0.2)                        #Simulate parameter from
   proposal distribution
    x<- rpois(100,lambda)                       #Simulate data using simulated
     parameter
    S<-sum(x)                                   #Calculate summary statistic
    d<-abs(S-Sast)                              #Calculate difference to
   observed statistic
    output<-rbind(output,c(lambda,S,d))         #Record
  }
  output                                        #Return values
}
sample<-Poisson.ABC(runs = 100000,528)
```

From the algorithm 57 of the simulations were exact matches for the sum, with $2811 (2.81\%)$ within 20 of the observed total. We take $\varepsilon = 20$, achieving the sample shown in Figure 6 which gives the estimate of the mean and standard deviation of the posterior distribution of λ as 5.273 and 0.252 respectively.
We can use the code below to apply local linear regression to the accepted sample. Since points with S outside of the interval $[508, 548]$ are given zero weight by the kernel, we decrease the required computational time by considering only points within this interval for the local linear regression.

```
sample2<- sample[sample[,3]<=20,]                #Matrix of accepted
   samples
sample2<- sample2[order(sample2[,3],decreasing = FALSE),]

Sast<- 528                                      #Observed summary
   statistic
m<- length(sample2[,1])                         #Number of accepted
   samples
                                                #(in this case 1055)

epikern<- function(t,delta){                    #Epanechnikov Kernel
```

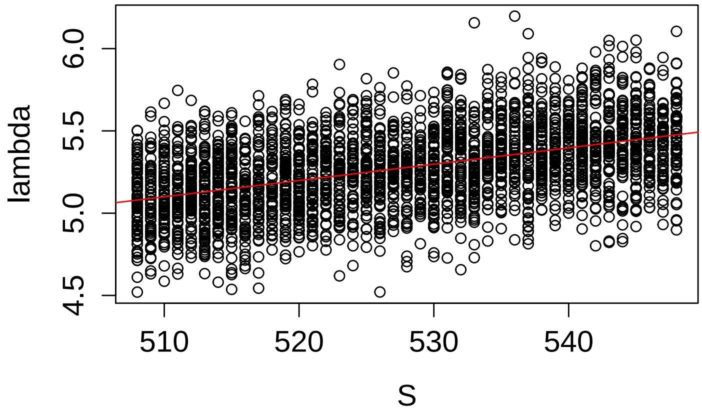


Figure 6: 2811 accepted samples from the ($\epsilon = 20$) approximate posterior distribution of mean computer failures, λ. Ordinary least squares line shown in red.

```
  y=0
  if(t<delta){
    y<- (1/delta)*(1-(t/delta)^2)
  }
  y
}

X<- matrix(1,ncol=2,nrow=m)                         #Construct matrix of
    differences
X[,2]<-sample2[,2]-Sast
Theta<-sample2[,1]                                  #Construct matrix of
    simulated
                                                    #parameters

W<- matrix(0,nrow=m,ncol=m)                         #Matrix of weights
for(i in 1:m){
  W[i,i]<- epikern(t=sample2[i,3],delta=20)
}

solve(t(X)%*%W%*%X)%*%t(X)%*%W%*%Theta              #Estimates of alpha and
    beta
```

This gives $(\hat{\alpha}, \hat{\beta}) = (5.281, 0.010)$, which is in line with what we would expect from a Poisson variable X. The estimate of alpha should be close to $E[100X] = 100E[X] = 100\lambda$ and beta should be close to $1/100$ because of the linear relationship between expectation of the sum and λ; $E[S|\lambda] = 100\lambda$. The standard deviation of the corrected posterior samples can be calculated as below, giving a value of 0.223.

```
corrected.samp<- c()
for (i in 1:m){
  corrected.samp[i]<- theta[i]-X[i,2]*gamma[2]
}
sd(corrected.samp)
```

The true posterior distribution is Gamma(528+1,100+0.2), with true mean 5.2794 and standard deviation 0.230. The standard deviation has been brought into line with the true value by using the local linear regression method, this demonstrates the benefits of utilising the local linear regression technique.

4.2 Kernel Density Estimation

A kernel was defined in Section 4.1 as a weighting function with expectation zero and finite variance. We will see in this section how kernels may be used to extend the idea of using a histogram to estimate density functions non-parametrically, using the methods described by Silverman [15]. Kernel density estimation is a non-parametric method of kernel density estimation, the density is not assumed to take any particular distributional form. This makes the method particularly well suited in application to cases where there is little or no prior knowledge about the distribution of the parameters being estimated.

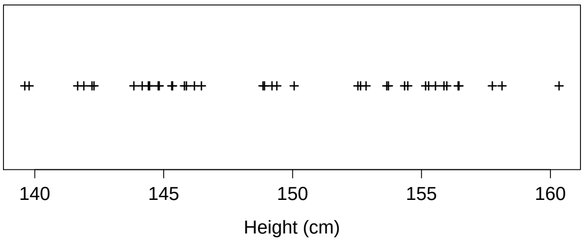


Figure 7: Stripchart of height data.

We will use a univariate example to illustrate the approach, though the techniques can be extended to multivariate densities [15]. Suppose we have the heights of 40 students in a class. We can display this data in a strip chart as in Figure 7 to visualise where the points fall, and estimate the density using one of the following methods.

Histograms

In the construction of a histogram, the plotted interval is divided into *bins* defined by an origin x_0 and a *binwidth* $2h$. The bins are then given, for positive and negative integers m by

$$[x_0 + (m-1)h, x_0 + (m+1)h).$$

Here the interval is closed on the left and open on the right so that each point may fall into at most one bin. The histogram estimating the density of a function $f(x)$ by data X_i, $i = 1, \dots, n$, is then defined as

$$\hat{f}(x) = \frac{1}{2nh}(\text{number of } X_i \text{ in the same bin as } x).$$

In general the binwidths need not be equal, though interpretation is simpler when this is the case. For non-constant binwidths, the histogram is defined by

$$\hat{f}(x) = \frac{1}{n} \times \frac{(\text{number of } X_i \text{ in the same bin as } x)}{(\text{width of bin containing } x)}.$$

Since the histogram is dependent on the choice of origin and the binwidth, the choice of these can make a great difference to how the diagram is interpreted. It is the binwidth which most impacts the diagram, as seen in Figure 8 which shows histograms of the height data with decreasing binwidth. In the left plot, we see that important features of the data, the bimodality for example, are not visible because the bins are too wide. In the right hand plot the features are again obscured because few points fall within each bin. The central plot shows a good degree of smoothing, showing the important features of the data without being too rough.

An issue with histograms is in extending to multivariate data, as presentation is unclear for even bivariate and trivariate data because it is not possible to construct contour plots from them. In addition to this, the problems of origin and binwidth selection are exacerbated by both the choice of coordinate orientation and the dependence of the binwidths, and so we explore further methods of density estimation.

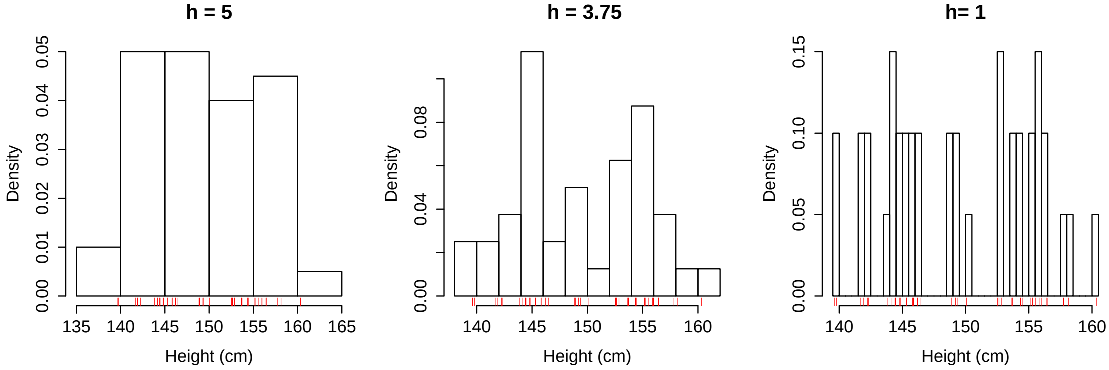


Figure 8: Histograms displaying, from left to right; over-smoothing, good smoothing and under-smoothing of the height data.

Naive Density Estimate

Rather than defining arbitrary centres for bins, we notice that for a random variable X with density f,

$$f(x) = \lim_{h \to 0} \left\{ \frac{1}{2h} P(x-h < X < x+h) \right\}.$$

We can estimate this density function by

$$\hat{f}(x) = \frac{1}{2hn}(\text{number of } X_i \text{ falling in } (x-h, x+h))$$

which is equivalent to 'box' of width $2h$ over each data point and then summing these boxes to obtain the density estimate, meaning that the density estimate is no longer dependent on the choice of x_0. This box analogy is made more clear by introducing the uniform kernel

$$w(x) = \begin{cases} \frac{1}{2} & \text{if } |x| < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The uniform kernel allows $\hat{f}(x)$ to be rewritten as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - X_i}{h}\right),$$

which more clearly shows the summation of boxes centered at each data point, of width $2h$ and height $(2nh)^{-1}$. Figure 9 shows this naive density estimate for the height data, using a uniform kernel. Once again, there is an issue with the selection of the correct width for the boxes to achieve the correct level of smoothing and the density estimate is still step-wise due to the nature of the kernel used.

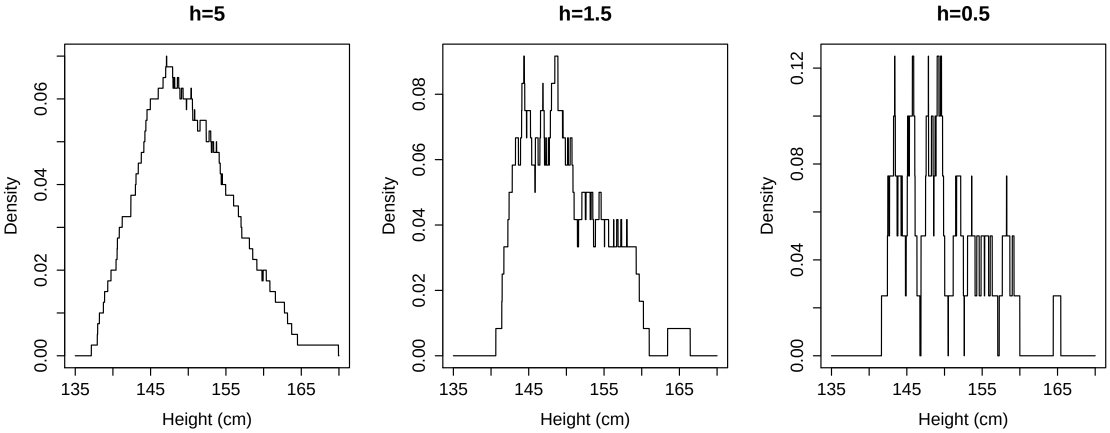


Figure 9: Naive density estimates for height data, using decreasing box widths h, to show over-smoothing and under-smoothing.

Kernel Density Estimate

The step-wise nature of the naive estimate can be corrected by using a continuous kernel, such as the Gaussian kernel described in Section 4.1 . In this case the density estimate for a general Kernel K with window width, or *bandwidth*, h is

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

Figure 10 shows the Gaussian kernels used in red. These are centered over 5 of the student heights from the heights data, and correspond to the boxes placed in the simple estimate. The density estimate is shown in black and is formed by summing the kernels. In the plot, a subset of the data was used for clarity. Figure 11 shows the density estimate constructed using this method on the full data set.

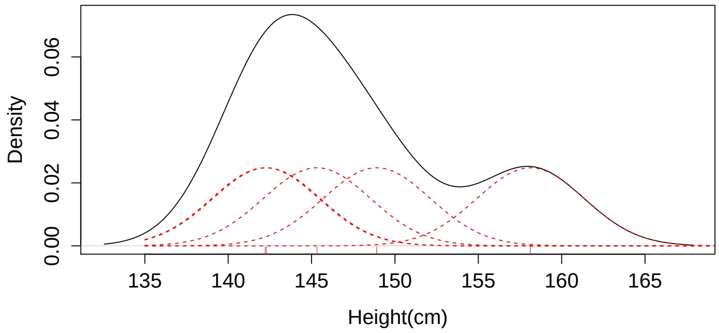


Figure 10: Kernel density estimate of heights using a sample of 5 students from the height data, with individual Gaussian kernels shown in red. $(h = 2.24)$

Selection of Bandwidth

We now have a continuous estimate of the density, however Figure 12 shows the dependence of this estimate on the selected bandwidth. If too large a bandwidth is selected, the features of the density function are smoothed away and if too small a bandwidth is selected the features are obscured by noise in the data. The selection of optimum bandwidth is described by Silverman [15]. The ideal value of h minimises the mean integrated square error

$$\text{MISE}(\hat{f}) = E\left[\int \{\hat{f}(x) - f(x)\}^2 \mathrm{d}x\right],$$

which is the expected area between the predicted and true density curves. The function we are estimating, f, is unknown and so unless we assume a particular distributional

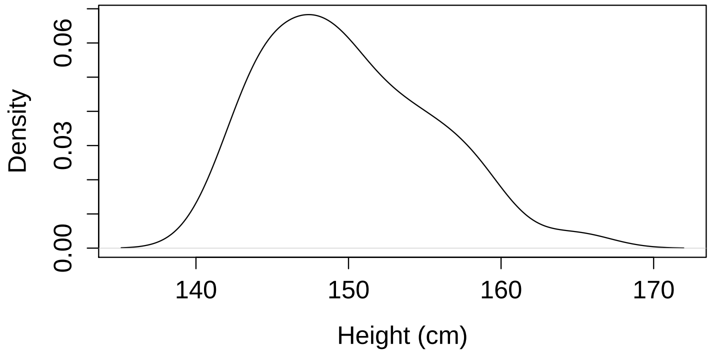


Figure 11: Kernel density estimate of heights using the full data set and Gaussian kernels. $(h = 2.06)$

form for f it is not possible to calculate $\int f''(x)^2 \mathrm{d}x$, this is an issue we will confront shortly. The MISE can be decomposed into the integrated squared bias and the integrated variance:

$$\mathrm{MISE}(\hat{f}) = \int \{E[\hat{f}(x)] - f(x)\}^2 \mathrm{d}x + \int \mathrm{var}\hat{f}(x) \mathrm{d}x.$$

Except in special cases these expressions are intractable, and so we approximate the integrated squared bias and integrated variance to find the value of h that minimises the approximate MISE, which is given by

$$\frac{1}{4} h^4 k_2^2 \int f''(x)^2 \mathrm{d}x \;+\; n^{-1} h^{-1} \int K(t)^2 \mathrm{d}t,$$

where $k_2 = \int t^2 K(t) \mathrm{d}t$. The value of h which minimises this can be shown to be h_{opt}, where

$$h_{\mathrm{opt}} = k_2^{-2/5} \left\{ \int K(t) \mathrm{d}t \right\}^{1/5} \left\{ \int f''(x)^2 \mathrm{d}x \right\}^{-1/5} n^{-1/5}.$$

We now consider methods for dealing with the fact that f is not known to us. If we assume a Gaussian distribution with variance σ, then $h_{\mathrm{opt}} = 1.06 \sigma n^{-1/5}$. A better result can be achieved using a robust measure of spread, such as the interquartile range R. Using R as a measure of spread gives $h_{\mathrm{opt}} = 0.79 R n^{-1/5}$ and lowers the MISE for long-tailed and skewed distributions. However, if the true distribution is multimodal then using this bandwidth increases the MISE because it leads to a greater level of smoothing. In this trade off oversmoothing is usually preferable so that we can be certain that any features are of the data and not the noise. In order to make best use of both of these situations it is advised that the value used should be

$$h_{\mathrm{opt}} = 0.9 A n^{-1/5} \quad \text{where} \quad A = \min\left\{ \sigma, \frac{R}{1.34} \right\}.$$

Using this choice of smoothing parameter gives an adequate choice of bandwidth for many applications, coming within 10% the optimum MISE for all t-distributions considered by Silverman [15]. This value is also a good starting value for further fine-tuning in cases where the density function is very heavy tailed or multimodal.

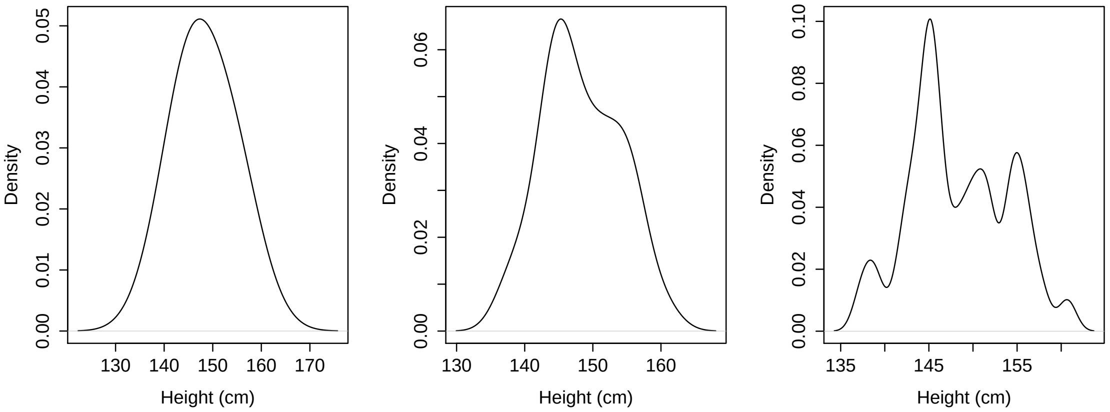


Figure 12: Kernel density estimate of heights using a full data set, bandwidths (left to right) h= 5, 2.445, 1 showing; over-smoothing, optimum choice of bandwidth and under-smoothing.

5 Extending the Reed-Frost Model

In Section 2.4 we considered the maximum likelihood estimation of parameters in the Reed-Frost model and saw that this becomes intractable as the number of individuals within the population increases. Now that Section 3 has introduced the computational methods associated with ABC which help us to navigate these issues, we consider two further extensions to the Reed-Frost model by looking at households that are no longer isolated and are comprised of individuals with varying infectious periods.

5.1 Introducing Community Structure

When the Reed-Frost model was defined in Section 2.2 the third assumption required that the population, in this case the household, to be closed, homogeneous and well-mixed. We now relax this assumption and consider a community of households, varying in size, which are independent of one another but all have a constant probability of infection for each individual from the surrounding community. The modified assumptions of the community Reed-Frost model are given below.

Assumptions of the Community Reed-Frost Model

1. The disease is transferred within a household from one individual to another only by some form of contact, described as *sufficient contact*. When this contact is made disease is always transmitted.
2. The infectious period of the disease is short in comparison to the latent period.
3. The population within households is homogeneous and well mixed.
4. The probability of sufficient contact occurring between two individuals within a household is proportional to the current number of infective individuals.
5. Sufficient contact with the community occurs with constant probability across idividuals in all households.
6. There are no births or deaths within the population during the outbreak of disease.

In this model we have two parameters to estimate, the probability of sufficient contact from within the household, p_h, and the probability of sufficient contact from the community p_c. The size of an outbreak within an initially healthy household of n individuals can be simulated using the following code. Since we are only interested in the final size of the outbreak, we can consider the interactions between individuals as a directed random graph with edges representing contact between the nodes which represent individuals within the household. Doing this means that we can then model infections from the community as if they occur initially.

```
RF.Com<- function(n,ph,pc){
  #      Simulates Reed-Frost progression through a
  #      household of size n = I_0 + S_0:
  #      pc  = probability of sufficient contact from community
  #      ph  = probability of sufficient contact from household
  #      I_0 = Number initially infective
  #      S_0 = Number initially susceptible

  #0: Set up.
  #0.1: Initial infectives and susceptibles
  I<- rbinom(n = 1,size = n,prob = pc)
  S<- n-I
  #0.2: qh=P(not infected by 1 household infective)
  qh<- 1-ph

  #1: looping to simulate discrete time progression.
  #while there are infectives now
  while(tail(I,1)>0){
    #1.1: Retrieve number of infectives and susceptibles.
    I_t<-tail(I,1)
    S_t<-tail(S,1)
    #1.2: calculate the number of susceptible next and add to S
    I_n<- rbinom(n = 1,size = S_t,prob = 1- qh^(I_t))
    I<- c(I,I_n)
    S_n<-S_t - I_n
    S<- c(S,S_n)
  }
  #2: Return the size of epidemic
  Size<- sum(I)
  return(Size)
}
```

Using this function to simulate the spread of disease within each household, we can build on this to simulate a vector of epidemic sizes in a community of N households , each of n individuals.

```
RF.Reps.Com<-function(n,ph,pc,N){
  out<- rep(0,N)
  for(i in 1:N){
    out[i]<- RF.Com(n = n,ph = ph,pc = pc)
  }
  return(out)
}
```

In order to generalise this to any community of households, where household size is allowed to vary, we define H, a community structure vector. Let l be one greater than the number of individuals in the largest household. H is a vector of length l such that H_i is the number of households of size $i-1$ within the community. We can then use a pre-specified H to simulate disease progression through a particular community, or else we can simulate the progression through a random community.
For the simulation of a random community structure, a Poisson distribution truncated to be without 0 is sampled from in order to generate the size of each of the N households. To do this, the mean European household size of $\mu = 2.3$ was used as the mean of the truncated distribution. This value was found by EuroStat using data from the Labour Force Survey, which was directed at private (non-communal) households across the 28 member states [6]. We wish to generate a sample from the random variable $Y \sim \text{TruncPois}(\mu)$,where $\mu = E[Y]$. We can do this by using the positive realisations of $X \sim Pois(\lambda)$, but first we must derive the relationship between the parameters μ and λ. We know that the expectation of X is given by:

$$
\begin{aligned}
E[X] = \lambda &= E[X|X=0]P(X=0) + E[X|X>0]P(X>0) \\
&= 0 \times \exp\{-\lambda\} + E[X|X>0] \times (1-\exp\{-\lambda\}) \\
&= \mu(1-\exp\{-\lambda\}),
\end{aligned}
$$

since $E[X|X>0] = E[Y] = \mu$. This can be solved numerically for λ. When we wish to sample from a truncated Poisson distribution with mean $\mu = 2.3$, by solving this equation numerically we find that we must we simulate from a Poisson(λ) distribution with $\lambda = 1.9836$. The following code simulates an outbreak of disease across a pre-specified or random community structure, H.

```r
# To sample from a truncated Poisson distribution.
 rtruncpois<-function(n,lambda){
    #n is number of realisations
    #lambda is mean of untruncated Poisson
    out<- c()
    while(length(out)<n){
      x=rpois(1,lambda)
      if(x>0) {
        out<- c(out,x)
      }
    }
    return(out)
  }


#To simulate community epidemic
RFCom.Sim<- function(H,pc,ph, random=FALSE){

  #If random = TRUE, H is number of households.
  #Generate the number of people in each household
  if(random==TRUE){
    Estate<- rtruncpois(n = H,lambda = 1.9836) #size of each of H
   households
    H<- c()
    for( i in 1:max(Estate)){
      H[i]<- sum(Estate==i) #H is vector of number of houses of each size.
    }
  }

  #Else random = FALSE, H is number of households of each size 1,2,...
  m<- max(H)
  l<- length(H)
  #For each household size, simulate the size of epidemic within household
  sizes<- matrix(NA,nrow= l,ncol=m)
  for(i in 1:l){
    if(H[i]!=0){
      sizes[i,]<- sort( c(RF.Reps.Com(i,ph,pc,H[i]),rep(Inf,m-H[i])) )
    }
    else{sizes[i,]<-Inf}
  }
  #Collect together the epidemics of given size
  out<- matrix(0,nrow=l,ncol=l+1)
  for(i in 1:l){
    for(j in 0:l){
      out[i,j+1]<- sum( sizes[i,]==j)
    }
  }
  return(out)
}
```

This returns a matrix which we shall call C, where entry C_{ij} corresponds to the number of households of i individuals $j-1$ of which become infected during the outbreak. Take as an example a random community of ten houses with probability of sufficient contact with the community $p_c = 0.2$ and probability of sufficient contact within the household $p_h = 0.4$.

```r
set.seed(100)
C<- RFCom.Sim(H=10,pc=0.2,ph=0.4,random = TRUE)
```

$$C = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 3 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Here $C_{2,3}$ denotes that in the simulation only one household of size 2 had $(3-1=)2$ individuals become infected, out of the $(3+1+1=)5$ households of size 2. We will see further application of this is Section 6.

5.2 Random Infectious Period

In the initial set up of the Reed-Frost model, it was explicitly assumed that the infectious period was short in comparison to the latent period, and that the probability of sufficient contact occurring between two individuals was proportional to the current number of infective individuals. We now consider that the infectious period may appreciably vary across all individuals, which would bring the model closer to reality, as for most diseases individuals have differing recovery times.

Consider again a single household of n individuals, and let T_i be the length of time for which individual i is infective, should they become infected. We will assume that the infectious period follows an exponential distribution so that $T_i \sim \text{Exp}(\mu)$ for some constant μ, though this could be substituted for any other suitable distribution. Under this assumption the probability of a susceptible individual not coming into sufficient contact with any individual i with infectious period T_i in the household is given by:

$$q_{h_i} = 1 - p_{h_i} = \exp(-\lambda_h T_i) = P(\text{Pois}(\lambda_h T_i) = 0),$$

where $\lambda_h \geq 0$ is the rate of attempted sufficient contact between individual i and all others.When within a single household there are I infectives, the probability that a susceptible individual does not make sufficient contact with any of these is

$$q_h = \prod_{i=1}^{I} \exp(-\lambda_h T_i) = \exp\left(-\lambda_h \sum_{i=1}^{I} T_i\right).$$

This is equivalent to

$$q_h = P(\text{Pois}(\lambda_h T) = 0), \quad \text{where} \quad T = \sum_{i=1}^{I} T_i \sim \text{Gamma}(I, \mu).$$

We incorporate this into our simulation of a household epidemic in the following way.

```r
RandIP<- function(n,lambda,mu,I_0){
  #Returns the total size of infection in a household of size n, with I initial
  #infectives. The random infectious period has mean length mu and lambda is the
  #rate at which adequate contact attempts are being made within the house.
  I<- I_0
  S<- n-I_0
  while(tail(I,1)>0){
    #Total time of infection
    TT<- rgamma(n = 1,shape = tail(I,1),rate = mu)
    #Probability of one individual avoiding infection
    qh<- exp(-lambda*TT)
    #Generate the number of susceptibles infected
    I<- c(I,rbinom(n = 1,size = tail(S,1),prob = 1-(qh^tail(I,1)) ))
    S<- c(S,tail(S,1)-tail(I,1))
  }
  #Calculate total size of epidemic
  size<- n-tail(S,1)
  return(size)
}
```

This model can be considered using ABC in a similar way to the model in Section 5.1, which includes community infection. In the earlier model we were interested in the posterior distributions of p_c and p_h, however in the model with random infectious period the probability of avoiding infection is given by $\exp\{-\lambda_h T\}$. In this expression it is not possible to distinguish whether λ_h is small and T is large or vice versa because in most instances we do not have longitudinal or individual level data available. We therefore

take the unit of time to be the mean infectious period, so that $E[T] = 1$ and λ_h represents the mean number of times sufficient contact is made between an infective and a given individual. This leaves us with the parameters λ_h and p_c to estimate.
In order to do this, we propose λ_h in accordance with the uniform prior on p_h, which we used previously. Let $q_h = 1 - p_h$. By definition, q_h is the probability of avoiding sufficient contact; which may also be given as $q_h = \exp(-\lambda_h)$. Since the prior on p_h is Uniform$[0,1]$, the prior on q_h is also Uniform$[0,1]$ and

$$P(q_h \leq u) = u \quad \text{so} \quad P(\exp(\lambda_h) \leq u) = u.$$

Therefore

$$P(\lambda_h \geq -\log u) = u,$$

and making the substitution $v = -\log u$ gives

$$P(\lambda_h \geq v) = \exp\{-v\}, \quad \text{meaning that} \quad P(\lambda_h \leq v) = 1 - \exp\{-v\}.$$

This is the cumulative density function of the Exp(1) distribution and because for random variables A and B

$$A \overset{D}{=} B \quad \text{iff} \quad P(A \leq x) = P(B \leq x) \quad \forall x \in \mathbb{R},$$

the proposal distribution of λ_h is the Exp(1) distribution.
This extension of the Reed-Frost model may be used alone or else in combination with that of Section 5.1. A model where individuals have a random infectious period as well as belonging to a household which is part of a larger community from which infection can occur is seen in more detail in Section 6.

6 Application of Methods

In this section we demonstrate the methods introduced thus far, by looking at application of these methods to two example data sets. The purpose here is to elucidate the methods, rather than on performing a full analysis of each data set. The data summarise outbreaks of influenza in Tecumseh, Michigan and in Seattle, Washington and are displayed in Table 3. This summary data are sourced from Clancy and O'Neill [5], and are shown here in the form of C, the realisation matrix introduced in Section 5.1. The Tecumseh data originate from Monto et al. [11] and detail the distribution of influenza A (H3N2) infections during the 1977-1978 epidemic in Washington; the data are summarised by size of household and number of people infected in each household. Similarly, the Seattle data originate from Fox et al. [9] and summarise an epidemic of influenza A (H1N1) in Washington during 1978-79 in the same way. The Seattle data only pertains to households of up to 3 individuals and was chosen for simplicity when constructing the necessary functions and because of the issues relating to local linear regression it

highlights well. The Tecumseh data are used to show that the functions extend to the larger problem , looking at households of up to 7 individuals. Little benefit is gained by using the model for larger households than this, firstly because of how infrequently they occur and secondly because as the household size increases the validity of the well-mixing assumption in the Reed-Frost model deteriorates.
When using Approximate Bayesian Computation for this problem there is no obvious choice of sufficient summary statistic, $S(\cdot)$. Therefore, we use a summary statistic which we hope to be close to sufficient, $T(\cdot)$, a vector detailing the total number of individuals households with an epidemic of each size.

	Number of individuals infected													
	Tecumseh (H3N2)									Seattle (H1N1)				
Household size	0	1	2	3	4	5	6	7	*Total*(H)	0	1	2	3	*Total (H)*
1	44	10	-	-	-	-	-	-	54	15	11	-	-	26
2	62	13	9	-	-	-	-	-	84	12	17	21	-	50
3	47	8	2	3	-	-	-	-	60	4	4	4	5	17
4	38	11	7	5	1	-	-	-	62					
5	9	5	3	1	0	1	-	-	19					
6	3	3	0	0	0	0	0	-	6					
7	2	0	0	0	0	0	0	0	2					
Total (T^*)	205	50	21	9	1	1	0	0		31	32	25	5	

Table 3: Observed distributions of influenza in communities of households; Tecumseh Michigan (1997-8) and Seattle, Washington (1978-9). In Tecumseh a small proportion of the population were infected, where as in Seattle a large proportion became infected.

6.1 Community Infection with Uniform Proposal Distribution

To begin, we consider the model with infection from both within the household and from the community with a constant infectious period for all individuals. The code for using this model is lengthy and can be found in the supplementary code document.
In the Approximate Bayesian Computation a simulated pair of parameter values were accepted if the absolute residuals of the simulated totals vector satisfy $|T_{sim} - T^*| \leq \varepsilon$. For the Tecumseh data ε was taken to be 33, giving an acceptance rate of 0.979%, and for Seattle ε was take to be 8 giving an acceptance rate of 1.17%. The 1000 samples from the approximate joint posterior distribution of p_c and p_h for each epidemic are shown in Figure 13. Figure 14 shows the estimated marginal posterior densities for p_c and p_h in each epidemic using these samples. We see in Figure 13 that the accepted parameter pairs in Tecumseh are both smaller and more concentrated than those for Seattle. We also notice that in both cases the accepted values fall in a small region of the

proposal space, the unit square. This suggests that the algorithm can be made more efficient by sampling from a distribution other than the prior distribution.

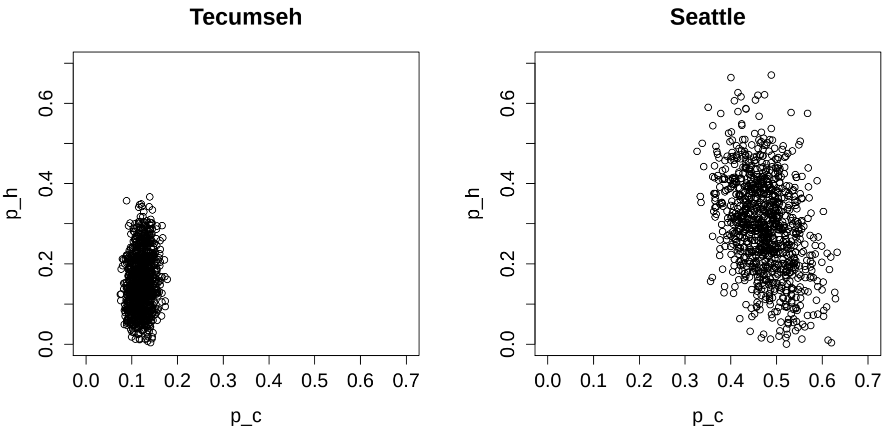


Figure 13: 1000 samples from the approximate joint posterior distributions of p_c and p_h in Tecumseh and Seattle epidemics, using ABC with Uniform[0,1] proposal distributions for both parameters. Tecumseh: p_c mean 0.121 , variance 0.0003 , p_h mean 0.154, variance 0.004. Seattle: p_c mean 0.473, variance 0.003, p_h mean 0.295 , variance 0.014 .

6.2 Community Infection with Beta Proposal Distribution

In order to make more efficient use of the parameter space we look at using a beta distribution as the proposal distribution for the ABC, so that we are proposing from a distribution that is closer to the posterior distribution. To do this we estimate the shape and scale parameters for each of the distributions using the mean and variance of the posterior samples we found using ABC with a Uniform[0,1] proposal distribution in the previous section.

Given a sample of n observations, $x_1, \cdots, x_n$, from a beta distribution with sample mean $\bar{x}$ and sample variance s^2, the estimates of the shape and scale parameters α and β satisfy:

$$\frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}} = \bar{x} \quad \text{and} \quad \frac{\hat{\alpha}\hat{\beta}}{(\hat{\alpha}+\hat{\beta})^2(\hat{\alpha}+\hat{\beta}+1)} = 2s^2.$$

Solving these equations simultaneously for $\hat{\alpha}$ and $\hat{\beta}$ gives:

$$\hat{\alpha} = \frac{\bar{x}^2 + 2s^2\bar{x}}{2s^2} \quad \text{and} \quad \hat{\beta} = \frac{\bar{x}(1-\bar{x})^2 + 2s^2(1-\bar{x})}{2s^2}.$$

The supplementary R code document contains the code used to estimate these parameters and to subsequently perform the ABC, proposing parameter values from the beta approximations to the densities found in Section 6.1. Using beta proposal

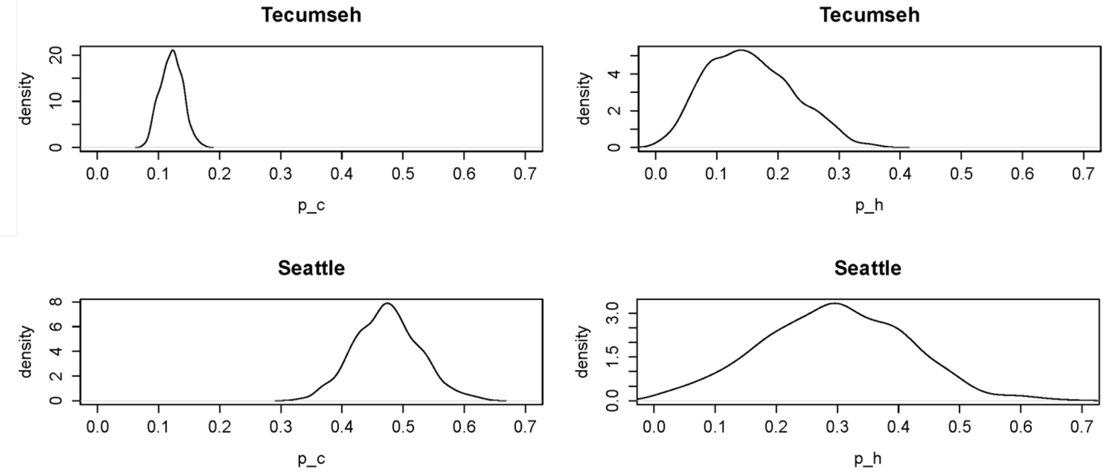


Figure 14: Approximate posterior density estimates of parameter values in Tecumseh and Seattle epidemics using uniform proposal distributions.

distributions increased the acceptance probability of both algorithms. The algorithm increased from accepting 0.979% to 36.7% of samples when applied to the Tecumseh data and from 1.17% to 1.75% in the Seattle data, for the same tolerances $\varepsilon = 33$ and $\varepsilon = 8$ as previously used.

Figure 15 and Figure 16 respectively show the approximate posterior samples and the marginal densities of the parameter estimates using the beta proposal distributions. The Tecumseh data shows the benefits of adopting a proposal distribution more similar to the posterior distribution, as the efficiency of the algorithm has been improved while producing an accepted sample similar to that obtained using the uniform proposal distribution. The Seattle data shows only very modest improvement to the efficiency of the algorithm; comparing Figures 14 and 16 provides some explanations for this. The distribution of accepted p_c values proposed from a uniform distribution is diffuse compared to that of those proposed using the beta distribution. This suggests that the density of the fitted beta distribution was concentrated too closely about the mean, and that parameter values far from the mean are not being accepted because they are being proposed with very low probability. This highlights an issue with the method; the posterior sample on which we base our proposal distribution may cause the new proposal distribution to not explore potentially viable regions of the parameter space. In low dimensional cases this can be identified by comparing the density curves as we have done here, and then corrected by increasing the number of samples taken in the first stage of the algorithm, or else by altering the distributional form of the second proposal distribution, so that it better matches the initial nonparametric posterior density estimate.

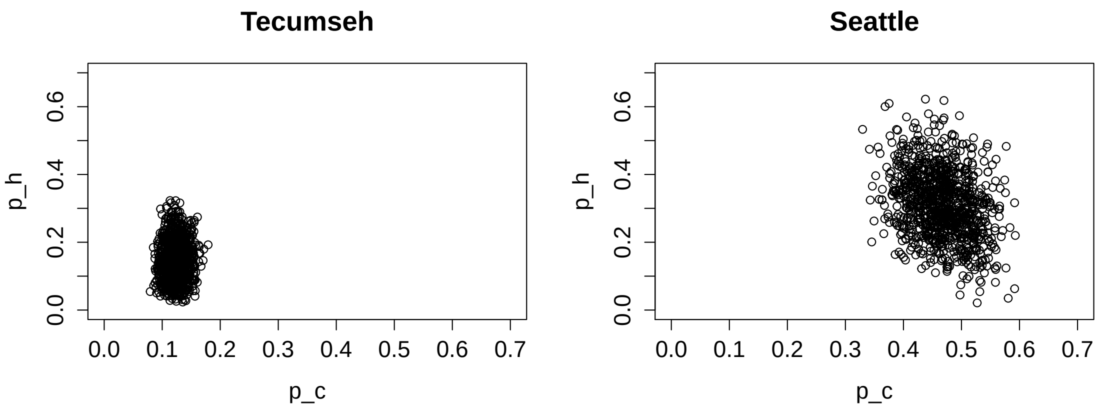


Figure 15: 1000 samples from the approximate joint posterior distributions of p_c and p_h in Tecumseh and Seattle epidemics, using ABC with beta proposal distributions for both parameters.

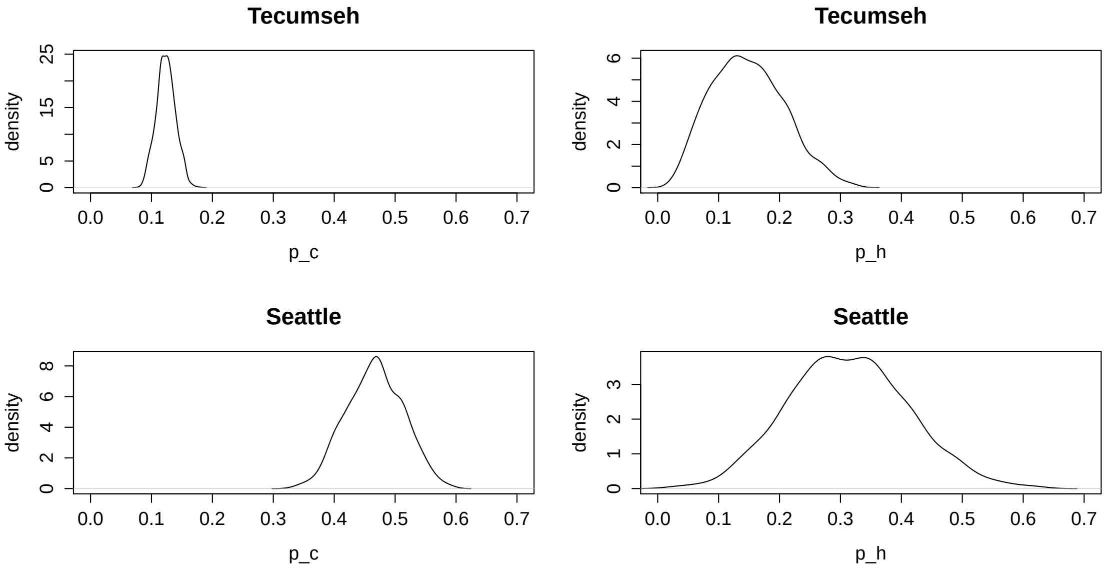


Figure 16: Approximate posterior density estimates of parameter values in Tecumseh and Seattle epidemics using beta proposal distributions. (Samples must be weighted by reciprocal of proposal likelihood before use)

6.3 Random Infectious Period

We now consider the model which allows the infectious period to differ between individuals, estimating the probability of infection from the community p_c and the expected number of contacts made by an individual during an average infectious period, λ_h. We also use this model to demonstrate the use of the local linear regression technique in practice. The R code used to generate the estimates in this section is given in The supplementary R code document.

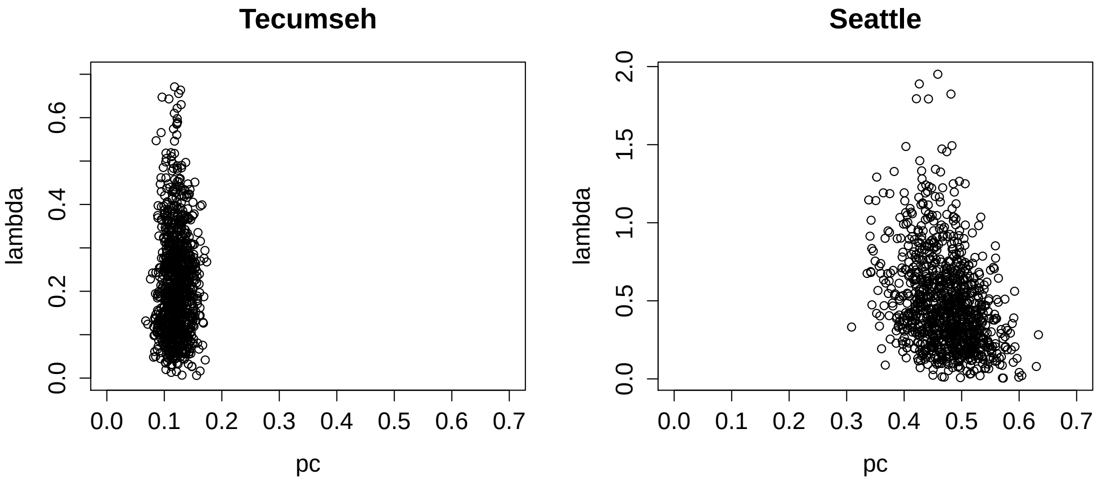


Figure 17: 1000 samples from the approximate joint posterior distributions of p_c and λ_h in Tecumseh and Seattle epidemics, using ABC with Uniform[0,1] proposal distribution for p_c and Exp(1) for λ_h.

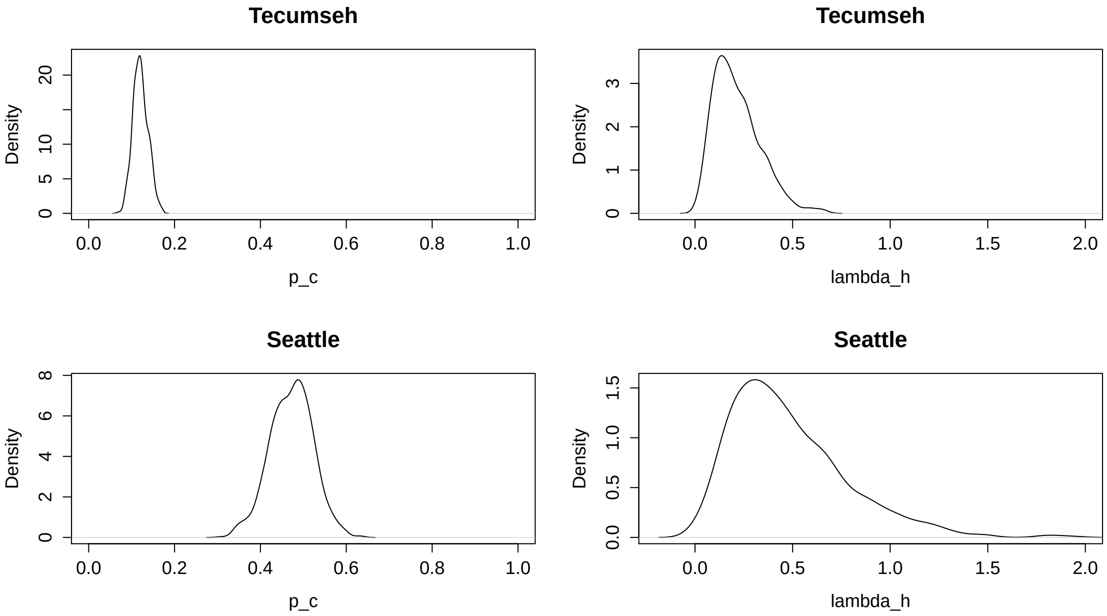


Figure 18: Approximate posterior density estimates of parameter values in Tecumseh and Seattle epidemics with random infectious period using Uniform[0,1] proposal distribution for p_c and Exp(1) for λ_h.

Approximate Bayesian Computation was used to simulate the approximate posterior samples in Figure 17, which were used to construct the approximate marginal posterior density estimates in Figure 18. The posterior density estimate on λ_h for Seattle highlights an issue with kernel density estimation; the estimated probability density of λ_h taking small negative values is greater than zero. This is because kernels placed over posterior samples of small positive values 'spill over' zero; this problem is discussed further in Section 7.

Figure 19 shows the estimated posterior density functions obtained having applied local linear regression to correct the posterior sample for non-matching summary statistics. This has resulted in posterior density estimates which are less variable, because the variability coming from non-matching summary statistics has been reduced. Applying local linear regression has most benefited the estimation of the parameters in Seattle, however it has exacerbated the issue of estimating a positive probability density for negative values of λ_h in both epidemics. This is caused by the local linear regression correction being applied across both parameters. A posterior sample with low λ_h and high p_c would overestimate the number of individuals being infected, and so both parameters are corrected and are reduced - causing there to be λ_h values which are near to or below zero in the posterior sample. Again, this problem is discussed further in Section 7.

We can compare the fixed and constant infectious period models by noticing that given the infectious period is exponentially distributed then the probability of coming into sufficient contact with an individual within the household is

$$p_h = \lambda_h/(1+\lambda_h).$$

This gives estimates p_h= 0.108 in Tecumseh and 0.325 in Seattle, which are in line with the estimates achieved in the earlier model, both falling within 1 standard deviation of the earlier means.

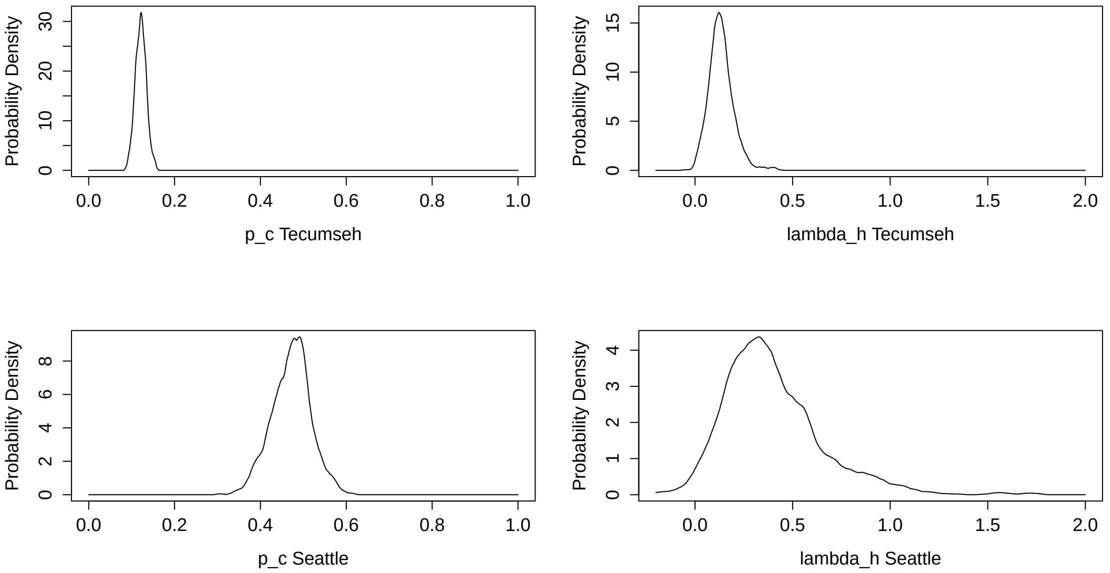


Figure 19: Approximate posterior density estimates of parameter values in Tecumseh and Seattle epidemics modelled with random infectious periods. Posterior sample achieved using ABC with a Uniform[0,1] proposal distribution for p_c and an Exp(1) proposal distribution for λ_h and local linear correction. Tecumseh: p_c mean 0.121 , variance 0.0003 , λ_h mean 0.216, variance 0.014. Seattle: p_c mean 0.472 , variance 0.002, λ_h mean 0.482, variance 0.089.

7 Conclusion

We have seen that the Reed-Frost model can be a useful tool in assessing the infectiousness of a disease, and that while it is a good model for small populations the assumptions it makes mean it is unsuitable for large populations and diseases which can not be modelled as generational. We then went on to extend the model so that it encompassed a community of households and to allow individuals to be infective for different lengths of time.

A variety of Approximate Bayesian Computation methods were introduced, progressing the method from exactly matching discrete data through to use on continuous data with no sufficient summary statistic. These computationally intensive methods allowed us to make parameter estimates in problems where the likelihood function was intractable, problems where Markov chain Monte Carlo methods were not sufficient, by producing a sample from the (approximate) posterior distribution of the joint parameter space.

When we were sampling from an ε-approximate posterior distribution, we saw that local linear regression could be used to improve our posterior sample by correcting for non-matching summary statistics, and that this reduced the variability in our approximate posterior sample. Kernel density estimation was also introduced to allow the construction of approximate posterior densities from the corrected or uncorrected posterior samples.

These techniques were seen in practice by application to the Tecumseh and Washington Influenza data. This highlighted the strengths of the techniques in allowing us to estimate and display the posterior distributions of our parameter estimates, as well as highlighting some of the weaknesses of the methods, particularly the potential to estimate parameter values outside of the parameter space as having positive probability density.

Extensions to Work

It should be clear that Approximate Bayesian Computation is an effective tool for many current areas of research where intractable likelihood functions occur, and as such there are many ways in which the work here could be extended. An initial extension, had time and space permitted, would be to look at transformations of the posterior samples so that they can take any real value and that negative values are not predicted, for example log transformation for λ_h and logit for p_c and p_h. Exploration of this was attempted, with issues arising in the back-transformation of the data. Another possible extension would be to look at Bayes Factors, and their use in model selection such as by Clancy and O'Neill [5]. These could be used to test whether there is evidence of variable infectious period or if there is risk of infection from the community at large. A third possibility would be to look further into the selection of summary statistics in cases where there is no known sufficient summary statistic. One way to do this would be to explore semi-automatic selection of summary statistics, as considered by Fearnhead and Prangle [7].

8 References

[1] Helen Abbey. An examination of the reed-frost theory of epidemics. *Human biology*, 24(3):201, 1952.

[2] Norman TJ Bailey et al. *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

[3] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[4] Daniel Bernoulli and Sally Blower. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. *Reviews in medical virology*, 14(5):275–288, 2004.

[5] Damian Clancy and Philip D O'Neill. Exact bayesian inference and model selection for stochastic models of epidemics among a community of households. *Scandinavian Journal of Statistics*, 34(2):259–274, 2007.

[6] EuroStat. Household composition statistics, 2015. URL `http://ec.europa.eu/eurostat/statistics-explained/index.php/Household_composition_statistics`.

[7] Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.

[8] Office for National Statistics. Sickness absence in the labour market, 2014. URL `http://www.ons.gov.uk/ons/rel/lmac/sickness-absence-in-the-labour-market/2014/rpt---sickness-absence-in-the-labour-market.html#tab-Sickness-Absence-in-the-Labour-Market`.

[9] John P Fox, Carrie E Hall, et al. *Viruses in families: surveillance of families as a key to epidemiology of virus infections.* PSG Publishing Company Inc., 545 Great Road, Littleton, Massachusetts 01460, USA, 1980.

[10] Major Greenwood. On the statistical measure of infectiousness. *Journal of Hygiene*, 31(03):336–351, 1931.

[11] Arnold S Monto, James S Koopman, and IRA M LONGINI. Tecumseh study of illness. xiii. influenza infection and disease, 1976–1981. *American Journal of Epidemiology*, 121(6):811–822, 1985.

[12] P Neal. Lecture notes in computationally intensive methods, 2012.

[13] The World Health Organisation. The top 10 causes of death, 2016. URL `http://www.who.int/mediacentre/factsheets/fs310/en/index2.html`.

[14] Philip D O'Neill and Gareth O Roberts. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):121–129, 1999.

[15] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[16] Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.

[17] Edwin B Wilson, Constance Bennett, Margaret Allen, and Jane Worcester. Measles and scarlet fever in providence, ri, 1929-1934 with respect to age and size of family. *Proceedings of the American Philosophical Society*, pages 357–476, 1939.

[18] Walter Zucchini. Applied smoothing techniques part 1: Kernel density estimation, 2003. URL `http://www.statoek.wiso.uni-goettingen.de/veranstaltungen/ast/ast_part1.pdf`.