AMSI
AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE
National Collaboration in the Mathematical Sciences

# Approximate Bayesian Computation and Summary Statistic Selection in Epidemic Models

Lydia Braunack-Mayer

May 9, 2013

## Contents

**1 Introduction** **2**

**2 Bayesian Statistical Theory** **2**

**3 Approximate Bayesian Computation** **4**
3.1 Approximate Bayesian Computation Algorithms . . . . . 4
3.2 Kernel Density Estimation . . . . . 5

**4 The SEIR Model** **8**
4.1 The CTMC SEIR Model . . . . . 8

**5 Summary Statistic Selection for ABC** **9**
5.1 Selection based on plots . . . . . 10
5.2 Selection based on semi-automatic ABC . . . . . 13

**6 Conclusion** **14**

**7 Works Cited** **16**

*Postal Address:* 111 Barry Street
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 9349 4106
Web: www.amsi.org.au

# 1 Introduction

Epidemic models are important for modeling and predicting the spread of a disease through a population. However, the application of an epidemic model to an emerging, or even past, epidemic can be extremely difficult. In this report I will develop a continuous-time Markov chain SEIR (CTMC SEIR) epidemic model in order to address some of the issues inherent in other epidemic models. I will describe Bayesian inference, which is a statistical approach to estimating the parameters of models such as the CTMC SEIR model. The method I will use in order to estimate parameters of an epidemic model is Approximate Bayesian Computation (ABC). I will discuss the application of the CTMC SEIR model to ABC. Finally, I will discuss the question of the best summary statistics to use when applying ABC to the CTMC SEIR model.

This project was completed in a group research environment with fellow AMSI vacation scholar Brock Hermans under the supervision of Dr Joshua Ross and Mr Jono Tuke. All coding in this project was done in the statistical program R (R core team 2012).

# 2 Bayesian Statistical Theory

Bayesian statistics, an approach to statistical inference, provided the theoretical framework for our project. Frequentist methods are familiar tools for an undergraduate statistician. However, frequentist methods are sometimes difficult to apply to complex models or large data sets. In particular, it is very difficult to apply frequentist methods to data where the likelihood function is difficult to sample from or intractable. Modern methods in Bayesian statistics can allow inference without the requirement of evaluating the likelihood function.

Bayesian inference is philosophically distinct from frequentist methods in two ways, both of which are expressed in Bayes's theorem. First, parameters are considered to be outcomes of random variables. Second, inference is based on both our beliefs about the parameter prior to observing data and on the observed data. These two philosophical differences between Bayesian and frequentist inference are captured by the following well known theorem, which forms the basis for Bayesian statistics.

**Theorem 1. Bayes's Theorem** *Let $A_1, \ldots, A_k$ be events that partition a sample space and let $B$ be an arbitrary event on that space for which $P(B) > 0$. Then*

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)}.$$

In Bayesian statistics this theorem is rephrased in terms of densities;

$$\pi(\theta|y) \propto \pi(\theta)f(y|\theta).$$

The prior density, $\pi(\theta)$, is chosen to summarise our beliefs about the parameter(s) of interest prior to observing data. The likelihood of the observed data is $f(y|\theta)$. The resulting function, $\pi(\theta|y)$, is called the posterior density and contains all the current knowledge of the parameter(s) (Christensen et al 2011, p. 30).

This reconstruction of Bayes's theorem highlights some important features of Bayesian inference. First, in Bayesian inference a parameter is an outcome of a random variable. Estimating a parameter is a matter of calculating the posterior density, rather than estimating an unknown constant.

Bayesian inference also incorporates the idea that we have information about parameters of interest before we conduct an experiment in the prior density. This contrasts sharply against the frequentist philosophy that all information comes from the observed data. The prior is a probability distribution that we choose to express beliefs that we have about the parameter of interest before the experiment. For example, if trying to predict the outcome of fair coin tosses we might choose the prior to be a symmetric function symmetric about $1/2$ taking values on $(0, 1)$. This expresses our knowledge that the parameter is a probability, between 0 and 1, whose most likely value is $1/2$. There are many different justifications for different priors. In our project we chose the prior to express that fact that we had very little information about the parameters of interest; the prior was usually the uniform PDF on $(0, 1)$.

A significant result of the inclusion of prior information, as represented in Bayes's theorem, is that we can update our model to include information from new observations. The posterior function is based on both our prior beliefs and evidence from the data. If we have new evidence we can treat our old posterior as the new prior density. The new likelihood is now based on new data and the model is updated to include new information. This approach has a significant advantage over frequentist inference in some situations. Inference no longer has to be repeated to include new information.

# 3 Approximate Bayesian Computation

## 3.1 Approximate Bayesian Computation Algorithms

The aim of our project was to develop a model for epidemic data with an intractable likelihood. Approximate Bayesian computation, or ABC, is a Bayesian method which avoids computation of or sampling from the likelihood. ABC is an algorithm that, given a model $M$ and an initial guess for the parameter $\theta$, generates observations from the posterior distribution of the data. Marjoram, Molitor, Plagnol and Tavaré describe variations on a general ABC algorithm in their paper Markov chain Monte Carlo without likelihoods (Marjoram et al 2003). Their general ABC algorithm is outlined in Table 1. This paper also outlines theoretical results justifying ABC, which I will not go into in this report.

A second ABC algorithm, outlined by Marjoram in their aforementioned paper, is more suitable for large data sets. The key difference between this algorithm and the general ABC algorithm is in the comparison of the simulated and observed data. In this second algorithm the comparison is between a set of summary statistics, $S$, where $S$ is a set of summary statistics chosen prior to running the algorithm. This variation is appropriate to large data sets, where calculating the distance between a simulated and real data set is not computationally efficient. This variation of the general ABC algorithm is detailed in Table 2.

There are a number of difficulties inherent in implementing any ABC algorithm. The first is choosing the acceptance threshold $\epsilon$. Choosing a threshold that is too large will

| | |
|---|---|
| **1** | Sample $\theta^*$ from the prior. |
| **2** | Simulate a data set $D^*$ from the model of interest M with parameter $\theta^*$. $D^*$ should have the same sampling schedule as the observed data D. |
| **3** | Compare the simulated data $D^*$ with the observed data D by computing the distance between the two data sets, $\lvert D - D^* \rvert$. |
| **4** | If the distance between $D^*$ and D is small, say $\lvert D - D^* \rvert < \epsilon$ then keep $\theta^*$ as a sample from the posterior. Otherwise discard $\theta^*$ and repeat steps 1 to 4. |
| **5** | Repeat for some pre-specified number of iterations. |

Table 1: The General ABC Algorithm, (Marjoram et al 2003)

result in accepting too many samples of $\theta$ and the resulting chain will not necessarily be a sample from the true posterior. On the other hand, generating a large enough sample for inference is computationally infeasible with a very small threshold. A balance is needed.

In order to implement ABC it is also necessary to generate data from the model $M$. This can sometimes be difficult. In our project this model was the SEIR epidemic model. I will discuss methods for generating data from the SEIR model in the next section.

A final difficulty in implementing ABC is in choosing the set of summary statistics $S$. I will discuss this problem in the final section of this report.

## 3.2 Kernel Density Estimation

In our project we used kernel density estimation to estimate the density of samples generated by an ABC algorithm. An ABC algorithm returns a sample from the posterior density. Generally we do not know the posterior density of data; in order to make inference we need to be able to approximate the posterior density of the data. A commonly used density estimator is the histogram. Histograms, however, are lacking in that they are not continuous and exact values cannot be read from them. In addition, the shape of a histograms is dependent on the choices of origin and interval width (Silverman 1986, p. 7-11). Kernel density estimation, or KDE, is an alternative method for estimating the density of data.

| | |
|---|---|
| **1** | Sample $\theta^*$ from the prior. |
| **2** | Simulate a data set $D^*$ from the model of interest M with parameter $\theta^*$. $D^*$ should have the same sampling schedule as the observed data D. |
| **3** | Calculate $S$, the values of the summary statistics on $D$, and $S^*$, the values of the summary statistics on $D^*$. Compare $S$ with $S^*$ by computing the distance between them, $\|S - S^*\|$. |
| **4** | If the distance between $S$ and $S^*$ is small, say $\|S - S^*\| < \epsilon$, then keep $\theta^*$ as a sample from the posterior. Otherwise discard $\theta^*$ and repeat steps 1 to 4. |
| **5** | Repeat for some pre-specified number of iterations. |

Table 2: A Variation on the General ABC Algorithm (Marjoram et al 2003)

An explanation for KDE can be found in Silverman's *Density Estimation for Statistics and Data Analysis* (Silverman 1986, p. 13-19). Suppose we have a sample of $n$ real observations $X_1, \ldots, X_n$. In order to estimate the density of this sample with KDE we first choose a *kernel function*. The kernel function is a function $K$ which satisfies the condition

$$\int_{-\infty}^{\infty} K(x)dx = 1.$$

Usually the kernel is a symmetric probability density function such as the Gaussian density.

The *kernel estimator* with kernel $K$ is defined to be

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

where $h$ is a smoothing parameter called the *bandwidth*. The kernel estimator can be understood to be the sum of kernel functions placed over each point in the sample. In other words, over each $X_i$ we place a symmetric density $K$. We then sum these densities and normalise this function so that it is a valid probability density. This resulting function is the probability density of the sample of $n$ real observations.

KDE resolves some of the problems inherent in histograms but introduces others. A difficulty in implementing KDE is in specifying the bandwidth $h$. The bandwidth governs the width of the kernel function and, hence, the smoothness of the kernel estimator. A large choice of bandwidth can lead to a kernel estimator that has been over-smoothed; the probability density of the sample may be obscured. A small choice of bandwidth can result in a kernel estimator that is under-smoothed or too bumpy.

In Figure 1 I have shown the influence that poor choices of bandwidth can have on a density estimated with KDE. Suppose that we toss a coin 20 times and observe two heads. I applied ABC to this problem and used KDE to estimate the density of the sample produced by ABC algorithm 2. The true density is a $beta(0.5, 0.5)$ probability density function.

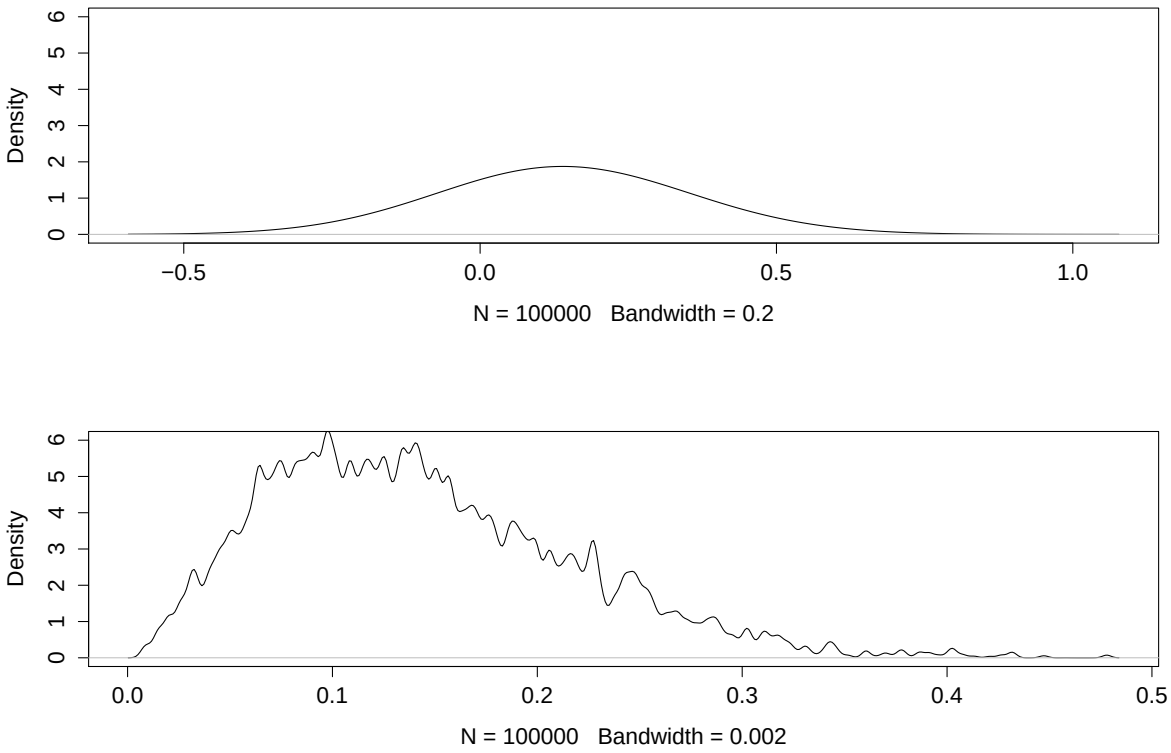


Figure 1: KDE used to estimate the density of ABC output for the coin problem, bandwidth = 0.2 (top) and bandwidth = 0.002 (bottom).

Scholars have suggested many methods for the selection of bandwidth. An automatic selection method based on the data has recently been developed by scholars Botev, Grotowski and Kroese (Botev 2010). However, we found the R functions developed by these scholars unsuitable for our data.

# 4 The SEIR Model

To implement the ABC algorithm we need a model, $M$. We applied an SEIR model, a compartmental epidemic model assigning individuals in a population to one of four categories: **S**usceptible to the disease, **E**xposed to the disease, **I**nfectious with the disease and **R**ecovered from the disease.

A SEIR model can be justified by the fact that diseases have different stages. A person with a cold might, for example, first be in a latent stage where they have no symptoms but are carrying the disease. The person might then experience symptoms such as a runny nose and cough. They might eventually recover from the cold. An epidemic model must be capable of capturing these behaviours. As individuals might experience the aforementioned stages of a cold, individuals move through the SEIR model from left to right. An individual may either stay in their category or move to the next. The SEIR model allows for individual modeling of this behaviour. Movement out of the S, E and I categories occurs with probabilities governed by the rate parameters $\beta$, $\sigma$ and $\gamma$. Once an individual is in the R category, they remain recovered. The epidemic ends when all individuals are recovered or when there are no more individuals in the exposed or infected categories.

$$S \to E \to I \to R$$

In order to apply ABC to the SEIR model we need to be able to simulate from the model. Formulating the SEIR model in terms of a continuous-time Markov chain is one way of doing this.

## 4.1 The CTMC SEIR Model

We can model the spread of a disease through a population by formulating the SEIR model as a continuous-time Markov chain. In the continuous-time Markov chain SEIR

model (CTMC SEIR) the state of the model at any time is the number of individuals in the susceptible, exposed and infectious categories, denoted by $(s, e, i)$ with constraints $s + e + i \leq N$ for $N$ the total population size, $0 \leq s$ and $e, i \leq N$. An individual leaving the infected category is considered to be removed from the model.

The transition rates from one state to another are given below.

$$\begin{aligned}
&1. -(s, e, i) \to (s-1, e+1, i) \text{ at rate } \frac{si}{N} \\
&2. -(s, e, i) \to (s, e-1, i+1) \text{ at rate } \sigma e \\
&3. -(s, e, i) \to (s, e, i-1) \text{ at rate } \gamma i
\end{aligned}$$

A CTMC remains in a state for a continuous amount of time. Let $T_1$, be the time to a change in state $(s, e, i) \to (s-1, e+1, i)$, $T_2$ to a change in state $(s, e, i) \to (s, e-1, i+1)$, and $T_3$ to a change in state $(s, e, i) \to (s-1, e+1, i)$. As a consequence of the Markov property we have $T_1 \sim exp(\frac{si}{N})$, $T_2 \sim exp(\sigma e)$ and $T_3 \sim exp(\gamma i)$.

Data can be simulated from the CTMC SEIR model by calculating the sequence of events that occur. To begin, specify the initial population state and parameter values $\beta$, $\gamma$ and $\sigma$. Calculate the minimum time to any change in state by sampling $min\{T_1, T_2, T_3\}$ and make the associated change in population state. This procedure is repeated until the end of the epidemic, or until there are no individuals in either of the exposed or infectious categories. However, this algorithm is computationally exhaustive. A more efficient algorithm for simulation for the SEIR model is the 'Gillespie' algorithm. The Gillespie algorithm was popularised by Daniel Gillespie in 1977 (Gillespie 1977). It is based on the fact that we can show that $min\{T_1, T_2, T_3\} \sim exp(\frac{si}{N} + \sigma e + \gamma i)$. The Gillespie algorithm is detailed in Table 3.

Algorithms such as the Gillespie algorithm describe the path of an epidemic. The Gillespie algorithm in particular can generate data from the CTMC SEIR model, which allows us to apply ABC to our CTMC SEIR model.

# 5 Summary Statistic Selection for ABC

In order to implement ABC, as described in section three, an appropriate set of summary statistics $S$ must be chosen. Summary statistics are extremely important when

dealing with large data sets. It is not always practical to deal with the data in its full form. A summary statistic can condense the data without compromising the accuracy of inference. For example, using summary statistics in an ABC algorithm avoids the computationally exhaustive requirement of calculating differences between data sets. Instead we compare the summary statistics.

Choosing sets of summary statistics can sometimes be very difficult. The ideal set is both small and contains as much information about the data as possible. A balance between these two requirements is needed. The ideal summary statistic is a *sufficient statistic*:

**Definition 1.** Suppose $y|\theta \sim f(y|\theta)$. A statistic $S(y)$ is said to be *s*ufficient if the distribution of $y$ given $S(y)$ does not depend on $\theta$ (Christensen et al 2011, p. 66).

A sufficient statistic and the data hold the same about of information about the parameter of interest. However, in many situations sufficient statistics cannot be found. Other methods must be used to determine a minimal set of summary statistics. I will describe two methods for summary statistic selection, a method based on plots of statistics against associated parameter values, and a semi-automatic ABC approach.

## 5.1 Selection based on plots

A simple way of assessing the usefulness of summary statistics is plotting values of the parameter against the observed value of a summary statistic on data simulated from the parameter. This method is outlined in Table 4. The stronger the relationship between the parameter and the associated summary statistic, the better the statistic.

| | |
|---|---|
| **1** | Set initial values for $\beta$, $\sigma$ and $\gamma$ and the initial state of the population $(s, e, i)$. |
| **2** | Calculate the minimum time to the next change in state, where $min\{T_1, T_2, T_3\} \sim exp(\frac{\beta si}{N} + \sigma e + \gamma i)$. |
| **3** | Choose the next event to occur:<br>Event 1 occurs with probability $\frac{\beta si}{N}$<br>Event 2 occurs with probability $\sigma e$<br>Event 3 occurs with probability $\gamma i$ |
| **4** | Update $(s, e, i)$ and repeat until $e = 0$ and $i = 0$. |

Table 3: The Gillespie Algorithm for the CTMC SEIR model (Gillespie 1977)

Summary statistics associated with plots showing no relationship are discarded; there is not a strong relationship between the statistic and the parameter.

Unfortunately plots of statistics for the SEIR models against parameters appeared to show, for the most part, no strong relationship. Plotting the statistics against ratios of parameters revealed a stronger relationship. These plots are shown in Figure 2. We suspect that generating these plots with larger initial populations might yield better results. This could be an avenue for further research.

| | |
|---|---|
| **1** | Generate a hypercube of values for $\beta$, $\sigma$ and $\gamma$. |
| **2** | Simulate data $x$ using the Gillespie algorithm with parameters $\beta$, $\sigma$ and $\gamma$. |
| **3** | Calculate the value of a set of summary statistics $S$ on $x$. |
| **4** | Repeat. |
| **5** | Choose appropriate summary statistics based on plots of $S$ against associated values of $\beta$, $\sigma$ and $\gamma$. |

Table 4: Summary statistic selection method one

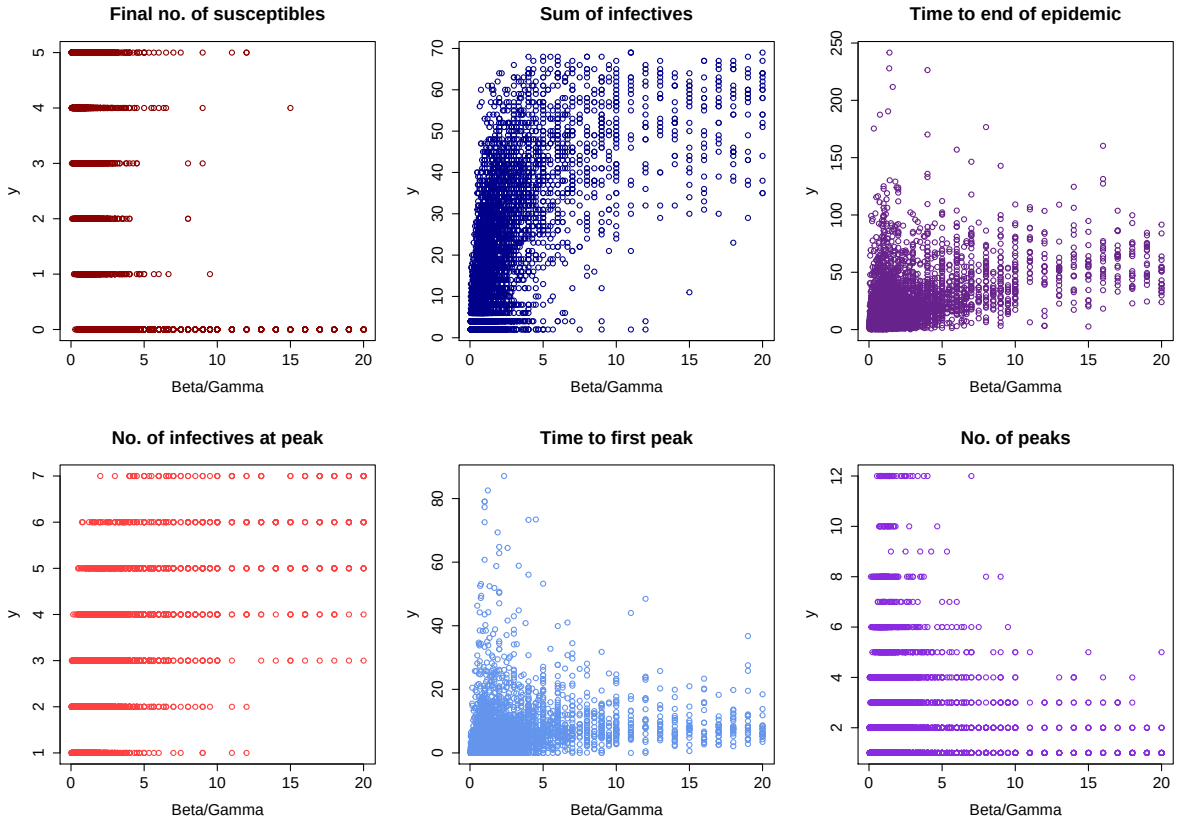


Figure 2: Plots of various summary statistics against associated parameter values, generated with $s = 5$, $e = 1$ and $i = 1$ (R core team 2012).

*Postal Address:* 111 Barry Street
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 9349 4106
Web: www.amsi.org.au

## 5.2 Selection based on semi-automatic ABC

Semi-automatic approximate Bayesian computation is a second approach to summary statistic selection developed by Paul Fearnhead and Dennis Prangle in 2012 (Fearnhead & Prangle 2012). In this method, summary statistic selection is built into an ABC algorithm. A linear regression is fitted to the data in order to estimate summary statistics. Their method is detailed in Table 5.

In step one we implement ABC as in Table 2. This is to reduce the region in which we suspect the parameter values lie. In step two we implement the Gillespie algorithm, as in Table 3. In step three we estimate appropriate summary statistics by fitting the following linear models, where $D^* = \{D_1, \ldots D_n\}$ is our simulated data and $(a_1, \ldots a_n)$, $(b_1, \ldots b_n)$, $(c_1, \ldots c_n)$ are estimated constants.

$$\begin{aligned}\beta_i &= a_0 + a_1 D_1 + \ldots + a_n D_n + \varepsilon_i \\ \sigma_i &= b_0 + b_1 D_1 + \ldots + b_n D_n + \varepsilon_i \\ \gamma_i &= c_0 + c_1 D_1 + \ldots + c_n D_n + \varepsilon_i.\end{aligned}$$

The real data $D$ is then substituted into each model and the resulting values are summary statistics for $\beta$, $\sigma$ and $\gamma$ respectively. Step four is again familiar; we run ABC as in Table 2 with the summary statistics generated in step three.

While this method worked for Fearnhead and Prangle, we found that fitting linear models to our data estimated unrealistic summary statistics. This seemed to be caused by the form of our data. The Gillespie algorithm simulates a string of population states, $(s, e, i)$ with an associated time of change in state. The algorithm stops when $e = 0$ and $i = 0$, which means that data generated by the Gillespie algorithm has no fixed length. Applying linear regression to data of wildly varying lengths resulted in linear models that were unusable.

| | |
|---|---|
| **1** | Use a pilot run of ABC to determine a region of non-negligible posterior mass. |
| **2** | Simulate sets of parameter values and data. |
| **3** | Use the simulated parameter values and data to estimate the summary statistics. |
| **4** | Run ABC with this choice of summary statistics. |

Table 5: Semi-automatic ABC (Fearnhead & Prangle 2012)

Our solution to this problem was to adapt the semi-automatic ABC algorithm by fixing the length of the simulated data. We adjusted the Gillespie to produce data in the form of a set of summary statistics $S = \{S_1, S_2, \ldots, S_n\}$. We implemented step three of the semi-automatic ABC algorithm with the following three linear models, where $(a_1, \ldots a_n)$, $(b_1, \ldots b_n)$ and $(c_1, \ldots c_n)$ are again estimated constants.

$$\begin{aligned}\beta_i &= a_0 + a_1 S_1 + \ldots + a_n S_n + \varepsilon_i \\ \sigma_i &= b_0 + b_1 S_1 + \ldots + b_n S_n + \varepsilon_i \\ \gamma_i &= c_0 + c_1 S_1 + \ldots + c_n S_n + \varepsilon_i.\end{aligned}$$

We assessed the fit of the three models and removed variables with large associated p-values. Summary statistics associated with large p-values were considered to be uninformative. This approach is advantageous in that it allowed us to implement Fearnhead and Prangle's semi-automatic ABC algorithm. However, it has a significant disadvantage in that our choice of a set of summary statistics $S$ is biased. Implementing a semi-automatic approach is designed to avoid the bias inherent in summary statistic selection. Our adjustment of the semi-automatic ABC algorithm re-introduces this bias.

We believe that disadvantages of our variation of Fearnhead and Prangle's semi-automatic ABC algorithm will be negligible. The models that we are fitting are functions of data and, as such, still contain information about the data. We were able to obtain results that gave weight to this idea. We believe, but have not yet shown, that very little information is lost by basing our estimation of summary statistics on functions of summary statistics. The next step in our research would be to attempt to show this. We could then fit this ABC algorithm to epidemic data and estimate the transition rates.

# 6 Conclusion

The aim of our AMSI vacation project was to apply a model to epidemic data with an intractable likelihood function. We began by investigating Bayesian inferential methods for data without the likelihood. We then worked to apply our chosen method, approximate Bayesian computation, to epidemic data. In order to do this we investigated a method for density estimation, KDE. We also developed a model, the CTMC SEIR model, for epidemic data. Within this we worked to apply the Gillespie algorithm

for simulation to our situation. A significant focus in our project was on the choice and use of summary statistics within the ABC algorithm. We investigated a number of methods for summary statistic selection, including plotting simulated statistics against associated values and the semi-automatic approximated Bayesian computation algorithm.

There are many avenues for further work on our project. The most immediate of these would be to try and show that our variation of the semi-automatic ABC algorithm does not lose information. It would also be of interest to repeat our first approach to summary statistic selection with larger initial populations. I would also like to explore the use of other ABC algorithms, including ABC MCMC and ABC rejection sampling. We would aim to complete our research by using our model to estimate the transition rates of a real epidemic.

I would like to thank AMSI for giving me the opportunity to explore this area of statistics and for funding my project. I would also like to thank CSIRO for allowing the opportunity to present my research at the Big Day In. I have very much appreciated the support and advice of my supervisors, Mr Jono Tuke and Dr Joshua Ross. Their input was invaluable throughout the project. Finally, I would like to thank my partner Brock Hermans. I would not have made as far without him.

# 7 Works Cited


Botev, I, Grotwoski, J and Kroese, D, 'Kernel Density Estimation Via Diffusion', *Annals fo Statistics*, vol. 38, no. 5, pp. 2916-2957.

Christensen, R et al, 2011, *Bayesian Ideas and Data Analysis*, ed. 1, Taylor & Francis Group, New York.

Fearnhead, P and Prangle, D, 2012, 'Constructing summary statistics for approximate Bayesian computation', *Journal of the Royal Statistical Society*, vol. 74, no. 3, pp. 419-474.

Gillespie, D, 1977, 'Exact Stochastic Simulation of Coupled Chemical Reactions', *The Journal of Physical Chemisty*, vol. 81, no. 25, pp. 2340-2361.

Majoram, P et al, 2003, 'Markov chain Monte Carlo without likelihoods', *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 26, pp. 15324-15328.

R Core Team (2012). R: A Language and Environment for Statistical Computing. (Version 2.15.1). (Software). [Downloaded March 2012]. Available from: ¡http://www.R-project.org¿.

Silverman, B W, 1986, *Density Estimation for Statistics and Data Analysis*, ed. 1, Chapman and Hall, London.