DEPARTMENT OF MATHEMATICS

Bayesian Modelling of Epidemic Processes

Dominic Hutchinson

Level - 7/M
Credits - 40
Supervisors - Dr Daniel Lawson & Dr Sam Tickle
Handed In - May 14, 2021

Acknowledgement of Sources

For all ideas taken from other sources (books, articles, internet), the source of the ideas is mentioned in the main text and fully referenced at the end of the report.

All material which is quoted essentially word-for-word from other sources is given in quotation marks and referenced.

Pictures and diagrams copied from the internet or other sources are labelled with a reference to the web page or book, article etc.

Signed ______________________

Date 14/05/2021

Abstract

Approximate Bayesian Computation (ABC) methods are a family of likelihood-free methods for estimating the posterior of model parameters. These methods have many features which users need to specify for each problem the methods are used on, and on which the success of these methods rests. In this project, I explore methods for automating the selection of one of these features: summary statistics. I provide an analysis of what creates a desirable summary statistic and discuss methods to automate their selection. Two types of selection method are discussed: those which choose an optimal set of statistics from a larger set of handcrafted statistics; and, those which generate their own summary statistics. I implement these methods and evaluate their ability to fit SIR models to data from the early stages of the SARS-CoV-19 (Covid-19) outbreaks in both France and Senegal, with results showing that these selection methods are able to select summary statistics which, when paired with an ABC method, are able to produce better fitting models than the naiv̈ely optimal identity function.

Contents

1 Introduction **1**

1.1 Motivation . . . 1

1.2 Objectives . . . 2

1.3 Structure . . . 2

1.4 Accompanying Resources . . . 3

2 Bayesian Modelling and Epidemic Processes **4**

2.1 Bayesian Modelling . . . 4

2.2 Epidemic Processes . . . 5

2.3 Compartmental Model . . . 6

2.3.1 Standard SIR Model . . . 7

2.3.2 Other Compartmental Models . . . 14

2.3.3 Stochastic SIR Model . . . 16

2.4 Beyond Compartmental Models . . . 19

3 Approximate Bayesian Computation **20**

3.1 Motivation and Background . . . 20

3.2 ABC Methods . . . 21

3.2.1 ABC-Rejection Sampling . . . 27

3.2.2 ABC-Importance Sampling . . . 29

3.2.3 ABC-MCMC . . . 31

3.2.4 ABC-SMC . . . 34

3.3 ABC for Model Choice . . . 38

3.4 Regression Adjustment in ABC . . . 40

4 Summary Statistic Selection **42**

4.1 Motivation . . . 42

4.2 Properties of Summary Statistics . . . 43

4.3 Methods for Summary Statistic Selection . . . 52

4.3.1 Approximate Sufficient Subset . . . 52

4.3.2 Minimising Entropy . . . 55

4.3.3 Two-Step Minimum Entropy . . . 57

4.3.4 Semi-Automatic ABC . . . 58

4.3.5 Non-Linear Projection . . . 61

5 ABC and Epidemic Events **63**

5.1 Comparison of ABC Methods . . . 63

5.2 Comparison of Summary Statistic Methods . . . 68

5.3 The Effect of Dimensionality . . . 72

5.4 Projection . . . 73

5.5 Case Studies 76
5.5.1 France Covid-19 78
5.5.2 Senegal Covid-19 81

6 Conclusion **85**
6.1 Future Work 86

Bibliography **i**

1 Introduction

1.1 Motivation

Diseases have plighted human history, bringing much suffering and death with them. The ability to accurately model the spread of diseases and the likely effects of preventative measures can lead to saved lives. Girolamo Fracastoro, in the 16^{th} century, was the first to present the idea that diseases are spread by microscopic organisms in the air [Nutton, 1990]. This "Germ Theory" for disease has led to many developments in the study of the spread of disease and its prevention.

The first mathematical modelling of epidemic events occurred in the 18^{th} century when Daniel Bernoulli sought to demonstrate the effectiveness of a smallpox vaccine [Colombo and Diamanti, 2015]. In his model Bernoulli grouped the population by birth year x and then identified those in each group who were susceptible to smallpox $S(x)$ and those who had already survived small pox and so assumed to be immune $R(x)$. Bernoulli theorised two differential equation (1) for the change in the susceptible population and removed population sizes with age.

$$\frac{dS}{dx} = -qS(x) - m(x)S(x) \tag{1a}$$
$$\frac{dR}{dx} = q(1-p)S(x) - m(x)R(x) \tag{1b}$$

where q is the probability of catching smallpox in a given year, p is the probability of dying from smallpox and $m(x)$ is the probability of dying at age x from anything other than smallpox. Parameters p and q are assumed constant for all ages.

Mortality rate by age $m(x)$ was estimated using data collected from new life insurance schemes. Manual calculations were then required to estimate p and q, and thus the total number of deaths due to smallpox each year for different age-ranges. Bernoulli proceeded to calculate the increase in life expectancy if everyone was inoculated at birth (i.e. $p = 0$) and found that, if this were the case then, life expectancy would rise from 26.57 years to 29.65 years.

Modelling of epidemic events has come on a long way since Bernoulli's first models. Compartmental models partition a population into disjoint compartments which represent some state of the individual (e.g. whether an individual is infectious, or not) and then proceeds to model the movement of individuals between these compartments. [Kermack *et al.*, 1927] presented the first compartmental model in the context of Epidemic events. Compartmental models have proved successful at modelling real epidemic events, including Measles [Verguet *et al.*, 2015], HIV-AIDS in Cuba [Blum and Tran, 2010] and the 2009 H1N1 (Swine-Flu) Pandemic [Coburn *et al.*, 2009].

Compartmental models have several parameters which govern the movements of individuals between populations, and these parameters need estimating. Moreover, being able to estimate these parameters accurately at different points in time can lead to inferences about the effect of any non-pharmaceutical interventions (NPIs) which have been introduced or to project the future course of an epidemic. There are two approaches to such estimation and inference: frequentist, and Bayesian. The frequentist approach assumes each parameter to be a fixed quantity and thus an estimate can be calculated directly from the available data, whereas the Bayesian approach treats each parameter as a random variable. The Bayesian approach enables for uncertainty to be incorporated into a model and is thus much more suitable to scenarios where either little or only imperfect data is available, such as real-world epidemics.

When modelling in the Bayesian paradigm the quantities of interest are typically the parameter posteriors, the probability of the parameters given some observed data. The difficulty in calculating these posteriors is the requirement to quantify a likelihood function. In modelling problems the likelihood function is the probability of observing some data given the current parameters of the model. In reality, the likelihood function is intractable and thus likelihood-free methods are required to estimate parameter posteriors.

A popular class of likelihood-free methods are known as Approximate Bayesian Computation methods (ABC). These methods negate the need to specify the likelihood function by instead seeking to generate samples from it. This is achieved by generating and comparing simulations of theorised models to observed data from the true model. This makes these methods computationally intensive and is the reason that these methods have only began to emerge at the turn of the Millennium, with the first recognised methods being presented by [Tavaré *et al.*, 1997]. This first method involved generating many-many simulations and until n were found which produced values sufficiently close to the observed data, according to some distance measure. The distributions of the parameters which were used in these n simulations is then used as an estimation of the parameter posteriors. This approach has since been extended upon with the introduction of popular Monte Carlo sampling methods, namely MCMC and SMC.

1.2 Objectives

The objectives of this project are to:

- Explore and understand models for epidemic events.
- Evaluate and implement several ABC methods.
- Explore ABC methods which minimise the number of hyper-parameters which users need to specify.
- Understand the importance of summary statistic selection in the success of ABC methods.
- Discuss and assess the effect of the dimensionality of summary statistics.
- Explore, evaluate and implement several methods for automating summary statistic selection for ABC methods. Including methods which generate their own summary statistics.
- Evaluate the ability of ABC methods to fit models to epidemic events when using the identity function, an informative, but naïve, summary statistic.
- Evaluate the ability of ABC methods to fit models to epidemic events when using summary statistics generated by automated methods.
- Discuss not only the theoretical qualities of these methods, but their computational practicalities as well.
- Apply the discussed methods to real-world data.

1.3 Structure

The structure of this project is as follows: In *Section 2* I provide an overview of Bayesian modelling and introduce some models which are often used for modelling epidemic events. In *Section 3* I present a class of computational methods known as Approximate Bayesian Computation methods (ABC) which are used to approximate the posterior distribution of model parameters in scenarios when the likelihood is intractable. In *Section 4* I perform an in-depth analysis of summary statistics and present methods for automating the selection of summary statistics for ABC methods. In *Section 5* I assess the performance of the methods presented in *Sections 3 & 4* on the epidemic models presented in *Section 2*, including fitting epidemic models to data from the SARS-Cov-19 outbreaks in France and Senegal. In *Section 6* I review the achievements of this project and discuss future work to build on those achievements.

1.4 Accompanying Resources

A git repository containing the code used during this project can be found at `https://github.com/dajhutchinson/Bayesian-Modelling-of-Epidemic-Processes`. Test code is provided in Jupyter Notebooks.

2 Bayesian Modelling and Epidemic Processes

In this section I discuss the Bayesian approach to statistical modelling, with a focus on how it can be applied to epidemic processes. I present an overview of the Bayesian paradigm and modelling within it *Section 2.1*. I define epidemic processes and discuss common considerations when producing mathematical models from them *Section 2.2*. I present a popular class of models for epidemic processes, compartmental models *Section 2.3*. And, close this section my mentioning other popular classes of model used in the modelling of epidemic processes *Section 2.4*.

2.1 Bayesian Modelling

Bayesian statistics is one of the two main statistical paradigms, with frequentist being the other. In frequentist statistics model parameters are considered as fixed quantities which can be estimated, whilst Bayesian statistics treats model parameters as random variables with their own distributions. Classical Bayesians believe in a "True Model" which is unknown, while Subjective Bayesians believe that no such model exists and rather that each distribution is only a prediction of future events.

Theorem 2.1 (Bayes' Rule)
Consider two random variables X and Y. Bayes' Rule is the following formulation for the conditional distribution of Y given X.

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X|Y)\mathbb{P}(Y)}{\mathbb{P}(X)}$$

where each component is known as

- $\mathbb{P}(Y|X)$, *the Posterior of Y given variable X.*
- $\mathbb{P}(X|Y)$, *the Likelihood of Y given event X.*
- $\mathbb{P}(Y)$, *the Prior distribution of Y.*
- $\mathbb{P}(X)$, *the Evidence for fixed event X.*

Proof. Bayes' Rule follows from the definition of conditional distributions and joint distributions

$$\begin{aligned} \mathbb{P}(Y|X) &= \frac{\mathbb{P}(X,Y)}{\mathbb{P}(X)} \\ &= \frac{\mathbb{P}(X|Y)\mathbb{P}(Y)}{\mathbb{P}(X)} \end{aligned}$$

□

The keystone of the Bayesian approach to statistics is Bayes' Rule (**Theorem 2.1**). For a model X with parameters θ, a relationship between the model and its parameters can be immediately defined by Bayes' Rule by setting $Y = \theta$.

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta)\mathbb{P}(\theta)}{\mathbb{P}(X)}$$

The starting point of Bayesian inference is the parameter prior $\mathbb{P}(\theta)$ which quantifies our existing beliefs about the distribution of the parameters, before seeing any data. These beliefs can be very loose with probability mass spread over a large proportion of the parameter space.

Bayes' Rule is used to update our beliefs, once data X has been observed, by calculating a posterior $\mathbb{P}(\theta|X)$. Typically, the evidence $\mathbb{P}(X)$ for the observed data is intractable, but this is not a limitation of Bayes' Rule as the evidence is only used as a normalising term and is constant with-respect-to the parameters θ. This means it can be calculated after the fact.

$$\mathbb{P}(\theta|X) \propto \mathbb{P}(X|\theta)\mathbb{P}(\theta)$$

In practice it is ideal if the posterior follows a standard distribution as inferences and computations become much easier, due to tractable closed-form functions for these distributions being known. This is where the concept of "conjugate priors" is useful. A prior is said to be conjugate if it has the same distribution as the posterior, this only occurs when certain pairs of distributions are defined for the prior and the likelihood $\mathbb{P}(X|\theta)$. Conjugate priors are a well studied area of Bayesian statistics and there are several resources which list popular ones, along with the computations required to calculate the parameters for the posterior (See [Fink, 1997]).

It is naturally preferable for priors to be informative and well motivated as this reduces the amount of data required for the posterior to resemble the true distribution for a model. However, given enough data, the posterior will always converge on the true distribution as long as the support of the prior does not truncate the true support of the posterior. This demonstrates a difficulty in defining priors as they introduce bias. For computational methods it is common to have to define a prior with a relatively small support for tractability, or for the priors to just be best guesses as a tractable prior needs to be defined.

A common problem in Bayesian modelling is model choice: The task of having to decide which to two, or more, models is the best fit for some given data. There are several options for comparing different models, including: Akaike Information Criterion (AIC); Deviance information criterion (DIC); and, Cross Validation. Bayes Factor is possibly the most popular due to its simplicity and direct relatedness to Bayes' Rule. I define and discuss the Bayes Factor in *Section 3.3*.

2.2 Epidemic Processes

[Goffman, 1965] characterises an epidemic process as "a time-dependent process of transition by the members of a population, where the state transitions are caused by exposure to some influence called 'infections material'.". Typically this 'infectious material' is an infectious disease (e.g. HIV, Ebola, Flu) but can be more abstract concepts such as a secret, learnable skill or drug addiction. A pandemic is an epidemic which has spread to multiple populations, however it is often still effective to model each epidemic separately due to geographical and political borders. I focus on the disease case of epidemics in this project due to its prevalence in the literature.

The study of modelling such processes is motivated from a public health stand point, with two main problems: predicting the future progress of the epidemic; or, evaluating the true effects of introducing different mitigation strategies. Being able to accurately complete either of these tasks can help control spread of infectious diseases and thus reduce human deaths and suffering (see [Ciofi degli Atti *et al.*, 2008]). This motivation extends to modelling the spread of diseases in animal populations (see [Brooks-Pollock *et al.*, 2014]) as many human diseases are zootonic[1]. Being able to compute accurate models for these processes allows for both qualitative and quantitative analysis to be performed, with the results being used to inform public health policies.

[1] A disease which originates in animal populations before jumping to human populations.

Definition 2.1 (Reproduction Number)
The basic reproduction number R_0 is the expected number of people each infected individual will pass a disease onto when the disease is uncontrolled. The effective reproduction number R_t is the mean number of people each infected individual will pass a disease onto at a given time t. There may be policies in place at time t which will effect R_t, but R_0 will remain constant. When $R_t < 1$ then the size of the infected population is decreasing at time-point t, when $R_t = 1$ the size of the infected population is stable, and when $R_t > 1$ the size of the infected population is increasing.

These definitions assume that every member of a population is susceptible and the population size is effectively unlimited.

As policy makers are rarely expert statisticians, several simple statistics have been developed to ease communication between statisticians and policy makers (and, policy makers and the public). These statistics typically summarise the full time-series into a few values which are readily interpretable. The most popular of these are the: basic reproduction number R_0; and, effective reproduction number R_t (**Definition 2.1**). How these values are calculated depend on the model being used and these might be useful summary statistics for the computational methods discussed *Section 3*, despite not being sufficient (see *Section 4*).

As epidemic processes naturally grow exponentially, it is pivotal to the success of public health programs to be able to respond quickly before the disease gathers momentum and becomes uncontrollable. This is often difficult as there will always be a delay between data being observed and it being incorporated into a model, as well as it being harder to make strong inferences from small sample sizes due to the high variability caused by natural noise. This delay is typically longer at the start of an epidemic of a novel disease due to lack of awareness. This allows the novel disease to spread unchecked, which is dangerous whilst the lethality of the disease is unknown.

The ideal theorised model will be a causal model of the underlying epidemic process, rather than just correlated with the process. These models are incredibly rare in practice due to the number of hidden variables, the complex nature of these processes and the lack of quality in the available data. This means that any well fitting model will likely only be correlated with the epidemic process and thus only very weak inferences can be made. In reality these models are rarely useless due to confounding variables which link the epidemic process to the predictor variables being used in these models, although there is no guarantee of this. As always, in reality it is impossible to know whether your theorised model is actually the true model. Moreover, the most suitable model for each epidemic will depend on the available data.

In the modern information age and with the current rise in "Big Data" the number of variables for which data is available has increased dramatically, as well as the amount of data available. Further, it is now much easier to collect individual-level data, rather than just population-level data. This allows for more complex models to be theorised. [Badr *et al.*, 2020] uses data collected from mobile phones to incorporate individuals mobility into their model for the spread of SARS-CoV-19 in the USA.

Although the quantity of data available has increased, much of the data is still of poor quality. As mentioned there is always a delay in the data, but, further that this, the data is often incomplete due to undetected or misdiagnosed cases/deaths.

2.3 Compartmental Model

Models for epidemic processes began with [Ross, 1916; Ross and Hudson, 1917] and the first recognisable compartmental model for epidemic processes is the Kermack-McKendrick model presented in [Kermack *et al.*, 1927] which considers a closed population.

Due to epidemic processes representing the transitions of individuals between groups within a population, compartmental models are a popular class of models for the problem. Compartmental models are defined by: several mutually-exclusive “compartments” which partition a population; and a set of equations which govern interactions between these compartments (i.e. How individuals move between compartments). Typically these equations are differential equations so they can capture the interactions inbetween time-periods. The Kermack-McKendrick model is a compartmental model with two states: Susceptible (S); and, Infectious (I). Under the Kermack-McKendrick model, once an individual has recovered from an infection they are removed from the population entirely. This model has been generalised to the standard SIR[1] model which has an extra compartment, Removed[2] (R), where individuals are moved to once they are no longer infectious.

2.3.1 Standard SIR Model

The standard SIR model [Huppert and Katriel, 2013] is a deterministic model compartmental model, with the following compartments: the Susceptible (S) compartment represents individuals who have not had the disease and could become infected if they came into contact with someone who is infected; the Infectious (I) compartment represents individuals who currently have the disease and are able to pass it one to members of the susceptible population; and, the Removed (R) compartment represents who have had the disease but are no longer able to spread it. For the standard SIR model we can consider the removed population to include both those who have died from the disease and those who have recovered. However, for more complex inferential problems, such as the affects of a vaccine program or a new treatment, it becomes necessary to separate these two groups. It is assumed that recovering from a disease infers immunity on the individual.

This model makes several assumptions: a constant population size N; that individuals are homogeneous, especially with-respect-to their health and movements; and, that individuals meet each other uniformly at random. These assumptions are reasonable for large populations as individual variations are averaged out by the law of large numbers and when all members of a population are equally susceptible to a disease. In practice, susceptibility will vary between individuals due to factors such as age, pre-existing health conditions and amount of human-to-human interactions an individual has. The nature of inter-person interactions is governed by two parameters β, γ (which are explored below).

The size of each population is mathematically represented by three time-dependent functions $S(t), I(t), R(t)$ where time is continuous. The time period between observations will vary between process, although in reality observations typically happen daily or less-often. This means the data these functions are representing is non-continuous with-respect-to time, which can cause difficulties when trying to fit continuous functions. As the total population is assumed to be constant $S(t) + I(t) + R(t) = N$ at all time points t. A disease has died out if the infectious population size ever falls to zero (i.e. $\exists\ t$ st $I(t) = 0$).

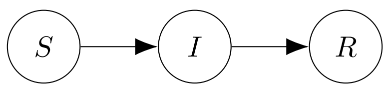


Figure 2.1: Diagram of interactions between compartments in the standard SIR model. S=Susceptible, I=Infectious, R=Removed.

[1] “Susceptible-Infectious-Removed”

[2] Sometimes referred to as Recovered.

The standard SIR model only allows for two interactions: Susceptible to Infected ($S \to I$); and, Infected to Removed ($I \to R$). These interactions are depicted in *Figure 2.1*.

These two interactions are governed by a system of three non-linear ordinary differential equations, given in (2). This system of equations represent the change in the total number of individuals in each compartment over time [Rodrigues, 2016].

$$\frac{dS}{dt} = -\frac{\beta}{N}S(t)I(t) \tag{2a}$$
$$\frac{dI}{dt} = \frac{\beta}{N}S(t)I(t) - \gamma I(t) \tag{2b}$$
$$\frac{dR}{dt} = \gamma I(t) \tag{2c}$$

where β is the average number of people infected by the a single infected individual in a single time-period and γ is the probability an individual is removed in a single-time period. Initial conditions imposed on this system are that $S(0) \geq 0, I(0) \geq 0, R(0) \geq 0$ and $S(0)+I(0)+R(0) = N$. Note that if $I(t) = 0$ at any point in time t then no new infections will occur after time-point t. Moreover, if $I(0) = 0$ then no infections will ever occur, which is distinctly uninteresting.

In the standard model, β and γ are non-negative constants with $\beta \in \mathbb{R}^{\geq 0}, \gamma \in (0,1]$[1]. Note that $\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$ which ensures the total population size is constant and since each differential equation only depends on the current values of $S(t), I(t)$ and $R(t)$, the standard SIR model is Markovian (as all are compartmental models).

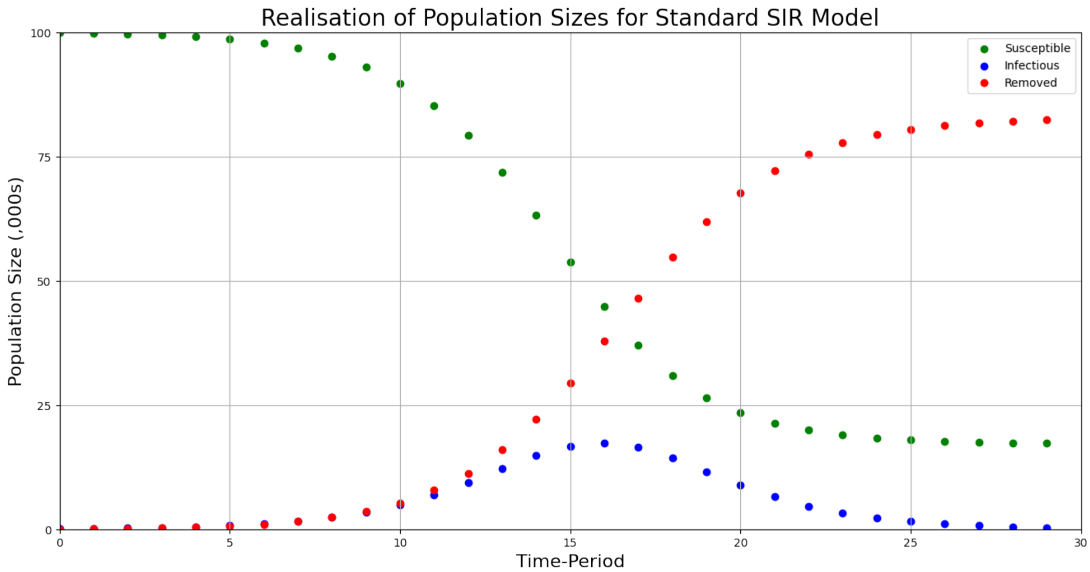


Figure 2.2: Realisation of a standard SIR model for a population of size $N = 100,000$ over 30 time-periods where $\beta = 1$ and $\gamma = 0.5$. ($R_0 = 2$)

Figure 2.2 provides a plot of how the different compartment sizes vary over time for a standard SIR model of a population of 100,000 individuals and a disease where each infected individual passes the disease to one susceptible individual each time-period ($\beta = 1$) and there is a 50% chance an infected individual is removed each time period ($\gamma = 0.5$).

In *Figure 2.2* we can observed that there is a point at which the size of the infectious population begins to decrease. We can mathematically identify this time-point as the time-point

[1] γ is specified be non-zero otherwise no-one ever transitions to the removed population and it is impossible for the epidemic to die out.

t at which the gradient $\frac{dI}{dt}$ becomes negative. As $\frac{dI}{dt} = I(t)\left(\frac{\beta S(t)}{N} - \gamma\right)$ and $I(t) \geq 0$ at all time-points t, the gradient becomes negative when $\frac{\beta S(t)}{N} - \gamma < 0$. Thus the maximum number of infectious individuals occur at the first time-period t where (3) is satisfied.

$$S(t) < \frac{N\gamma}{\beta} \tag{3}$$

This result can be written as (4), in terms of the total population which is either infected or removed. This result is intuitively useful as it provides a value for what proportion of the total population need to be immune from the disease for “Herd-Immunity” to be reached and the rate of infection to decrease and die-out naturally. This can be used to set targets for vaccination rollouts. Although the standard SIR model is too basic for these sort of decisions, as it does not allow for β or γ to vary over time, it is useful for building the intuition which motivates more complex models.

$$\begin{aligned} & N - I(t) - R(t) < \frac{N\gamma}{\beta} \\ \iff \quad & N\left(1 - \frac{\gamma}{\beta}\right) > I(t) + R(t) \end{aligned} \tag{4}$$

An alterative way to intuitively understand β and γ is to consider that $1/\beta$ is the mean time between an infected individual infecting someone who is susceptible and $1/\gamma$ is the mean time before an infected individual becomes removed. This means that $\frac{1/\gamma}{1/\beta}$ is the mean number of people each infectious individual infects before being removed, assuming all individuals are susceptible. This is the definition of the basic reproduction number R_0 and by simplify we have a simple result (5) for the R_0 value of any standard SIR model.

$$R_0 = \frac{\beta}{\gamma} \tag{5}$$

As β and γ are constant with-respect-to time in the standard SIR model, the effective reproduction number R_t is simply the basic reproduction number R_0 at all time-points t. We can use (5) to restate the tipping point equation (4) in terms of the R_0 value.

$$N\left(1 - \frac{1}{R_0}\right) > I(t) + R(t) \tag{6}$$

Many inference problems focus around the parameter estimation. Given a realisation of the model $\{(S(t), I(t), R(t))\}_t$, it is straightforward to estimate these parameters using the difference equations given in (7).

$$\hat{\beta} = \frac{S(t_i) - S_{t_{i+1}}}{S(t_i)I(t_i)} \tag{7a}$$

$$\hat{\gamma} = \frac{R(t_{i+1}) - R(t_i)}{I(t_i)} \tag{7b}$$

SIR Model with Vaccinations

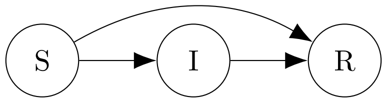


Figure 2.3: Diagram of interactions between compartments of an extended SIR model which models vaccinations.

As SIR models are used to model epidemic events, a common inferential questions is what effects introducing a vaccine to a population will have. For the SIR model it is only necessary to give vaccinations to members of the susceptible population as the removed population are already immune and the infectious population actively have the disease. This means than modelling a vaccination program only requires the inclusion of one extra interaction, from the susceptible population to the removed population ($S \to R$), as shown in *Figure 2.3*.

Typically this new interaction is implemented by moving a constant number of individuals α from the susceptible population to the removed population each time-period. (8) is an extension of (2) which encodes this interaction. Modelling this as a constant is reasonable for real world scenarios as most vaccination programs aim to vaccinate as many people as possible until the whole susceptible population has been vaccinated, and thus the number of daily vaccinations is typically very stable. Although in real-world scenarios vaccinations are often also given to members of the removed population too. This can be modelled by making α depend on the proportion of the population which is susceptible $\alpha(S(t)/N)$. It would be straightforward to define $\alpha(t)$ as a time-dependent function in order to represent variations in the capacity to vaccinate over time, accounting factors such as weekday vs. weekend variations in work patterns and increased vaccination capacity. This model assumes that vaccinations take effect immediately.

$$\frac{dS}{dt} = -\frac{\beta}{N} S(t)I(t) - \alpha \tag{8a}$$

$$\frac{dI}{dt} = \frac{\beta}{N} S(t)I(t) - \gamma I(t) \tag{8b}$$

$$\frac{dR}{dt} = \gamma I(t) + \alpha \tag{8c}$$

where α is the number of vaccinations administered each time-period.

SIR Model with Demography

The standard SIR model is very simple and forms a good basis from which to gain intuition for more complex models. A natural advancement is to incorporate births and natural deaths into the model. These processes are known collectively as "Demography". It is reasonable to ignore demography for many epidemic processes as the rate of infection and removal from the epidemic process is significantly greater than that the rates natural demography.

Demography easily extends beyond births and deaths to include immigration and emigration by simply considering immigration and births to be equivalent, and deaths and emigration to be equivalent. If we assume that immigrants cannot carry the disease then these processes are truly equivalent.

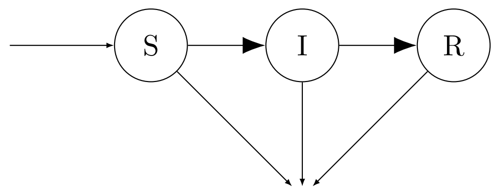


Figure 2.4: Diagram of interactions between compartments in an SIR model with demography.

Typically birth is not assumed to bring immunity to diseases, so each birth causes an increase in the size of the susceptible population. While deaths can occur to any individual, regardless of the compartment they are in, and are modelled as completely removing an individual from the population meaning each death leads to a decrease in the total population size. The specification of the two interactions from the standard SIR model are unchanged. *Figure 2.4* outlines these interactions. This is a good model for diseases, such as VZV (Chickenpox), which are endemic in a population but non-fatal.

Two new parameters are introduced for the encoding of demography: λ the number of births per time-period; and, μ the probability of an individual dying naturally in a given time-period. It is assumed that an individual is equally likely to die naturally, independent of which compartment they are in. The reciprocal of the death rate $1/\mu$ is the average life-span of each individual in the population. Defining λ and μ to be independent of the total population size is reasonable when the overall timespan of the epidemic is relatively short and when recovery from an infection rarely results in death, as the total population is almost constant. The three differential equations (9) extended (2) to model these new interactions [Rodrigues, 2016]:

$$\frac{dS}{dt} = \lambda - \frac{\beta}{N}S(t)I(t) - \mu N S(t) \tag{9a}$$
$$\frac{dI}{dt} = \frac{\beta}{N}S(t)I(t) - \gamma I(t) - \mu N I(t) \tag{9b}$$
$$\frac{dR}{dt} = \gamma I(t) - \mu N R(t) \tag{9c}$$

where β, γ are as defined in (2) and the same initial conditions are imposed.

(9) is very similar to (2) except each equation has a term which subtracts the number of natural deaths each time-period, which is proportional to the current population size of the associated compartment. And, (9a) has an additional term for introducing natural births.

For the same removal rate γ, the mean time an individual is infected for is reduced compared to an SIR model without demography due to the possibility of said individual dying naturally before they are removed. The mean time an individual is infected is now $1/(\gamma + \mu)$. This means the R_0 for an SIR model with demographics is (10).

$$R_0 = \frac{\beta}{\gamma + \mu} \tag{10}$$

The inclusion of demography in a model allows for two possible equilibria to occur: the disease dies out; or, for the disease to persist endemically in the population due to the constant influx of new susceptible individuals. *Remark 2.1* presents the events where each of these outcomes occurs, and the resulting equilibria. Both of these results hold as time tends to infinity.

Remark 2.1 (Equilibria of SIR model with Demography)
Note that $\lambda, \beta, \gamma, \mu \geq 0$. An equilibrium for the SIR model with demography is achieved when

$$\left(\frac{dS}{dt}, \frac{dI}{dt}, \frac{dR}{dt}\right) = (0, 0, 0)$$

There are two cases to consider: $R_0 \geq 1$; and, $R_0 < 1$. For each case, I derive values (S^*, I^*, R^*) for the population sizes which produce an equilibrium.

Case 1 - $R_0 \geq 1$.

$$\begin{array}{rrcl} & \frac{dI}{dt} &=& 0 \\ \implies & \frac{\beta}{N}S^*I^* - \gamma I^* - \mu I^* &=& 0 \\ \implies & \frac{\beta}{N}S^* - \gamma - \mu &=& 0 \\ \implies & S^* &=& \frac{N(\mu+\gamma)}{\beta} \\ & &=& \frac{N}{R_0} \end{array}$$

$$\begin{array}{rrcl} & \frac{dS}{dt} &=& 0 \\ \implies & \lambda - \frac{\beta}{N}S^*I^* - \mu S^* &=& 0 \\ \implies & \lambda - \frac{\beta I^* + \mu N}{R_0} &=& 0 \\ \implies & I^* &=& \frac{\lambda R_0 - \mu N}{\beta} \end{array}$$

$$\begin{array}{rrcl} & \frac{dR}{dt} &=& 0 \\ \implies & \gamma I^* - \mu R^* &=& 0 \\ \implies & R^* &=& \frac{\gamma}{\mu}I^* \\ & &=& \frac{\gamma\lambda R_0}{\mu\beta} - \frac{\gamma N}{\beta} \end{array}$$

Thus, when the $R_0 \geq 1$ an equilibrium is reached when the compartment populations fulfil

$$(S, I, R) = \left(\frac{N}{R_0},\ \frac{\lambda R_0 - \mu N}{\beta},\ \frac{\gamma\lambda R_0}{\mu\beta} - \frac{\gamma N}{\beta}\right)$$

This state is known as an "Endemic Equilibrium" as the disease maintains a constant level of infection.

Case 2 - $R_0 < 1$.

As $R_0 < 1$ the size of the infected population will eventually decrease to zero, after which point no new infections can occur ($I^* = 0$). This also means that the removed population can not increase after this point, moreover it will decrease to zero ($R^* = 0$) due to those in the population dying. This means the population equilibrium occurs when $\frac{dS}{dt} = 0$ and $I = 0$.

$$\begin{array}{rrcll} & \frac{dS}{dt} &=& 0 & \\ \implies & \lambda - \frac{\beta}{N}S^*I^* - \mu S^* &=& 0 & \\ \implies & \lambda - \mu S^* &=& 0 & \textit{since } I^* = 0 \\ \implies & S^* &=& \frac{\lambda}{\mu} & \end{array}$$

Thus, when the $R_0 \geq 1$ an equilibrium is reached when the compartment populations fulfil

$$(S, I, R) = \left(\frac{\lambda}{\mu}, 0, 0\right)$$

This state is known as a "Disease Free Equilibrium".

For some diseases immunity is passed from mother to child due to antibodies from the mother passing across the placenta. This means that a certain proportion of births are born with immunity, with the proportion typically being the ratio of the size of the removed population to the size of the rest of the population. It would be straightforward to model this by having these children be placed straight into the removed population. However it often makes more sense to create a separate compartment for these children, or to not add these births to the population at all, as many inferential questions focus around the total number of infections.

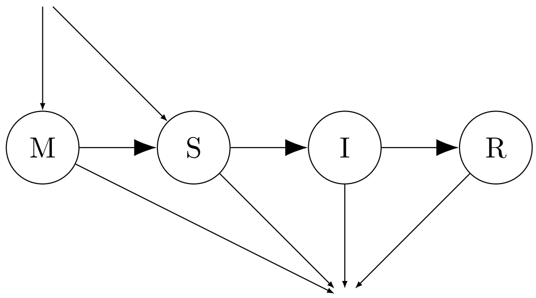


Figure 2.5: Diagram of interactions between compartments in an MSIR model. M=Maternally Derived Immunity, S=Susceptible, I=Infectious and R=Removed.

For some of these diseases, such as measles, the immunity passed from the mother is only temporary and wears off after a period of time. The MSIR[1] model [Hethcote, 2000] was introduced to model this scenario, see *Figure 2.5.* The MSIR model introduces a new compartment which is placed before the susceptible compartment, commonly referred to as "Maternally Derived Immunity" and denoted by an M. A proportion of new births enter the M compartment each time-period, while the rest enter the susceptible population. Members of the M compartment move to the susceptible compartment at a rate α, where $1/\alpha$ is the mean time between birth and immunity wearing off, or they die and are removed from the system.

SIR Model with Non-Constant Parameters

The standard SIR model has limited uses in practice mainly due to its assumptions of constant parameter values for β and γ, as well as the assumption that all interactions between individuals are equal. These assumptions may be reasonable over short periods of time where there are no changes in the efforts being made to suppress the spread of a disease[2], such as a single flu season. [Mahaffy, 2018] fits a standard SIR model to data from the 2004-05 flu season in American and shows how it produces a good fit of the data.

In the real-world the true values for the parameters β and γ do not exist due to noise which occurs from factors such as variability in human interactions, variability in individuals health and the weather. It is much more realistic to model these parameters using non-negative probability distributions and to then sample from these distributions each time-period. Due to the removal parameter γ needing to be constrained to $(0, 1)$ a Beta-distribution is a common choice. Alternatively, a distribution can be defined for its reciprocal $1/\gamma$ and then γ can be calculated after each sample.

Using distributions for model parameters does increase the complexity of analysing such models as β and γ now have their own hyper-parameters. However, much of the analysis above can be performed using the expected value of each parameter. Further, the use of distributions

[1]"Maternal Immunity-Susceptible-Infectious-Removed"

[2]This would likely be due to a disease not being particularly deadly

allows for further analysis into the uncertainty of a model and whether changes seen are due to changes in policy, or just random noise.

SIR Model with Time-Dependent Parameters

A common modelling problem is to model the effects of introducing a better treatment for those who are infected. This treatment would reduce the mean time each member of the infectious population is infected, and thus increase the value of the removal parameter γ. This is implemented into the SIR model by changing γ from being a constant to being a time-dependent step function of the form (11) where t' is the time-period at which the new treatment is implemented. Naturally, more steps can be added to model several advancements in treatment.

$$\gamma(t) = \begin{cases} \gamma_0 & t \leq t' \\ \gamma_1 & t > t' \end{cases} \tag{11}$$

The same extensions can be made to the infection parameter β when seeking to model the effects of public health policies which seek to control the rate of infection.

This concept can be generalised to $\beta(t)$ and $\gamma(t)$ being continuous time-dependent functions to account for other factors, such as weather. Implementing such a formulation is practically impossible in practice due to lack of available data and only approximate relationships known between these factors and the nature of the epidemic.

In (2) the parameters β and γ represent very general concepts: the average number of people infected by a single infectious individual and the probability of an infectious individual being removed in a single period, respectively. And, as such, it is reasonable to consider them as functions of other variables. For example, we could define β as (12a) the product of the mean number of interactions an individual has each time-period b and the mean probability of someone becoming infected after an interaction with an infectious individual c; And, γ as (12b) the empirical mean for time of an infection across different strains of the disease where $\mathbf{p}$ is the distribution of likelihood of each strain and $\mathbf{s}$ is the mean time of infection for each strain. The possible formulations are endless. Each formulation introduces more parameters which need to be fitted or observed, increasing the modelling difficulty.

$$\beta = f(b, c) = b \cdot c \tag{12a}$$

$$\gamma = g(\mathbf{p}, \mathbf{s}) = \mathbf{p}^T \mathbf{s} \tag{12b}$$

Being able to respecify these abstract parameters in terms of real world quantities helps us make a leap from correlation to causation. Helping ascertain the relative role of different real world events in the life-cycle of an epidemic.

2.3.2 Other Compartmental Models

The SIR models described so far cover a very narrow, although prevalent, set of diseases. Here I briefly describe a selection of alternative compartmental models. To avoid tedium I do not perform much analysis of these models, as it is broadly similar to that of the SIR model, nor do I state the system of differential equations which govern each model. It should be apparent that the extensions discussed for the SIR model (demography, vaccinations, maternally derived immunity etc.) are readily applicable to the below models.

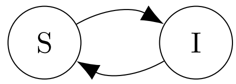


Figure 2.6: Diagram of interactions between compartments in an SIS model. S=Susceptible and I=Infectious.

The SIR model works on the assumption that recovering from an infection confers immunity from future infection on the individual. This is not always the case, especially for diseases with a high mutation rate such as influenza. Once someone has recovered from an infection from one of these diseases they should be returned to the susceptible group. This means the removed group can be removed completely from the model. This is how we reach the SIS[1] model where individuals only move from the susceptible to the infectious group ($S \to I$), and back again ($I \to S$). *Figure 2.6* represents these interactions.

It is straightforward to define differential equations (13) for the SIS model using those from the standard SIR model (2). These equations are subject to the restriction that $S(0) \geq 0,\ I(0) \geq 0$ and $S(0) + I(0) = N$. The R_0 value for an SIS model is calculated using the same formula (5) as for the standard SIR model without demography.

$$\frac{dS}{dt} = \gamma I(t) - \beta S(t)I(t) \tag{13a}$$

$$\frac{dI}{dt} = \beta S(t)I(t) - \gamma I(t) \tag{13b}$$

(13c)

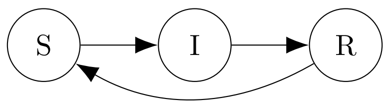


Figure 2.7: Diagram of interactions between compartments in an SIRS model.

A similar class of diseases are those where individuals loose immunity after some period of time. This can be modelled by extending the SIR with one additional dynamic, as individuals can now move from the removed compartment back to the susceptible compartment ($R \to S$). This is formalised as the SIRS[2] model, see *Figure 2.7*.

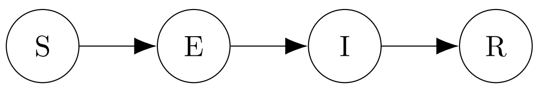


Figure 2.8: Diagram of interactions between compartments in an SEIR model. S=Susceptible, E=Exposed, I=Infectious and R=Removed.

For many diseases an individual does not become infectious to others the moment they become infected with the disease. Rather there is an incubation period where an individual has the disease but cannot yet infect others. In the real world this scenario causes problems for epidemiologist and policy makers as it is difficult to ascertain how many people have the disease. A problem which is compounded if individuals are asymptomatic during the incubation period. Alternatively, individuals may be highly symptomatic during the incubation period and thus naturally isolate themselves from the population before they before infectious, heavily reducing the spread of the disease.

[1]"Susceptible-Infectious-Susceptible"

[2]"Susceptible-Infectious-Removed-Susceptible".

The SEIR[1] model is a formalisation of this scenario and introduces a new compartment, Exposed (E), between the susceptible and infectious compartments of the standard SIR model. See *Figure 2.8.*

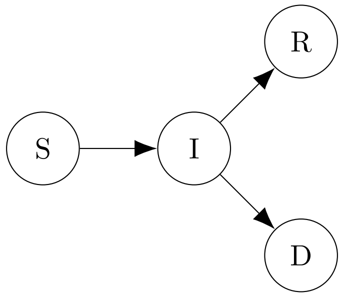


Figure 2.9: Diagram of interactions between compartments in an SIRD model. S=Susceptible, I=Infectious, R=Removed and D=Dead.

The final variation of the standard SIR model I will mention is the SIRD[2] model. This model is useful for inferences as it separates the two scenarios by which someone becomes no longer infectious: either recovering and thus gaining immunity to future infections; or, by succumbing to the disease and dying. This is implemented by adding an additional compartment, deceased (D), to the end of the model and creating a fork which allows for individuals who are infectious to move either the removed or deceased compartments. Different parameters are assigned for each of these dynamics. See *Figure 2.9*

Separating the removed population into these two classes is useful for inferential problems which concern the death rate of a disease. Public health policy makers should wish to implement policies which minimise death, and suffering, whilst also minimising the restrictions placed on the lives of individuals.

From these extensions to the standard SIR model, it is evident that the concept of the SIR model can be extended almost endlessly by adding more dynamics and compartments. It is reasonable to want to create compartments which group people by intrinsic characteristics (e.g. age, gender, geography etc.) as it is likely that the parameters governing the dynamics for how each group interacts with a disease will vary. Although these models may better represent a given disease it becomes very difficult to fit them due to their very high degrees of freedom. This is especially true when a population size is relatively small, as the sample size for each compartment will be even smaller, allowing random noise to dominate observations. And, these are issues which arise before considering impurity in the available data.

2.3.3 Stochastic SIR Model

Definition 2.2 (One-Dimensional Brownian Motion[a])
A stochastic process $\{W_t\}_{t\geq 0}$ is called Standard Brownian Motion if it fulfils the following four criteria

1. $W_0 = 0$, *almost surely.*
2. *Increments of W are independent:* $(W_{t+u} - W_t)$ *are independent of the filtration* $\mathcal{F}_t$ *for all* $t, u \geq 0$.

[1]"Susceptible-Exposed-Infectious-Removed"
[2]"Susceptible-Infectious-Removed-Deceased"

3. *Increments of W have a stationary Gaussian distributions:* $(W_{t+u} - W_t) \sim \mathcal{N}(0, u)$ *for all* $t, u \geq 0$.

4. W_t *is continuous with-respect-to* t.

[a] Also known as a Wiener Process as its existence was proved by Norbert Wiener.

Definition 2.3 (Itô Process)
A stochastic process $\{X(t)\}_{t \in [0,T]}$ *is called an Itô Process if it has the following form*

$$X(t) = X(0) + \int_0^t b(u, X(u))d_u + \int_0^t \sigma(u, X(u))dW_u$$

where b, σ *are functions and* W_t *is standard one-dimensional Brownian motion. Note that the first integral is a standard integral whilst the second is a stochastic integral. This form can be stated as the following differential equation*

$$dX(t) = b(t, X(t))dt + \sigma(t, X(t))dW_t$$

The compartmental models discussed in *Sections 2.3.1 & 2.3.2* are deterministic[1]. This is highly limiting in real-world applications where random noise is guaranteed to exist. [Maki and Hirose, 2013] present a formulation of the SIR model which replaces (2) with a system of equations (14) which include a stochastic differential equation (14b).

$$dS = -\frac{\beta}{N}S(t)I(t)dt \tag{14a}$$

$$dI = \left(\frac{\beta}{N}S(t)I(t) - \gamma I(t)\right)dt + \sigma(t, I(t))dW_t \tag{14b}$$

$$R = N - I(t) - S(t) \tag{14c}$$

where W_t is a one-dimensional Brownian Motion and $\sigma(t, I(t)) = \alpha I(t)$ for some diffusion constant $\alpha \in \mathbb{R}$. This system of equations is subject to the initial conditions $I(0) > 0$ and $S(0) \geq 0$.

The system of equations defined in (14) is very similar to (2) for the standard SIR model. Moreover, the equations (2a,14a) for $\frac{dS}{dt}$ are identical and the definition of (14c) is necessary to ensure the population size is constant. The only differences occurs in (14b) which, due to the inclusion of the stochastic term $\sigma(t, I(t))dW_t$, is now an Itô Process with $b(t, I(t)) = \beta S(t)I(t) - \gamma I(t)$. This is the equation which makes this system of equations stochastic.

(14) has one additional parameter, the diffusion parameter α, which needs to be estimated when compared to (2). The parameters β and γ can be estimated using difference equations (7) as for the standard SIR model. [Maki and Hirose, 2013] recommends the use of (15) as an estimator for the diffusion parameter α. This result follows from the quadratic variance of (14b)

$$\hat{\alpha}^2 = \frac{\sum_i \left(I(t_{i+1}) - I(t_i)\right)^2}{\sum_i \left(t_{i+1} - t_i\right) I(t_i)^2} \tag{15}$$

[1] Disregarding the brief mention of using probability distributions in place of constants.

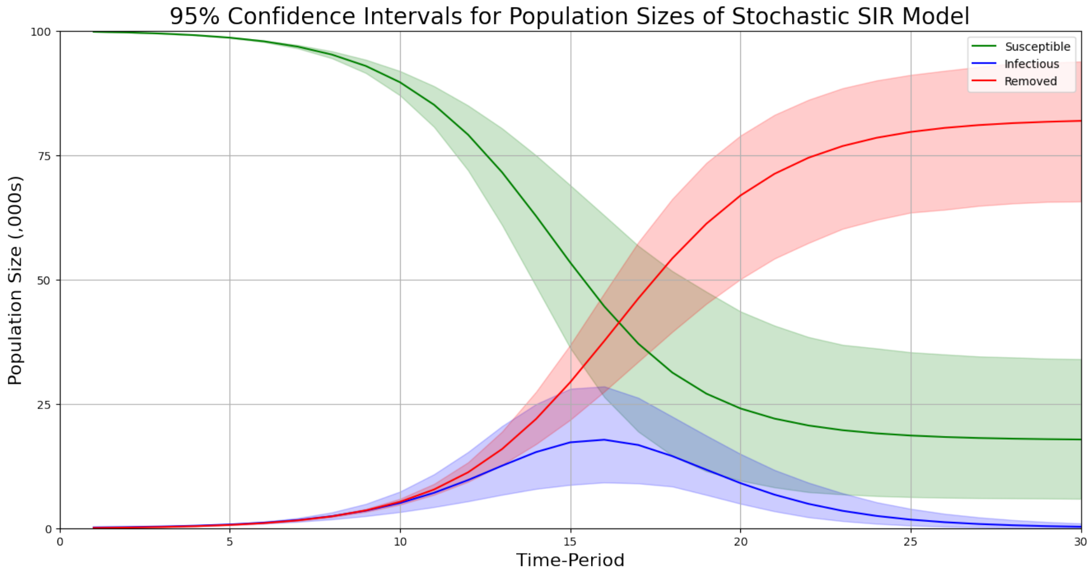


Figure 2.10: 95% credible intervals for the population size of the different compartments of a Stochastic SIR model with population size $N = 100,000$ over 30 time-periods where $\beta = 1, \gamma = .0.5$ and $\alpha = 0.1$. ($R_0 = 2$)

Figure 2.10 provides a plot of an example Stochastic SIR model. The example model use has the same parameters as the standard SIR model depicted in *Figure 2.2* except with the diffusion parameter $\alpha = 0.1$. Notice how the mean for each series is identical to the realisation in *Figure 2.2*.

A numerical evaluation of the deterministic system of differential equations (2) is straightforward as there is only one set of possible outcomes. Whereas, a numerical evaluation of (14) is not quite so straightforward. The Euler-Maruyama formula [Kayode *et al.*, 2016] provides a straightforward method for calculating an approximate numerical solution (16) to a stochastic differential equation such as (14b).

$$\begin{array}{rrl} & I(t_{i+1}) & = I(t_i) + \left\{\frac{\beta}{N}S(t_i)I(t_i) - \gamma I(t_i)\right\}\Delta t_i + \alpha I(t_i)\Delta W_i \\ \text{where} & \Delta t_i & := t_{i+1} - t_i \\ & \Delta W_i & := W_{t_{i+1}} - W_{t_i} \end{array} \tag{16}$$

Note that $\Delta W_i \sim \mathcal{N}(0, t_{i+1} - t_i)$ by the definition of Brownian motion having stationary Gaussian increments, while Δt_i can be treated as constant when assuming observations occur at regular intervals in time. We can therefore deduce the expectation and variance of $I(t_{i+1})$ given a filtration up to time-period t_i $\mathcal{F}_{t_i}$[1], as shown below.

$$\begin{aligned} \mathbb{E}\left[I(t_{i+1})|\mathcal{F}_{t_i}\right] &= \mathbb{E}\left[I(t_i) + \left\{\frac{\beta}{N}S(t_i)I(t_i) - \gamma I(t_i)\right\}\Delta t_i + \alpha I(t_i)\Delta W_i\right] \\ &= I(t_i) + \left\{\frac{\beta}{N}S(t_i)I(t_i) - \gamma I(t_i)\right\}\Delta t_i + \alpha I(t_i)\mathbb{E}\left[\Delta W_i\right] \\ &= I(t_i) + \left\{\frac{\beta}{N}S(t_i)I(t_i) - \gamma I(t_i)\right\}\Delta t_i \end{aligned}$$

$$\begin{aligned} \text{Var}\left[I(t_{i+1})|\mathcal{F}_{t_i}\right] &= \text{Var}\left[\alpha I(t_i)\Delta W_i|\mathcal{F}_{t_i}\right] \\ &= (\alpha I(t_i))^2\text{Var}[\Delta W_i] \\ &= (\alpha I(t_i))^2(t_{i+1} - t_i) \end{aligned}$$

By linearity, we can deduce the distribution of $I(t_{i+1})$ given a filtration $\mathcal{F}_{t_i}$ to be (17).

[1]i.e. We know the values of $\{(S(t), I(t), R(T))\}_{t\in[0,t_i]}$.

This distribution allows us to calculate credible intervals for the size of the population at each time-period.

$$(I_{t_{i+1}}|\mathcal{F}_{t_i}) \sim \mathcal{N}\left(I(t_i) + \left\{\frac{\beta}{N}S(t_i)I(t_i) - \gamma I(t_i)\right\}\Delta t_i,\ (\alpha I(t_i))^2(t_{i+1} - t_i)\right) \tag{17}$$

This shows that the expected result of (16) is identical to the guaranteed value of (2b) at every time-point. This is an important result as it shows (14) describes the same scenarios as (2), but with stochastic variation included. This means much of the analysis performed on the standard deterministic SIR model in *Section 2.3.1* holds for the expected results of the stochastic SIR model discussed hear.

2.4 Beyond Compartmental Models

The compartmental models discussed in *Section 2.3.2* are powerful models when only population-level data is available. This data is typically readily available for epidemics; although the accuracy of the data will depend on the reliability and pervasiveness of the testing strategy being used, especially for diseases where infectious people can be asymptomatic. In the modern world we now have access to large amounts of individual-level data which concern human behaviours and interactions, namely mobility and personal network data. These data sources are likely useful in modelling an epidemic. It is not obvious how this data could be incorporated in a compartmental model, so we look to other models.

Agent Based Models

Agent-Based models seek to capture the unique behaviour of each individuals in a population (referred to as “agents”) by separately parameterising each individual and specifying a set of rules which define the behaviours of agents. For epidemics, agent-based models typically have three main components: a model of agent movements; a model of agent-to-agent interactions; and, a model of the disease which provides a probability of an interaction resulting in an infection.

There are a number of sources for movement data, [Frias-Martinez *et al.*, 2011] use cell phone data in an agent-based model to model the 2009 H1N1 (Swine-Flu) outbreak in Mexico. These approaches generally work best in urban areas there is greater density of cell towers and thus a more accurate location can be triangulated for each individual.

Models of agent-to-agent generally rely on sources such as censuses to determine the close contacts of individuals (e.g. family members and colleagues). These are the individuals an infectious person is most likely to pass the disease onto, or have caught the disease off. This data is used in tandem with the movement data as two close contacts who are in the same area at the same time are likely to interact with each other.

Modelling of the disease generally accounts for factors such as age, sex, general health, weather and location (i.e. indoor or outdoor) to determine the probability that an interaction results in an infection. Close contacts are will have a greater probability as it is more likely they will have a longer interaction, such as a conversation rather than passing in the street.

The high complexity of these models is clearly a limitation as it is much harder to estimate parameter values. Moreover, running a simulation of an agent-based method is significantly more computationally intensive than doing so for the compartmental models discussed above. Further, it is very difficult to assess the validity of an agent-based model due to their complexity.

3 Approximate Bayesian Computation

In this section I: Motivate and provide the mathematical background for Approximate Bayesian Computation (ABC) methods *Section 3.1*; Present the general approach of ABC methods *Section 3.2* and discuss four flavours of ABC algorithm *Section 3.2.1-3.2.4*; and, close this section by exploring how ABC methods can be used for model choice *Section 3.3* and how regression adjustment can be used to improve the results of ABC methods *Section 3.4*.

3.1 Motivation and Background

Consider a model X with parameters θ. The centre-point of Bayesian inference is the posterior distribution $\mathbb{P}(\theta|X)$ for the parameters θ given observations of X. Using Bayes' Rule we can formulate this posterior as (18).

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta)\mathbb{P}(\theta)}{\mathbb{P}(X)} \tag{18}$$

For Bayesian inference we are only concerned with the relative weight the posterior assigns to each parameter value θ, so we can discard the evidence $\mathbb{P}(X)$ as it is just a normalising constant with-respect-to θ. Meaning we can simplify the expression for the posterior as being proportional to the product of the likelihood $\mathbb{P}(X|\theta)$ and the prior $\mathbb{P}(\theta)$ (19).

$$\mathbb{P}(\theta|X) \propto \mathbb{P}(X|\theta)\mathbb{P}(\theta) \tag{19}$$

As the prior is defined by the user, the only remaining task is to deduce an expression for the likelihood. However, for most real-world processes an explicit expression of the likelihood is computationally intractable due to the complex nature of the systems which govern them and their high degrees of freedom. Moreover, there are often so many parameters that it is intractable to specify all of them and thus we generally theorise a simpler model $\hat{X}$ and seek to calibrate this model to the true model by fitting its parameters. This motivates the need for likelihood-free inference methods such as the Approximate Bayesian Computation methods.

Suppose you have a sequence of n of observations $x_{obs} := (x_{obs,1}, \dots, x_{obs,n})$ from our model X where each observation may be multi-dimensional, $x_{obs,i} \in \mathbb{R}^p$ for $p \in \mathbb{N}$. Let $K_\varepsilon(\cdot)$ denote a kernel density function with bandwidth $\varepsilon \geq 0$ and $\|\cdot\|$ denote a distance measure between observations of model X. I discuss kernel density functions and distance measures in *Section 3.2*.

As the bandwidth tends to zero the value of the kernel density function for the distance between two points $K_\varepsilon(\|x - x_{obs}\|)$ tends to the Dirac delta function $\delta_{x_{obs}}(x)$. This result (20) follows trivially from the definition of a kernel density function.

$$\lim_{\varepsilon \to 0} K_\varepsilon(\|x - x_{obs}\|) = \delta_{x_{obs}}(x) := \begin{cases} 1 & \text{if } x_{obs} = x \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

This result can be used to restate the likelihood function in terms of a kernel density function and distance measure as (21).

$$
\begin{aligned}
\mathbb{P}(x_{obs}|\theta) &= \int \delta_{x_{obs}}(x)\mathbb{P}(x|\theta)dx \\
&= \lim_{\varepsilon\to 0}\int K_\varepsilon(\|x - x_{obs}\|)\mathbb{P}(x|\theta)dx
\end{aligned}
\tag{21}
$$

Consider the definition (22) of π_{ABC} and note that it tends to, within a normalising constant, of the true posterior (23).

$$
\pi_{ABC}(\theta|x_{obs}) := \int K_\varepsilon(\|x - x_{obs}\|)\mathbb{P}(x|\theta)\pi_0(\theta)dx \tag{22}
$$

$$
\begin{aligned}
\lim_{\varepsilon\to 0}\pi_{ABC}(\theta|x_{obs}) &= \lim_{\varepsilon\to 0}\int K_\varepsilon(\|x - x_{obs}\|)\mathbb{P}(x|\theta)\pi_0(\theta)dx \\
&= \int \delta_{x_{obs}}(x)\mathbb{P}(x|\theta)dx \cdot \pi_0(\theta) \\
&= \mathbb{P}(x_{obs}|\theta)\pi_0(\theta) \\
&\propto \mathbb{P}(\theta|x_{obs})
\end{aligned}
\tag{23}
$$

(23) shows that π_{ABC} is an approximation of the true posterior, with it being a good approximation when ε is small.

Typically, due to the observations x being of high dimension, a summary statistic $s(\cdot)$ is applied to them first and then the quantities $s := s(x)$ and $s_{obs} := (x_{obs})$ are used in place of x and x_{obs} in (22), respectively. The analysis of (23) is unchanged when using summary statistics as long as the summary statistics are sufficient, if the summary statistics are not sufficient then π_{ABC} can only ever be an approximation of the true posterior regardless of the bandwidth used. *Section 4* is dedicated to the topic of how to approach summary statistic selection, with sufficiency being discussed in *Section 4.2*.

$$
\pi_{ABC}(\theta|s_{obs}) := \int K_\varepsilon(\|s - s_{obs}\|)\mathbb{P}(s|\theta)\pi_0(\theta)ds \tag{24}
$$

The formulation (24) for the ABC approximation of the posterior is the one found in the standard ABC framework (e.g. [Sisson *et al.*, 2018; Beaumont, 2019]).

The utility of being able to use (24) to approximate the true posterior is apparent when you consider the implied joint distribution of parameters and summary statistics $\pi_{ABC}(\theta, s|s_{obs})$, as shown in (25).

$$
\begin{aligned}
\pi_{ABC}(\theta|s_{obs}) &= \int \pi_{ABC}(\theta, s|s_{obs})ds \\
\text{where}\quad \pi_{ABC}(\theta, s|s_{obs}) &:= K_\varepsilon(\|s - s_{obs}\|)\mathbb{P}(s|\theta)\pi_0(\theta)
\end{aligned}
\tag{25}
$$

We can define Monte Carlo algorithms which target sampling from joint distribution 25 without needing to specify the likelihood $\mathbb{P}(s|\theta)$. These samples become samples from the posterior by simply ignoring the summary statistic values s which are sampled.

3.2 ABC Methods

Approximate Bayesian Computation (ABC) methods are a family of computational methods which can be used to approximate posteriors for the parameters of models where the likelihood is intractable. This is achieved by simulating from the likelihood, rather than evaluating it explicitly.

The first algorithm to use the concept which would later became known as ABC was presented in [Tavaré *et al.*, 1997], although this algorithm does not include the use of summary statistics, nor does it use a distance measures and kernel density functions to determine whether to accept a simulation. The algorithm presented in [Pritchard *et al.*, 1999] is much more recognisable as an ABC method and is consider by many as the first true ABC algorithm. This algorithm would later be generalised to becomes the Rejection Sampling approach to ABC (See *Section 3.2.1*). Both of these papers were studies of population genetics, a field in which ABC methods are still popularly used.

The key feature that simulation based methods exploit is that we only know the response values x_{obs} from the true model, but for each simulation we know both the parameter values and the response values $(\tilde{\theta}, \tilde{x})$. Thus we can inspect the parameter values for accepted simulations and draw inferences about the parameter values of the true model. Moreover, it is generally easier to simulate from a model than to reconstruct it.

Algorithm 3.1 (Generic Approximate Bayesian Computation)
Require: *Observed values x_{obs}; Summary statistics $s(\cdot)$; Theorised model $f(X|\cdot)$; Acceptance Kernel $K_\varepsilon(\cdot)$; Distance Measure $\|\cdot\|$; Parameter Priors $\pi_0(\theta)$.*

1. *Calculate summary statistic values $s_{obs} = s(x_{obs})$.*
2. *Until stopping condition reached:*
   (a) *Sample a set of parameters $\tilde{\theta}$.*
   (b) *Run the theorised model with sampled parameter $\tilde{x} = f\tilde{\theta}(X|\tilde{\theta})$.*
   (c) *Calculate summary statistic values $\tilde{s} = s(\tilde{x})$.*
   (d) *Accepted parameters $\tilde{\theta}$ with probability $K_\varepsilon(\|\tilde{s} - s_{obs}\|)$.*
3. *Return all accepted parameter sets $\hat{\Theta}$.*

The central concept for all ABC methods is that the likelihood function can be approximated by comparing simulated observations to observations from a true model. ABC methods require a set of observations from the true model x_{obs}; a theorised model f for which parameters θ can be set and observations generated; and a set of priors π_0 for the parameters of the theorised model. ABC methods then generate many simulations of the theorised model and, by comparing the summary statistic values of the simulated observations to those of the true observations, inferences are made about which parameter values are most likely to be closely represent the true parameter values. **Algorithm 3.1** outlines this basic flow which ABC methods follow. The general idea being that the parameter sets which cause the theorised model to generate observations which are closest to true observations are more likely to be the true parameter values.

Algorithm 3.1 demonstrates the simplicity of the underlying algorithm for ABC methods. Most ABC methods are straightforward to implement as they follow this basic structure and then change how certain parts of performed in practice (Typically how new samples are drawn and how the acceptance criteria are defined). This allows for a high level of modularity which has motivated innovations in ABC methods.

There are two sources of approximation in the standard ABC algorithm: The use of summary statistics; and, using a bandwidth on the acceptance criteria. The first source can be removed by using sufficient summary statistics (See *Section 4.2*). The second is eliminated if the bandwidth is set to zero $\varepsilon = 0$, but in general this leads to the algorithms becoming intractable.

The ideal ABC methods are those which run efficiently and perform well with small bandwidths ε. Computationally efficient methods are important as more simulations can be processed in a given time-period, making convergence of the estimated posterior more likely. A method being able to handle smaller bandwidths means the posterior it produces will be a better approximation of the true posterior (See result (20)). All ABC methods will run with any value of the bandwidth, however those that use an informed search method for generating samples (e.g. ABC-SMC) require fewer simulations to achieve good estimates of the posterior.

Monte Carlo methods are a family of algorithms which use repeated random simulations to evaluate a model. These form the basis of how ABC methods approach exploring the parameter space. Monte Carlo methods are a class of methods which seek to generate samples from a space in a way which mimics sampling from the true model. They do this by running many, many simulations and use some degree of randomness to determine how each simulation is generated and which are accepted.

Here is an overview of classes of Monte Carlo methods which are commonly used in ABC methods:

- *Rejection-Sampling methods* calculate a probability p that a given set of simulated values came from the true model. A value $u \sim U[0,1]$ is sampled from standard uniform distribution and if the sampled value u is less than the acceptance-probability p then the simulation is accepted as a sample. This procedure is run on a large number of simulations with each simulation being generated and assessed independently.

- *Importance-Sampling methods* extend rejection-sampling by, instead of only accepting a subset of simulated values; all simulations are accepted, but each is assigned a weight which indicates the perceived probability that that simulation could be generated by the true model. Typically this weight is proportional to the acceptance probability p calculated by rejection-sampling methods.

- *Markov Chain Monte Carlo (MCMC) methods* extend rejection-sampling by, instead of generating each simulation independently, the parameters of the last accepted simulation are slightly perturbed and then used to generate a new simulation. This creates a search process rather than random simulation due to the dependency between consecutive samples.

- *Sequential Monte Carlo (SMC) methods*[1] extend importance-sampling by repeatedly resampling from the same set of samples, with the weights of each parameter determining the probability it is sampled. Each iteration the acceptance criteria are tightened. This means the estimated posterior will become more concentrated each iteration.

The use of Monte Carlo methods means that ABC methods are inherently computationally inefficient due to the need to perform many random simulations. This inefficiency means ABC methods perform badly for models which generate a lot of data as it takes longer to assess each simulation. In the most extreme cases, part of this data needs to be omitted for computational efficiency which adds another layer of approximation. Being able to increase the acceptance rate of simulations means less simulations are required and thus more complex models can be assessed. ABC-MCMC generally achieves the greatest acceptance rates for a given bandwidth.

Monte Carlo methods introduce a high degree of randomness into ABC methods which further motivates the need to perform lots of simulations as the strong law of large number is required to obtain consistent results. This limitation is mitigated due to the simplicity of most ABC algorithms meaning they are capable of processing millions of simulations an hour on modern computers. Moreover, many Monte-Carlo methods are readily parallelisable due to the

[1] Also known as Particle-Filter methods.

indepenence of each simulation. This improves computation time when the algorithms are run on super-computers with many parallel processes.

The set of accepted parameters $\hat{\Theta}$ returned by ABC methods can be used for Bayesian inference. Estimating properties of the distributions, such as the mean, mode and quantiles, is straightforward. Producing a discretised estimate of the posterior for each parameter can be achieved by calculating a histogram of the accepted values for each parameter, again straightforward. Kernel density functions can be used to produce a continuous estimates of the posteriors (See [Zambom and Dias, 2012]).

Remark 3.1 (Posterior Mean is Minimum Mean-Square Error Estimator)
Let θ denote the quantity we wish to estimate, A denote an arbitrary estimator of θ and suppose we have observed x_{obs} from model X. Then

$$\begin{array}{rrcl} & MSE_\theta(A) & = & \mathbb{E}\left[(\theta - A)^2|X = x_{obs}\right] \\ & & = & \mathbb{E}\left[\theta^2 - 2A\theta + A^2|X = x_{obs}\right] \\ & & = & \mathbb{E}\left[\theta^2|X = x_{obs}\right] - 2A\mathbb{E}\left[\theta|X = x_{obs}\right] + A^2 \\ \implies & \frac{\partial}{\partial A}MSE_\theta(A) & = & -2\mathbb{E}[\theta|X = x_{obs}] + 2A \\ \implies & A & = & \mathbb{E}[\theta|X = x_{obs}] \end{array}$$

This shows that mean-square error is minimised when the posterior mean of θ given x_{obs} is used as an estimator.

ABC methods are commonly used to calibrate or compare models. Typically calibration is done by setting parameter values to the estimated posterior mean since the posterior mean minimises mean-square error (see **Remark 3.1**). ABC methods are used for model comparison as they can directly estimate Bayes factor, I discuss model comparison further in *Section 3.3*.

The key advantage of ABC methods, over other approaches to Bayesian inference, is that it produces a distribution, rather than a point-estimate, for parameter values. This allows for analysis into uncertainty around the parameter values. Additionally, as the strictness of the acceptance criteria is a parameter of ABC methods, ABC methods can fit or compare a large range of theorised models by loosening the acceptance criteria. Being able to use simpler models has the advantage of reducing issues which occur due to curse-of-dimensionality.

A limitation of using ABC methods is the large number of hyper-parameters they have (Distance measures, summary statistics, bandwidths, perturbation kernels, etc.) and that the choices the user makes for how these parameters are set can drastically effect the algorithm's performance. It is trivial to realise that if an uninformative distance measure (such as $\|x\| = 0\ \forall\ x$) , or an acceptance kernel which accepts all simulations, is used then the returned set of parameters will resemble their priors, and no meaningful inferences can be drawn. Moreover, these hyper-parameters need to be tuned for each model these methods are applied to, which is laborious. This has motivated the innovation of adaptive ABC algorithms which automate the setting of some of these parameters.

As stochastic processes determine whether a simulation is accepted, or not, ABC methods incur information loss. This can mean that promising areas of the parameter space are not explored. This issue is mitigate by increasing the number of simulations assessed by the method, which will require more computing time.

Summary Statistics

See *Section 4*.

Kernel Density Functions

Definition 3.1 (Kernel Density Functions $K_\varepsilon(\cdot)$, Epanechnikov [1969])
Kernel density functions are functions $K_\varepsilon : \mathbb{R} \to [0,1]$ with a single smoothing parameter ε geq0 and the following properties:

1. *Non-negative*
$$K_\varepsilon(x) \geq 0 \ \forall \ x \in \mathcal{X}$$
where $\mathcal{X}$ is the range of values x can take.

2. *Symmetric*
$$K_\varepsilon(x) = K_\varepsilon(-x) \ \forall \ x \in \mathcal{X}$$

3. *Normalised*
$$\int_{\mathcal{X}} K_\varepsilon(x)dx = 1$$

4. $K_\varepsilon(x) = \frac{1}{\varepsilon}K_1(x/\varepsilon)$.

The choice of kernel density function does not play a notable role in the asymptotic behaviour of ABC methods, however the bandwidth chosen for them does. A high bandwidth means that the weight of the kernel is spread much more evenly across its support, meaning there is less discrimination between values close to the mean and those further away.

It is standard to define kernel density functions such that they have zero mean. This property means that that $\max_x K_\varepsilon(x) = K_\varepsilon(0)$, following immediately from the kernel being symmetric. This is a useful property in the context of ABC methods as we pass the distance between two points $\|x - x_{obs}\|$ to the kernel density function to determine the probability we accept a simulation and this property means that simulations x closest to the observed values x_{obs} are more likely to be accepted.

In practice, when implementing ABC methods we typically scale up the values returned by the kernel such that $K_\varepsilon(0) = 1$. This is straightforward to do for well-known kernels as it only requires the discarding of the normalising term. As the relative weights given to each value are maintained this does not affect the asymptotic behaviour of the algorithms, but will increase the acceptance rate. This also has the desirable effect that every time an exact match is found it will definitely be accepted.

Name	**Formula**
Uniform Kernel	$K_\varepsilon(x) = \dfrac{1}{2\varepsilon}\mathbb{1}\{x \leq \varepsilon\}$
Gaussian Kernel	$K_\varepsilon(x) = \dfrac{\varepsilon}{\sqrt{2\pi}} \exp\left\{-\dfrac{1}{2}x^2\varepsilon^2\right\}$
Epanechnikov Kernel	$K_\varepsilon(x) = \dfrac{3}{4}\left(1 - x^2\varepsilon^2\right)\mathbb{1}\left\{\lvert x\rvert \leq \varepsilon\right\}$

Table 3.1: Definitions of common kernel density functions for ABC methods. $\mathbb{1}\{A\} := \begin{cases} 1 & \text{if } A \\ 0 & \text{otherwise} \end{cases}$

Table 3.1 provides a table of the most commonly used kernel density functions for ABC, as recommended by [Beaumont, 2019]. The Epanechnikov kernel is asymptotically optimal for kernel density estimation when seeking to minimise mean-square error (See [Epanechnikov, 1969]), although this theoretical result is disputed in [Tsybakov, 2008].

The uniform kernel is popular in ABC as it is equivalent to accepting all simulations whose distance from the true observation is no greater than ε. This creates spherical acceptance regions when using the Euclidean distance or rectangular ones when using Manhattan distance. The Gaussian kernel is more commonly used as it has an infinite support.

For ABC methods, when choosing which kernel to use, it is intuitive to match the kernel to the theorised distribution of the noise in the theorised model. This further motivates the popularity of a gaussian kernel as many models assume gaussian noise.

Distance Measures

Distance measures quantify how far apart two multi-dimensional vectors are from each other, with greater values indicating the vectors are further away. A value of zero means that the two vectors are identical under the given measure.

Name	**Formula**
Manhattan Distance	$L_1(\mathbf{x}) := \sum_{i=1}^{m} \lvert x_i \rvert$
Euclidean Distance	$L_2(\mathbf{x}) := \sqrt{\sum_{i=1}^{m} x_i^2}$
L_p Norm	$L_p(\cdot) := \left(\sum_{i=1}^{m} x_i^p\right)^{1/p}$
L_∞ Norm	$L_\infty(\mathbf{x}) := \max\{x \in \mathbf{x}\}$

Table 3.2: Definitions of common distance measures for ABC methods.

The choice of distance measure is integral to the performance of an ABC method as it determines whether a set of simulated observations are deemed to be representative of the true model by quantifying how similar they are to true observations. **Table 3.2** provides a list of popular distance measures for ABC methods. The Euclidean distance is most commonly used as minimising Euclidean distance is clearly related to minimising the sum of squares error (SSE) of a model.

It is important to note that the value passed to distances measures in ABC methods is the difference of two sets of summary statistics, not the difference of two observations. Well chosen summary statistics will extract meaningful information and perform a certain level of pre-processing of this data, such as standardisation and weighting different dimensions. This means we do not need to consider these problems during selection of distance measure.

An issue which arises when specifying distance measures is the "Curse of Dimensionality"[1]. This is the phenomenon that as the dimensionality of vectors being compared increases, it becomes harder to distinguish between different pairs of vectors. This is an issue for ABC methods as the success of these methods relies on being able to accurately identify which simulations most closely resemble observed data. Using summary statistics which introduce a high level of dimensionality reduction will help.

Different demonstrations of the "Curse of Dimensionality" phenomenon are required for different distance measures, but for Euclidean distance it is generally demonstrated by comparing of the volume of a hyper-sphere with radius r and the volume of a hyper-cube with side length $2r$. The volume of the hyper-cube quickly dwarfs that of the hyper-sphere as the number of dimensions are increased.

[1]Term first coined in [Bellman, 1961] in reference to how many algorithms may work well when applied to low-dimensions, but are intractable when for higher-dimensions.

The “Curse of Dimensionality” is a big short-coming of the Euclidean distance as it was only every conceived for real-world spaces (i.e. two or three dimensions). Which distance measure is best ultimately depends on the data being used. Using the L_p norm has shown promise, but adds the additional problem of which value of p is optimal. [Schnitzer *et al.*, 2014] present an approach to choosing an optimal p which assesses the “hub-ness” of a dataset.

There is a wealth of literature on the “Curse of Dimensionality” in the machine learning space, particularly for nearest-neighbour problems which are very relevant to the problem being addressed by distance measures in ABC (See [Beyer *et al.*, 1999; Hinneburg *et al.*, 2000; Radovanovic; *et al.*, 2010])

3.2.1 ABC-Rejection Sampling

ABC-Rejection Sampling is a generalisation of the sampling algorithms presented in [Tavaré *et al.*, 1997; Pritchard *et al.*, 1999]. The general idea is to keep simulating from the theorised model until a predefined number of simulations M have been accepted by the acceptance kernel. Each simulation involves: sampling a set of parameters $\tilde{\theta}$ from the predefined parameter priors $\pi_0(\theta)$; initialising the theorised model $f(X|\tilde{\theta})$ with the sampled parameters; Observing values $\tilde{x}$ from the initialised model; and then comparing these simulated observations to the true observations using summary statistics s, a distance measure $\|\cdot\|$ and an acceptance kernel $K_\varepsilon(\cdot)$. This approach is stated formally in **Algorithm 3.2**.

ABC-Rejection Sampling is most suitable in problems where it is believed that the posterior is not very different from the defined priors, as this algorithm has very limited search capabilities. These are typically problems which have already been studied heavily and a general understanding of the priors is known.

Algorithm 3.2 (ABC-Rejection Sampling “Fixed Sample Size”)
Adapted from [Beaumont et al.*, 2002].*

require: *Observed values x_{obs}; Summary statistics $s(\cdot)$; Theorised model $f(X|\cdot)$; Prior Distributions $\pi_0(\theta)$; Acceptance Kernel $K_\varepsilon(\cdot)$; Distance Measure $\|\cdot\|$; Target Number M.*

```
s_obs ← s(x_obs).
Θ̃ ← {}.
t ← 0.
while t < M do
    θ̃_t ← sample π_0(θ).
    x̃ ← f(X|θ̃_t).
    s̃ ← s(x̃).
    with probability K_ε(‖s_obs − s̃‖)
        θ̂^(t) ← θ̃.
        Add θ̂^t to θ̂.
        t ← t + 1
    otherwise Pass;
return Θ̃ = {θ^(1), ..., θ^(M)}
```

The approach in **Algorithm 3.2** is intuitive and straightforward to implement. A limitation of this simplicity is that each simulation is completely independent and no “learning” is incorporated from information about which parameter sets have previously been accepted, or rejected. However, this independence does mean that it is straightforward to implement **Algorithm 3.2** in a parellisable fashion, allowing for more simulations to be analysed in a given time-period.

A practical difficulty in using **Algorithm 3.2** is in setting the bandwidth of the acceptance kernel as the bandwidth dictates the acceptance rate. A lower acceptance rate means more meaningful inferences can be drawn from the set of accepted parameters, as the average distance between accepted simulations and the true observations will be lower. However, there is no clear relationship between bandwidth and the acceptance rate of the algorithm so run-times are near-impossible to predict to any meaningful degree of accuracy.

Algorithm 3.3 (ABC-Rejection Sampling "Best Samples")

require: *Observed values* x_{obs}*; Summary statistics* $s(\cdot)$*; Theorised model* $f(X|\cdot)$*; Prior Distributions* $\pi_0(\theta)$ *Distance Measure* $\|\cdot\|$*; Number of Simulations* M*; Simulations to Accepted* N*.*

```
s_obs ← s(x_obs).
Θ̃ ← {}.
t ← 0.
for i = 0, ..., M do
    θ̃^(i) ← sample π0(θ).
    x̃^(i) ← f(X|θ̃^(i)).
    s̃^(i) ← s(x̃^(i)).
    d^(i) ← ||s(x̃^(i)) − s_obs||.
    Add (d^(i), θ̃^(i)) to Θ̃.
return N elements with smallest distance values d^(i).
```

The problem of setting a bandwidth (and acceptance kernel) can be avoid completely by instead running a fixed number M of simulations and then accepting a predefined number N of these simulations which are closest to the observed values. This enforces an acceptance rate to be N/M. Additionally, this removes the risk of the algorithm running indefinitely. This approach is outlined in **Algorithm 3.3**. This algorithm runs in linear time with-respect-to the number of simulations $O(M)$, as the set of simulated values $\tilde{\Theta}$ is unsorted with-respect-to the distance values $d^{(i)}$ and thus finding the N^{th} order-statistic takes linear time.

The space requirements for **Algorithm 3.3** grow linearly with-respect-to the number of simulations being run $O(M)$. This creates a practical limit of the number of simulations which can be run. The space requirements can be reduced to $O(N)$ (grow linearly with-respect-to the number of accepted simulations) by, instead of storing all simulations in the set $\tilde{\Theta}$, we instead only store the N closest. This requires keeping the set $\tilde{\Theta}$ ordered and thus increases the time complexity of the algorithm to $O(M \log_2 N)$.

There is no requirement for the number of simulations to be generated, nor the number of simulations to accept, to be defined before running the algorithm. Rather the algorithm can be allowed to run and evaluated simulations until a time limit is reached. Then either a predefined proportion of the simulations can be accepted, or the distribution of distances can be inspected to choose an appropriate acceptance rate. The disadvantage of this approach is that it has high space requirements due to the need to store the distance and parameters for all simulations until the very end of the algorithm. This creates a cap on how long this version of the algorithm could be run, but this can be mitigated by dropping the very worst simulations (among other mitigation approaches).

Equal weight is assigned to each of the parameters in the set of accepted parameters returned by ABC-Rejection Sampling. A natural extension is to place greater weight on parameters which produce observations which are closer to those produced by the true model. [Beaumont *et al.*, 2002] proposes a technique of weighted local-linear regression to adjust parameter values by weights which are determined by the distance value associated with the parameter set.

Rejection Sampling techniques can be used to estimate the probability of a given set of observations under different models $f(X|M)$. This is useful in model choice as it immediately leads to estimations of Bayes' factor. I discuss this further in *Section 3.3*.

3.2.2 ABC-Importance Sampling

Importance Sampling methods use a tractable distribution to sample from an intractable distribution. Importance Sampling is an exact method as, given enough iterations, it will always converge on the target distribution in an unbiased fashion. The theory behind Importance Sampling is presented in **Remark 3.2**.

Remark 3.2 (Importance Sampling)
Let X be a model $X \sim h(X;\theta)$ with parameters θ. Consider two distributions for parameter values $f(\theta)$ and $g(\theta)$ where f is a target distribution, from which sampling is intractable, and g is a distribution we can sample from. Then the expected value of X under distribution f is the same as the expected value of $X \cdot w(\theta)$ under distribution g, where $w(\theta) := \frac{f(\theta)}{g(\theta)}$ is a weight measure.

$$\begin{aligned}
\mathbb{E}_f[X] &= \int h(X;\theta)f(\theta)d\theta \\
&= \int h(X;\theta)f(\theta)\frac{g(\theta)}{g(\theta)}d\theta \\
&= \int h(X;\theta)g(\theta)\frac{f(\theta)}{g(\theta)}d\theta \\
&= \int h(X;\theta)g(\theta)w(\theta)d\theta \text{ where } w(\theta) := \frac{f(\theta)}{g(\theta)} \\
&= \mathbb{E}_g[h(X;\theta)w(\theta)]
\end{aligned}$$

This means we can estimate the expected value of the model under f by weighting samples from g using $w(\theta)$. This is the likelihood ratio of an observation coming from the two models.

In ABC context the target distribution is the approximate joint distribution posterior $f = \pi_{ABC}(\theta, s|s_{obs})$ and the distribution we can sample from is the joint distribution of summary statistics and parameters under the importance distribution $g(s,\theta) = p(s|\theta)g(\theta)$. The importance weighting $\tilde{w}$ for each simulation in ABC is derived as (26).

$$\begin{aligned}
\frac{\pi_{ABC}(\theta, s|s_{obs})}{g(\theta, s)} &\propto \frac{K_\varepsilon\left(\|s - s_{obs}\|\right)p(s|\theta)\pi_0(\theta)}{p(s|\theta)g(\theta)} \\
&= \frac{K_\varepsilon\left(\|s - s_{obs}\|\right)\pi_0(\theta)}{g(\theta)} \\
&=: \tilde{w}
\end{aligned} \tag{26}$$

ABC-Importance Sampling should be used in cases where we have a good idea of what the distribution for the posterior will be so that the parameter priors π_0 and importance distribution g are informative.

Algorithm 3.4 (ABC-Importance Sampling)
Adapted from [Fan and Sisson, 2018].

require: *Observed values* x_{obs}*; Summary statistics* $s(\cdot)$*; Theorised model* $f(X|\cdot)$*; Prior Distributions* $\pi_0(\theta)$ *Distance Measure* $\|\cdot\|$*; Number of Simulations* M*; Importance Kernel* $g(\cdot)$*.*

1 $s_{obs} \leftarrow s(x_{obs})$.

2 $\tilde{\Theta} \leftarrow \{\}$.

3 **for** $i = 0, \dots, M$ **do**

4 $\quad \tilde{\theta}^{(i)} \leftarrow$ *sample* $g(\theta)$.

5 $\quad \tilde{x}^{(i)} \leftarrow f(X|\tilde{\theta}^{(i)})$.

6 $\quad \tilde{s}^{(i)} \leftarrow s\left(\tilde{x}^{(i)}\right)$.

7 $\quad \tilde{w}^{(i)} \leftarrow \frac{\pi_0\left(\theta^{(i)}\right)}{g\left(\theta^{(i)}\right)} K_\varepsilon\left(\left\|s^{(i)} - s_{obs}\right\|\right)$.

8 $\quad$ *Add* $\tilde{\theta}^{(i)}$ *to* $\tilde{\Theta}$ *with weight* $\tilde{w}^{(i)}$.

9 **return** $\tilde{\Theta} := \left\{\left(\tilde{\theta}^{(1)}, \tilde{w}^{(1)}\right), \dots, \left(\tilde{\theta}^{(M)}, \tilde{w}^{(M)}\right)\right\}$

An Importance Sampling approach to ABC is an extension of the Rejection Sampling approach which replaces calculating acceptance probabilities with calculating importance weightings for each simulation. All simulations are accepted and their importance weight is used to weight them during Bayesian inference. An acceptance kernel K_ε and distance measure $\|\cdot\|$ still need to be specified as they are required to calculate the importance weights. This approach is given in **Algorithm 3.4**.

Similarly to ABC-Rejection Sampling, **Algorithm 3.4** is straightforward to implement in a parallelisable fashion due to the independence of each simulation. However, it is much less space efficient than the ABC-Rejection Sampling approaches as it requires the storage of every simulation. This is mitigated by an approach presented by [Fearnhead and Prangle, 2011] which combines the rejection and Importance Sampling approaches to ABC. I discuss this approach more further down.

The approach to ABC-Importance Sampling given in **Algorithm 3.4** requires the specification of priors $\pi_0(\theta)$ and an importance distribution $g(\theta)$. If these distributions are the same, or are proportional to each other, then $\frac{\pi_0(\theta)}{g(\theta)} \approx 1 \ \forall \ \theta$ meaning the acceptance probability $K_\varepsilon(\|s - s_{obs}\|)$ is the only factor weighting each simulation.

An issue with all sampling approaches which weight their results is that it is possible for a small subset of accepted samples to dominate the weight space. This can lead to results becoming unstable. This can naturally be tackled by increasing the number of simulations, this is inefficient and does not inform us as to when a sufficient number of simulations have been made. The Effective Sample Size (ESS) (27) is a useful metric in these cases as it quantifies how many equally weighted samples our set of parameters is equivalent to. The stopping condition of the algorithm should be updated such that the algorithm terminates once the effective sample size of the accepted set of parameters $\tilde{\Theta}$ has reached some threshold N.

$$ESS := \frac{\sum_{i=0}^{M} w^{(i)}}{\sum_{i=0}^{M} \left(w^{(i)}\right)^2} \tag{27}$$

where $(w^{(0)}, \dots, w^{(M)})$ is the weights assigned to each simulation.

[Fearnhead and Prangle, 2011] propose an algorithm which combines ABC-Rejection Sampling and ABC-Importance Sampling by, rather than accepting every simulation (Line 8), each simulation is accepted with probability $K_\varepsilon(\|s - s_{obs}\|)$ and is assigned weight $\tilde{w} = \pi_0(\theta)/g(\theta)$. This reduces the accepted set of simulations to only those that produce reasonably similar observations as the model. This improves the effective sample size of the set of accepted parameters as fewer simulations are given very small weights, and is more space efficient than **Algorithm 3.4** as it does not require every simulation to be stored.

3.2.3 ABC-MCMC

Definition 3.2 (Markov Chain)
A Markov Chain is a Stochastic Process $\{X_t\}_t$ with the Markov Property. This means that the current state of the process solely depends on its state in the time-period immediately before.

$$\mathbb{P}(X_{t+1}|X_t, \ldots, X_1) = \mathbb{P}(X_{t+1}|X_t)$$

The transitions a Markov Chain can make can be summarised in a square matrix P_t, known as the "Transition Matrix", where $[P_t]_{ij} = \mathbb{P}(X_{t+1} = j|X_t = i)$. The transition matrix can be time invariant.

A Markov chain is said to be "irreducible" if it is possible to go from any state to any other state, in some finite period of time.

$$\exists\ n\ st\ \mathbb{P}(X_{t+n} = x|X_t = y) > 0\ \forall\ x, y$$

The Stationary Distribution of a Markov Chain is a probability distribution π which is invariant under a time-invariant transition matrix P.

$$\pi = \pi P$$

The stationary distribution represents the asymptotic proportion of time the chain spends in each state. The stationary distribution is unique if the Markov chain is irreducible.

Markov Chains (**Definition 3.2**) are sequences of events where the probability of which event occurs next only depends on the current event. When targeting a probability distribution the transition matrix for a Markov chain will be stationary, this means it will have a stationary distribution which can be approximated. Many models of epidemic processes have the Markov property, including the deterministic compartmental models discussed in *Section 2*. Most notably SIR models do as each set of values only depend on the number of members in each group in the previous time period.

Markov Chain Monte Carlo (MCMC) methods are sampling methods which exploit Markov chains in order to perform a more informed search procedure through the parameter space. The Markov chain is used to determine which set of parameters to simulate with next, with the next choice being dependent upon the most recently accepted set of parameters. An acceptance step, similar to ABC-Rejection Sampling, is then used to evaluate the simulated observations against the true model observations and thus whether to accept the new set of parameters. The distribution of accepted parameter sets is an approximation of the stationary distribution of the Markov chain, and thus of the target distribution. These algorithms are ideally run until the distribution of accepted samples satisfies some convergence criteria, although in practice it is more practical to stop the algorithm once the chain has reached a given length.

This more informed search procedure has the advantage of increasing the acceptance rate of simulations. This is due to the reduced variation between simulated observations, compared to when parameters are chosen truely randomly. In ABC methods we harness this advantage by creating stricter acceptance criteria, improving the level of approximation.

A popular class of MCMC algorithms are Metropolis-Hastings algorithms [Metropolis *et al.*, 1953; Hastings, 1970] which seeks to produce a Markov chain whose stationary distribution is unique and thus converges on the target distribution (The parameter posterior in the case of ABC methods). This approach requires the specification of a perturbation kernel $K^*(\theta)$ which generates a new set of parameters by slightly perturbing a given set of parameters. The perturbance kernel needs to be implemented in such a way that the probability (28) of generating a given set of parameters θ' given the input θ is calculatable.

$$\mathbb{P}\left(K^*(\theta) = \theta'\right) \tag{28}$$

A simple set of perturbation kernels are those which apply additive gaussian noise to the input (29), the variance on the noise is a hyper-parameter which would require tuning. More complex perturbation kernels consider the correlation between parameters and then nudge correlated parameters in the same/opposite direction. Fisher Information can be incorporated into perturbation kernels in order to determine which parameters have a greater effect and thus should be explored more. [Filippi *et al.*, 2012] explores selecting perturbation kernels for ABC-SMC but many of the themes are relevant to ABC-MCMC too.

$$K^*(\theta) = \theta + \mathcal{N}(0, \sigma_0^2) \text{ for some } \sigma_0^2 \geq 0 \tag{29}$$

As each sample is drawn using the previously accepted sample, there is dependence between samples leaving MCMC methods open to auto-correlation issues. Auto-Correlation is a measure of correlation between the current value of a sample and its previous values. Auto-Correlation can be reduced by increasing the size of steps the perturbation kernel is expected to produce but this will have adverse effects on the acceptance rate. Auto-Correlation can be particularly high if the chain becomes stuck in a region where there is very concentrated probability mass as it will struggle to escape. The problem with auto-correlation is that most analysis assumes that parameters are independent, but the presence of high auto-correlation can contradict this assumption.

A limitation of MCMC methods is that they are only able to search one region of the sample space at any given time and they struggle to move between disjoint areas of high density. In the context of Bayesian inference, this causes an issue when wishing to model multi-modal distributions as MCMC will typically only be able to find one of the modes. The solution to this is to run multiple chains at once and then to merge their results. This does, however, require greater computational resources and typically means that each chain is made shorter to compensate.

MCMC methods have limited scope for being parallelised as each iteration depends on the previous iteration. If multiple chains are being run, then they can be parallelised.

Algorithm 3.5 (ABC-MCMC)
Adapted from [Marjoram et al.*, 2003].*

require: *Observed values* x_{obs}; *Summary statistics* $s(\cdot)$; *Theorised model* $f(X|\cdot)$;
Prior Distributions $\pi_0(\theta)$ *Distance Measure* $\|\cdot\|$; *Chain length* M;
Acceptance Kernel $K_\varepsilon(\cdot)$; *Perturbation Kernel* $K^*(\cdot)$.

1 $s_{obs} \leftarrow s(x_{obs})$.
2 $\tilde{\Theta} \leftarrow \{\}$.
3 # *Burn-In Step*
4 **while** $K_\varepsilon\left(\|\tilde{s}^{(0)} - s_{obs}\|\right)$ *is not accepted* **do**
5     $\tilde{\theta}_0 \leftarrow$ *sample* $\pi_0(\theta)$.
6     $\tilde{x}^{(0)} \leftarrow f(X|\tilde{\theta}^{(0)})$.
7     $\tilde{s}^{(0)} \leftarrow s\left(\tilde{x}^{(0)}\right)$
8 # *MCMC Step*
9 **for** $t = 1, \dots, M$ **do**
10     $\theta^* \leftarrow K^*\left(\tilde{\theta}^{(t-1)}\right)$.
11     $x^* \leftarrow f\left(X|\theta^*\right)$.
12     $s^* \leftarrow s\left(x^*\right)$.
13   **with probability** $\min\left\{1, \frac{K_\varepsilon(\|s^*-s_{obs}\|)\pi(\theta^*)\mathbb{P}\left(K^*(\tilde{\theta}^{(t-1)})=\theta^*\right)}{K_\varepsilon(\|s^{(t-1)}-s_{obs}\|)\pi(\theta^{(t-1)})\mathbb{P}\left(K^*(\theta^*)=\tilde{\theta}^{(t-1)}\right)}\right\}$
14         $\tilde{\theta}^{(t)} \leftarrow \theta^*$.
15         $s^{(t)} \leftarrow s^*$.
16   **otherwise**
17         $\tilde{\theta}^{(t)} \leftarrow \tilde{\theta}^{(t-1)}$.
18         $s^{(t)} \leftarrow s^{(t-1)}$.
19   *Add* $\tilde{\theta}^{(t)}$ *to* $\tilde{\Theta}$.
20 **return** $\tilde{\Theta} := \left\{\tilde{\theta}^{(1)}, \dots, \tilde{\theta}^{(M)}\right\}$

[Marjoram *et al.*, 2003] presents the first ABC method to have an MCMC approach, using the popular Metropolis-Hastings. **Algorithm 3.5** presents their algorithm. This approaches has two main stages: An initial burn-in (Lines 4-7) where random sets of parameters are evaluated until one is found which is accepted by the same acceptance criteria used in ABC-Rejection Sampling; and, the MCMC step (Lines 9-19) which starts at the first accepted parameter set $\tilde{\theta}_0$ and proceeds to generate new parameter sets θ^* by perturbing the last accepted parameter set. These new parameters sets are then used to simulate observations x^* and are accepted with probability (30)[1].

$$\min\left\{1, \frac{K_\varepsilon(\|s^* - s_{obs}\|)\pi(\theta^*)\mathbb{P}\left(K^*(\tilde{\theta}^{(t-1)}) = \theta^*\right)}{K_\varepsilon(\|s^{(t-1)} - s_{obs}\|)\pi(\theta^{(t-1)})\mathbb{P}\left(K^*(\theta^*) = \tilde{\theta}^{(t-1)}\right)}\right\} \tag{30}$$

The approach in **Algorithm 3.5** chooses to stop the MCMC step after a set number of iterations. This is not a good choice as it does not consider whether the stationary distribution of the Markov chain has converged. There are a few empirical methods which can be implemented to assess convergence. [Gelman and Rubin, 1992] propose running multiple chains, with different starting locations, and assessing the ratio of intra-chain to inter-chain variance for each parameter. When this ratio is close to one then convergence has been achieved. This method is not always practical due to its requirement for multiple chains and in practice we often choose to run the algorithm until some time-limit is reached.

[1]This probability is known as the "Metropolis Acceptance Ratio" and was derived so that the stationary distribution of the Markov chain will converge on the target distribution.

The burn-in period (Lines 4-7) is equivalent to running ABC-Rejection Sampling until the first set of parameters is accepted. Thus it is liable to running for an indeterminant amount of time (potentially indefinitely). The solution to this problem is the same as for ABC-Rejection Sampling: run a fixed number of simulations and choose the best one. This approach can be extended to automate the setting of the bandwidth used in the MCMC step (Lines 9-19), which can otherwise be a difficult task during tuning. The burn-in period is a crucial part of the algorithm as, if the Markov chain does not start in an area of high posterior density, then the rest of the algorithm will perform very badly. It is often necessary to run multiple burn in simultaneously in order to chose a more informed starting location.

Algorithm 3.5 can be made more adaptable by having it actively update the perturbation kernel K^* to maintain a target acceptance rate. In the case of an additive gaussian noise kernel, increasing the variance will lead to a decrease in acceptance rate as it is more likely that large steps will be taken. The acceptance rate can also be managed by adaptively setting the bandwidth on the acceptance kernel used in calculating (30).

The acceptance rate of an MCMC methods controls the rate of convergence, with both too high and too low values leading to slow convergence. An ideal acceptance rate will achieve a good level of mixing so that the parameter space is explored efficiently. It was shown in [Gelman *et al.*, 1997] that the asymptotically optimal acceptance rate for a Metropolis-Hasting sampler is 0.234, as the number of dimensions tends to infinity, when the target distribution is Gaussian. This result does rely on each dimension being independent identically distribution gaussian distributions, which is not always reasonable. Studies into some more general in-homogeneous target distributions have also shown 0.234 to be the asymptotically optimal acceptance rate (See [Roberts and Rosenthal, 2001]) but a general result has yet to be found. This research does motivate the use of adaptive MCMC methods which target an acceptance rate of 23.4%.

Due to its more informed search procedure, ABC-MCMC significantly outperforms ABC-Rejection and Importance Sampling in cases where the prior and posterior are very different. This makes ABC-MCMC a better choice in cases were informative priors are not known. However, the ABC-MCMC approach performs very poorly with mixtures models, which are becoming increasingly popular, due to the "Label Switching Problem" [Jasra *et al.*, 2005]. The "Label Switching Problem" occurs when two, or more, parameters are nonidentifiable when assigned the same priors[1] and thus the estimated posteriors produced for them will coverge on a combination of all of their true posteriors. [Jasra *et al.*, 2005] explores the "Label Switching Problem".

3.2.4 ABC-SMC

Sequential Monte Carlo (SMC) methods[2] approximate a probability distribution by collecting an initial sample which creates a rough approximation of the distribution; and then iteratively refining this approximation by resampling under ever tighter acceptance criteria (Referred to as improving the "Resolution" of the approximation). The acceptance criteria are tightened by defining a set of bandwidths $\{\varepsilon_0, \dots, \varepsilon_T\}$ such that $\varepsilon_0 \geq \dots \geq \varepsilon_T$ and iterating through this set to determine the bandwidth used in each resampling step. A extension of this approach is to incorporate Importance Sampling such that the resampling step also involves reweighting accepted samples. This extension is known as Population Monte Carlo (PMC).

The main advantage of SMC methods is that they iteratively make their acceptance criteria stricter. This is ideal for problems where it is hard to predict a good set of acceptance criteria beforehand. There is still an issue of having to define a set of bandwidths $\{\varepsilon_0, \dots, \varepsilon_T\}$ to be

[1] In a Gaussian mixtures model with two mixtures, the parameters associated with each mean can be swapped without affecting the fit of the model. This means that under identical priors it is impossible to separate these two parameters.

[2] Originally coined Particle Filters in [Del Moral, 1997].

used, towards the end of this subsection I discuss how this can be automated using an adaptive version of ABC-SMC.

SMC methods are susceptible to "Loss of Opportunity". This phenomenon occurs when part of the parameter space is not included in one of the sample sets, as this means that part of the parameter space can never be sampled from in the future. This mainly occurs to regions of the parameter space where little probability mass is placed, but can occur to denser areas if the sample size is too small. This issue can never be eliminated, except for very simple distributions, due to the practical limits on the sample size but can be mitigated by increasing the sample size.

[Sisson *et al.*, 2007] presents the first SMC approach to ABC, but this approaches produces a biased approximation of the posterior, mainly due to it underestimating the tails of the distributions caused by how they originally proposed to evaluate the likelihood ratio. [Beaumont *et al.*, 2009] presents an SMC approach to ABC which incorporates Importance Sampling and an optimised adaptive strategy. This is the version of ABC-SMC I discuss in this section.

Algorithm 3.6 (ABC-SMC)
Adapted from [Beaumont et al.*, 2009].*

require: *Observed values* x_{obs}*; Summary statistics* $s(\cdot)$*; Theorised model* $f(X|\cdot)$*; Prior Distributions* $\pi_0(\theta)$*; Distance Measure* $\|\cdot\|$*; Acceptance Kernel* $K_\varepsilon(\cdot)$*; Set of Bandwidths* $\{\varepsilon_0, \dots, \varepsilon_T\}$*; Number of Iterations* T*; Sample Size* N*.*

$s_{obs} \leftarrow s(x_{obs})$.

 # Initial Sample Step

$\tilde{\Theta}_0 \leftarrow \{\}$.

$i \leftarrow 0$

while $i < N$ **do**

  $\tilde{\theta}_0^{(i)} \leftarrow$ *sample* $\pi_0(\theta)$.

  $\tilde{x}_0^{(i)} \leftarrow f(X|\tilde{\theta}_t^{(i)})$.

  $\tilde{s}_0^{(i)} \leftarrow s\left(\tilde{x}_t^{(i)}\right)$.

  with probability $K_{\varepsilon_0}\left(\left\|\tilde{s}_0^{(i)} - s_{obs}\right\|\right)$

    $w_0^{(i)} \leftarrow \dfrac{1}{N}$.

    Add $\tilde{\theta}_0^{(i)}$ *to* $\tilde{\Theta}_0$ *with weight* $\tilde{w}_0^{(i)}$.

    $i \leftarrow i + 1$

  otherwise *Pass;*

 # Resampling Step

for $T = 1, \dots, T$ **do**

  $\sigma_{t-1}^2 \leftarrow$*Sample variance of each parameter dimension in* $\tilde{\Theta}_{t-1}$.

  $K^* \leftarrow$ *Normal*$(\theta, 2 \cdot \sigma_{t-1}^2)$.

  $\tilde{\Theta}_t \leftarrow \{\}$.

  $i \leftarrow 0$

  while $i < N$ **do**

    $\tilde{\theta}_t^{(i)} \leftarrow$ *sample* $\tilde{\Theta}_{t-1}$.

    $\theta^* \leftarrow K_t^*\left(\tilde{\theta}_t^{(i)}\right)$.

    $\tilde{x}_t^{(i)} \leftarrow f(X|\theta^*)$.

    $\tilde{s}_t^{(i)} \leftarrow s\left(\tilde{x}_t^{(i)}\right)$.

    with probability $K_{\varepsilon_t}\left(\left\|\tilde{s}_t^{(i)} - s_{obs}\right\|\right)$

      $\tilde{\theta}_t^{(i)} \leftarrow \theta^*$.

      $\tilde{w}_t^{(i)} \leftarrow \dfrac{\pi_0\left(\tilde{\theta}_t^{(i)}\right)}{\sum_{j=1}^{N} w_{t-1}^{(j)} \mathbb{P}\left(K_t^*\left(\tilde{\theta}_{t-1}^{(j)}\right) = \tilde{\theta}_t^{(i)}\right)}$.

      Add $\tilde{\theta}_t^{(i)}$ *to* $\tilde{\Theta}_t$ *with weight* $\tilde{w}_t^{(i)}$.

      $i \leftarrow i + 1$.

    otherwise *Pass;*

  # Normalise Weights

  for $i = 1, \dots, N$ **do**

    $w_t^{(i)} \leftarrow \dfrac{\tilde{w}_t^{(i)}}{\sum_{i=1}^{N} \tilde{w}_t^{(i)}}$.

    Update weight of $\tilde{\theta}_t^{(i)}$ *in* $\tilde{\Theta}_t$ *to be* $w_t^{(i)}$.

return $\tilde{\Theta}_T := \left\{\left(\tilde{\theta}_T^{(1)}, w_T^{(1)}\right), \dots, \left(\tilde{\theta}_T^{(N)}, w_T^{(N)}\right)\right\}$

Algorithm 3.6 is the algorithm presented in [Beaumont *et al.*, 2009]. This algorithm has two phases: First, (Lines 3-13) generating an initial sample of parameters $\tilde{\Theta}_0$ of size N using standard ABC-Rejection Sampling methods. Each sample is assigned the same importance weight $1/N$; And, second, the *Resampling Step* (Lines 15-34). The resampling step involves resampling from the previously set of accepted parameter samples $\tilde{\Theta}_{t-1}$ with the probability of sampling each parameter being equal to its importance weight. Each sample $\tilde{\theta}$ is perturbed using a perturbation kernel K^* to generate a new set of parameters θ^*. The new parameter set θ^* is used to simulate a set of observations to which summary statistic are applied $\tilde{s}$ and a Rejection Sampling step is used to accepted the new parameter set with probability $K_{\varepsilon_t}(\|\tilde{s} - s_{obs}\|)$. Note that the acceptance criteria are tightened each iteration. Each accepted parameter set is assigned an importance weight $\tilde{w}$. The importance weights are normalised after each resampling phase so that they sum to one and thus represent a probability distribution, which is important for sampling from this set.

The importance weight $\tilde{w}_t^{(i)}$ assigned in Line 27 is calculated as (31) and is the prior probability of the accepted parameter set divided by the probability of that parameter set under the posterior $\hat{\pi}_t$ generated by the previous step. This is just the standard importance weighting of the likelihood ratio. Note that each resampling step is aiming to produce a more refined version of the posterior distribution generated by the previous step, and thus the previous distribution is the target distribution and the prior is the originally proposed distribution.

$$\tilde{w}_t^{(i)} := \frac{\pi_0\left(\tilde{\theta}_t^{(i)}\right)}{\hat{\pi}_t\left(\theta_t^{(i)}\right)} \text{ where } \hat{\pi}_t\left(\theta_t^{(i)}\right) = \sum_{j=1}^{N} w_{t-1}^{(j)} \mathbb{P}\left(K_t^*\left(\tilde{\theta}_{t-1}^{(j)}\right) = \tilde{\theta}_t^{(i)}\right) \tag{31}$$

The adaptive feature of **Algorithm 3.6** is the setting of the perturbance kernel K^*. The perturbance kernel used in **Algorithm 3.6** is a component-wise random walk kernel which perturbs each component of a sampled parameter set independently by adding additive Gaussian noise to them. The variance of this Gaussian noise is equal to twice the component-wise sample variance (32) of the accepted samples from the previous iteration.

$$\left[\sigma_{t-1}^2\right]_i = \frac{1}{N-1} \sum_{j=1}^{N} \left(\left[\tilde{\theta}_{t-1}^{(j)}\right]_i - \left[\bar{\theta}_{t-1}\right]_i\right)^2 \tag{32}$$

where $\bar{\theta}_{t-1}$ is the sample mean of the previous set of accepted samples.

Using a component-wise random walk kernel is ideal for an adaptive algorithm as it is straightforward to implement and is computationally efficient as simple closed-form expressions are known for the probabilities required to calculate the importance weight for each accepted parameter set.

The variance is set to twice the sample variance of the previously accepted set as this minimises the Kullback-Leibler divergence between the target distribution and proposed distribution for the component-wise random walk kernel being used [Beaumont *et al.*, 2009]. Minimising Kullback-Leibler divergence means the two distributions are increasingly similar. See [Filippi *et al.*, 2012] for discussion of other optimal perturbation kernels for ABC-SMC.

The calculation of the importance weight (31) for each accepted parameter set, during resampling, requires summing over all the parameter sets from the previously accepted sample set. This means the resample stage takes $O(N^2)$ time and thus the overall run time of the algorithm is $O(TN^2)$ where T is the number of resampling iterations and N is the sample size. In practice the run-time of the algorithm will be dominated by assessing and generating samples, as a high proportion will be rejected, rather than by calculating the weight for each accepted set.

Each resampling step is dependent on the previous step as it requires the previous set of accepted samples $\tilde{\Theta}_{t-1}$ in order to generate samples. This means that this part of the algorithm cannot be parallelised. However, the simulations within each resampling step can be parallelised. As well as the initial sample generation step, as discussed in *Section 3.2.1*.

Algorithm 3.6 requires the specification of a set of bandwidths $\{\varepsilon_0, \dots, \varepsilon_T\}$. This can be difficult to do in an informed way, and would rather be avoided. Firstly, it is important to note that it is not strictly necessary for the algorithm to use the whole set of bandwidths and rather the algorithm can be stopped after it has reached a certain level of convergence (or number of simulations). Further, there is no need to define a full set of bandwidths at the start of the algorithm, instead an initial bandwidth ε_0 can be defined and then future bandwidths are chosen adaptively such that a target percentage $\Delta\%$ of previously sampled parameters would be accepted. Implementing this is straightforward for most standard acceptance kernels. For example, if a uniform acceptance kernel is being used it simply requires setting the next bandwidth to be the Δ^{th} percentile distance among the previously accepted parameter sets.

The need to set the initial bandwidth can be removed too by simply accepting all simulations into the initial sample set $\tilde{\Theta}_0$. This would make the algorithm significantly more inefficient as the initial sample with simple resemble the prior. Further, unless the sample size is very large there will be a high level of "Loss of Opportunity". A better approach would be to use the "Best Samples" variation of ABC-Rejection Sampling (**Algorithm 3.3**).

Incorporating this adaptive approach to bandwidth selection removes the need to define a set of bandwidths $\{\varepsilon_0, \dots, \varepsilon_T\}$ or the number of iterations T; and replaces them with defining an acceptance rate and number of simulations for the "Initial Sample Step"; a target acceptance rate between resampling iterations Δ, and a maximum number of simulations. These are significantly easier hyperparameters to define as their effects are much more apparent and predictable.

3.3 ABC for Model Choice

Definition 3.3 (Bayes Factor, Kass and Raftery [1995])
Consider two models M_1 and M_2, and some observed data x_{obs}. The Bayes Factor $B_{1,2}$ for data x_{obs} coming from model M_1 rather than model M_2 is the ratio of the likelihoods of x_{obs} coming from M_1 rather than M_2.

$$B_{1,2} := \frac{\mathbb{P}(x_{obs}|M_1)}{\mathbb{P}(x_{obs}|M_2)}$$

[Jeffreys, 1961] gives a qualitative assessment of Bayes Factor: "1 to 3 is barely worth a mention, 3 to 10 is substantial, 10 to 30 is strong, 30 to 100 is very strong and over a 100 is decisive evidence in favour of model M_1. Values below 1 take the inverted interpretation in favour of model M_2."

Bayes Factor (**Definition 3.3**) is a metric used to determine which of two models is more likely to have generated some observed data x_{obs}. Bayes Factor can be restated in terms of posteriors, using Bayes' Rule, as (33).

$$B_{1,2}(x_{obs}) := \frac{\mathbb{P}(x_{obs}|M_1)}{\mathbb{P}(x_{obs}|M_2)} = \frac{\frac{\mathbb{P}(x_{obs})\mathbb{P}(M_1|x_{obs})}{\mathbb{P}(M_1)}}{\frac{\mathbb{P}(x_{obs})\mathbb{P}(M_2|x_{obs})}{\mathbb{P}(M_2)}} = \frac{\mathbb{P}(M_2)\mathbb{P}(M_1|x_{obs})}{\mathbb{P}(M_1)\mathbb{P}(M_2|x_{obs})} \tag{33}$$

where $\mathbb{P}(M_i)$ is the prior weight given to model M_i.

It is generally reasonable to assume equal prior likelihood for both models. Under this assumption Bayes factor is the same as the posterior ratio (34) which is readily estimatable from ABC methods as the ratio of probabilities that the models generate x_{obs}.

$$B_{1,2}(x_{obs}) = \frac{\mathbb{P}(M_1|x_{obs})}{\mathbb{P}(M_2|x_{obs})} \tag{34}$$

Algorithm 3.7 (ABC Model Choice "Rejection Sampling")
Adapted from [Grelaud et al.*, 2009].*

require: *Observed values x_{obs}; Summary statistics $s(\cdot)$; Model Priors $\pi_M(M)$; Theorised models $M_1(X|\theta_{M_1})$ & $M_2(X|\theta_{M_2})$; Parameter Priors $\pi_{M_1}(\theta_{M_1})$ & $\pi_{M_2}(\theta_{M_2})$; Acceptance Bandwidth ε; Distance Measure $\|\cdot\|$; Target Number M.*

$s_{obs} \leftarrow s(x_{obs})$.
$\mathcal{M} \leftarrow \{\}$.
$t \leftarrow 1$.
while $t \leq M$ **do**
  $m_t \leftarrow$ *sample* $\pi_M(M)$.
  $\tilde{\theta}_t \leftarrow$ *sample* $\pi_{m_t}(\theta)$.
  $\tilde{x} \leftarrow f(X|\tilde{\theta}_t)$.
  $\tilde{s} \leftarrow s(\tilde{x})$.
  if $\|s_{obs} - \tilde{s}\| \leq \varepsilon$ **then**
    $\hat{\theta}^{(t)} \leftarrow \tilde{\theta}$.
    Add m_t *to* $\mathcal{M}$.
    $t \leftarrow t + 1$
  otherwise *Pass;*
return $\mathcal{M} = \{m_1, \dots, m_M\}$

[Grelaud *et al.*, 2009] presents an algorithm which uses an alteration of the ABC-Rejection Sampling algorithm (**Algorithm 3.2**), using a uniform kernel, to estimate Bayes Factor. Their approach is outlined in **Algorithm 3.7**. This approach defines a meta-model $M = (M_1, M_2)$ which is a mixtures model which uses model M_1 or M_2 according to some distribution π_M. The distribution π_M indicates our prior belief of the likelihood of each model. The algorithm then proceeds as a standard ABC-Rejection Sampling algorithm except, during the parameter sampling step it also samples which model to use (this defines which set of parameter priors to use too). Each time a simulation is accepted, the model which generated it is recorded in the set $\mathcal{M}$. The returned set $\mathcal{M}$ provides the ratio of the number of times simulations from each model were accepted which estimates Bayes Factor as (35).

$$\hat{B}_{1,2} = \frac{\sum_{i=1}^{N} \mathbb{1}\{m_i = M_1\}}{\sum_{i=1}^{N} \mathbb{1}\{m_i = M_2\}} \tag{35}$$

The results of **Algorithm 3.7** are sensitive to how informative the priors are for each model and thus can be used to compare different specifications of parameter priors too.

Algorithm 3.7 is based on the ABC-Rejection Sampling algorithm and thus does not gain any of the advantages of the ABC-MCMC or ABC-SMC algorithms, namely being effective when the prior and posterior are significantly different. [Toni *et al.*, 2009] presents a model selection algorithm which uses ABC-SMC, but requires the use of a meta-model which incorporates the models being tested, as in [Grelaud *et al.*, 2009]. [Didelot *et al.*, 2011] presents an approach which estimates the evidence for each model independently, using ABC-SMC.

3.4 Regression Adjustment in ABC

Regression adjustment for ABC methods is an innovation first suggested by [Beaumont *et al.*, 2002] where regression methods are applied to the accepted parameter sets in order to reduce the distance between the observed summary statistic value s_{obs} and the simulated summary statistic values $\tilde{s}$. Reducing this distance results in an improved approximation. This step is applied after a set of accepted simulations $\tilde{\Theta}$ has been produced.

The approach suggested in [Beaumont *et al.*, 2002] performs weighted local-linear regression of the accepted parameter values $\tilde{\Theta}$ on the difference between simulated summary statistic values and true summary statistic values $(s - s_{obs})$. Standard linear regression assumes the linear relationship (36) between parameters and summary statistic values exists.

$$\theta = \alpha + s^T\beta + \varepsilon \quad \text{with} \quad \alpha \in \mathbb{R}^\rho, \beta \in \mathbb{R}^{\phi\times\rho}, \varepsilon \in \mathbb{R}^\rho \tag{36}$$

where $\rho := |\theta|$ is the number of parameters, $\phi := |s|$ is the dimensionality of the summary statistics and ε is independent additive gaussian noise with zero mean and constant variance.

It is generally unrealistic to assume that such a relationship, or properties of ε, hold for the whole sufficient-statistic space. However, they may hold in the locality of s_{obs}. This is why [Beaumont *et al.*, 2002] suggests using local-linear regression, which assumes relationship (37) exists.

$$\theta = \alpha + (s - s_{obs})^T\beta + \varepsilon \tag{37}$$

When taking a least-squares approach to this regression problem, the objective function is (38).

$$\hat{\alpha}, \hat{\beta} = \text{argmin}_{\alpha,\beta} \sum_{i=1}^{N} \left(\tilde{\theta}_i - \left\{\alpha - (\tilde{s}_i - s_{obs})^T\beta\right\}\right)^2 K_\varepsilon\left(\|\tilde{s}_i - s_{obs}\|\right) \tag{38}$$

where N is the number of accepted parameter sets, $\tilde{\theta}_i$ is the i^{th} accepted parameter set, $\tilde{s}_i$ is the summary statistic values associated with $\tilde{\theta}_i$, $\|\cdot\|$ is a distance function and $K_\varepsilon(\cdot)$ is a kernel density function with bandwidth ε.

[Beaumont *et al.*, 2002] recommends using the Epanechnikov kernel as doing so means that very few simulations will be assigned small, but non-zero, weights. Having lots of small, non-zero weightings is computationally inefficient as they still need to be assessed but offer little insight. The solution to (38) is the following, where $\hat{\alpha}_{LSE}$ is the first column of the resulting matrix and $\hat{\beta}_{LSE}$ is the remaining columns.

$$\begin{aligned}
\left(\hat{\alpha}_{LSE}, \hat{\beta}_{LSE}\right) &:= (X^TWX)^{-1}X^TW\boldsymbol{\theta} \\
\text{where} & \\
X &:= \begin{pmatrix} 1 & (\tilde{s}_{1,1} - s_{obs,1}) & \dots & (\tilde{s}_{1,M} - s_{obs,M}) \\ 1 & (\tilde{s}_{2,1} - s_{obs,1}) & \dots & (\tilde{s}_{2,M} - s_{obs,M}) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (\tilde{s}_{N,1} - s_{obs,1}) & \dots & (\tilde{s}_{N,M} - s_{obs,M}) \end{pmatrix} \\
W &:= \begin{pmatrix} K_\varepsilon(\|\tilde{s}_1 - s_{obs}\|) & 0 & \dots & 0 \\ 0 & K_\varepsilon(\|\tilde{s}_2 - s_{obs}\|) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & K_\varepsilon(\|\tilde{s}_N - s_{obs}\|) \end{pmatrix} \\
\boldsymbol{\theta} &:= (\tilde{\theta}_1, \dots, \tilde{\theta}_N)^T
\end{aligned}$$

These solutions implies the general relationship (39) between summary statistic values s and the expected parameter set to have generated them, for values of s in the region of s_{obs}.

$$\hat{\mathbb{E}}[\theta|s] = \hat{\alpha}_{LSE} + \hat{\beta}_{LSE} \tag{39}$$

Relationship (39) is used to define a mean-corrected parameter set θ_i^* (40) for each accepted parameter set θ_i by subtracting the expected difference between the parameters which produce s_i and s_{obs}.

$$\begin{aligned} \theta_i^* &:= \theta_i - \left(\hat{\mathbb{E}}[\theta|s_i] - \hat{\mathbb{E}}[\theta|s_{obs}]\right) \\ &= \theta_i - (\hat{\alpha}_{LSE} + s_i^T \hat{\alpha}_{LSE}) + (\hat{\alpha}_{LSE} + s_{obs}^T \hat{\alpha}_{LSE}) \\ &= \theta_i + (s_{obs} - s_i)^T \hat{\beta}_{LSE} \end{aligned} \tag{40}$$

Note that it is possible for the corrected parameter sets θ_i^* to fall outside of the priors. There are ways to address this, but this should be taken as a sign that the priors are ill-specified as the regression is extrapolating rather than interpolating.

In [Beaumont *et al.*, 2002] simulation experiments on the Growing Population Model of [Pritchard *et al.*, 1999] are run to compare methods which implement their regression adjustment, to those which don't. They found that using regression adjustment meant that ABC-Rejection Sampling had less of a tendency to estimate posteriors which were similar to the prior. And, that acceptance kernel bandwidth had less of an effect when regression adjustment was applied. This suggests that using regression adjustment increases the rate of convergence.

Implementing this regression method is fairly straightforward as it is applied after all other steps in standard ABC methods. There are many programming packages available which can perform local-linear weighted regression efficiently. There is an increased space requirement for algorithms which use this approach as they now need to store the summary statistic values for each accepted parameter set. This should approximately double the space used as it is generally efficient to use one summary statistic per parameter (See *Section 4*). This is not a major concern as space requirements are rarely a limiting factor for ABC methods, but it is worth being aware of.

[Fagundes *et al.*, 2007] required 10 CPU months to estimate the parameters of their complex model using the regression method of [Beaumont *et al.*, 2002]. This is unsurprising as this approach requires the inversion of a matrix whose size is dependent on the number of summary statistics being used and the number of parameter sets which has been accepted.

[Blum and François, 2010] presents using a shallow neural network for nonlinear-heteroscedastic regression of parameter values on the summary statistic values. They compare an adaptive and a non-adaptive approach, with their aim being to reduce the computational load required to perform this regression when compared to the approach in [Beaumont *et al.*, 2002]. In their worked examples their approaches require significantly less computational time.

In their examples, the bandwidth had little effect on their neural network approaches, compared to the approach of [Beaumont *et al.*, 2002]. And, that the posterior quantiles from their neural network methods were typically tighter and better centred around the true parameter value. This suggests the neural network methods produce a more accurate fit.

The limitations to the approach of [Blum and François, 2010] are the same as with most uses of neural networks: overfitting and lack of interpretability. I discuss these issues in more detail in *Section 4.3.5*.

4 Summary Statistic Selection

In this section I: Motivate research into summary statistic selection *Section 4.1* and discuss features to consider when selecting summary statistics *Section 4.2*: Present and discuss five methods for summary statistic selection: three which use hand-crafted summary statistics *Sections 4.3.1-4.3.3*; and two which automatically generate summary statistics *Sections 4.3.4-4.3.5*. These approaches are covered in the chronological order in which they were original proposed.

4.1 Motivation

The study of summary statistics has relevance beyond ABC methods, largely due to the recent "Big-Data" Revolution which has seen the rate at which data can be collected and stored significantly outpace improvements in computational power. This has motivated research into effective methods to reduce the size of datasets so that more computationally intensive algorithms can be used to analyse the data.

$$s : \mathbb{R}^m \to \mathbb{R}^p \text{ with } m > p$$

A summary statistic s is a statistic which reduces the dimensionality of some sampled data, in a deterministic fashion, whilst retaining as much information about the sampled data as possible. Reducing the dimensionality of data is desirable as it reduces the computational requirements to analyse the data and makes comparing data significantly easier. Ideally, a summary statistic would compress the sampled data without any information loss (A property known as "sufficiency"). However, low-dimension sufficient summary statistics are rare in practice and we often have to trade-off information retention against dimensionality reduction.

In most cases, each dimension of the output of a summary statistic is the result of an independent calculation. As such, it is often conceptually easier to consider each dimension as an independent summary statistics when selecting summary statistics. This idea of each dimension having independence also makes it conceptually easy to combine summary statistics by appending the results of one statistic onto the end of the other, as new dimensions. As long as the sum of the dimensions of the outputs from the summary statistics in the set is less than that of the sampled data (41), then using a set of summary statistics still produces effective dimensionality reduction.

$$m > \sum_{i=1}^{k} p_i \text{ where } s_i : \mathbb{R}^m \to \mathbb{R}^{p_i} \tag{41}$$

The success of ABC methods depends mainly on three user choices: the summary statistics; the distance measure; and the acceptance kernel. Of these, summary statistic selection is arguably the most important as the other two mainly effect the rate at which the algorithm converges on the posterior mean. Whereas, choosing summary statistics which are uninformative can lead to the estimated parameter posteriors returned by the algorithm being drastically different from the true parameter posteriors. This is trivial to realise if you consider a scenario where $s(x) = c$, for some constant $c \in \mathbb{R}$, is used as the sole summary statistic. This would result in all (or none) of the simulations being accepted and so the returned posterior will be the same as the supplied prior.

In practice, the quality of the estimated posteriors returned by an ABC method is limited by the amount of computational time which is dedicated to running the algorithm. For some problems, such as the SIR model fittings performed in *Section 5*, it is reasonable to dedicate the majority of your computing time on summary statistic selection, rather than on model fitting, as it is clear that even the simplest ABC methods (e.g. ABC-Rejection Sampling) will be sufficient to fit the model when given a good choice of summary statistics.

Traditional Thinking

Traditionally, summary statistics for ABC methods are chosen manually using expert, domain-specific knowledge. Utilising this expert knowledge is desirable as these statistics will incentivise exploring regions of the parameter space which have been scientifically shown to be relevant to the given problem. Thus, it is more likely that these regions will contain the true parameter values. Similarly, these statistics will disincentivise exploring regions which have been shown to not be of little interest.

However, relying on expert knowledge to choose summary statistics limits the scenarios where ABC methods can be applied to only those where there has already been significant research. And, leads to statistics being chosen due to their prevalence in a field rather than their suitability to computational methods. Moreover, the use of hand-crafted summary statistics means that any limitations in current understanding of a field will be encoded into the model fitting process, possibly leading to misspecification and bias.

When using a set of summary statistics, expert knowledge is generally not sufficient to determine how best to weight each summary statistic. The methods I present in *Sections 4.3.1-4.3.3* can be used to automate the process of determining these weights by specifying multiple versions of the same summary statistic, with each version having a different weighting. Weighting summary statistics is a task which whould be approached with much care for reasons discussed in these sections.

4.2 Properties of Summary Statistics

When evaluating a summary statistic for use in ABC there are several properties, both practical and mathematical, to consider.

A useful mathematical property, observed in [Wood, 2010], is that asymptotically, summary statistics typically have a normal distribution $s \sim \text{Normal}(\mu_s, \Sigma_s)$ with mean μ_s and covariance matrix Σ_s. This property can be used to motivate a regression adjustment approach, as explored in *Section 3.4*, using the sample estimates of mean and co-variance.

Practical Properties

The key reason for using summary statistics is for the computational efficiencies which result from their dimensionality reduction. Reducing the size of a dataset means less operations need to be performed to analyse it, meaning more simulations can be processed in the same time-period. This naturally means summary statistics which result in greater dimensionality reduction are preferable, but similarly means that a summary statistic which is computationally inefficient to calculate is less desirable.

For a model which produces data of dimension $n \times m$ (i.e. n observations each with m features), most standard summary statistics are calculated in $O(n \cdot m)$ time. However, this is only a asymptotic result, to with a constant multiplier, and in practice there are meaningful differences in the computational requirements of each summary statistics. Calculating the mean and maximum values for each feature takes $O(n \cdot m)$ time in theory. But, since calculating the mean relies on arithmetic operations and the maximum on comparison operations, they will take different amounts of time in practice. Statistics which rely on search or sort operations (most notably order statistics) are variable in the their time complexity for different data sets, which will effect the reliability of models which use them. Integer overflow is a possible issue for some summary statistics, although this is often easy to avoid when actively being considered during the implementation of an algorithm. Moreover, for statistics with non-linear computational complexity (e.g. correlation between each pair of features), the size of the dataset being analysed needs to be considered when evaluating summary statistic choice.

ABC methods rely on distance measures to determine whether a simulated observations are a good approximation of true observations, or not. This means that the range and scale of values a summary statistic will likely produce will have an effect on how influential that summary statistic is to the final model fit. In most cases it is reasonable to standardise all statistics to have the same mean and variance, effectively giving the same weighting to each statistic. This can be implemented to occur adaptively within the ABC method. There may be cases where assigning different weights to different summary statistics makes sense[1], and produces a better model fit, but these are hard to justify from a theoretical approach. The selection methods I discuss which compare hand-crafted statistics (Sections 4.3.1-4.3.3) can be used to compare possible weightings by including several versions of the same summary statistic, each with a different scaling, in the set of statistics being compared. This will however increase computation time due to the increase size of the set of statistics and may make the results harder to interpret[2].

For real-world modelling problems, the interpretability of summary statistics used in the final model is a key factor in how useful the solution is. Senior stakeholders[3] in a problem will want to use the final model to justify their future (and past) decisions. This is much easier to do when the features the model is considering and the relative weight it assigns to them, are readily interpretable. Hand-crafted statistics are almost always the most interpretable statistics available, and, as such, generated statistics are rarely used in commercial problems[4]. In cases where it is chosen to use automatically generated statistics; one can develop an intuition for their model by varying the model parameters, or removing certain features, and observing how the output varies.

Sufficiency

Definition 4.1 (Sufficient Statistic Casella and Berger [2001])
Let $s : \mathbb{R}^m \to \mathbb{R}^n$ be a statistic and X be a model with parameters θ. The statistic s is said to be sufficient for the parameters θ if the conditional distribution of the model X, given the value of the statistic $s(X)$, is independent of the model parameter.

$$\mathbb{P}(X|s(X)) = \mathbb{P}(X|s(X), \theta)$$

Verbosely, a statistic is sufficient for a model parameter(s) if it captures all the information which a sample of the model carries about said parameter(s). This means that knowing the value of a sufficient statistic is as informative as knowing the true model parameters. This is clearly a desirable property as in practice we can always calculate the value of the summary statistic using the sampled data, but cannot know the true parameter values (otherwise we would not be trying to predict them). Sufficient statistics exist for all models as, trivially, the identity function

[1] Typically these occur when seeking to answer a specific inferential question, such as when will the peak of the infectious population occur, rather for more general model fitting problems.

[2] Multiple sets of weighted summary statistics will be equivalent due to having the same ratio of weights

[3] Individuals who have a vested interested in the project but are not responsible for its execution. e.g. Public Health policy makers in the case of epidemic modelling.

[4] The current popularity of using "Neural Networks" in commercial settings does buck this trend. I hope this fad will subside soon in favour of more interpretable alternatives. I believe it is worth noting that the new European Union Payment Services Directive (PSD2) requires that certain models used by financial institutions be "explainable" in order to improve the customer experience and to ensure no one is discriminated against due to their protected characteristics.

is a sufficient statistic for all models, but this is the only sufficient statistic for non-exponential models (**Theorem 4.4**).

It can be intuitively useful to consider a sufficient statistic as a data reduction method. Moreover, a sufficient summary statistic provides a loss-less compression of sampled data as it reduces the dimensionality of the data but retains all relevant information.

Remark 4.1 (Supersets of Sufficient Statistics)
Let $s_{1:k-1}(\cdot) := \{s_1(\cdot), \ldots, s_{k-1}(\cdot)\}$ be a collection of $k-1$ summary statistics and suppose that $s_{1:k-1}$ is sufficient for the parameters θ of some model X. Then $s_{1:k-1} \cup \{s_k\}$ is also sufficient for the parameters θ, for all summary statistics s_k.

Proof. Consider a model with parameters θ and let $s_1, \ldots, s_k$ be summary statistics where the set $s_{1:k-1} := \{s_1, \ldots, s_{k-1}\}$ is sufficient for parameter θ. Note that the likelihood of set $s_k := s_{1:k-1} \cup \{s_k\}$ given the model parameters θ can be stated as

$$\mathbb{P}(s_{1:k}(X)|\theta) = \mathbb{P}(s_k(X)|s_{1:k-1}(X), \theta)\mathbb{P}(s_{1:k-1}|\theta)$$

Now consider the following decomposition of the posterior for the model parameters θ given summary statistics $s_{1:k}$

$$\begin{aligned}\mathbb{P}(\theta|s_{1:k}(X)) &= \frac{\mathbb{P}(s_{1:k}(X)|\theta)\mathbb{P}(\theta)}{\mathbb{P}(s_{1:k}(X))} \\ &= \frac{\mathbb{P}(s_k(X)|s_{1:k-1}(X), \theta)\mathbb{P}(s_{1:k-1}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(s_k(X)|s_{1:k-1}(X))\mathbb{P}(s_{1:k-1}(X))}\end{aligned}$$

Since the set $s_{1:k-1}$ is sufficient for θ we have that

$$\mathbb{P}(s_k(X)|s_{1:k-1}(X)) = \mathbb{P}(s_k(X)|s_{1:k-1}(X), \theta)$$

Applying this result to the decomposition above, we deduce that the posterior for the model parameters θ given $s_{1:k}$ or $s_{1:k-1}$ are identical.

$$\begin{aligned}\mathbb{P}(\theta|s_{1:k}(X)) &= \frac{\mathbb{P}(s_k(X)|s_{1:k-1}(X), \theta)\mathbb{P}(s_{1:k-1}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(s_k(X)|s_{1:k-1}(X), \theta)\mathbb{P}(s_{1:k-1}(X))} \\ &= \frac{\mathbb{P}(s_{1:k-1}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(s_{1:k-1}(X))} \\ &= \mathbb{P}(\theta|s_{1:k-1}(X))\end{aligned}$$

Thus the set $s_{1:k}$ is sufficient for model parameters θ. Due to the arbitrary nature of $s_{1:k-1}$ and s_k, this result holds for all supersets of sufficient summary statistics. □

Remark 4.1 states that if we have a set of summary statistics which are sufficient for a set of parameters, then adding more summary statistics will never increase (or decrease) the amount of relevant information being extracted from the sampled data. This means there is an optimally minimal number of summary statistics required to achieve sufficiency.

I demonstrate in **Example 4.1** that the sample mean is a sufficient summary statistic for a normal distribution with unknown mean, but known variance.

Example 4.1 (Sufficient Statistic for Normal Distribution with Unknown Mean)
Let $X \sim$ Normal(μ, σ_0^2), with $\mu \in \mathbb{R}$ unknown and $\sigma_0^2 \in \mathbb{R}$ known, and $\mathbf{x}$ be n independent observations of X.

We have that

$$f_{\mathbf{X}}(\mathbf{X}) = \prod_{i=1}^{n} f_X(X_i) = \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i-\mu)^2\right\}$$

Let $s = s(\mathbf{X})$ be an arbitrary statistic of n observations from the model. We will build up the conditional distribution of $\mathbf{X}$ given $s(\mathbf{X})$, by first considering their joint distribution

$$\begin{aligned}
f_{\mathbf{X},s(\mathbf{X})}(\mathbf{X},s) &= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i+s-s-\mu)^2\right\} \\
&= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}((X_i+s)-(\mu-s))^2\right\} \\
&= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\left((X_i-s)^2+(\mu-s)^2-2(\mu-s)(X_i-s)\right)\right\} \\
&= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \cdot \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i-s)^2\right\} \cdot \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(\mu-s)^2\right\} \\
&\quad \cdot \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}-2(\mu-s)(X_i-s)\right\} \\
&= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \cdot \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i-s)^2\right\} \cdot \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(\mu-s)^2\right\} \\
&\quad \cdot \exp\left\{\frac{\mu-s}{\sigma_0^2}\left(\sum_{i=1}^{n}(X_i)-ns\right)\right\}
\end{aligned}$$

If we define $s(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$, the sample mean, then the third exponential disappears. Note that $s(\mathbf{X}) \sim$ Normal $\left(\mu, \frac{1}{n}\sigma_0^2\right)$.

Now consider the conditional distribution of $\mathbf{X}$ given $s(\mathbf{X})$.

$$\begin{aligned}
f_{\mathbf{X}|s(\mathbf{X})}(\mathbf{X}|s) &= \frac{f_{\mathbf{X},s(\mathbf{X})}(\mathbf{X},s)}{f_{s(\mathbf{X})}(s(\mathbf{X}))} \\
&= \frac{\sqrt{\frac{1}{(2\pi\sigma_0^2)^n}} \cdot \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i-s)^2\right\} \cdot \exp\left\{-\frac{n}{2\sigma_0^2}(\mu-s)^2\right\}}{\sqrt{\frac{n}{2\pi\sigma_0^2}} \cdot \exp\left\{-\frac{n}{2\sigma_0^2}(\mu-s)^2\right\}} \\
&= \sqrt{\frac{1}{n(2\pi\sigma_0^2)^{n-1}}} \cdot \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i-s)^2\right\}
\end{aligned}$$

This shows that the conditional distribution of $\mathbf{X}$ given $s(\mathbf{X})$ is independent of μ, the unknown parameter, and thus the sample mean is a sufficient statistic for a normal distribution with unknown mean but known variance.

Example 4.1 demonstrates that finding sufficient summary statistics can be highly manual and did require us to "guess" at the possible formulation of a summary statistic, then verify that it was sufficient. The Fisher-Neyman Factorisation Criterion (**Theorem 4.1**) [Fisher, 1922; Neyman, 1935], first recognised by Fisher in [Fisher, 1922], specifies a property which all sufficient statistics have. This property is used as the basis of a more formulaic approach to finding sufficient statistics by separating the terms of the conditional probability of a model given the summary statistic value into those which depend on the summary statistic and those which do not.

Theorem 4.1 (Fisher-Neyman Factorisation Criterion Casella and Berger [2001])
Let $X \sim f(\cdot;\theta)$ be a model with parameters θ and $s(\cdot)$ be a statistic.

$s(\cdot)$ is a sufficient statistic for the model parameters θ iff there exist non-negative functions $g(\cdot;\theta)$ and $h(\theta)$ where $h(\cdot)$ is independent of the model parameters[a] and

$$f(X;\theta) = h(X)g(s(X);\theta)$$

This formulation shows that the distribution of the model X only depends on the parameter θ through the information extracted by the statistic s. A consequence of the sufficiency of s.

Proof. [Roussas, 1998]

$\Longrightarrow$ First, consider the forwards direction of the theorem and suppose s is a sufficient summary statistic. Define functions

$$h(x) = \mathbb{P}(X = x|s(X) = s(x)) \quad \text{and} \quad g(s(x);\theta) = \mathbb{P}(s(X) = s(x);\theta)$$

Note that $h(\cdot)$ is independent of the model parameter θ due to the sufficiency of s. Then

$$\begin{aligned} f_X(x) &= \mathbb{P}(X = x) \\ &= \mathbb{P}(X = x, s(X) = s(x)) \\ &= \mathbb{P}(X = x|s(X) = s(x))\mathbb{P}(s(X) = s(x)) \\ &= h(X)g(s(X)) \end{aligned}$$

$\Longleftarrow$ Now, consider the reverse direction of the theorem and suppose there exists some functions $h(\cdot), g(\cdot;\theta)$, with $h(\cdot)$ independent of model parameter θ, such that

$$f(x;\theta) = h(x)g(s(x);\theta) \text{ for all } x \in \mathcal{X},\ \theta \in \Theta$$

where $\mathcal{X}$ is the support of X and Θ the set of possible parameters.

Then, for an arbitrary $c \in \mathbb{R}$

$$\begin{aligned} \mathbb{P}(X = x|s(X) = c) &= \frac{\mathbb{P}(X = x, s(X) = c)}{\mathbb{P}(s(X) = c)} \\ &= \frac{\mathbb{1}\{s(x) = c\}f(x;\theta)}{\sum_{y\in\mathcal{X};s(y)=c} f(y;\theta)} \\ &= \frac{\mathbb{1}\{s(x) = c\}h(x)g(s(x);\theta)}{\sum_{y\in\mathcal{X};s(y)=c} h(y)g(s(y);\theta)} \\ &= \frac{h(x)g(c;\theta)}{\sum_{y\in\mathcal{X};s(y)=c} h(y)g(c;\theta)} \\ &= \frac{h(x)}{\sum_{y\in\mathcal{X};s(y)=c} h(y)} \end{aligned}$$

This final expression is independent of the model parameter θ.

The result holds in both directions. □

[a] i.e. $h(\cdot)$ only depends on the sampled data

Example 4.2 below demonstrates how the Fisher-Neyman Factorisation Criterion can be used to find a sufficient summary statistic for a Poisson model where the mean λ is unknown.

Example 4.2 (Using Fisher-Neyman Factorisation Theorem to find sufficient statistics for a Poisson distribution with unknown mean)
Let $X \sim Poisson(\lambda)$, with $\lambda \in \mathbb{R}^{>}$ unknown, $\mathbf{x}$ be n independent observations of X and $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$ be the sample mean of these n observations.

Consider the joint distribution of these n observations

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^{n} f_X(x_i) \\ &= \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} \\ &= \frac{1}{\prod_{i=1}^{n} x_i!} \cdot \theta^{\sum_{i=1}^{n} x_i} e^{-n\theta} \\ &= \underbrace{\left\{\frac{1}{\prod_{i=1}^{n} x_i!}\right\}}_{(1)} \cdot \underbrace{\left\{\theta^{\sum_{i=1}^{n} x_i} e^{-n\theta}\right\}}_{(2)} \end{aligned}$$

The last step shows how the terms can be collected into: (1), those which are independent of model parameter θ; and, (2), those which are dependent on model parameter θ. We can how derive the conditions of the Fisher-Neyman Factorisation theorem by inspecting the final expression.

It is apparent that we should define the function $h(\mathbf{x})$ as

$$h(\mathbf{x}) = \frac{1}{\prod_{i=1}^{n} x_i!}$$

In order to define the function $g(s(\mathbf{x});\theta)$ we first need to define the summary statistic $s(\mathbf{x})$. This is straightforward as all the sampled data $\mathbf{x}$ only occurs in a sum in (2), so we define $s(\mathbf{x}) = \sum_{i=1}^{n} x_i$. Meaning we can define $g(\mathbf{x};\theta)$ as

$$g(\mathbf{x};\theta) = \theta^{s(\mathbf{x})} e^{-n\theta}$$

With these definitions we fulfil the conditions of the Fisher-Neyman Factorisation theorem, meaning $s(\mathbf{X}) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for the mean for a Poisson distribution.

In most cases sufficient statistics for a parameter are not unique. Moreover, each sufficient statistic does not necessarily produce the same level of compression. Consider a normal distribution with unknown mean, here both the sample sum and identity function are both sufficient statistics. However the sample sum is a much more desirable statistic to use as it provides compression down to a single dimension. This lack of uniqueness motivates the concept of minimal sufficiency (**Definition 4.2**).

Definition 4.2 (Minimally Sufficient Statistic, Dodge *et al.* [2006])
Let $s(\cdot)$ be a sufficient statistic for parameter θ of model X. $s(\cdot)$ is minimally sufficient if for any other sufficient statistic $t(\cdot)$ of parameter θ there exists a function f which maps $t(x) \mapsto s(x)$.

$$s(X) = f(t(X))$$

Minimally sufficient statistics have lower (effective) dimensionality than their non-minimal counterparts. This makes minimally sufficient statistics desirable as they produce the greatest level of compression and, in doing so, maximally reduce the computational resources required to analyse the sampled data.

As with identifying sufficient statistics; determining whether, or not, a sufficient statistic is minimally sufficient is not a trivial task. I demonstrate this in **Example 4.3**.

Example 4.3 (Minimally Sufficient Statistic for IID Bernoulli Random Variables)
Let $X_1, \ldots, X_n$ are independent and identically distribution Bernoulli random variables. Note that the identity function $s_1(\mathbf{X}) = \mathbf{X}$ and the sum function $s_2(\mathbf{X}) = \sum_{i=1}^{n} X_i$ are both sufficient statistics.

We can map from s_1 to s_2 as follows

$$s_2(\mathbf{X}) = \sum_{i=1}^{n} [s_2(\mathbf{X})]_i$$

However, there is no function which can map from s_2 to s_1 as it would have to map the value 1 to both $(1, 0, \ldots, 0)$ and $(0, 1, \ldots, 0)$. This proves that the identity function s_1 is not a minimally sufficient statistic, but does not prove that the sum function s_2 is a minimally sufficient statistic as we have not considered all possible sufficient statistics for this distribution.

Theorem 4.2 (Condition for Minimal Sufficiency, Balakrishnan [2019])
Consider a model with parameters θ. Let $\mathbf{x}, \mathbf{y}$ be two samples from this model and $s(\cdot)$ be a statistic.

> *If $\frac{\mathbb{P}(\mathbf{y};\theta)}{\mathbb{P}(\mathbf{x};\theta)}$ is independent of θ iff $s(\mathbf{x}) = s(\mathbf{y})$, then statistic s is minimally sufficient.*

Proof. Let $s(\cdot)$ be a statistic for model X with parameters θ and assume that $\frac{\mathbb{P}(\mathbf{y};\theta)}{\mathbb{P}(\mathbf{x};\theta)}$ is independent of θ iff $s(\mathbf{y}) = s(\mathbf{x})$. I first show that this s is sufficient and then that it is minimally sufficient.

Note that this statistic s produces a partition of the sample space $A = \{A_c : \exists\ \mathbf{x} \in \mathcal{X},\ s(\mathbf{x}) = c\}$. For each set A_c of the partition A fix a point $\mathbf{x}_c \in \mathcal{X}$ to represent it.

Let $\mathbf{x}$ be a sample of X and define $\mathbf{y} = \mathbf{x}_{s(\mathbf{x})}$. Note that sample $\mathbf{y}$ is a function of $s(\mathbf{x})$ only and $s(\mathbf{x}) = s(\mathbf{y})$. Consider the joint distribution of $\mathbf{x}$

$$\mathbb{P}(\mathbf{x};\theta) = \mathbb{P}(\mathbf{x};\theta)\frac{\mathbb{P}(\mathbf{y};\theta)}{\mathbb{P}(\mathbf{y};\theta)} = \mathbb{P}(\mathbf{y};\theta)\frac{\mathbb{P}(\mathbf{x};\theta)}{\mathbb{P}(\mathbf{y};\theta)}$$

By our assumptions of s, we have that $\frac{\mathbb{P}(\mathbf{x};\theta)}{\mathbb{P}(\mathbf{y};\theta)}$ is independent of θ. Thus, we can produce the following decomposition

$$\begin{aligned} \mathbb{P}(\mathbf{x};\theta) &= h(\mathbf{x})g(s(\mathbf{x});\theta) \\ \text{where} \\ h(\mathbf{x}) &= \tfrac{\mathbb{P}(\mathbf{x};\theta)}{\mathbb{P}(\mathbf{y};\theta)} \\ g(s(\mathbf{x});\theta) &= \mathbb{P}(s(\mathbf{y}); theta) \end{aligned}$$

By the Fisher-Neyman factorisation criterion we can deduce that s is sufficient.

Now, let t be another sufficient statistic for θ and let $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ st $t(\mathbf{x}) = t(\mathbf{y})$. By the

Fisher-Neyman factorisation criterion, we have

$$\begin{array}{rcl}\mathbb{P}(\mathbf{x};\theta) & = & h(\mathbf{x})g(t(\mathbf{x});\theta) \\ & = & \frac{h(\mathbf{x})}{h(\mathbf{y})}h(\mathbf{y})g(t(\mathbf{y});\theta) \\ & = & \frac{h(\mathbf{x})}{h(\mathbf{y})}\mathbb{P}(\mathbf{y};\theta) \text{ by Fisher-Neyman factorisation} \\ \implies \frac{\mathbb{P}(\mathbf{x};\theta)}{\mathbb{P}(\mathbf{y};\theta)} & = & \frac{h(\mathbf{x})}{h(\mathbf{y})}\end{array}$$

This shows that $\frac{\mathbb{P}(\mathbf{x};\theta)}{\mathbb{P}(\mathbf{y};\theta)}$ is independent of θ, meaning $s(\mathbf{x}) = s(\mathbf{y})$ by our assumptions of s. This result means there exists a function f st $s(\mathbf{x}) = f(t(\mathbf{x})) \ \forall \ \mathbf{x} \in \mathcal{X}$. Moreover, due to the arbitrary definition of t, for each sufficient statistic of θ there exists a function which maps from it to our statistic s, fulfilling the definition of s being minimally sufficient. $\square$

Theorem 4.2 states that if the ratio of the marginal distributions of two samples from a model are independent of the model parameters, only when the samples map to the same value under some statistic s, then s is minimally sufficient. This property can be used to identify minimally sufficient summary statistics, either by assisting in deduction or by verifying a proposed statistic.

Statistics carry information about sampled data, but in Bayesian modelling most problems centre around estimating parameter values. In some cases a sufficient statistic may be a good estimator of a model parameter too, in **Example 4.1** it was shown that the sample mean is a sufficient statistic for the population mean of a normal distribution. This is not always the case, in **Example 4.2** it was shown that the sum of sampled values is a sufficient statistic for the mean of a Poisson distribution but this is not a good estimator.

Theorem 4.3 (Rao-Blackwell Theorem, Rao [1945]; Blackwell [1947])
Let X be a model with parameters θ, $U = u(X)$ be an unbiased estimator for function $g(\theta)$ and $s(X)$ is a sufficient statistic for θ.

> *The statistic $v(X) := \mathbb{E}[u|T = t(X)]$ is an unbiased estimator of $g(\theta)$ and $Var(v(X)) \leq Var(u(X))$.*

The statistic $v(X)$ is known as the Rao-Blackwell Estimator.

Proof. The proof that $v(X)$ is unbiased is immediate from the Tower Law

$$\begin{array}{rcl}\mathbb{E}[v(X)] & = & \mathbb{E}[\mathbb{E}[u|T = t(X)]] \\ & = & \mathbb{E}[u] \\ & = & g(\theta)\end{array}$$

Now consider the variance of $v(X)$

$$\begin{array}{rcl}\text{Var}(v(X)) & = & \text{MSE}[v(X)] - \text{Bias}[v(X)]^2 = \text{MSE}[v(X)] \\ & = & \mathbb{E}[(v(X) - g(\theta))^2] \\ & = & \mathbb{E}[(\mathbb{E}[v|T = t(X)] - g(\theta))^2] \\ & = & \mathbb{E}[(\mathbb{E}[v - g(\theta)|T = t(X)])^2] \\ & {}^a\leq & \mathbb{E}[(v - g(\theta))^2|T = t(X)] \\ & = & \text{Var}(u(X)) \\ \implies \text{Var}(v(X)) & \leq & \text{Var}(u(X))\end{array}$$

□

[a] $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \implies \mathbb{E}[X^2] \geq \mathbb{E}[X]^2$

The Rao-Blackwell Theorem (**Theorem 4.3**) provides a general relationship between estimators and sufficient statistics by demonstrating a transformation of an unbiased estimator, using a sufficient statistic, which produces an unbiased estimator with decreased variance and thus reduced mean-squared error. This is desirable as it is often straightforward to derive a crude estimator and then apply this transformation in order to improve its performance. A Rao-Blackwell transformation is idempotent as applying it to an already transformed estimator returns the same estimator, the proof of this follows immediately from the Tower Law.

The Lehmann-Scheffe Theorem [Lehmann and Scheffé, 1950] states that if the statistic used in a Rao-Blackwell transformation is both sufficient and complete, then the resulting estimator is in fact the unique minimum-variance unbiased-estimator. This result is independent of how good the initial estimator was.

Sufficiency In Practice

In Bayesian modelling problems we want to deduce the posterior for some model parameters to as high a degree of accuracy as possible. Let $f^*(\theta|X(\theta) = x_{obs})$ be the true posterior for model parameters θ and $\hat{f}(\theta|s(X(\theta)) = s(x_{obs}))$ be the estimated posterior produced by our modelling method, given x_{obs} was observed from the true model and summary statistics $s(\cdot)$ were used. If the summary statistics $s(\cdot)$ are sufficient then the estimated posterior $\hat{f}$ will converge towards the true posterior f^*, given enough simulations, however, if $s(\cdot)$ are not sufficient then $\hat{f}$ can never (consistently) converge on the true posterior f^*, and rather will always be an approximation. Thus, finding sufficient statistics for our models is highly desirable in Bayesian modelling.

Theorem 4.4 (Pitman–Koopman–Darmois Theorem, Andersen [1970])
Among families of probability distributions whose domain does not vary with the parameter being estimated, only in exponential families are there sufficient statistics whose dimension are bounded as the sample size increases.

Proof. See [Darmois, 1935; Pitman, 1936; Koopman, 1936] for the original proofs. □

However, although sufficient statistics do exist for all models, as the identity function is a sufficient statistic for all models, they are not necessarily the best choice of summary statistic when implementing computational methods as they may provide very little dimensionality reduction relative to other statistics which still manage to retain a large about of the relevant data from a sample. Moreover, the Pitman-Koopman-Darmois theorem **Theorem 4.4** states that sufficient summary statistics which provide a high level of dimensionality reduction only exist for probability distributions from exponential families.

This lack of computationally efficient sufficient statistics, for most models, motivated the concept of "approximate sufficiency" in [Joyce and Marjoram, 2008] which aims to balance the number of summary statistics with the amount of information being retained from a sample. I discuss this concept more when I present the summary statistic selection algorithm from [Joyce and Marjoram, 2008] in **Section 4.3.1**.

It is demonstrated in [Didelot *et al.*, 2011] that the using summary statistics which are sufficient for parameters produces unreliable results when performing model selection. This is due to it being impossible to distinguish between models which have the same sufficient statistics for their parameters. For example, the sum of sampled values is a sufficient statistics for the means of both geometric and Poisson distribution and so cannot be used to compare these two models. Rather, cross-model sufficient statistics would be required to distinguish between these models in practice, which is impossible in practice.

To close this section, I shall mention the Ewens' Sampling formula Ewens [1972] which illustrates a real-world scenario where useable and useful sufficient statistics have been found. The Ewens' Sampling formula provides, under certain conditions, a parametric probability distribution for the frequencies of unique types of allele observed in a sample of gametes when using the Infinite Alleles model. The mutation rate is the only parameter of this distribution and it is notable that the total number of types is a sufficient statistic for the mutation rate [Joyce, 1998]. This is especially appealing as ABC methods are used widely in population genetics research (See [Wegmann and Excoffier, 2010; Beaumont *et al.*, 2002; Marjoram and Tavaré, 2006] among many others).

4.3 Methods for Summary Statistic Selection

When thinking about summary statistic selection it is useful to consider the summary statistics themselves as a feature of your theorised model. This makes the process of selecting summary statistics analogous to model selection, with each combination of summary statistics being considered as a unique model. This is the motivation behind many summary statistic selection methods.

4.3.1 Approximate Sufficient Subset

[Joyce and Marjoram, 2008] presents the first algorithm for automating the selection of summary statistic. The key idea of their approach is to find a subset of summary statistics, from a large set of hand-crafted statistics, such that ABC methods perform approximately as well when using the subset. This requires a method for empirically evaluating the information extracted by sets of summary statistics. The use of hand-crafted statistics, as discussed above, comes with its own advantages and limitations.

Remark 4.2 (Difference of Log-Likelihood)
Let $s_1,\dots,s_k$ be summary statistics for a model X with parameters θ. Define sets $s_{1:k-1} := \{s_1,\dots,s_{k-1}\}$, $s_{1:k} := \{s_1,\dots,s_k\}$ and consider the likelihood of the set $s_{1:k}$ with-respect-to the model parameters θ

$$\begin{array}{rrcl} & \mathbb{P}(s_{1:k}(X)|\theta) &=& \mathbb{P}(s_k(X)|s_{1:k-1}(X),\theta)\cdot\mathbb{P}(s_{1:k-1}(X)|\theta) \\ \implies & \ln\mathbb{P}(s_{1:k}(X)|\theta) &=& \ln\mathbb{P}(s_k(X)|s_{1:k-1}(X),\theta)+\ln\mathbb{P}(s_{1:k-1}(X)|\theta) \\ \implies & \ln\mathbb{P}(s_{1:k}(X)|\theta)-\ln\mathbb{P}(s_{1:k-1}(X)|\theta) &=& \ln\mathbb{P}(s_k(X)|s_{1:k-1}(X),\theta) \end{array}$$

For the theoretical basis of their algorithm, Joyce & Marjoram first show that the difference in log-likelihood value between two sets of summary statistics can be directly quantified as $\ln\mathbb{P}(s_k(X)|s_{1:k-1}(X),\theta)$ (**Remark 4.2**). It is worth noting that if the set $s_{1:k-1}$ is sufficient for model parameter θ then the quantity $\ln\mathbb{P}(s_k(X)|s_{1:k-1}(X),\theta)$ would be independent of θ and thus mean s_k does not contribute to inferences about θ. This result reduces the problem of comparing sets of statistics to calculating or estimating a single value and motivates Joyce & Marjoram use of log-likelihood in their definition of score. Score quantifies how much extra information is extracted when a single extra statistic is added to a set with greater score values meaning more extra information is extracted. Thus we want to find the statistics with the greatest scores. Moreover, if the score of a statistic differs significantly from 0 then it should be accepted.

Definition 4.3 (Score δ_k, Joyce and Marjoram [2008])
Let $s_1, \ldots, s_k$ be k summary statistics. The score of s_k relative to the set $s_{1:k-1} := \{s_1, \ldots, s_{k-1}\}$ is defined as

$$\delta_k := \sup_\theta \{\ln \mathbb{P}(s_k | s_{1:k-1})\} - \inf_\theta \{\ln \mathbb{P}(s_k | s_{1:k-1})\}$$

Definition 4.4 (ε-Approximate Sufficiency, Joyce and Marjoram [2008])
Let $s_1, \ldots, s_k$ be k summary statistics. The set $s_{1:k-1} := \{s_1, \ldots, s_{k-1}\}$ ε-sufficient for statistic s_k if the score of s_k relative to $s_{1:k-1}$ is no greater that ε.

$$\delta_k \leq \varepsilon$$

ABC methods are applied in scenarios where likelihoods are intractable. This means that the score of a statistic is intractable too. Thus, Joyce & Marjoram only use the score to motivate their algorithm and in practice use different approaches to compare statistics. I discuss this in more detail later when I explore the practicalities of their algorithm.

Algorithm 4.1 (Approximately Sufficient Subset of Summary Statistics)
Adapted from [Joyce and Marjoram, 2008].

```
require: Set of summary statistics S; Score threshold ε
S' ← ∅
while true do
    Calculate the score for each statistic in S with-respect-to S'
    δ_max ← max_{s∈S} Score(s; S')
    s_max ← argmax_{s∈S} Score(s; S')
    if δ_max > ε then S' ← S' ∪ {s} ;
    else return S' ;
```

Joyce & Marjorams' algorithm (**Algorithm 4.1**) starts with an empty set and proceeds to, each iteration, add the summary statistic with the greatest score with-respect-to the set of already selected statistics, until it believes that the none of the remaining unselected summary statistics extracts a significant amount of extra information about the model parameters. They define the concept of ε-approximate sufficient sets to formalise this stopping condition, with the algorithm running until the set of accepted summary statistics S' is ε-approximate sufficient for each unchosen summary statistic, individually. This makes ε is a parameter of the algorithm, with smaller values likely leading to more summary statistics being accepted as the threshold for the amount of extra information extracted by each new statistic is lower. Alternatively, we could fix or cap the number of summary statistics we want to be accepted from the superset.

As mentioned, in practice the score cannot be calculated. Joyce & Marjoram instead determined that a proposed statistic introduces significant extra information if the posterior of parameters accepted under its usage was significantly different from the posterior when it was not used. This approach, set out in **Algorithm 4.2**, consists of estimating the expected value and standard deviation for the number of occurrences of each parameter value; and then accepting the proposed statistic if any of the observed number of occurrences is more than four standard deviations away from its expected value[1]. For this approach to be computationally tractable the posterior space is discretised into M bins whose counts can be compared. When this approach is applied the stopping condition of the main algorithm is changed to be *"Stop if*

[1] In [Joyce and Marjoram, 2008] it is recommended to use a value of between one and four standard deviations

no proposed statistics were accepted in the last cycle". There are alternative stopping conditions which could be used, it is reasonable to place a cap on the number of statistics allowed to be accepted[1].

Algorithm 4.2 (Evaluate Proposed Statistic)
Adapted from [Joyce and Marjoram, 2008].

require: *Sets of accepted parameters* $\Theta_{1:k-1}, \Theta_{1:k}$*; Number of bins* M

1 $N_{1:k} \leftarrow |\Theta_{1:k}|$

2 $N_{1:k-1} \leftarrow |\Theta_{1:k-1}|$

3 $C_{1:k-1} \leftarrow \Theta_{1:k-1}$ *discretised into* M *bins*

4 $C_{1:k} \leftarrow \Theta_{1:k}$ *discretised into* M *bins*

5 $E \leftarrow \dfrac{C_{1:k-1} \cdot N_K}{N_{K-1}}$; `// Expected value of each bin`

6 $sd \leftarrow \sqrt{\dfrac{E(N_{K-1} - C_{1:k-1})}{N_{K-1}}}$; `// Standard deviation of each bin`

7 **if** *Any* $|C_{1:k} - E| > 4sd$ **then return** *Accept proposed statistic ;*

8 **else return** *Reject proposed statistic;*

The expected values E (Line 5), the standard deviations sd (Line 6) and the condition of the if statement (Line 7) are each evaluated piece-wise.

Algorithm 4.2 requires sets of parameters which were accepted under each set of summary statistics in order to compare posteriors. These sets are acquired by generating a large number of simulations of the theorised model, using parameters sampled from the model priors, and then running ABC-Rejection Sampling to determine which parameters would be accepted under each set of summary statistics[2]. This approach has the desirable property that we only need to generate simulations once, and can then use the same set of samples each time we run **Algorithm 4.2**. This property allows us to justify generating a very large number of simulations which will make the posterior estimates more accurate. Using this approach means the approximation factor ε is no longer a parameter of the algorithm, but the distance measure, acceptance kernel and bandwidth used in the ABC-Rejection Sampling step are now parameters, as well as the number of bins M and number of model simulations. Implement caching to avoid having to run ABC-Rejection Sampling multiple times for the same set of statistics will dramatically improve the computational efficiency of this approach, especially when a large super-set of statistics is being used.

A limitation of **Algorithm 4.2** is that it does not produce a numerical value which can be used to rank each proposed statistic[3], as the theoretical score would. This means we cannot choose to keep adding the highest scoring statistic, as in **Algorithm 4.1**, and instead have to consider statistics in a somewhat arbitrary order. This means that the order in which statistics are considered will effect the result of the algorithm. An imperfect solution to this is to consider statistics in a random order and whenever a statistic is accepted, consider removing each statistic which has already been chosen. Implementing this is not trivial as considerations need to be made to avoid infinite loops where the same statistics keep getting added and removed.

Algorithm 4.2 performs poorly when the supplied set of statistics include uninformative statistics. This can be seen by noticing that a summary statistic which maps to a constant

[1] A leave-one-out cross-validation could be used to determine the optimal number of statistics to use.

[2] Considerations need to be made for how the bandwidth of the kernel scale with the number of parameters. The simplest solution is for it to scale linearly.

[3] You could compare each possible subset but this would highly inefficient as it potentially requires $\binom{K!}{2}$ executions of Algorithms 4.2, where K is the number of statistics being considered, and there is no guarantee this would produce a definitive best set, due to the complex relationships between statistics.

will almost always produce a posterior which is significantly different from an informative set of statistics and therefore be accepted as a statistic despite.

4.3.2 Minimising Entropy

[Nunes and Balding, 2010] explores using the set of summary statistics which minimise the entropy of the approximate posterior distribution returned by an ABC method. [Nunes and Balding, 2010] proposes two algorithms: the first I discus in this section; and the second, a two-step approach, I discuss in section 4.3.3. Both methods consider sets of handcrafted statistics.

Definition 4.5 (Entropy H, Shannon [1948])
The entropy $H(X)$ of a probability distribution X is a measure of the information and uncertainty in distribution.

$$\begin{aligned} \textit{Discrete} \quad H(X) &:= -\sum_{x\in\mathcal{X}} \mathbb{P}(X=x)\cdot\ln\mathbb{P}(X=x) \\ \textit{Continuous} \quad H(X) &:= -\int_{\mathcal{X}} f_X(x)\cdot\ln f_X(x)dx \end{aligned}$$

where $\mathcal{X}$ is the support of distribution X.
The joint-entropy of probability distributions $X_1,\dots,X_n$ is defined as

$$\begin{aligned} \textit{Discrete} \quad H(X_1,\dots,X_n) &:= -\sum_{x_1\in\mathcal{X}_1}\cdots\sum_{x_n\in\mathcal{X}_n}\mathbb{P}(x_1,\dots,x_n)\cdot\ln\mathbb{P}(x_1,\dots,x_n) \\ \textit{Continuous} \quad H(X_1,\dots,X_n) &:= -\int f_{X_1,\dots,X_n}(x_1,\dots,x_n)\cdot\ln f_{X_1,\dots,X_n}(x_1,\dots,x_n)dx\dots dx_n \end{aligned}$$

where $\mathcal{X}_i$ is the support of distribution X_i

A greater entropy value indicates a lower amount of information in the distribution, and visa-versa. This motivates approaches which seek to minimise entropy as they will in turn maximise information. Nunes & Balding's use of entropy is equivalent to Joyce & Marjoram's use of score, the advantage of entropy is that there are well-studied methods for estimating its value. Entropy may appear to be an equivalent measure to variance, but this is only true for unimodal distributions. Entropy measures the spread of probability mass whereas variance measures the spread of the data values. The difference can be seen by considering how the values of entropy and variance change for a bimodal distribution if the distance between the two peaks is increased; entropy will not change, whilst variance will increase.

Definition 4.6 (k^{th}-Nearest Neighbour Estimator of Entropy, Singh *et al.* [2003])
Consider a distribution X with ρ different parameters and a set of parameter values Θ which were accepted during some ABC method, with $n=|\Theta|$. Singh et al. *[2003] define the k^{th}-nearest neighbour estimator of entropy as*

$$\hat{H} = \ln\left(\frac{\pi^{\rho/2}}{\Gamma\left(1+\frac{\rho}{2}\right)}\right) - \frac{\Gamma'(k)}{\Gamma(k)} + \ln(n) + \frac{\rho}{n}\sum_{i=1}^{n}\ln D_{i,k}$$

where $D_{i,k}$ is the Euclidean distance between the i^{th} accepted parameter set and its k^{th} nearest neighbour and $\Gamma(\cdot)$ is the gamma function.

In the context of summary statistic selection we want to calculate the entropy of the posterior distribution of model parameters given summary statistic values. We only ever have an approximation of this distribution and thus can only estimate its entropy. For computational efficiency it is common to discretise the approximated distribution. There are many techniques for estimating the entropy of a distribution from samples, see [Beirlant *et al.*, 1997] for an overview. Due to most models of interest in Bayesian modelling having multiple parameters and thus the posterior being multivariate, [Nunes and Balding, 2010] suggests using the asymptoticly exact k^{th}-Nearest Neighbour estimator of entropy Singh *et al.* [2003] (**Definition 4.6**).

When implementing **Definition 4.6**, determining the k^{th} nearest neighbour in an efficient manner is not trivial. A truncated insertion sort is a good approach, but has time complexity $O(kn)$ so does not scale efficiently for large values of k. [Singh *et al.*, 2003] recommends using $k = 4$ as their experiments found that greater values of k did not decrease the root-mean square error (RMSE) significantly, and so were not worth the increased computational complexity.

Using the "Best Samples" ABC-Rejection Sampling algorithm to acquire the approximate posterior used in **Definition 4.6** is advisable as it does not require the specification of an acceptance kernel and thus the same configuration can be used for all sets of summary statistics. Also, as we specify the number of simulations this step should have the same run-time each time it is called, regardless of the set of statistics being analysed, assuming that the summary statistics take trivial time to calculate. Moreover, a single set of simulations can be generated and Rejection Sampling used to determine the best simulations under each set of statistics. This saves having to generate new simulations for each set of statistics.

Algorithm 4.3 (Minimum Entropy Summary Statistic Selection)
Adapted from [Nunes and Balding, 2010].

require: *Set of summary statistics* S
1 **for** $S' \in 2^S$ **do**
2 $\quad\Theta \leftarrow$ *Parameter sets accepted from ABC-Rejection Sampling using* S'
3 $\quad\hat{H}_{S'} \leftarrow \hat{H}(\Theta)$
4 $S^*_{ME} \leftarrow argmin_{S' \in 2^S} \hat{H}_{S'}$
5 **return** S^*_{ME}

The first algorithm proposed by [Nunes and Balding, 2010] is given in **Algorithm 4.3** and is very straightforward. It calculates the entropy for each subset of the supplied set of summary statistics S and returns whichever set has the lowest entropy. A limitation of **Algorithm 4.3** is how its computational complexity scales with-respect-to the size of the set of supplied summary statistic S. As the for-loop (line 1) considers every subset, the computational complexity of the algorithm scales exponential with the size of S. The simplest mitigation of this is to only consider subsets whose size is in some specified range, this could be implemented adaptively. A more complex procedure would be to introduce a pruning algorithm which does not evaluate sets whose subsets produce high entropy values.

The estimated entropy value for a set of statistics will vary each time the algorithm is run, due to the random nature of the parameter set Θ returned by the ABC-Rejection Sampling step (Line 2). This means the set of parameters returned by **Algorithm 4.3** will vary each time it is executed. Allowing more simulations to be performed in this step will reduce the variability in the entropy results. Alternatively, you could instead run the algorithm multiple times, keeping the number of simulations performed in line 2 relatively low, and use the results to generate a mixtures model.

Algorithm 4.3 only returns the best performing set, and no other information. It could be extended to instead return the best m sets along with their entropy values so that a mixtures model could be generated.

Algorithm 4.3 only uses entropy to evaluate the sets of summary statistics. However, as justified above, having a smaller set of statistics is preferable. This preference can be encoded into the algorithm by inflating the entropy value of larger sets. How much the value should be inflated is not a trivial matter.

As each subset is assessed independently, **Algorithm 4.3** can be readily implemented using parallelisation. This will dramatically improve run time for this algorithm and is not something which can be done with Joyce & Marjorams' approximately sufficient subset approach.

4.3.3 Two-Step Minimum Entropy

The second algorithm in [Nunes and Balding, 2010], given in **Algorithm 4.4**, is an extension of the first. It uses the set of statistics S^*_{ME} returned by **Algorithm 4.3** to simulate parameter sets Θ_{acc} which are treated as if they came from the true model. Each subset of statistics is then re-assessed using these parameter sets Θ_{acc}, with the subset which optimises some error measure returned as the recommended set.

Definition 4.7 (Mean Residual Sum of Squares Error, Nunes and Balding [2010])
Let $\mathbf{X} := \{X_1, \dots, X_n\}$ be a set of observations and X^* be a target value. Residual sum of squares error (RSSE) measures the difference between the observed values and the target value by calculating the mean of the square of the residuals. A smaller RSSE value indicates less error as the observed values do not deviate much from the target value.

$$RSSE(\mathbf{X}, X^*) := \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|X_i - X^*\|}$$

where $\|\cdot\|$ is the Euclidean distance.

Now define $\mathbf{X}^* := \{X^*_1, \dots, X^*_m\}$ to be a set of target values. The mean residual sum of squares error (MRSSE) is the mean RSSE value for each target value with-respect-to the observed data X.

$$MRSSE(\mathbf{X}, \mathbf{X}^*) := \frac{1}{m}\sum_{i=1}^{m} RSSE(\mathbf{X}, X^*_i)$$

The accepted parameter sets Θ_{acc} are treated as if they are the true parameter space distribution, this means the re-assessments now considers the error between a simulated distribution and Θ_{acc}. There are various measures which could be used, including Kolmogorov–Smirnov statistic [Chakravarti *et al.*, 1967] and cross-entropy. [Nunes and Balding, 2010] chooses to use the Mean Residual Sum of Squares Error (MRSSE, **Definition 4.7**) with the set of statistics which minimises MRSSE with-respect-to Θ_{acc} returned as the recommended set of statistics.

MRSSE is a worthwhile statistic to use in the context of Bayesian modelling as there are theoretical results which prove that minimising MRSSE is a good metric for estimating the mean of a distribution and that posterior means are optimal summary statistics. MRSSE is straightforward to compute and can be applied to multivariate distributions. Note that the scale of parameter values will effect the MRSSE and thus parameter values should be standardised before computation. A limitation of MRSSE is its sensitivity of outlier values, which is not mitigated by standardisation.

Algorithm 4.4 (Two-Step ME Summary Statistic Selection)
Adapted from [Nunes and Balding, 2010].

require: *Observations from true model x_{obs}, Set of summary statistics S, Number of simulations to run n_{run}, Number of simulations to accept n_{obs}*

1 $S_{ME} \leftarrow$***Algorithm 4.3(S)***
2 $\hat{\Theta}_{ME} \leftarrow$*Parameter sets accepted from "Best Samples" ABC-RS*$(x_{obs}, S_{ME}, n_{run}, n_{acc})$
3 *Standardise* $\hat{\Theta}_{ME}$
4 **for** $S' \in 2^S$ **do**
5 $\quad\Theta_{acc} \leftarrow$*Parameter sets accepted from "Best Samples" ABC-RS*$(x_{obs}, S', n_{run}, n_{acc})$
6 $\quad$*Standardise* Θ_{acc}
7 $\quad MRSSE_{S'} \leftarrow MRSSE(\Theta_{acc}, \hat{\Theta}_{ME,i})$
8 $S^* \leftarrow argmin_{S' \in 2^S} MRSSE_{S'}$
9 **return** S^*

Algorithm 4.4 inherits many of the limitations of the **Algorithm 4.3**, namely those concerning how its performance scales with the size of S and the use of minimum entropy. The mitigations for these are the same as discussed in Section 4.3.2. Additionally, to reduce the number of subsets being evaluated in the for-loop (line 4). As **Algorithm 4.4** requires the running of **Algorithm 4.3** it will always have greater computational complexity.

4.3.4 Semi-Automatic ABC

[Fearnhead and Prangle, 2011] presents the first algorithm which constructs its own summary statistics for ABC methods, rather than choose from a set of hand-crafted ones. Their approach (**Algorithm 4.5**) perfroms a pilot run of an ABC method to generate a naïve approximation of the parameter posterior which is then used to generate summary statistics. The approximate posterior is used to generate a "training set" from which a regression model is fitted. Model parameters are assumed to be independent and one summary statistic is generated per each model parameter. The generated summary statistics target the posterior mean, an optimal summary statistic, and should be used in a proper running of ABC to generate parameter posteriors.

This approach is referred to as semi-automatic as it requires the user to specify the summary statistics used in the pilot run of ABC. The identity function would be an appropriate choice, although inefficient.

Algorithm 4.5 (Semi-Automatic ABC)
Adapted from [Fearnhead and Prangle, 2011].

require: *Observations from true model x_{obs}, Set of summary statistics S, Number of simulated parameter sets m, Theorised model X*

1 $f_\theta \leftarrow$*Posterior from pilot run of an ABC method using x_{obs} and S*
2 $\hat{\Theta} \leftarrow m$ *simulations from* f_θ
3 $X_{\hat{\theta}} \leftarrow X\left(\hat{\theta}\right)$ *for each* $\hat{\theta} \in \hat{\Theta}$
4 *Generate summary statistics using* $\hat{\Theta}$ *and* $\{X_{\hat{\theta}}\}_{\hat{\theta} \in \hat{\Theta}}$

Regression methods are used in line 4 with the goal of creating mappings from the simulated response data $X_{\hat{\theta}}$ and the generated parameter values $\hat{\Theta}$. The best regression methods are those which target the expected value of the parameter as the posterior mean is an optimal summary statistic. There are several approaches which can be taken, I outline three here

1. Linear regression [Fearnhead and Prangle, 2011] assumes that the model can be expressed

as $\mathbf{y} = \alpha + \boldsymbol{\beta}^T X + \varepsilon$ where X is the explanatory variables, $\mathbf{y}$ is the response variables[1], $\alpha \in \mathbb{R}, \beta \in \mathbb{R}^{|\theta|}$ are coefficients to be fitted and ε is some zero-mean additive noise which can be modelled by a random variable. Linear regression seeks to find the values $\hat{\alpha}, \hat{\boldsymbol{\beta}}$ which optimises some loss function

$$\begin{aligned}\hat{\alpha}, \hat{\boldsymbol{\beta}} &= \operatorname{argmin}_{\alpha,\boldsymbol{\beta}} \sum_i L\left(\mathbb{E}[y|\mathbf{x}_i, \alpha, \boldsymbol{\beta}] - y_i\right) \\ &= \operatorname{argmin}_{\alpha,\boldsymbol{\beta}} \sum_i L\left(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i - y_i\right)\end{aligned}$$

Linear regression works well when each response variable is independent and can easily be extended to projections of X by replacing all X terms with $f(X)$ where $f(\cdot)$ is a (potentially non-linear) function. This is useful in the context of ABC methods as we can define $f(\cdot)$ to be our summary statistics.

Linear regression is a well study problem and there any many tractable solutions with least-squares estimation being perhaps the most popular. In ordinary least-squares estimation the quadratic loss function L_2 is used meaning the problem is to find

$$\begin{aligned}\hat{\alpha}_{LSE}, \hat{\boldsymbol{\beta}}_{LSE} &= \operatorname{argmin}_{\alpha,\boldsymbol{\beta}} \sum_i \left(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i - y_i\right)^2 \\ &= \operatorname{argmin}_{\alpha,\boldsymbol{\beta}} \sum_i \left(\alpha + \boldsymbol{\beta} \mathbf{x}_i - y_i\right)^T \left(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i - y_i\right)\end{aligned}$$

A closed-form estimator for these quantities is known [Hayashi, 2000].

$$(\hat{\alpha}_{LSE}, \hat{\boldsymbol{\beta}}_{LSE}) = \left(\tilde{X}^T \tilde{X}\right)^{-1} \tilde{X}^T \mathbf{y}$$

where $\tilde{X}$ is X with a column of 1s at the start for the constant term. There are extensions of ordinary least-squares which allow for weighting of variables and for the model to be heteroscedasticity. These extensions are not relevant to the problems being covered in this project.

2. Lasso regression [Hastie *et al.*, 2009] seeks the vector $\hat{\beta}$ which satisfies the following expression

$$\begin{aligned}\hat{\beta} &= \operatorname{argmin}_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{\rho} x_{ij}\beta_j\right)^2 \\ \text{subject to} \quad & \sum_{j=1}^{\rho} |\beta_j| \le t\end{aligned}$$

   where X are the explanatory variable values, $\mathbf{y}$ are the response variable values, $\rho := |X_i|$ is the number of model parameters and t is a restriction on the size of regression coefficients.

   Lasso and Ridge regression have the same objective function, but ridge regression uses an L_2 penalty function rather than lasso's L_1 function. An L_1 penalty function is preferable for feature selection as it shrinks coefficient values to zero more aggressively than an L_2 function, this is useful if the coefficient for a feature is (near) zero then the feature can be dropped.

3. Canonical correlation analysis (CCA) [Mardia *et al.*, 1979] splits variables into two sets $\mathbf{X}, \mathbf{Y}$[2] and basis vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are sought such that the linear combinations $\psi := \boldsymbol{\alpha}^T \mathbf{X}$, $\phi := \boldsymbol{\beta}^T \mathbf{Y}$ are as correlated as possible.

$$\boldsymbol{\alpha}, \boldsymbol{\beta} = \operatorname{argmax}_{\boldsymbol{\alpha},\boldsymbol{\beta}} \operatorname{Corr}(\boldsymbol{\alpha}\mathbf{X}, \boldsymbol{\beta}\mathbf{Y})$$

[1] In Bayesian modelling context typically X is set to the observed values x_{obs} and y are set to the model parameters θ.

[2] For Bayesian modelling you typically set $\mathbf{X}$ to be the model parameters and $\mathbf{Y}$ to be observed values.

Solutions to this are known and readily calculatable.

$$\begin{aligned}\boldsymbol{\alpha} &= \Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} \\ \boldsymbol{\beta} &= \Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\end{aligned}$$

where Σ_{UV} is the cross-covariance matrix of random vectors U, V. R provides an inbuilt function `cancor`.

As Lasso uses the L_1 penalty function, which is non-linear, there is no closed expression of Lasso regression. Meaning that computing a solution to Lasso has $O(N^2)$ time-complexity. [Fearnhead and Prangle, 2011] recommends the use of linear regression as it is straightforward to implement and does not perform notably worse than the other approaches in general.

For the specific implementation of linear least-squares regression in [Fearnhead and Prangle, 2011], each model parameter θ_i is treated completely separately and mappings $f(\cdot)$ of the response data are allowed. This means they are fitting $\rho = |\theta|$ different models to the relationship (42).

$$\theta_i = \alpha^{(i)} + (\boldsymbol{\beta}^{(i)})^T f(\mathbf{x}) + \varepsilon_i \tag{42}$$

Since ABC methods only consider the distance between summary statistic values, the constant terms $\alpha^{(i)}$ in (42) can be neglected from our generated summary statistics. This means the summary statistic s_i for the i^{th} model parameter is defined as (43).

$$s_i(\mathbf{x}) = \hat{\beta}^{(i)} f(\mathbf{x}) \tag{43}$$

The mapping $f(\cdot)$ is a parameter of this **Algorithm 4.6** and should be used to encode likely relationships between observations and parameters. It can just be set to the identity function for simplicity. As the mapping is part of the generated summary statistic s_i, it is important for it to be computationally efficient in order for the summary statistic to be efficient.

Algorithm 4.6 (Semi-Automatic ABC - Least Squares)

require: *Observations from true model x_{obs}, Set of summary statistics S, Number of simulated parameter sets m, Theorised model X, Mapping $f(\cdot)$*

1 $f_\theta \leftarrow$ *Posterior from pilot run of an ABC method using x_{obs} and S*
2 $\hat{\Theta} \leftarrow m$ *simulations from f_θ*
3 $X_{\hat{\theta}} \leftarrow X\left(\hat{\theta}\right)$ *for each $\hat{\theta} \in \hat{\Theta}$*
4 $\hat{X} \leftarrow \{X_{\hat{\theta}_1}, \dots, X_{\hat{\theta}_m}\}$
5 $F \leftarrow f(\hat{X})$
6 $\tilde{F} \leftarrow F$ *with a preceding column of 1s*
7 **for** $i = 1, \dots, \rho$ **do**
8 $\quad A_i \leftarrow i^{th}$ *element of each set in $\hat{\Theta}$*
9 $\quad (\alpha^{(i)}, \boldsymbol{\beta}^{(i)}) \leftarrow (\tilde{F}^T\tilde{F}^{-1})\tilde{F}^T A_i$
10 $\quad s_i(\mathbf{x}) := \boldsymbol{\beta}^{(i)}\mathbf{x}$
11 **return** $\{s_1, \dots, s_\rho\}$

$\rho := |\theta|$, the number of model parameters.

Algorithm 4.6 is a restatement of the general algorithm **Algorithm 4.5** using linear least-squares regression. Any ABC method can be used for the pilot run (Line 1), using "Best Samples" ABC-Rejection Sampling is the simplest as it has simpliest acceptance criteria to define and the most predictable run-time. Further, any set of summary statistics S can be used to. The pilot run is an opportunity for expert knowledge to be encoded into the model by hand-crafted statistics, but, as this algorithm will mainly be run when such statistics are not known, the identity function can be used for simplicity and guaranteed sufficiency. The closer the posterior produced by the pilot run, the more representative the generated values (lines 2-3) will be and thus the more informative the regression fit will be, creating better summary statistics. The other opportunity to encode expert knowledge is in the specification of mapping $f(\cdot)$.

The least-squares approach used in **Algorithm 4.6** treats each model parameter as fully independent. This may not be true and ignoring this may lead to missed insights. Different regression approaches can be used to maintain dependencies between parameters (e.g. CCA). The generated summary statistics offer little insight or interpretability, on their own, but can be viewed intuitively as posterior mean estimators due to how they generated. This approach generates one summary statistic for each model parameter, if it could incorporate dependencies between model parameters then the total number of summary statistics could be reduced, increasing the compression level.

Using the generated summary statistics in ABC methods is not straightforward as we lack the intuition required to defined acceptance criteria. The use of adaptable versions of the ABC methods mitigates this issue as you only have to specify target acceptance rates.

4.3.5 Non-Linear Projection

The semi-automatic approach of [Fearnhead and Prangle, 2011] does allow for non-linear projections from the response data x to the parameter values θ but the user needs to specify the non-linear functions. More specifically, **Algorithm 4.6** produces non-linear projections if, and only if, the mapping $f(\cdot)$ is non-linear.

Being able to generate non-linear projections is desirable as it is not guaranteed that an (accurate) linear projection from response variables to model parameters exists. [Wong *et al.*, 2018] presents the first attempt at using a deep neural-network[1] to construct summary statistics. The general approach to ABC is the same as [Fearnhead and Prangle, 2011]: Perform a pilot run to generate training data; Train a neural network to fit response values to parameter values; And, then use the trained network to calculate summary statistic values for a proper run of ABC. Due to the flexibility of DNNs the number of outputs (i.e. the dimensions of the summary statistic) can be specified to any value, although more outputs require more training time and potentially a larger network.

The network used to demonstrate this approach in [Wong *et al.*, 2018] is fairly small with three hidden layers, with 5-5-3 nodes each, and takes all the observed data as an input. The model was trained to fit to parameter values, resulting in summary statistics which approximate the posterior mean. They demonstrate their method on an Ising model and moving-average model and show it to outperform the usage of hand-crafted summary statistics and semi-automatic ABC. The trade-off is that their DNN approach requires significantly more time than the other approaches, requiring twenty minutes when the semi-automatic approach of [Fearnhead and Prangle, 2011] required less than one.

A natural extension to this approach is to apply a mapping to the observed data before it is passed to the network, as in semi-automatic ABC. This would allow for the encoding of expert knowledge into the network which would mean a smaller network is required, reducing training time.

This use of a neural network is liable to the same issues as many other uses, with the most prevalent being overfitting. Overfitting occurs when a neural network models the training

[1] A feedforward neural-network is presented too, but these cannot model non-linear relationships unless they use a non-linear activation function.

data too closely and therefore does not perform well with more general data. Early stopping and regularisation are standard practices to mitigating overfitting. Additionally, improving the training set can help too. The training set can be improved by increasing its size and its diversity so that it is more representative of the general space. In this particular context, as the training set is generated from a posterior from a pilot run of ABC, we can improve the quality of the training set by improving this posterior. Allowing the pilot run to complete more simulations is guaranteed to improve the posterior when using the identity function as the only summary statistic, due to the sufficiency of the identity function. Alternatively, the use of less naïve statistics will help too but it can be hard to identify these in practice. Using neural networks offers no interpretability of what inferences are being made, without a very intricate investigation of the network.

5 ABC and Epidemic Events

In this section I: Compare the performance of ABC methods discussed in *Section 3* by applying them to toy examples of the compartmental models discussed in *Section 2 Section 5.1*; And, assess the performance of summary statistic selection methods discussed in *Section 4* when applied to compartmental models *Section 5.2*. I use these results to present what effect dimensionality of summary statistics has on the ability of ABC methods to fit models *Section 5.3*; and, present how these ABC and selection methods perform when tasked with projecting the future course of an epidemic *Section 5.4*. I close this section by applying these methods to two sets of real data *Section 5.5*.

5.1 Comparison of ABC Methods

To compare the ABC methods discussed in *Section 3* I begin with a toy example of the standard SIR model. This is to justify that these methods are effective at fitting such models and to create a benchmark from which to compare the summary statistic selection methods. More specifically, the ABC methods are used to fit an SIR model to data generated by an SIR model with parameters $\beta = 1$, $\gamma = 0.5$ and constant population size $N = 100,000$. A realisation of this SIR model is given in *Figure 2.2*. This specification of the model was chosen arbitrarily, but it does have some desirable features such as the full epidemic being completed in less than 30 time-periods and that not all of the population becomes infected.

I perform two experiments to assess the performance of the algorithms: The first supplies the algorithms with the full 30-day time-series; and, the second is a leave-one-out cross-validation (LOO-CV) where each algorithm is run thirty times each time with a different day's data missing. This cross-validation allows for an assessment of the general performance of each algorithm by training on incomplete data and then assessing the predictive performance of the algorithm for the missing data point. As the cross-validation assess the algorithms under several scenarios it is practically impossible to tune the hyper-parameters of the algorithms to maximise performance, and rather one should aim for generalised hyper-parameter configurations.

It is very difficult to set up these experiments to make a fair comparison between the algorithms. In practice, the limiting factor when implementing these algorithms is time meaning a proper assessment should set a time-limit for each algorithm. However, the model I am fitting on is relatively simple and in practice this time-limit would likely be several days, meaning any half-decent algorithm will produce an accurate result by the cutoff-time. Rather, I cut each algorithm after 5,000 simulations[1] (except for ABC-Rejection Sampling which performs 50,000 and is still by far the quickest algorithm). This means my experiments focus on assessing the algorithms ability to extract information from each simulation, rather than on the asymptotic behaviour of each algorithm.

Experiment Methodology

I assess five different ABC methods: "Best Samples" ABC-Rejection Sampling, *Algorithm 3.3*; ABC-Importance Sampling, *Algorithm 3.4*; ABC-MCMC, *Algorithm 3.5*; ABC-SMC, *Algorithm 3.6*; and, an adaptive version of ABC-SMC, specified below.

Each ABC method has different hyper-parameters which need to be specified, many of which are unique and some are shared. Priors, summary statistics, and distance measures are

[1] 5,000 was chosen as pilot experiments showed that it allowed all algorithms to perform well, in a relatively short period of time (minutes rather than hours), without any algorithm converging completely. The quick run-times of each algorithm allows for multiple runs of each experiment to be made and the variability between each to be assessed.

required by all the methods and I specify each of these the same for each method. An acceptance kernel is required by all methods except "Best Samples" ABC-Rejection Sampling.

The specification of a set of priors is required for all the ABC methods. Relatively uninformative priors[1] were chosen for both model parameters, with $\mathcal{U}[0, 2.5]$ being specified for β and $\mathcal{U}[0, 0.8]$ for γ. These priors were chosen in part because their means are not equal to the true value of the parameters they represent. If this was not case then, in the case where the methods perform poorly and the generated posterior is very close to the prior, a model with posterior means would have low error despite no learning having occurred.

The identity function $s(x) = x$ was used as the summary statistic for each algorithm due to it being sufficient and thus meaning the algorithms would converge on the true posterior. The use of the identity function is not necessarily ideal as it means the distance measure has to compare 90 data-points, as the same distance measure is used by all algorithms the issues should be shared. However, the choice of distance measure will affect ABC-IS and ABC-SMC more as distances are used to weight accepted parameter sets.

The Euclidean distance of the logarithm of each value (44) was used as the distance measure. Logarithms were taken due to the nature of the values in the SIR model. If raw values are compared, rather than logarithms, then much greater emphasis would be placed on fitting to the greatest values (The susceptible population in the early time-steps and the removed population in the later time-steps) which is not ideal as it means a loser fit is required for the lower values, such as the infectious population size.

$$\|x - y\| = \sqrt{\sum_{i=1}^{n} \big(\ln(x_i) - \ln(y_i)\big)^2} \text{ where } n = |x| = |y| \tag{44}$$

I only assess the "Best Samples" version of ABC-Rejection Sampling as it reduces the effect that user choices have and allows for a more generally assessment of the algorithms performance. Moreover, I specified the algorithm to generate 50,000 simulations and to accepted the closest 500 (1%). For the other algorithms I specified a Gaussian acceptance kernel and roughly tuned their bandwidths to achieve recommended acceptance rates. An acceptance rate of 22%-25% was targeted for ABC-MCMC (As justified in *Section 3.2.3*). For ABC-SMC a sample size of 100 and 20 bandwidth steps were used. The bandwidths were evenly spaced on a $\log_1 0$ scale so that they tightened by a given percentage each step, creating a consistent tightening of the acceptance space each step.

ABC-MCMC and ABC-SMC require the specification of perturbance kernels. These were set to be additive gaussian noise with variance 0.1 for both parameters. Pilot experiments showed that this value of the variance produced good mixing and the desired acceptance rates, although little difference is seen when using values of a similar order of magnitude.

In *Section 3.2.4* an adaptive version of ABC-SMC is discussed which requires the fewest hyper-parameter specifications (Hereby known as "Adaptive ABC-SMC"). This method automatically sets the perturbance kernel variance to be twice the dimension-wise variance of the previously accepted set of particles and the acceptance kernel bandwidth is automatically set such that only the first $\alpha\%$ of the previously accepted set of particles would have been accepted, were $\alpha \in [0, 100]$ is defined by the user. For ease-of-implementation, I defined the acceptance kernel to be the uniform kernel and $\alpha = 90$ as this means the acceptance kernel bandwidth is just the 90^{th} percentile of the distance values of particles accepted in the previous step.

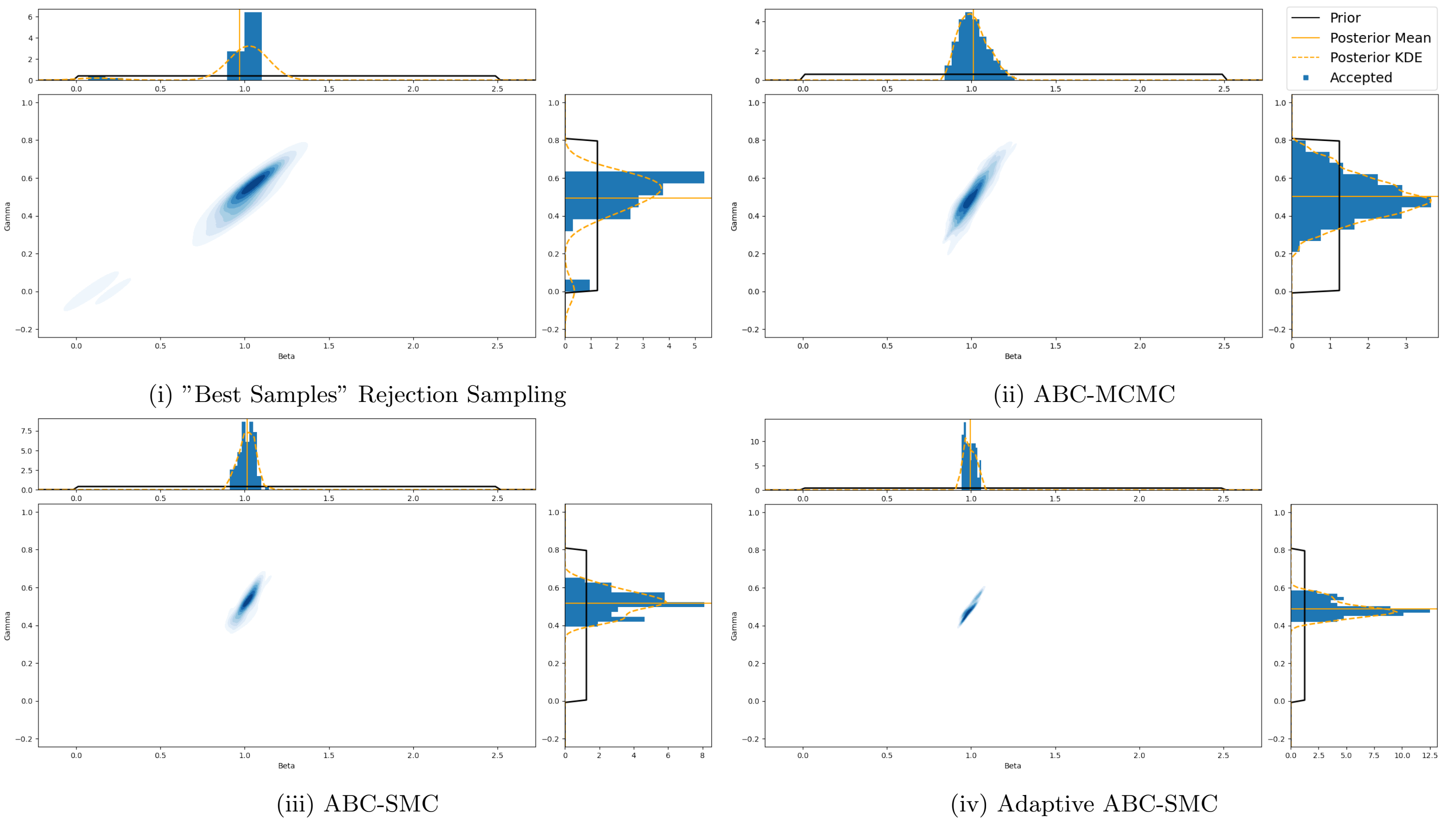
[1]Note that β can be any non-negative value and γ can only take values in the interval $[0, 1]$.

Results - Experiment One

Algorithm	**RSSE**	**Time (s)**
"Best Samples" ABC-Rejection Sampling	3,721	5
ABC-Importance Sampling	3,541	11
ABC-MCMC	3,021	37
ABC-SMC	2,393	342
Adaptive ABC-SMC	2,563	1,451

Algorithm	**95% CI R_0**	**95% CI β**	**95% CI γ**
"Best Samples" ABC-Rejection Sampling	[1.763,2.481]	[0.133,1.098]	[0.000,0.621]
ABC-Importance Sampling	[1.803,2.318]	[0.410,1.102]	[0.206,0.654]
ABC-MCMC	[1.612,2.927]	[0.871,1.194]	[0.290,0.736]
ABC-SMC	[1.723,2.318]	[0.921,1.089]	[0.411,0.632]
Adaptive ABC-SMC	[1.852,2.235]	[0.947,1.057]	[0.422,0.572]

Table 5.1: The Residual Sum of Squares Error (RSSE) of the natural logarithm of values; 95% credible intervals for R_0, β & γ; and, the execution time of different ABC methods when given the full time-series and using the identity function as the summary statistic. Values are the mean of 50 independent runs of each algorithm. True values: $R_0^* = 2,\ \beta^* = 1,\ \gamma^* = 0.5$.



(i) "Best Samples" Rejection Sampling

(ii) ABC-MCMC

(iii) ABC-SMC

(iv) Adaptive ABC-SMC

Figure 5.1: Joint and marginal posterior distributions of β, horizontal axis, and γ, vertical axis, parameters generated by different ABC methods when provided with the full time-series of data.

The first experiment ran each algorithm 50 times on the full time-series data, with the configurations described above. *Table 5.1* provides a summary of these results and *Figure 5.1* provides the marginal and joint posteriors returned by each algorithm.

It is apparent from the distributions in *Figure 5.1* that all methods have achieved a significant level of learning as the priors (black) and estimated posteriors (blue/orange) are highly dissimilar. This is clear when you observe the plots of the joint distribution as the limits of the plot include the whole prior space but the vast majority of each plot is white, indicating effectively no posterior mass is placed in those regions. Moreover, each algorithm has estimated a posterior which significant posterior mass around the true parameter values.

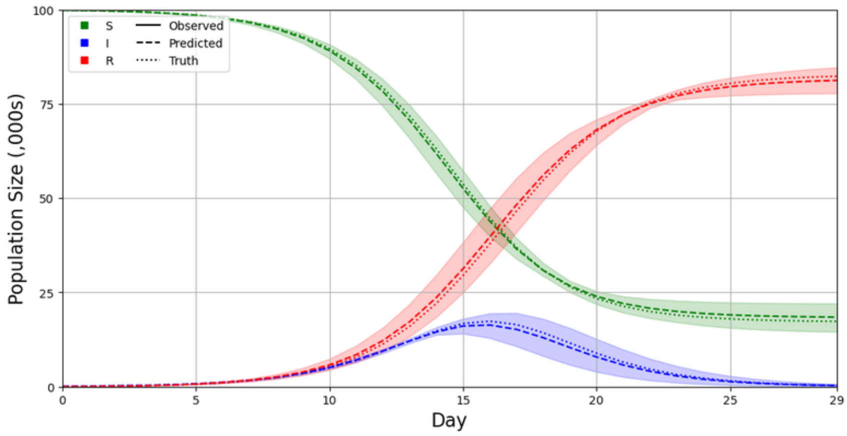


Figure 5.2: 95% Confidence Intervals for model fitted to the SIR model described in *Figure 2.2* using adaptive ABC-SMC with 5,000 simulations and the identity function as the summary statistic. 95% credible interval for R_0 is [1.784,2.266].

The best fitting model is produced by the ABC-SMC algorithm (RSSE=2,393) with the Adaptive ABC-SMC producing only a slightly worse fit (RSSE=2,563, See *Figure 5.2*) . This is highly encouraging for the Adaptive ABC-SMC algorithm as it requires significantly less tuning then all the other algorithms, except for "Best Samples" ABC-Rejection Sampling which does not perform nearly as well (RSSE=3,721). However, the Adaptive ABC-SMC algorithm took on average 4.24 times as long to be executed than the ABC-SMC algorithm. This is unsurprising as many of the first iterations of Adaptive ABC-SMC have almost a 100% acceptance rate whilst the acceptance kernel bandwidth is still large. This high acceptance rate is inefficient as significant computation time occurs at the end of each step of the Adaptive ABC-SMC algorithm in order to calculate perturbance kernel variances and acceptance kernel bandwidths, meaning the average computation time per simulation is notably greater early on in the algorithms run.

By inspecting *Table 5.1* we can note that all the algorithms produce 95% credible intervals which include the true value of both model parameters. The credible intervals produced for β are notably tighter than those produced for γ in all cases except for "Best Samples" ABC-Rejection Sampling. This is surprising as the magnitude of β is greater than that of γ so we would expect small changes in γ to have a greater affect on the resulting model fit. This is likely a result of the peak of infections occurring in the later half of the time-series (day 17 of 30) and thus there being more days in which infections are increasing.

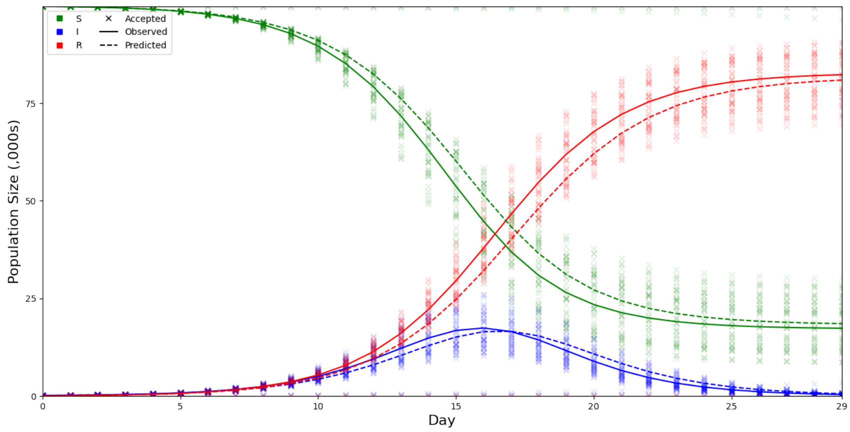


Figure 5.3: The 1% best generated samples which were accepted by ABC-Rejection Sampling (crosses) when given the full time-series and using the identity function as the summary statistic. Includes the observed data (solid line) and the data from a model fitted with the posterior mean values for each parameter (dashed line).

The "Best Samples" ABC-Rejection Sampling algorithm is the only one to produce a bi-modal posterior. The majority of the estimated posterior mass is placed around the true values of the parameters, but a non-negligible amount is also placed around $(\beta, \gamma) = (0, 0)$. This pair of parameter values is the unique case where no infections ever occur. *Figure 5.3* plots all the simulated values which were accepted by a particular run of the "Best Samples" ABC-Rejection Sampling algorithm, along with the observed data (solid line) and the data generated by a model fitted with posterior means (dashed line). In *Figure 5.3* we can clearly see these special cases where $(\beta, \gamma) \simeq (0, 0)$ as their simulated values never leave the horizontal axes. This plot motivates the ability to set the acceptance rate of the "Best Samples" flavour of ABC-Rejection Sampling after all samples have been generated as we clearly see that these special cases will have distinctly greater distance values that the other accepted samples and thus would have just snuck into the top 1% of samples. We can justify tightening the acceptance criteria to exclude these simulations as they are demonstrably incorrect.

Results - Experiment Two

Algorithm	**LOO-CV Score**
"Best Samples" ABC-Rejection Sampling	6,135
ABC-Importance Sampling	5,001
ABC-MCMC	3,023
ABC-SMC	643
Adaptive ABC-SMC	439

Table 5.2: Leave-One-Out Cross-Validation Scores for different ABC algorithms when applied to the SIR model described in *Figure 2.2*. Each algorithm performed approximately 5,000 simulations each iteration, had rough tuning to achieve appropriate acceptance rates and used the identity function as the summary statistic.

The second experiment assessed the general performance of the ABC methods when applied to standard SIR models by performing a leave-one-out cross-validation of each algorithm. A leave-one-out cross-validation was performed on each algorithm when applied to the SIR model depicted in *Figure 2.2*, with each algorithm configured as described at the start of this section. The score of each algorithm is given in *Table 5.2*.

The Adaptive of ABC-SMC achieves the lowest LOO-CV score, suggesting it performs the best in general scenarios. This is unsurprising as each iteration of the cross-validation sees a different piece of data missing and thus the optimal configuration of the hyper-parameters will be different each time. As Adaptive ABC-SMC automatically updates the majority of its hyper-parameters it should perform well in this test, and does.

The difference in scores between each algorithm in *Table 5.2* than in *Table 5.1*, when the full time-series was supplied to each algorithm. This demonstrates the effect that tuning of parameters has on the performance of each algorithm.

A closer inspection of the results from the cross-validation showed that the majority (40-60%) of the total error from each algorithm occurred when one of the data-points from days 14-18 was removed. This is due to these days coincide with the peak-infectious population size which is a key point in the lifespan of an epidemic and is the point at which more information is available about parameter β than γ.

5.2 Comparison of Summary Statistic Methods

Section 4 presented four methods for automating the process of choosing summary statistics for ABC methods: Approximately Sufficient Subset (ASS); Minimum Entropy (ME); Two-Step Minimum Entropy (2-ME); and, Semi-Automatic ABC (SA-ABC). Here I assessed these algorithms by running 50 instances of each algorithm; Running the Adaptive ABC-SMC algorithm, for 5,000 simulations with target acceptance rate $\alpha = 90\%$, using the most recommended set of summary statistics from each algorithm; and then calculating the mean-square error produced by a model fitted with the estimated posterior mean of each parameter.

- (S1) *Peak size of infectious population (1).
- (S2) Date of peak infectious population (1).
- (S3) *Final size of susceptible population (1).
- (S4) *Final size of infectious population (1).
- (S5) *Final size of removed population (1).
- (S6) *Mean size of susceptible population (1).
- (S7) *Mean size of infectious population (1).
- (S8) *Mean size of removed population (1).
- (S9) *Maximum day-on-day increase in infectious population (1).
- (S10) *Maximum day-on-day increase in removed population (1).
- (S11) *Population sizes on days 1,. . . ,30 as different statistics (3).
- (S12) *Net weekly change in susceptible population (4).
- (S13) *Net weekly change in infectious population (4).
- (S14) *Net weekly change in removed population (4).
- (S15) *Net daily change in susceptible population (29).
- (S16) *Net daily change in infectious population (29).
- (S17) *Net daily change in removed population (29).
- (S18) *Cummulative sum of daily susceptible population (29).
- (S19) *Cummulative sum of daily infectious population (29).
- (S20) *Cummulative sum of daily removed population (29).
- (S21) Uniform random variable in [10,22] (1).
- (S22) Constant mapping to 16 $s(x) = 16$ (1).

Table 5.3: Comprehensive set of 99 summary statistics supplied to the ASS, ME and 2-ME algorithms. An asterisk (*) denotes that a summary statistic of the natural logarithm of this value was supplied as well. In brackets is the dimensionality of each summary statistic when applied to the SIR-model depicted in *Figure 2.2* which runs for 30 time-periods.

Three of the algorithms presented in *Section 4* (ASS,ME and 2-ME) choose an optimal set of summary statistics from a larger set. *Table 5.3* provides the set of summary statistics supplied to these algorithms. ((S1),(S2)) are taken from [Blum and Tran, 2010] which found the peak size of the infectious population and the date on which this peak occurred to be the best performing summary statistics when applied to data regarding HIV in Cuba[1]. ((S21),(S22)) are totally uninformative and were supplied to test whether the algorithms would ever choose such summary statistics. These two summary statistics were specified to both have an expected value of 16 so that they coincide with (S2) for the true model.

All four algorithms were configured to use the log-Euclidean distance (44). The ASS algorithm was configured to generate 100,000 simulations and to use a Gaussian acceptance kernel with bandwidth .5. These simulations and kernel were used to determine which sets of parameters to accept for each set of summary statistics, and these sets were then supplied to *Algorithm 4.2* in order to determine whether to accept the newly proposed summary statistic.

Algorithm ME was configured to generate 100,000 simulations from which the best 1,000 were chosen (using "Best Samples" ABC-Rejection Sampling) and entropy estimated from. Algorithm 2-ME was configured to run ME, as just described, to identify the initial set of summary statistics S_{ME} and to retain the best 100 simulations $\hat{\Theta}_{ME}$ according to S_{ME}. 100,000 fresh simulations were then generated and the 1,000 best simulations Θ_{acc} were identified for each set of summary statistics (using "Best Samples" ABC-Rejection Sampling) and the MRSSE was calculated between Θ_{acc} and $\hat{\Theta}_{ME}$. For time-tractability both algorithms were restricted to only considering subsets of summary statistics up to size four, this still

[1]It is noteworthy that they were working with yearly data, so this result may be less applicable to data collected at shorter intervals.

involved the consideration of 92,170 different subsets[1]. It is shown in *Section 5.3* than considering higher dimensional summary statistics would have been futile anyway.

The pilot posterior f_θ for SA-ABC was generated by "Best Samples" ABC-Rejection Sampling using the identity function and retaining the best 1,000 of a set of 100,000 simulations. 5,000 sets of parameters were then generated from the pilot posterior f_θ and used in the regression to calculate the summary statistics. As the SIR model has two parameters SA-ABC generates a summary statistic with two dimensions.

To compare the performance of each algorithm I ran each algorithm 50 times and identified the most commonly returned set of statistics. Adaptive ABC-SMC[2]. was then run 50 times with each set of statistics and the average MSE was calculated between models fitted with the estimated posterior means and the true model. I chose to only calculate the MSE for the most commonly returned set of statistics, rather than for each set of returned statistics and taking the mean, as each the tested methods only takes a few minutes to run and thus it is reasonable to run each algorithm many times. Moreover, for each algorithm there was a set of statistics which was chosen significantly more often than any other.

Results

Algorithm	**Returned Statistics**	**Dimensions**	**% Runs Returned**
Control	Identity Function	90	N/A
Approximately Sufficient Subset	[Log-Final Susceptible Population (S3)*]	1	44%
Minimum Entropy	[Log-Mean Infectious Population (S7)*, Log-Mean Removed Population (S8)*]	2	30%
2-Step ME	[Log-Peak Infectious Population Size (S1)*, Log-Mean Infectious Population (S7)*, Log-Mean Removed Population (S8)*]	3	34%
Semi-Automatic ABC	N/A	2	N/A

Algorithm	**ABC-SMC MSE**	**Time (s)**
Control	121,777	N/A
Approximately Sufficient Subset	101,730,336	25
Minimum Entropy	1,131,712	150
2-Step ME	228,150	302
Semi-Automatic ABC	643,255	102

Table 5.4: The most commonly returned set of summary statistics returned by each algorithm and the mean square error when using Adaptive ABC-SMC to fit the SIR model described in *Figure 2.2* with said set of statistics. Additionally, the dimensionality of each summary statistic when applied to the SIR-model depicted in *Figure 2.2*; the percentage of instances in which this set of statistics was returned; and, the mean run-time of each selection algorithm are given. All values are the means over 50 instances of each selection algorithm. Note that in all cases the natural logarithm version of a summary statistic was chosen, rather than its raw value.

Table 5.4 provides the summary of the results of running each selection algorithm 50 times and then running Adaptive ABC-SMC with the most commonly returned set of statistics. None of the returned set of statistics performs as well as the identity function, although it would be very surprising if this was the case. Both 2-ME and SA-ABC returned sets of summary statistics which produced MSE of the same order of magnitude as the identity function despite only having 3 and 2 dimensions, respectively, compared to the 90 of the identity function. The ASS algorithm

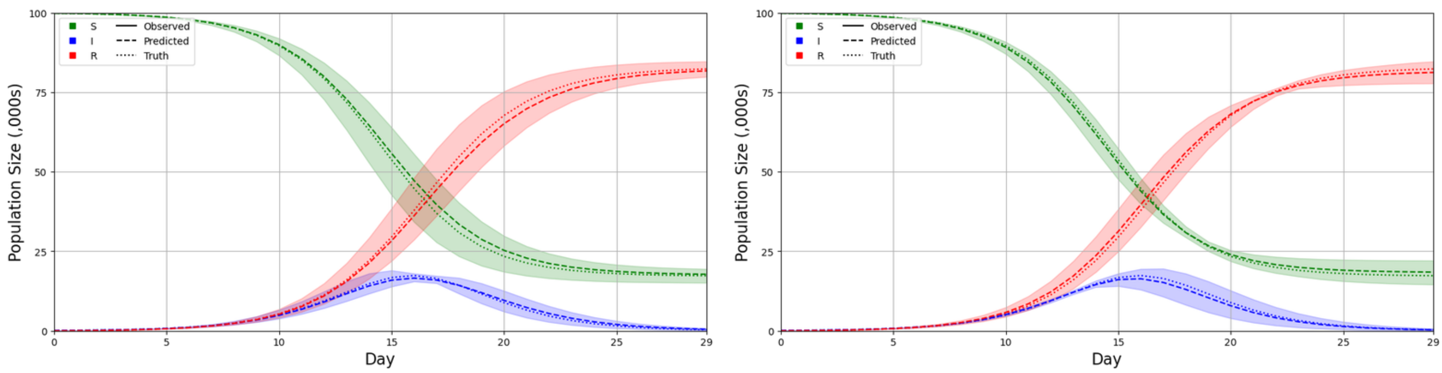
[1](S11) were removed from the set of possible summary statistics for these two algorithms as it represents 60 unique summary statistics and thus would bring the total number of sets up to size four to 3,926,175.

[2]Adaptive ABC-SMC was used as it performed best in the general setting in *Section 5.1*. Adaptive ABC-SMC was configured as described in *Section 5.1* so that the results from *Table 5.1* can be used as a benchmark.

performs very poorly, this is likely due to difficulties in implementation which mean it heavily favours low dimension summary statistics. ME produces a well performing set of summary statistics, but it is notably slower and produces a greater MSE than SA-ABC. It is noteworthy that 2-ME returns a set of summary statistics which is a superset of those returned by ME. This indicates that the second step of 2-ME does draw more nuance than ME, the motivation behind adding the second step.

These results demonstrate that their are methods for selecting summary statistics which produce summary statistics which perform similarly to the identity function despite having significantly fewer dimensions. Moreover, if these algorithms were supplied with more simulations then we can expect for the difference in performance between the identity function and generated summary statistics to decrease. This is more relevant to SA-ABC as each extra simulation means its regression step is more informed, while 2-ME can only ever return a set of statistics from a predefined set and thus more simulations would only make its returned results more consistent.

None of the algorithms returned the same set of summary statistics a majority of the time. An inspection of the distribution of returned sets shows that the sets given in *Table 5.4* were all accepted as least 3 times more often than the next most commonly returned sets. If the algorithms were given more simulations then this ratio should increase as the estimated values (i.e. entropy) will tend towards their true value, by the law of large numbers. Further, none of the algorithms returned any sets which included either of the totally uninformative summary statistics (S21),(S22).



(i) Two-Step Minimum Entropy. 95% credible interval for R_0 is [1.957,2.073].

(ii) Semi-Automatic ABC. 95% credible interval for R_0 is [1.754,2.132].

Figure 5.4: 95% CI when running adaptive ABC-SMC using summary statistics generated by two-step minimum entropy and semi-automatic ABC respectively.

Figure 5.4 provides plots of the 95% credible intervals for each population size produced by ABC-SMC when using the summary statistics returned by 2-ME (5.4i) and SA-ABC (5.4ii). (See *Figure 5.2* for the results when the identity function is used). A notable difference between the two plots its their confidence around the peak of the infection, with 2-ME having much more concentrated certainty around the true peak than SA-ABC. This is due to the set of statistics returned by 2-ME including the logarithm of the peak number of infections. Moreover, *Figure 5.4ii* perfectly fits the infectious population until the peak is reached at which point it typically expected the peak to occur later than it actually does.

This difference in the resulting fit of the two sets of summary statistics suggests that different selection algorithms should be used for different inference problems. With SA-ABC being best suited for general problems regarding the full time-series, whilst 2-ME is better when you have a specific problem (e.g. will hospitals be over-run when the peak of infections occurs) as you can supply it summary statistics which it is reasonable to expect to perform well for this problem (e.g. ((S1),(S2)) for problems regarding peak infections).

5.3 The Effect of Dimensionality

The Minimum Entropy methods for summary statistics are easily adjustable such that the dimensions of the returned summary statistics is within a certain bound. This means these methods can be used to determine the best performing summary statistics at different degrees of dimensionality, and thus discern how great an effect the dimensionality of a summary statistic affects its performance. Ideally, this will identify a point at which dimensionality is highly minimised but a good level of performance is still achieved. As discussed in *Section 4* lower-dimensional summary statistics are desirable as they typically decrease the time taken to analyse each simulation in an ABC method and thus increase the number of simulations which can be assessed in a given time-frame, leading to more exact results. Here I evaluated how the dimensionality of summary statistics affects the ability of ABC methods to fit the SIR model depicted in *Figure 2.2*.

Dimensions	**Best Statistics**	**Estimated Entropy**	**ABC-SMC MSE**
1	[Log-Mean Susceptible Population (S6)*]	-9.48	2,272,479
2	[Log-Mean Infectious Population (S7)*, Log-Mean Removed Population (S8)*]	-10.74	1,131,712
3	[Log-Final Removed Population (S5)*, Log-Mean Infectious Population (S7)*, Log-Mean Removed Population (S8)*]	-10.59	228,150
4	[Log-Net Weekly Change in Removed Population (S14)*]	-10.60	207,124
5	[Log-Net Weekly Change in Removed Population (S14)*, Log-Final Susceptible Population (S3)*]	-10.54	198,888
90	Identity Function	-11.70	121,777

Table 5.5: The set of summary statistics which produce the lowest estimated entropy value at different levels of dimensionality; Along with the MSE of a an SIR model fitted with the estimated posterior mean from the Adaptive ABC-SMC algorithm targeting the model in *Figure 2.2*, when using said set of summary statistics. Adaptive ABC-SMC was configured to run for 5,000 simulations, with a sample size of 100, using the log-Euclidean distance measure and a target acceptance rate of 90%. Entropy was estimated using the 4^{th}-Nearest Neighbour Estimator of Entropy with the 1,000 parameter sets, from a superset of 1,000,000 parameter sets, which minimise log-Euclidean distance. The stated dimensionality of each summary statistic is with-respect-to an SIR model over 30 time-periods.

To determine the best set of summary statistics at each level of dimensionality I estimated the entropy of each subset of the set of statistics given in *Table 5.3* using the k^{th}-Nearest Neighbour Estimator of Entropy formula (*Definition 4.6*). Specifically, I set $k = 4$ and use "Best Samples" ABC-Rejection Sampling, with the log-Euclidean distance, on the SIR model in *Figure 2.2* to generate 1,000,000 simulations and identify the best 1,000 sets of parameters for each subset of statistics, which was then used to estimate the entropy. *Table 5.5* provides the set of summary statistics which minimised estimated entropy at each level of dimensionality, along with the MSE of a an SIR model fitted with the estimated posterior mean from the Adaptive ABC-SMC algorithm targeting the model in *Figure 2.2*, when using said set of summary statistics. Adaptive ABC-SMC was configured to run for 5,000 simulations, with a sample size of 100, using the log-Euclidean distance measure and a target acceptance rate of 90%.

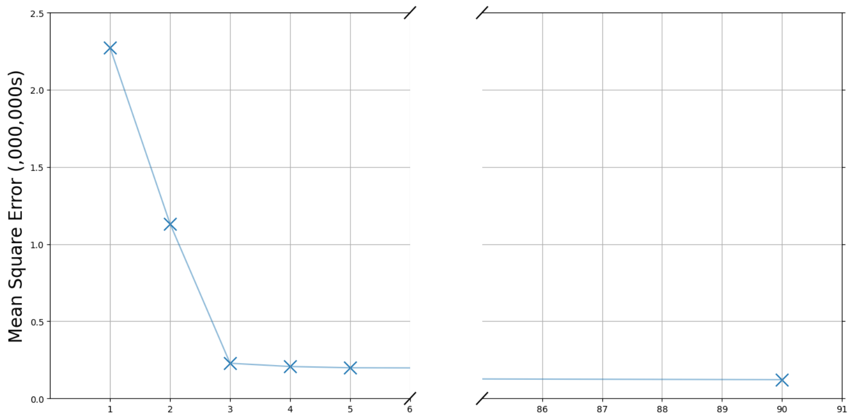


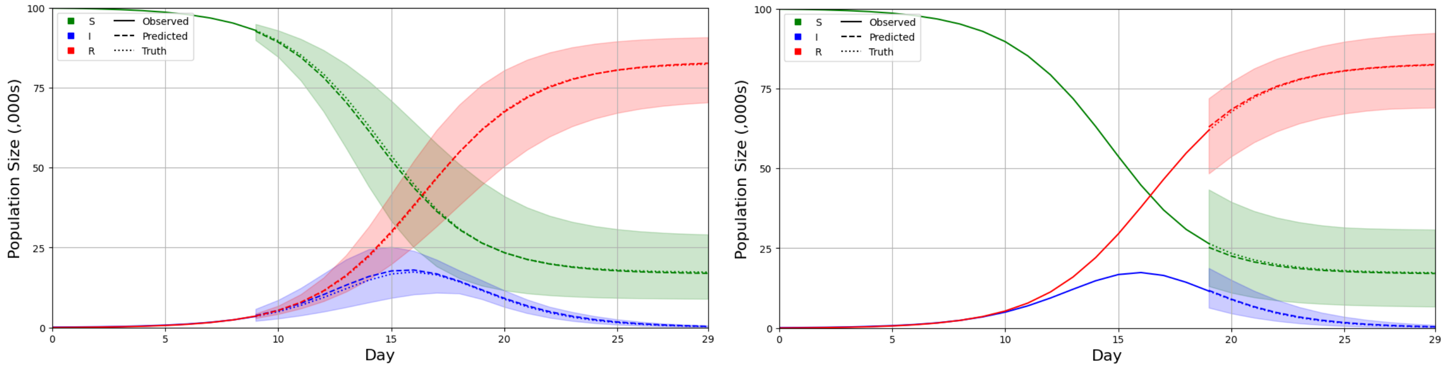
Figure 5.5: Plot of MSE against dimensionality from *Table 5.5*.

Table 5.5 shows that the MSE of fitted models reduces as the dimensionality of the summary statistics increases. This is to be expected as each extra dimension should extract at least some new information. However, the MSE values do plateau for summary statistics with 3, or more, dimensions. This is made clear in *Figure 5.5* which plots these MSE values against the dimensionality of the used summary statistics. This result shows that even though increasing dimensionality of summary statistics does improve model fit, the level of improvement quickly diminishes after a certain level of dimensionality. Further, we can conclude that although the identity function is likely the only sufficient summary statistic for the SIR model, there are plenty of lower-dimensional summary statistics which perform approximately as well.

5.4 Projection

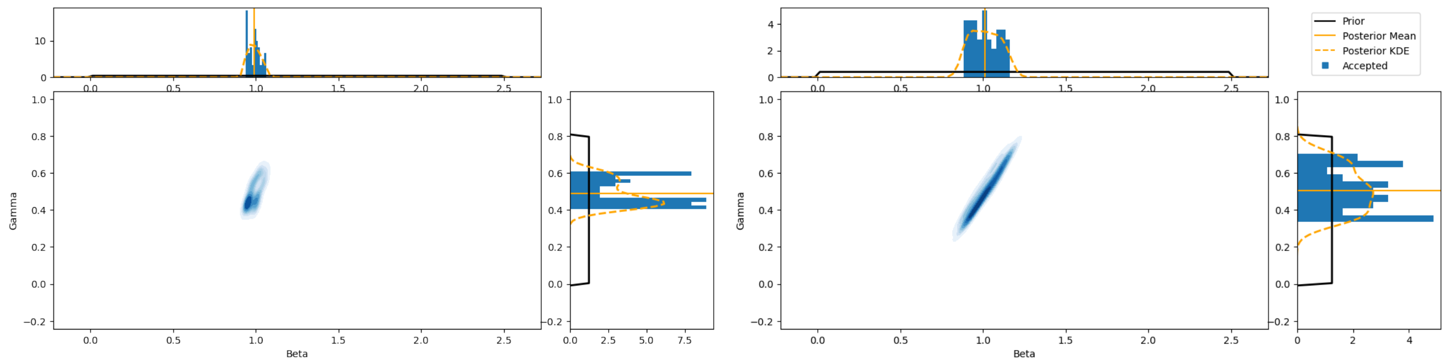
A common area of interest regarding epidemic events is projecting the future evolution of the disease. Here I assess how the automated methods I have presented so far perform in this field. This is done by only providing the first 10 and the first 20 days of the epidemic to the SA-ABC algorithm in order to generate summary statistics, and then running Adaptive ABC-SMC with said summary statistics.

The same priors as have been throughout *Section 5* for β and γ were defined. SA-ABC was initialised as described in *Section 5.2*, except with now considering more simulations. The pilot posterior f_θ is generated by "Best Samples" ABC-Rejection Sampling using the identity function and retaining the best 5,000 of a set of 500,000 simulations. 5,000 sets of parameters were then generated from the pilot posterior f_θ and used in the regression to calculate the summary statistics. Adaptive ABC-SMC was configured to run for 5,000 simulations, with a sample size of 100, using the log-Euclidean distance measure and a target acceptance rate of 90%. For a comparison, I also ran the same configuration of Adaptive ABC-SMC using the identity function as the summary statistic.



(i) First 10 days of data given. 95% CI for $R_0 = [1.686, 2.419]$.

(ii) First 20 days of data given. 95% CI for $R_0 = [1.632, 2.655]$.



(iii) Posterior when first 10 days of data given. 95% CI for $\beta = [0.942, 1.060]$, $\gamma = [0.405, 0.605]$

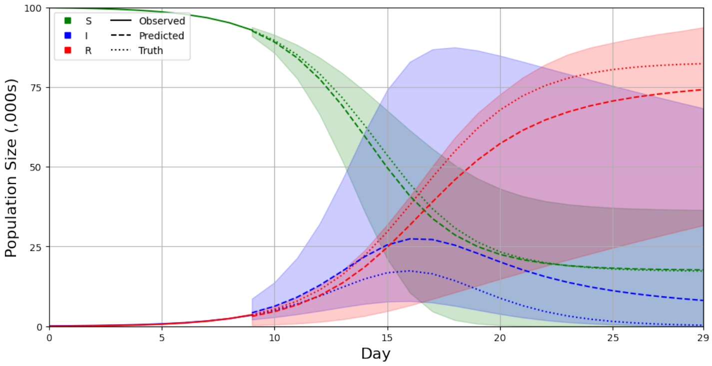
(iv) Posterior when first 20 days of data given. 95% CI for $\beta = [0.891, 1.164]$, $\gamma = [0.340, 0.703]$

Figure 5.6: Estimated posteriors and 95% CI for population size for SIR Model fitted using the posterior means generated by Adaptive ABC-SMC using the identity function as a summary statistic, with only the first 10 or 20 days of data available.

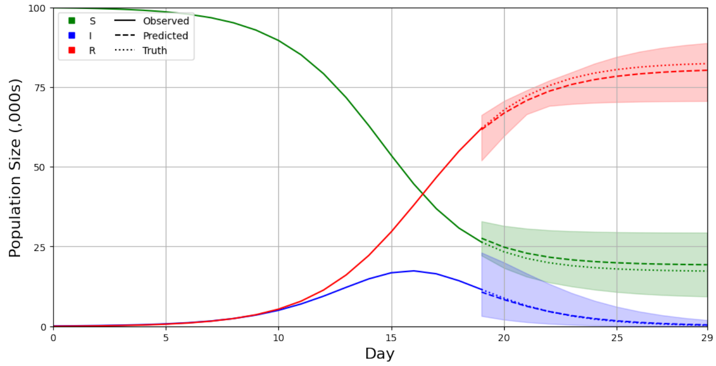
Figure 5.6 shows the estimated posteriors generated by the described Adaptive ABC-SMC method when supplied with the identity function as the summary statistic, with either the first 10 days or the first 20 days of the time-series being supplied. Additionally, the figure includes the 95% projection for the evolution of the population sizes calculated using the estimated posterior. *Figure 5.7* shows the same, but when the summary statistics have been automatically generated by SA-ABC.

The posterior generated using the identity function and 10 days of data (*Figure 5.6i*) is notably more concentrated than when 20 days of data (*Figure 5.6ii*) is supplied. This is true for both model parameters. The 10 day posterior (*Figure 5.6iii*) has a much more similar shape to and occupies the same parameter-space as the posterior when the full time-series is supplied (*Figure 5.1iv*), than the posterior after 20 days. The posterior produced when given the full time-series is the most representative of the true posterior, and thus suggests that a better fit has been achieved when only 10 days have been supplied. This is likely due to the first 10 days being dominated an increase in the infectious population, whereas the later part of the first 20 days sees a decay in the infectious population making this a harder data set to fit to. Moreover, this suggests that the four data-points (days 16-19) which represent the decrease in infectious population size are insufficient to make meaningful inferences with. This is unsurprising as the penalty imposed by the distance function for misfitting these 4 days will be less than for misfitting the first 16 days.

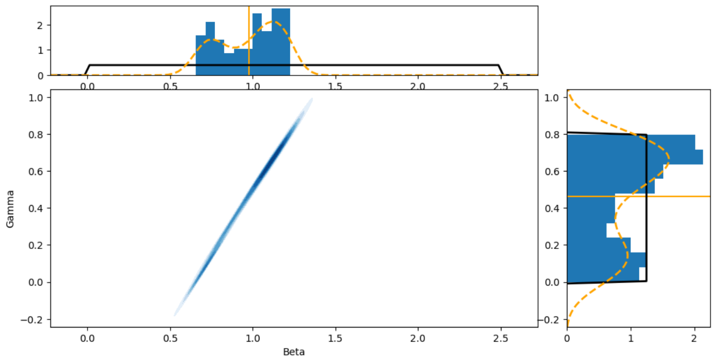
Despite a more concentrated posterior given 10 days of data, models fitted with posterior means from either implementation produce almost perfect fits. The credible intervals from both implementations are very similar and both implementations predict the disease will, at least 95% of the time, will die out within the 30 day window and that not all of the population will have been infected.



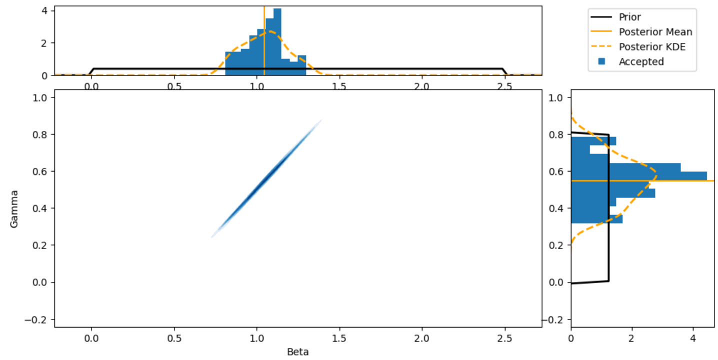
(i) First 10 days of data given. 95% CI for $R_0 = [1.536, 26.923]$. $RSSE = 3,469$



(ii) First 20 days of data given. 95% CI for $R_0 =$. $RSSE = 3,2131$.



(iii) Posterior when first 10 days of data given. 95% CI for $\beta = [0.665, 1.225]$, $\gamma = [0.024, 0.795]$.



(iv) Posterior when first 20 days of data given. 95% CI for $\beta = [0.824, 1.296]$, $\gamma = [0.333, 0.783]$

Figure 5.7: Estimated posteriors and 95% CI for population size for SIR Model fitted using the posterior means generated by Adaptive ABC-SMC and the summary statistics generated by SA-ABC, with only the first 10 or 20 days of data available. The between the fitted model and true model is calculated over the full time-series.

The results when using summary statistics generated by SA-ABC to fit a model with just the first 10 days of data are notably worse than when using the identity function. Inspecting the estimated posterior (*Figure 5.7iii*) shows that very little learning has occurred for the β parameter, which represents the recovery rate. As mentioned, this is unsurprising as the first 10 days are dominated by increasing infections. The credible intervals produced from 10 days of data (*Figure 5.7i*) are extremely wide, especially for the infectious population where the credible interval spans approximately 90% of the total population size on day 17. Notably, this implementation predicts the peak of infections to occur a few days after it actual does, whilst the implementation which used the identity function predicted the peak of infectious to occur slightly earlier than it actually does.

Importantly for inferential questions about the future course of the disease, the implementation which uses 10 days of data does not predict the disease to have died out within the 30 day.

The posterior produced when SA-ABC and Adaptive ABC-SMC are supplied with 20 days of data (Figure 5.7iv) is clearly a concentration of that produced when 10 days of data is supplied. Notably, the posterior from 20 days of data for β is unimodal, unlike the posterior produced from 10 days which was bimodal. Additionally, the posterior from 20 days of data is a concentration of the posterior produced when the identity function was used (*Figure 5.6iv*). This is due to greater learning having occured for the γ parameter when using the SA-ABC summary statistics rather than the identity function. This shows that SA-ABC was able to balance the penalty incurred by misfitting γ to be much more similar to that of misfitting β.

The credible intervals for population size given 20 days of data are notably narrow when using SA-ABC parameters rather than the identity function, due to the more concentrated estimated posterior for the γ parameter.

Overall, these results show that it is possible to automate the summary statistic selection process using SA-ABC and still produce well-fitting projections of SIR models. These predictions are better the more data SA-ABC is supplied with, which is unsurprising given it data driven approach.

5.5 Case Studies

Here I present two implementations of the ABC and summary statistic selection methods discussed in this project on two sets of real-world data: Covid-19 outbreak in France between February 17^{th} 2020 and March 17^{th} 2020 (*Section 5.5.1*); And, the SARS-CoV-2 (Covid-19) outbreak in Senegal between 2^{nd} March 2020 and 31^{st} March 2020 (*Section 5.5.2*). Both datasets are collected from the Covid-19 data repository maintained by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University University [2021]. I use the ABC and summary statistic selection methods to fit standard SIR models to both sets of data.

Both countries were chosen as they published the number of active case and recovered cases each day, rather than the number of new cases and new deaths as the UK has. This makes these datasets readily applicable to the standard SIR model discussed in this project. When fitting the models, both confirmed deaths and confirmed recoveries are grouped into the Removed compartment of the SIR model. The data from Senegal represents the first 30 days since the first confirmed case in the country, a period in which very few restrictions were imposed except for foreign travel bans. The data for France represents the 30 days leading up to the imposition of the first national lockdown, a period in which the government's own models predicted a sufficiently dire future for a lockdown to be deemed necessary.

For all implementations discussed below I used the following uniform priors for β and γ

$$\pi_{0,\beta} \sim \mathcal{U}(0, 0.5),\ \pi_{0,\gamma} \sim \mathcal{U}(.01, .1)$$

In pilot tests I specified a less informative prior for β, $\pi_{0,\beta} \sim \mathcal{U}(0, 2)$, and through these tests found that a narrower prior was appropriate as none placed significant posterior mass outside of $[0, 0.5]$. This narrower prior significant increased the convergence rate of the methods for γ was informed by the assumption that an individual is infectious for at least 10 days after initial infection Byrne *et al.* [2020].

The initial number of cases was fixed to the true value for both national populations [Bank, 2021] and the susceptible population was set to the national population size, less the initial number of cases.

For each dataset I run five implementations

(I1) "Best Samples" ABC-Rejection Sampling using the identity function as the summary statistic;

(I2) Adaptive ABC-SMC using the identity function as the summary statistic;

(I3) Adaptive ABC-SMC using the summary statistics recommended by the Minimum Entropy (ME) algorithm;

(I4) Adaptive ABC-SMC using the summary statistics recommended by the Two-Step Minimum Entropy (2-ME) algorithm; and, (I5) Adaptive ABC-SMC using the summary statistics generated by Semi-Automatic ABC (SA-ABC).

In implementation (I1) the ABC-Rejection Sampling method was configured to generate 500,000 simulations and to accepted the best 0.1% (500) according to the log-Euclidean distance measure. For implementations (I2)-(I5) Adaptive ABC-SMC was configured to run for 5,000 simulations, with a sample size of 100, using the log-Euclidean distance measure and a target acceptance rate of 90%. Adaptive ABC-SMC was used for all these implementations as it was shown in *Section 5.1* to perform best in general scenarios and it minimises user influence on these implementations. For implementations (I3)-(I4) ME and 2-ME were supplied with all summary statistics specified in *Table 5.3* (except for (S11) for time-tractability) and were configured to generate 100,000 simulations and to accept the best 100 for each set of summary statistics, up to size 4. (I3)-(I4) were both run 100 times for each dataset. For (I5), SA-ABC was configured to generated 1,000,000 simulations; keeping the best .1% (1,000) to generate a set of 5,000 parameters to use for the regression calculations.

For each implementation I produce plots of the 95% credible intervals for the sizes of the infectious and removed populations. Along with residual sum of squares error (RSSE)[1] between the true data and an SIR model fitted with the predicted posterior means for β and γ. I provide the 95% credible intervals for the model parameters β and γ, and for R_0. [Alimohamadi *et al.*, 2020] estimated a 95% credible interval for the R_0 of Covid-19 to be [2.81,3.82].

As discussed in *Section 2.3.1*, the standard SIR model does make a number of assumptions which limits its applicability. Namely, it assumes a constant rates of infection β and removal γ, and a constant population. Due to this assumption of constant rates I am only using data before major non pharmaceutical interventions (NPIs) had time to take effect. A worthwhile extension to the SIR model used in this section would be to fit the parameters β and γ as step functions of time with steps occurring when new NPIs were introduced. As the time-periods being considered are short, only 30 days, the assumption of a constant population is reasonable.

At the start of each time-period being considered the number of active cases is very small (¡10) and are geographically concentrated, by the end of the time-periods the number of active cases and the geographic spread of the disease has increase rapidly. It is likely that the spread of the disease will vary between geographical regions. Moreover, at the start of the datasets the majority of the susceptible population is not truly susceptible as they live many-many miles away from any infectious person. By the end of the datasets this is no longer true. This could be accounted for changing the size of the susceptible population with each confirmed case, such that it includes all those who live a region with confirmed cases. I have not implemented this due to the necessary complexity it would entail.

The standard SIR model also assumes that all cases of the disease are detected and recorded accurately. This is unrealistic in most real-world scenarios, especially for novel diseases which are not being actively screened for. Senegal has been praised for its approach to testing. Due to it not confirming its first case until March 2^{nd}, months after many other countries, it was better prepared and quickly introduced temperature check policies and free 24 hour test results. Under-reporting can be incorporated into the standard SIR model by creating two extra compartments: undetected infections; and, undetected deaths. The natural difficult with this is validating the sizes of these compartments and whether undetected deaths are even possible, as someone who is seriously ill enough to die from a disease is likely showing enough symptoms for them to be diagnosed correctly with said disease. The rate of under-reporting can be approximated by evaluating flucations in the positivity rate over time[2]. [Deo and Grover, 2021] present an

[1]Only the sizes of the infectious and removed populations at each time-step are compared, not the susceptible population. The natural logarithms of each value is used in the calculation due to the exponential nature of SIR models.

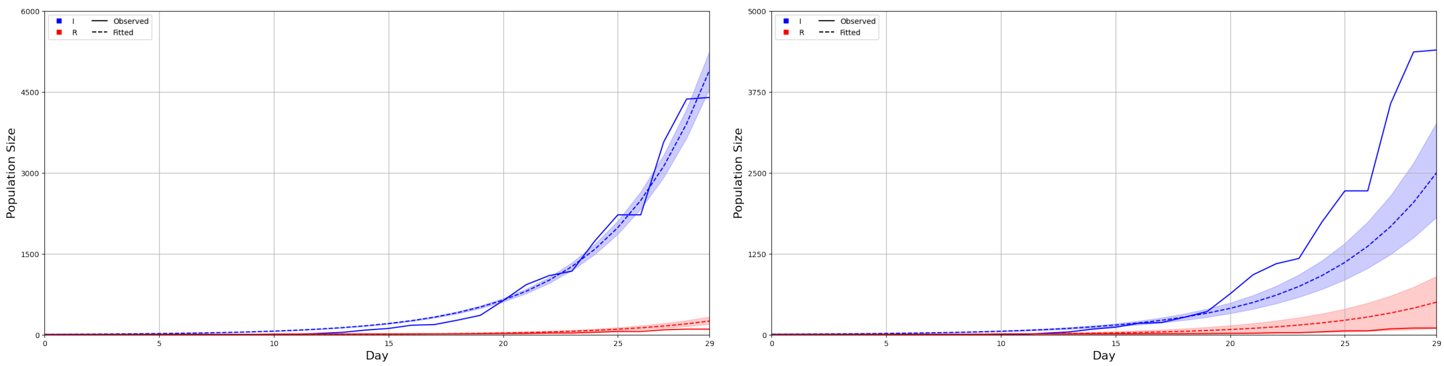
[2]The percentage of tests which return positive results.

implementation of an SIR model which accounts for under-reporting and apply it to Covid-19 data from California and Florida.

5.5.1 France Covid-19

February 17^{th} 2020 up to 17^{th} March marks the 30 day period leading upto the imposition of the first national lockdown in France. This is an interesting period to study as although France recorded it's first confirmed case of Covid-19 on the 24^{th} January 2020, the total number of active cases have only reached 7 by the 17^{th} Feb. Over the next 30 days the total number of active cases increased to 4,396.

Leading up to the national lockdown France did introduce some restrictions on public gatherings. Specifically, banning indoor gatherings of more than 5,000 people (5^{th} March); banning all gatherings of more than 1,000 people (10^{th} March); banning all gatherings of more than 100 people (14^{th} March); Closer of non-essential shops and services (15^{th} March); and, the closer of schools (16^{th} March). The last two restrictions would be believed to have the greatest effect on transmission rates, although some papers have shown the inverse to be true [Rice *et al.*, 2020]. Regardless, these restrictions would have little impact on the period of data I am considering.



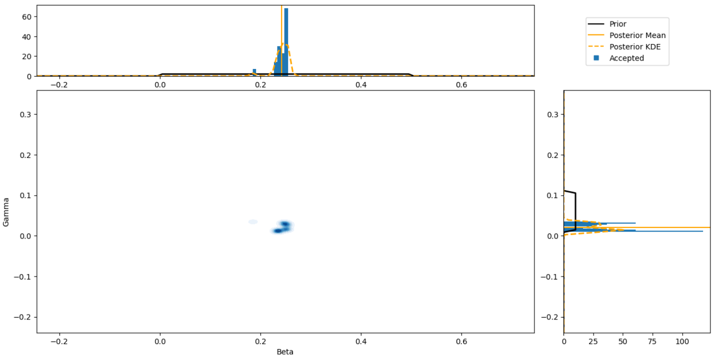
(i) (I1) "Best Samples" ABC-Rejection Sampling. 95% CI for $R_0 = [15.937, 25.627]$. RSSE=1,267.

(ii) (I2) Adaptive ABC-SMC. 95% CI for $R_0 = [3.990, 18.976]$. RSSE=10,395.

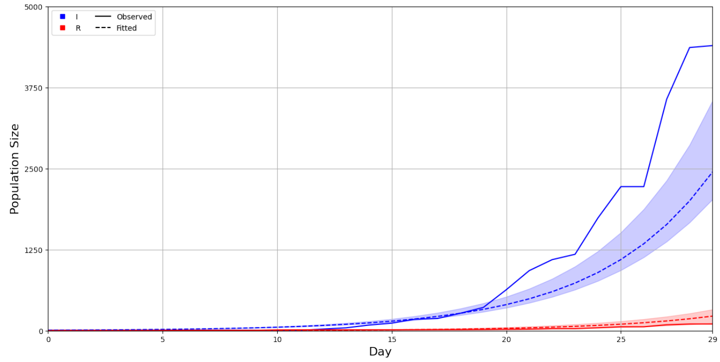
Figure 5.8: 95% credible intervals for infectious (blue) and removed (red) populations for Covid-19 in France between February 17^{th} 2020 and March 17^{th} 2020 using "Best Samples" ABC-Rejection Sampling (I1) and Adaptive ABC-SMC (I2), respectively. Both methods use the Log-Euclidean distance as their distance measures and the identity function of the infectious and removed populations as their summary statistic. The size of the susceptible population is not considered.

Figure 5.8 plot 95% credible intervals for the size of the infectious and removed populations of an SIR model fitted with the estimated posterior produced by "Best Samples" ABC-Rejection Sampling (I1) and Adaptive ABC-SMC (I2), respectively, when using the identity function as the summary statistic. The plot from (I1) (*Figure 5.8i*) has a significantly lower RSSE than the plot from (I2) (*Figure 5.8ii*). A visual inspection of the plots tells us that this is due to (I2) significantly underestimating the number of cases and overestimating the number of removals. *Figure 5.8i* shows a very good fit for both populations, although it begins to over-estimate the size of the removed population at the end of the time-series.

(I1) estimated a 95% credible interval for $R_0 = [15.9, 25.6]$ which is significantly greater than that estimated by [Alimohamadi *et al.*, 2020]. This is despite producing a good fit of the true data, suggesting significant underreporting of cases.



(i) (I3) Joint and marginal posteriors for β and γ.



(ii) (I3) 95% credible intervals for population sizes. RSSE=1,440.

Figure 5.9: Estimated posteriors for SIR model parameters and resulting 95% credible intervals for the sizes of the infectious (blue) and removed (red) populations for Covid-19 in France between February 17^{th} 2020 and March 17^{th} 2020 when using Adaptive ABC-SMC with summary statistics generated by Minimum Entropy (I3). 95% CI for $R_0 = [7.532, 21.216]$, $\beta = [0.186, 0.255]$, $\gamma = [0.011, 0.034]$.

When the ME algorithm was applied to the France data (I3), the following set of summary statistics was recommended:

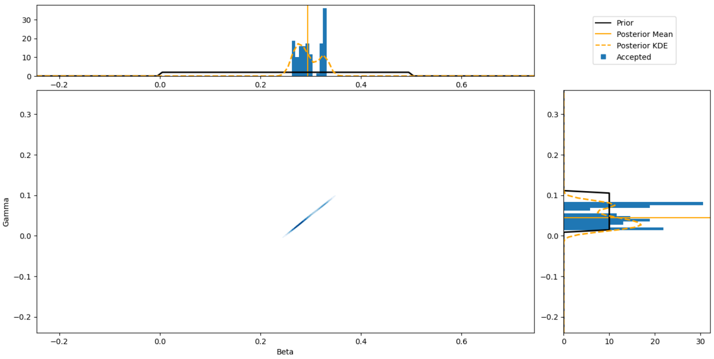
(S5)* Log-Final Size of Removed Population.

(S7)* Log-Mean Size of Infectious Population.

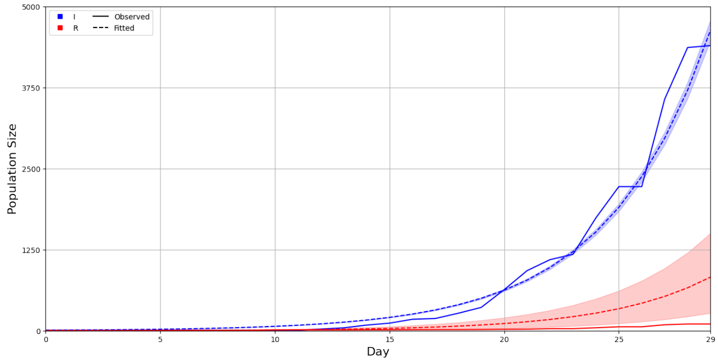
(S8)* Log-Mean Size of Removed Population.

Each of these statistics is one dimensional, thus the recommended set of summary statistics has 3 dimensions which is significantly less than the 90 dimensions of the identity function. It is notably that this set of statistics is the set of statistics recommended by ME in *Section 5.2* with statistics (S8)* included as well.

Figure 5.9 plots the estimated posteriors and the 95% credible intervals for population sizes from (I3) on the France dataset. Both posteriors are very concentrated compared to their supplied priors. (I3) produces an RSSE of 1,440 which is highly comparable to, although worse than, that from (I1). A visual inspection of *Figure 5.9ii* shows that (I3) fits the removed population extremely well, whilst is significantly under-predicts the size of the infectious population.



(i) (I4) Joint and marginal posteriors for β and γ.



(ii) (I4) 95% credible intervals for population sizes. RSSE=1,459.

Figure 5.10: Estimated posteriors for SIR model parameters and resulting 95% credible intervals for the sizes of the infectious (blue) and removed (red) populations for Covid-19 in France between February 17^{th} 2020 and March 17^{th} 2020 when using Adaptive ABC-SMC with summary statistics generated by Two-Step Minimum Entropy (I4). 95% CI for $R_0 = [4.062, 18.373]$, $\beta = [0.263, 0.331]$, $\gamma = [0.014, 0.082]$.

When the 2-ME algorithm was applied to the France data (I4), the following set of summary statistics was recommended:

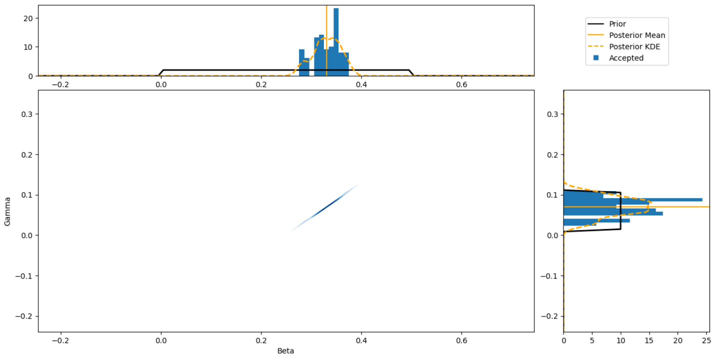
(S5)* Log-Final Size of Removed Population.

(S17)* Log-Net Daily Change in Removed Population.

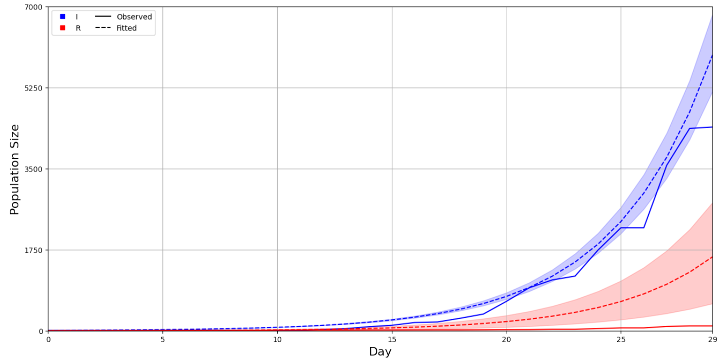
(S19)* Log-Cumulative Sum of Daily Infectious Population.

Statistic (S17)* has 29 dimensions when applied to the France data-set, meaning the total dimensionality of this recommended set of summary statistics is 31. This is significantly more than the set recommended by (I3), but still less than the identity function.

Figure 5.10 plots the estimated posteriors and the 95% credible intervals for population sizes from (I4) on the France dataset. Both marginal posteriors are markedly less concentrated than those from (I3) (*Figure 5.9i*), with the posterior for γ showing very little learning to have occured. The fitted model from (I4) produces an RSSE of 1,459 which is practically identical to that of the fitted model from (I3). A visual inspection of these credible intervals shows that a very good fit has been achieved for the infectious population, whilst the removed population is significantly overpredicted. This is the inverse of the results from (I3).



(i) (I5) Joint and marginal posteriors for β and γ.



(ii) (I5) 95% credible intervals for population sizes. RSSE=1,131.

Figure 5.11: Estimated posteriors for SIR model parameters and resulting 95% credible intervals for the sizes of the infectious (blue) and removed (red) populations for Covid-19 in France between February 17^{th} 2020 and March 17^{th} 2020 when using Adaptive ABC-SMC with summary statistics generated by Semi-Automatic ABC (I5). 95% CI for $R_0 = [3.469, 9.009]$, $\beta = [0.281, 0.374]$, $\gamma = [0.028, 0.108]$.

Figure 5.11 plots the estimated posteriors and the 95% credible intervals for population sizes from (I5) on the France dataset. The estimated posterior is highly comparable to that produced by (I4) in both concentration and the part of the parameter space which it occupies. This is an ideal result as the 2-ME algorithm used in (I4) recommended a set of summary statistics with 31 dimensions, whilst the SA-ABC algorithm use din (I5) generates a summary statistic with only 2 dimensions. The posterior for γ is still very similar to its prior, although this is to be expected as the data being fitted is dominated by increased infection rates and only 103 removals have been recorded by the end of the time-series (compared to 4,396 active cases).

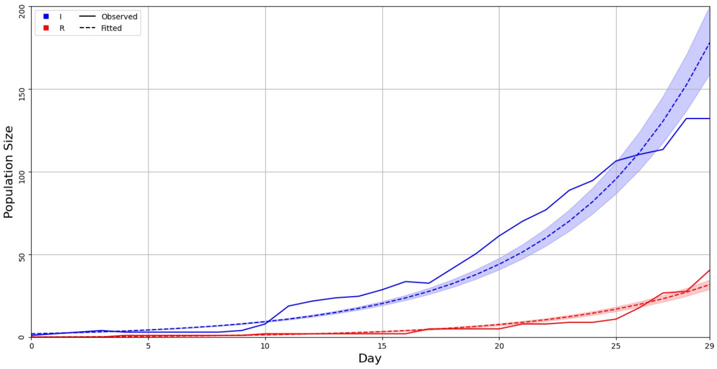
The 95% credible interval of R_0 from (I5) is [3.469,9.009]. This is still significantly higher than that of [Alimohamadi *et al.*, 2020] but it is the first implementation to overlap with their prediction. It is likely that given further data-points, especially those where removals take greater prominence, this method would converge towards the credible interval estimate of [Alimohamadi *et al.*, 2020].

The model fitted with the posterior means calculated by (I5) has the lowest RSSE (1,131) of any previous method.

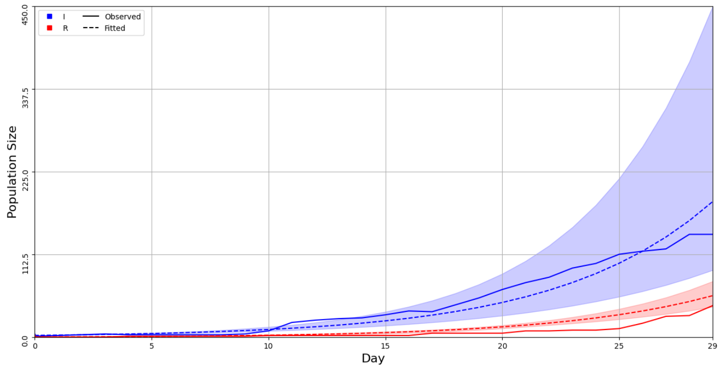
Each of the five implementations I have just presented have managed to produce fitted models and credible intervals which appear to closely match the observed data. This is promising for these models to be applied to real world data sets. Moreover, the best performing method (I5) was the one which automated the summary statistic selection process and used summary statistics with only two dimensions. This shows that there do exist low-dimension projections of real-world data which perform very well.

5.5.2 Senegal Covid-19

On 2^{nd} March 2020 Senegal confirmed it's first two infectious cases of Covid-19, by the end of March they had 135 confirmed active cases and 40 confirmed removed cases. On the 15^{th} March, with 24 active cases, the Senegalese government imposed travel restrictions, banned public gatherings and closed schools for a month. This is much swifter actions than many other countries up to this point. Senegal never ended up imposing a national lockdown and the only other NPIs they imposed were mandatory mask wearing in indoor spaces from 8^{th} August 2020. This lack of using the severe NPIs make Senegal a good case study for Covid-19.



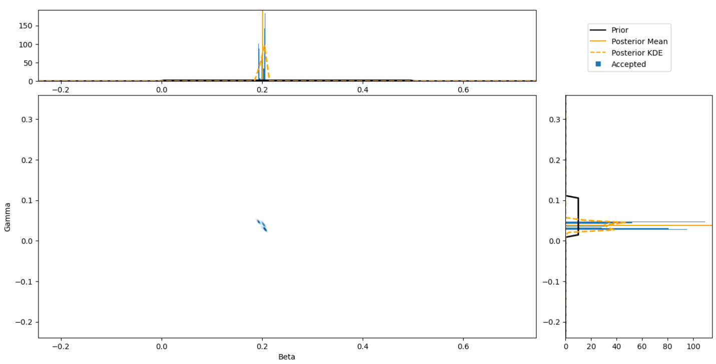
(i) (I1) "Best Samples" ABC-Rejection Sampling. 95% CI for $R_0 = [5.996, 7.172]$. RSSE=344.



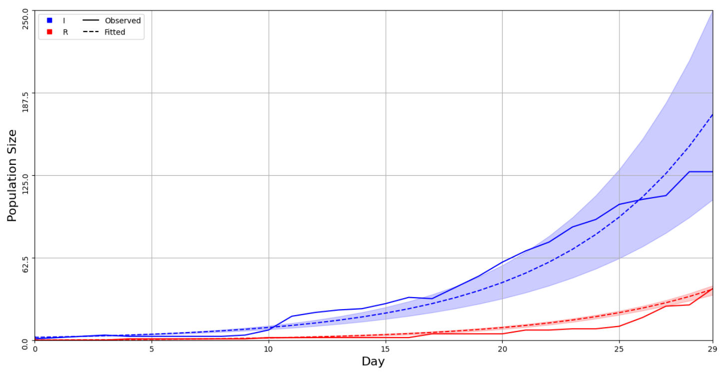
(ii) (I2) Adaptive ABC-SMC. 95% CI for R_0 =[3.160,8.107]. RSSE=376.

Figure 5.12: 95% credible intervals for infectious (blue) and removed (red) populations for Covid-19 in Senegal between March 2^{nd} 2020 and March 31^{st} 2020 using "Best Samples" ABC-Rejection Sampling (I1) and Adaptive ABC-SMC(I2), respectively. Both methods use the Log-Euclidean distance as their distance measures and the identity function of the infectious and removed populations as their summary statistic. The size of the susceptible population is not considered.

Figure 5.12 plots the 95% credible intervals for the sizes of the infectious and removed populations for an SIR model fitted with the estimated posterior produced by "Best Samples" ABC-Rejection Sampling (I1) and Adaptive ABC-SMC (I2), respectively, when using the identity function as the summary statistic. The fitted model for (I1) has a lower RSSE (344) than the fitted model for (I2), however the credible interval for R_0 produced by (I2) ([3.160,8.107]) is significantly closer to that estimated by [Alimohamadi *et al.*, 2020].



(i) (I3) Joint and marginal posteriors for β and γ.



(ii) (I3) 95% credible intervals for population sizes. RSSE=324.

Figure 5.13: Estimated posteriors for SIR model parameters and resulting 95% credible intervals for the sizes of the infectious (blue) and removed (red) populations for Covid-19 in Senegal between March 2^{nd} 2020 and March 31^{st} 2020 when using Adaptive ABC-SMC with summary statistics generated by Minimum Entropy (I3) . 95% CI for $R_0 = [4.103, 7.566]$, $\beta = [0.185, 0.211]$, $\gamma = [0.049, 0.012]$.

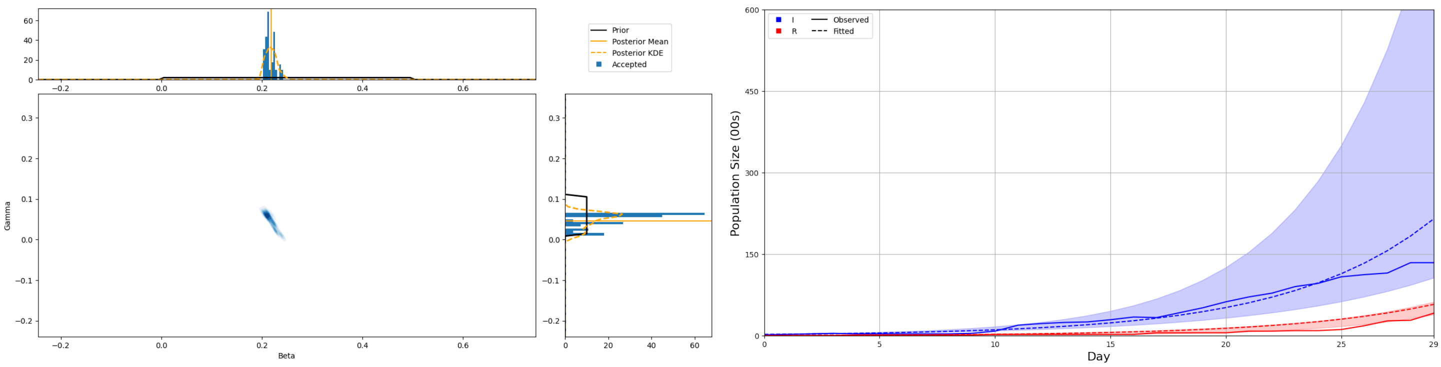
When the ME algorithm applied to the Senegal data (I3), the follow set of summary statistics were recommended:

(S5)* Log-Final Size of Removed Population.

(S13)* Log-Net Weekly Change in Infectious Population.

Statistic (S13)* has 4 dimensions when applied to the Senegal Covid data-set, so the total dimensionality of this set is 5. The results of using this set of statistics are shown in *Figure 5.13*. The estimated posterior *Figure 5.13i* is notably more concentrated that any of the posteriors generated on the France dataset. Moreover, a significant amount of learning has occurred for the γ parameter. This is liekly due to the ratio of infections to removals being much lower in the Senegal set than the France set, meaning less of a penalty is incurred by fitted to removes at the behest of infections.

The model fitted using the results of (I3) achieves a lower RSSE (324) than either model fitted using the identity function. So we have already found a set of statistics which have lower dimensions than the identity function, but perform better at fitting a model to the Senegal dataset.



(i) (I4) Joint and marginal posteriors for β and γ.

(ii) (I4) 95% credible intervals for population sizes. RSSE=290.

Figure 5.14: Estimated posteriors for SIR model parameters and resulting 95% credible intervals for the sizes of the infectious (blue) and removed (red) populations for Covid-19 in Senegal between March 2^{nd} 2020 and March 31^{st} 2020 when using Adaptive ABC-SMC with summary statistics generated by Two-Step Minimum Entropy algorithm (I4). 95% CI for $R_0 = [3.242, 21.665]$, $\beta = [0.203, 0.240]$, $\gamma = [0.011, 0.066]$.

When the 2-ME algorithm applied to the Senegal data (I4), the follow set of summary statistics were recommended:

(S7)* Log-Mean Size of Infectious Population.

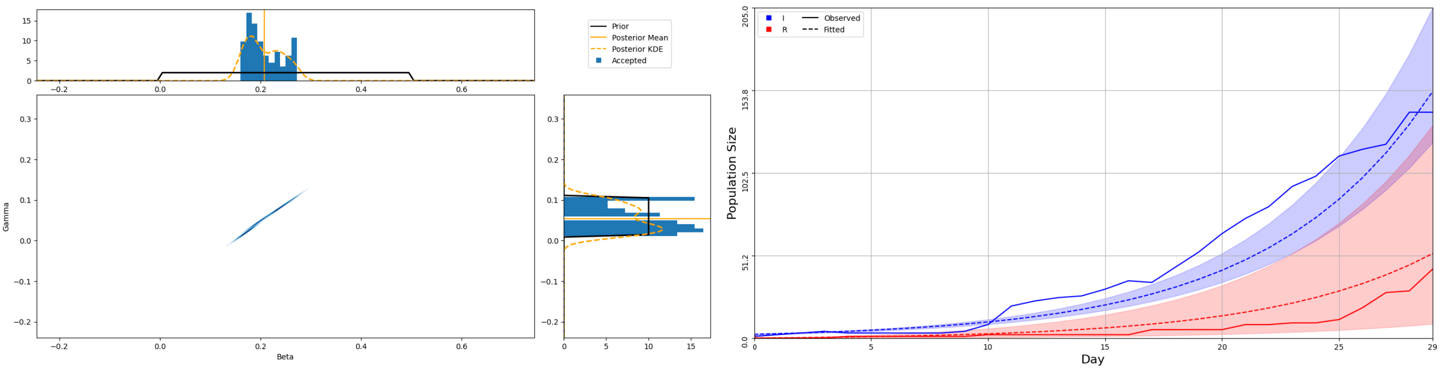
(S8)* Log-Mean Size of Removed population.

(S13)* Log-Net Weekly Change in Infectious Population.

As (S13)* has four dimensions when applied to the Senegal data set, thus this set has a total dimensionality of 6. This is the only time on either data-set that an ME method has not recommended (S5)*.

Figure 5.14 plots the estimated posterior and the 95% credible intervals for population size when Adaptive-ABC is applied with this set of recommended summary statistics (I4). The joint posterior is not as concentrated as the posterior estimate of (I3), this is due to less learning having been successful achieved for γ.

The model fitted with posterior means in *Figure 5.14ii* achieves the lowest RSSE so far. However, it also has the widest credible intervals for the size of the infectious population. If more simulations were performed this credible interval would decrease in size, but this sugests that the recommended set of statistics is not as efficient at extracting information as the set recommended by ME.



(i) (I5) Joint and marginal posteriors for β and γ.

(ii) (I5) 95% credible intervals for population sizes. RSSE=240.

Figure 5.15: Estimated posteriors for SIR model parameters and resulting 95% credible intervals for the sizes of the infectious (blue) and removed (red) populations for Covid-19 in Senegal between March 2^{nd} 2020 and March 31^{st} 2020 when using Adaptive ABC-SMC with summary statistics generated by Semi-Automatic ABC (I5). 95% CI for $R_0 = [2.547, 11.937]$, $\beta = [0.163, 0.270]$, $\gamma = [0.013, 0.107]$.

Figure 5.15 plots the estimated posterior and the 95% credible intervals for population size when Adaptive-ABC is applied with the set of summary statistics generated by SA-ABC (I5). This set of statistics has only two dimensions, lower than any of the sets of statistics recommended by ME or 2-ME.

The posteriors depicted in *Figure 5.15i* show significant learning has occurred for β, although less than in implementations (I3) and (I4). But, almost no learning has occurred for γ with no parts of the prior space being given zero posterior mass. This is very similar to when SA-ABC was applied to the France data-set (*Figure 5.11i*) and the reasoning for why this occurs is the same.

The model fitted with posterior means from (I5) achieves the best RSSE score (240) of any method discussed in this section. The credible intervals for the population sizes are wider than earlier methods, especially for the removed population. These intervals will narrow as more simulations are performed.

Again, each of the five implementations I have present have achieved well fitting models. Morever, the most automated implementation (I5) has achieved the best scoring model for both the France and Senegal dataset. This implementation does produce wide credible intervals, but those can be reduced by increasing the number of simulations being performed. The results from these case studies show that there are methods available which can automate the selection of low dimensional summary statistics, and methods which can generate their own summary statistics, such that well fitting models can be produced efficiently.

6 Conclusion

In this project I set out to evaluate methods for fitting models to epidemic processes in the Bayesian paradigm of statistics. This began with a discussion of Bayesian modelling and popular models for epidemic processes, focusing on compartmental models. I discussed how compartmental models are very versatile due to their modular nature, but that in the age of "Big Data" their assumptions of homogeneity are no longer necessary.

ABC methods were then presented as an approach for estimating the parameters for said models. I presented the mathematical motivation behind them and four flavours of ABC methods. With each, I discussed alterations which could either improve computational efficiency, and thus the rate of convergence, or could make the methods more adaptable and thus require the user to specify future features of the algorithm each run.

The discussion of ABC methods showed that there were three key, user-specified features which the success of the method depended upon: summary statistics; distance measure; and, choice of acceptance kernel. I chose to take an in-depth look into summary statistic selection as this feature is arguably the most important of the three just mentioned, as the use of uninformative summary statistics can cause the ABC methods to converge on demonstrably incorrect models.

As part of this analysis of summary statistics I discussed what features make up good summary statistics in practice; namely sufficiency and low-dimensionality. From this discuss the identity function was identified as a naïvely optimal summary statistic due to it being sufficient but of high dimension. I presented in-depth four methods for automating summary statistic selection, one of which generated its own summary statistics.

The crescendo of this project was the evaluation of the ABC methods to fit an SIR model and of the summary statistic selection methods to select well-performing statistics for this problem. During initial tests, using the identity function as the summary statistic, ABC-SMC was found to be the best method when supplied with the full time-series of data but its Adaptive version outperformed it during a cross-validation which assessed the more general performance of each algorithm. The clear downside to the Adaptive ABC-SMC method was its long run-times, taking 5 times as long as standard ABC-SMC and ~ 500 times longer than ABC-Rejection Sampling. All methods performed well, generating posteriors with significant mass placed around the true parameter values when given sufficiently long to run.

The assessment of the four summary statistic selection methods found that none managed to produce summary statistics which performed as well as the identity function. However, all methods, except for ASS, were able to generate summary statistics that performed similarly to the identity function despite having less than a thirtieth of the number of dimensions. Notably, the method which generates its own summary statistics from scratch, SA-ABC, produced the second-best set of summary statistics.

An assessment of the effect of dimensionality of summary statistics on the success of ABC methods to fit an SIR model found that although more dimensions always led to improved performance, there was a point where the gains became negligible. In the scenario tested this was after three dimensions, which is one more than the number of parameters in the SIR model.

This project concluded with an evaluation of ABC methods and summary statistic methods to fit SIR models to real-world data which was collected during the early phases of the Covid-19 breakouts in both France and Senegal. In all cases, good model fits were achieved. This is despite the very simple SIR model being used, which did not account for changes in parameter values over time nor for the under-reporting of cases. For both datasets, the best fitting models were produced by an implementation that used SA-ABC to generate summary statistics and Adaptive ABC-SMC to estimate parameter posteriors. These implementations also produced credible intervals for R_0 which most closely matched those from more sophisticated research of [Alimohamadi *et al.*, 2020]. This implementation was the one that required the fewest user specifications to run and thus this result is promising for the possible automation of algorithm specification.

6.1 Future Work

To close this project I will mention a few areas for future work which this project has thrown up.

Optimisation of Adaptive ABC-SMC

This project has shown Adaptive ABC-SMC to be a powerful ABC method despite being one of the easiest to initialise. This method does still have parameters which need to be initialised, namely the target acceptance rate which is used to determine the bandwidth of the acceptance kernel after each iteration. Discovering a practically optimal value for this acceptance rate would be an interesting area to study as it would need to trade-off theoretical and computational results. As a significant proportion of the computational complexity of this algorithm occurs between iterations, when calculating new perturbance and acceptance kernels, as optimal acceptance rate would minimise the mean time per simulation whilst also minimising the acceptance kernel being used (generally achieved by completing more iterations). As [Roberts and Rosenthal, 2001] showed 23.4% to be an optimal acceptance rate for MCMC mixing, a similar result for SMC would be useful. An evaluation of the specification of the perturbance kernels would likely need to be performed in tandem with that of the target acceptance rate.

Utilise the Entropy Value

The entropy methods presented in [Nunes and Balding, 2010] only return the set of summary statistics which achieve the lowest entropy and MRSSE values respectively. No consideration is made for how much the best scoring set outscores the next best set, or any other set for that matter, and in practice it was found that although one set of statistics did get returned notably more often than any other, several different sets were returned. This ambiguity and variability is reduced as more simulations are performed, but I believe there are more efficient ways to achieve this. Either, after the intitial ranking has been produced, generate more simulations and re-evaluate the best n; or, produce a mixtures model where sets of statistics are used with frequencies defined by their relative entropy or MRSSE values.

More Complex Models

I only assessed the discussed methods on the standard SIR model. This showed promising results but these are naturally limited in their scope. ABC methods, especially ABC-SMC, have been shown to run successfully on significantly more complex models (such as in [Lenormand *et al.*, 2013] which successfully modelled the "SimVillages" model [Huet and Deffuant, 2011] which models the lives of 3,000 villagers). However the same analysis has not been performed on the summary statistic selection methods discussed in this project. Such an assessment would be interesting and I believe fruitful for the two entropy algorithms and SA-ABC as they rely on the same principles as ABC methods and thus can be expected to perform well in scenarios where ABC methods perform well.

Bibliography

Alimohamadi, Y., Taghdir, M. and Sepandi, M. (2020). Estimate of the basic reproduction number for COVID-19: A systematic review and meta-analysis. *Journal of Preventive Medicine and Public Health* **53**(3), 151–157.

Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association* **65**(331), 1248–1255.

Ciofi degli Atti, M. L., Merler, S., Rizzo, C., Ajelli, M., Massari, M., Manfredi, P., Furlanello, C., Scalia Tomba, G. and Iannelli, M. (2008). Mitigation measures for pandemic influenza in italy: an individual based model considering different scenarios. *PloS one* **3**(3), e1790–e1790, 18335060[pmid].

Badr, H., Du, H., Marshall, M., Dong, E., Squire, M. and Gardner, L. (2020). Association between mobility patterns and covid-19 transmission in the usa: a mathematical modelling study. *The Lancet Infectious Diseases* **20**.

Balakrishnan, S. (2019). Lecture notes in 36-705: Intermediate statistics, lecture 12. http://www.stat.cmu.edu/ siva/705/lec12.pdf.

Bank, T. W. (2021). World bank open data. `https://data.worldbank.org/`.

Beaumont, M. A. (2019). Approximate bayesian computation. *Annual Review of Statistics and Its Application* **6**(1), 379–403.

Beaumont, M. A., Zhang, W. and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics* **162**(4), 2025–2035.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M. and Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika* **96**(4), 983–990.

Beirlant, J., Dudewicz, E., Györfi, L. and Dénes, I. (1997). Nonparametric entropy estimation. an overview. *INTERNATIONAL JOURNAL OF MATHEMATICAL AND STATISTICAL SCIENCES* **6**(1), 17–39.

Bellman, R. E. (1961). *Adaptive Control Processes: A Guided Tour*. First introduction of curse of dimensionality.

Beyer, K., Goldstein, J., Ramakrishnan, R. and Shaft, U. (1999). When is "nearest neighbor" meaningful? , 217–235.

Blackwell, D. (1947). Conditional Expectation and Unbiased Sequential Estimation. *The Annals of Mathematical Statistics* **18**(1), 105 – 110.

Blum, M. and François, O. (2010). Non-linear regression models for approximate bayesian computation. *Statistics and Computing* **20**, 63–73.

Blum, M. G. B. and Tran, V. C. (2010). HIV with contact tracing: a case study in approximate Bayesian computation. *Biostatistics* **11**(4), 644–660.

Brooks-Pollock, E., Roberts, G. O. and Keeling, M. J. (2014). A dynamic model of bovine tuberculosis spread and control in great britain. *Nature* **511**(7508), 228–231.

Byrne, A. W., McEvoy, D., Collins, A. B., Hunt, K., Casey, M., Barber, A., Butler, F., Griffin, J., Lane, E. A., McAloon, C., O'Brien, K., Wall, P., Walsh, K. A. and More, S. J. (2020). Inferred duration of infectious period of sars-cov-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic covid-19 cases. *BMJ Open* **10**(8).

Casella, G. and Berger, R. (2001). *Statistical Inference*. Duxbury Resource Center.

Chakravarti, I., Laha, R. and Roy, J. (1967). Handbook of methods of applied statistics (v. 1), 392–394.

Coburn, B. J., Wagner, B. G. and Blower, S. (2009). Modeling influenza epidemics and pandemics: insights into the future of swine flu (h1n1). *BMC medicine* **7**, 30–30, 19545404[pmid].

Colombo, C. and Diamanti, M. (2015). The smallpox vaccine: the dispute between bernoulli and d'alembert and the calculus of probabilities. *Lettera Matematica* **2**(4), 185–192.

Darmois, G. (1935). Sur les lois de probabilité à estimation exhaustive. *Comptes Rendus de l'Académie des Sciences* , 1265–1266.

Del Moral, P. (1997). Nonlinear filtering: Interacting particle resolution. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* **325**(6), 653–658.

Deo, V. and Grover, G. (2021). A new extension of state-space sir model to account for underreporting – an application to the covid-19 transmission in california and florida. *Results in Physics* **24**, 104182.

Didelot, X., Everitt, R. G., Johansen, A. M. and Lawson, D. J. (2011). Likelihood-free estimation of model evidence. *Bayesian Anal.* **6**(1), 49–76.

Dodge, Y., Institute, I. S. and Commenges, D. (2006). *The Oxford Dictionary of Statistical Terms*. Oxford University Press.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications* **14**(1), 153–158.

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**(1), 87–112.

Fagundes, N. J. R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L. and Excoffier, L. (2007). Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences* **104**(45), 17614–17619.

Fan, Y. and Sisson, S. A. (2018). Abc samplers.

Fearnhead, P. and Prangle, D. (2011). Constructing summary statistics for approximate bayesian computation: Semi-automatic abc .

Filippi, S., Barnes, C., Cornebise, J. and Stumpf, M. P. H. (2012). On optimality of kernels for approximate bayesian computation using sequential monte carlo .

Fink, D. (1997). A compendium of conjugate priors. *Technical Report.* .

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A* **222**, 309–368.

Frias-Martinez, E., Williamson, G. and Frias-Martinez, V. (2011). An agent-based model of epidemic spread using human mobility and social network information. In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 57–64.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* **7**(4), 457 – 472.

Gelman, A., Gilks, W. R. and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolisalgorithms. *The Annals of Applied Probability* **7**(1), 110 – 120.

Goffman, W. (1965). An epidemic process in an open population. *Nature* **205**(4973), 831–832.

Grelaud, A., Marin, J.-M., Robert, C. P., Rodolphe, F. and Taly, J.-F. (2009). ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis* **4**(2), 317 – 335.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2nd edition.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1), 97–109.

Hayashi, F. (2000). *Econometrics*. Princeton Univ. Press, Princeton, NJ [u.a.].

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Rev.* **42**(4), 599–653.

Hinneburg, A., Aggarwal, C. C. and Keim, D. A. (2000). What is the nearest neighbor in high dimensional spaces? , 506–515.

Huet, S. and Deffuant, G. (2011). Common framework for the microsimulation model in prima project. *Cemagref LISC* .

Huppert, A. and Katriel, G. (2013). Mathematical modelling and prediction in infectious disease epidemiology. *Clinical Microbiology and Infection* **19**(11), 999–1005.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science* **20**(1), 50 – 67.

Jeffreys, H. (1961). *Theory of Probability*. Oxford, Oxford, England, 3rd edition.

Joyce, P. (1998). Partition structures and sufficient statistics. *Journal of Applied Probability* **35**(3), 622–632.

Joyce, P. and Marjoram, P. (2008). Approximately Sufficient Statistics and Bayesian Computation. *Statistical Applications in Genetics and Molecular Biology* **7**(1), 1–18.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**(430), 773–795.

Kayode, S., Ganiyu, A. and Ajiboye, A. (2016). On one-step method of euler-maruyama type for solution of stochastic differential equations using varying stepsizes. *OALib* **03**, 1–15.

Kermack, W. O., McKendrick, A. G. and Walker, G. T. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **115**(772), 700–721.

Koopman, B. O. (1936). On Distributions Admitting a Sufficient Statistic. *Transactions of the American Mathematical Society* **39**(3).

Lehmann, E. L. and Scheffé, H. (1950). Completeness, similar regions, and unbiased estimation: Part i. *Sankhyā: The Indian Journal of Statistics (1933-1960)* **10**(4), 305–340.

Lenormand, M., Jabot, F. and Deffuant, G. (2013). Adaptive approximate bayesian computation for complex models. *Computational Statistics* **28**(6), 2777–2796.

Mahaffy, J. (2018). Math 636 - mathematical modeling - discrete sir - models. https://jmahaffy.sdsu.edu/courses/f17/math636/beamer/sir.pdf.

Maki, Y. and Hirose, H. (2013). Infectious disease spread analysis using stochastic differential equations for sir model. In: *2013 4th International Conference on Intelligent Systems, Modelling and Simulation*, 152–156.

Mardia, K., Kent, J. and Bibby, J. (1979). *Multivariate analysis*. Probability and mathematical statistics, Acad. Press, London [u.a.].

Marjoram, P. and Tavaré, S. (2006). Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet* **7**, 759–770.

Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences* **100**(26), 15324–15328.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**(6), 1087–1092.

Neyman, J. (1935). Sur un teorema concernente le cosidette statistiche sufficienti. *Giorn. Ist. Ital. Att., 6* , 320–334.

Nunes, M. and Balding, D. (2010). On optimal selection of summary statistics for approximate bayesian computation. *Statistical Applications in Genetics and Molecular Biology* **9**(1).

Nutton, V. (1990). The reception of fracastoro's theory of contagion: The seed that fell among thorns? *Osiris* **6**, 196–234.

Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. *Proceedings of the Cambridge Philosophical Society* , 567–579.

Pritchard, J., Seielstad, M., Perez-Lezaun, A. and Feldman, M. (1999). Population growth of human y chromosomes: A study of y chromosome microsatellites. *Molecular biology and evolution* **16**, 1791–8.

Radovanovic;, M., Nanopoulos;, A. and Ivanovic, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research* **11**(86), 2487–2531.

Rao, C. R. (1945). *Information and accuracy attainable in the estimation of statistical parameters*. Bulletin of the Calcutta Mathematical Society, 81–91.

Rice, K., Wynne, B., Martin, V. and Ackland, G. J. (2020). Effect of school closures on mortality from coronavirus disease 2019: old and new predictions. *BMJ* **371**.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science* **16**(4), 351 – 367.

Rodrigues, H. S. (2016). Application of sir epidemiological model: new trends .

Ross, R. (1916). An application of the theory of probabilities to the study of a priori pathometry. part i. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **92**(638), 204–230.

Ross, R. and Hudson, H. P. (1917). An application of the theory of probabilities to the study of a priori pathometry.—part ii. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **93**(650), 212–225.

Roussas, G. (1998). *A Course in Mathematical Statistics*. Academic Press, 2nd edition, 263.

Schnitzer, D., Flexer, A. and Tomasev, N. (2014). Choosing the metric in high-dimensional spaces based on hub analysis. .

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal* **27**(3), 379–423.

Singh, H., Misra, N., Hnizdo, V., Fedorowicz, A. and Demchuk, E. (2003). Nearest neighbor estimates of entropy. *American Journal of Mathematical and Management Sciences* **23**(3-4), 301–321.

Sisson, S. A., Fan, Y. and Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences* **104**(6), 1760–1765.

Sisson, S. A., Fan, Y. and Beaumont, M. A. (2018). Overview of approximate bayesian computation .

Tavaré, S., Balding, D. J., Griffiths, R. C. and Donnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics* **145**(2), 505–518.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. P. (2009). Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface* **6**(31), 187–202.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition.

University, J. H. (2021). Covid-19 data repository by the center for systems science and engineering (csse). `https://github.com/CSSEGISandData/COVID-19`.

Verguet, S., Johri, M., Morris, S. K., Gauvreau, C. L., Jha, P. and Jit, M. (2015). Controlling measles using supplemental immunization activities: a mathematical model to inform optimal policy. *Vaccine* **33**(10), 1291–1296, 25541214[pmid].

Wegmann, D. and Excoffier, L. (2010). Bayesian Inference of the Demographic History of Chimpanzees. *Molecular Biology and Evolution* **27**(6), 1425–1435.

Wong, W., Jiang, B., Wu, T.-y. and Zheng, C. (2018). Learning summary statistic for approximate bayesian computation via deep neural network. *Statistica Sinica* .

Wood, S. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466**(7310), 1102–1104.

Zambom, A. Z. and Dias, R. (2012). A review of kernel density estimation with applications to econometrics .