Bayesian Modelling of Epidemic Processes

D. Hutchinson

April 11, 2021

Dedication

Accompanying Resources

Abstract

Contents

1 Introduction

What is a model? A (simple) mathematical formulation of a process which incorporates parameters of interest and likely some stochastic processes. Models need to be computational tractable (i.e. fairly simple)

"All models are wrong, some are useful".

What to use models for? check intuition, explanation & prediction.

What is "posterior estimation"?

The problem - Posterior estimation when likelihood is intractable. "Likelihood-free" estimation. (Classical example of determining most recent common ancestor of two DNA strands. Likelihood is intractable due to number of branches growing factorially. ([Burr and Skurikhin, 2013])

Motivation

Bayes Rule? Describe each component & why is likelihood intractable?

Why now? More, better data. Greater computational power.

What can posterior be used for?

Motivating Examples

DNA mutation ([Marjoram and Tavaré, 2006])

History

Traditional parameter estimation methods - "Maximum Likelihood".

Neutrality testing - (Hypothesis testing), compare results against a null hypothesis for a parameter value.

Successful Applications of these Methods

2 Bayesian Modelling

Bayes' Rule

vs. Frequentist modelling

Stochastic vs deterministic models

Consistency

In general, we never know if our calculated posterior is actually close to the true posterior.

3 Approximate Bayesian Computation

3.1 Motivation

3.2 Background

3.3 ABC Methods

3.3.1 ABC-Rejection Sampling

3.3.2 ABC-Importance Sampling

3.4 Importance Sampling

3.4.1 ABC-MCMC

3.4.2 ABC-SMC

3.4.3 Comparison

Which algorithm to use in different scenarios - complexity of model, amount of data available.

3.5 ABC for Model Choice

3.6 Regression Adjustment in ABC

Beaumont et al - Local Linear Regressions (LOCL)

Blum and Francois' - Nonlinear Conditional heteroscedastic regressions (NCH). (Uses neural networks)

4 Summary Statistic Selection

In this chapter I motivate the research into summary statistic selection *Section 4.1* and discuss features to consider when selecting summary statistics *Section 4.2*. I then describe five methods for summary statistic selection methods: 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. To close the section I use a toy example of an SIR model to compare these methods *Section 4.3.6*.

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.

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. 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.

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

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 of independence also makes it conceptually easy to combine summary statistics by appending the result 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, then using a set of summary statistics still produces effective dimensionality reduction.

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

The success of ABC methods depends mainly on three user choices: choice of summary statistic; choice of distance measure; and choice of acceptance kernel. Of these, summary statistic choice is arguably the most important as the other two mainly affect the rate at which the algorithm converges on the posterior mean. Whereas, choosing summary statistics which are uninformative can lead to the 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 as this would result in all (or none) of the simulations being accepted as thus the returned posterior will be the same as the supplied prior.

In practice, the quality of the posteriors returned from an ABC method is limited by the amount of computational time which is dedicated to running the algorithm. For some problems,

such as ..... , 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, 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 and thus more likely to contain the true parameter values (Similarly, these statistics will disincentivise exploring regions which have been shown to not be of 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.

When using a set of summary statistics, expert knowledge is generally not sufficient to determine how best to weight each summary statistic. Some of the methods I describe below 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 weight.

4.2 Properties of Summary Statistics

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

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 readings, each with m features) most standard summary statistics are calculated in $O(n \cdot m)$ time. However, this is only a theoretical result 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 sorting operations (most notably order statistics) are variable in the their time complexity for different data sets which will affect 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 simulation is good, or not. This means that the range and scale of values a summary statistic will likely produce will have an affect 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, 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[1].

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 in a problem will want to use the final model to justify their future decisions, this is much easier to do when the factors the model is considering, and the weights it assigns to them, are readily understandable. Hand-crafted statistics are almost always the most readily understandable statistics, as such generated statistics are rarely used in commercial problems[2]. In cases where it is chosen to use automatically generated statistics; one can develop an intuition for their model by varying the inputs, or removing certain features, and observing how the output varies. This is naturally harder to

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 is a sufficient statistic for all models.

It can be intuitively helpful 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

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

[2] 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.

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.

Example 4.1 (Sufficient Statistic for Normal Distribution with Unknown Mean)
Let $X \sim Normal(\mu, \sigma_0^2)$, 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 $s(\mathbf{X})=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}X_i$*, the sample mean, then the third exponential disappears. Note that* $s(\mathbf{X})\sim$ *Normal*$\left(\mu,\dfrac{1}{n}\sigma_0^2\right)$*.*

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*

Example 4.1 shows that finding sufficient summary statistics can be a highly manually 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[3] *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]

$\implies$ 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

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

$\impliedby$ 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. □

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

Example 4.2 below demonstrates how the Fisher-Neyman Factorisation Theorem 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(λ), 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 (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

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

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. □

Theorem 4.2 states that if the ratio of the marginal distributions of two samples from a model are independent of the model parameters if, and only if, 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

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

Now consider the variance of $v(X)$

$$\begin{aligned}\mathrm{Var}(v(X)) &= \mathrm{MSE}[v(X)] - \mathrm{Bias}[v(X)]^2 = \mathrm{MSE}[v(X)] \\ &= \mathbb{E}[(v(X) - g(\theta))^2] \\ &= \mathbb{E}[(\mathbb{E}[v|T = t(X)] - g(\theta))^2] \\ &= \mathbb{E}[(\mathbb{E}[v - g(\theta)|T = t(X)])^2] \\ &\overset{[4]}{\leq} \mathbb{E}[(v - g(\theta))^2|T = t(X)] \\ &= \mathrm{Var}(u(X)) \\ \implies \mathrm{Var}(v(X)) &\leq \mathrm{Var}(u(X))\end{aligned}$$

□

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

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 straight-forward 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 [Ruli, 2018] 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 $s_1, \ldots, s_k$ be summary statistics for a model X with parameters θ. Define sets $s_{1:k-1} := \{s_1, \ldots, s_{k-1}\}$, $s_{1:k} := \{s_1, \ldots, s_k\}$ and consider the likelihood of the set $s_{1:k}$ with respect to the model parameters θ

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

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, Joyce and Marjoram [2008])

require: *Set of summary statistics S; Score threshold ε*

$S' \leftarrow \emptyset$
while *true* **do**
  Calculate the score for each statistic in S wrt S'
  $\delta_{max} \leftarrow \max_{s \in S} Score(s; S')$
  $s_{max} \leftarrow argmax_{s \in S} Score(s; S')$
  if $\delta_{max} > \varepsilon$ **then** $S' \leftarrow S' \cup \{s\}$;
  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 wrt 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[5]. 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 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[6].

Algorithm 4.2 (Evaluate Proposed Statistic)

[5] In [Joyce and Marjoram, 2008] it is recommended to use a value of between one and four standard deviations
[6] A leave-one-out cross-validation could be used to determine the optimal number of statistics to use.

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[7]. 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[8], 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 affect 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 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. In the paper

[7] 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.

[8] 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.

Nunes & Balding propose 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 $X_1, \ldots, X_n$ is defined as

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

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 usage of entropy is equivalent to Joyce & Marjoram's usage 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 & Balding suggest using the asymptoticly 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 straightforward approach but has time

complexity $O(kn)$ so does not scale efficiently for large values of k. [Singh *et al.*, 2003] recommend 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" version of the 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 the number 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.

Algorithm 4.3 (Minimum Entropy Summary Statistic Selection, 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 & Balding **Algorithm 4.3** is very straight-forward. 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 wrt 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 evaluate sets whose subsets produce high entropy values.

The estimated entropy value for a set of statistics will vary each time 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 varability 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] 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 were observed. Each subset of statistics is then reassessed 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 $\mathbf{X} := \{X_1, \ldots, X_n\}$ *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 $\mathbf{X}^* := \{X_1^*, \ldots, X_m^*\}$ *to be a set of target values. The mean residual sum of squares error (MRSSE) is the mean RSSE value for each target value wrt 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 reassessments 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 & Balding choose to use the mean residual sum of squares error (MRSSE, **Definition 4.7**) with the set of statistics which minimises MRSSE wrt Θ_{acc} is return as the recommended set of statistics.

MRSSE is a desirable 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 is sensitive to outlier values. Note that the scale of parameter values will affect 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 the standardisation.

Algorithm 4.4 (Two-Step ME Summary Statistic SelectionNunes 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}

```
S_ME ← Algorithm 4.3(S)
Θ̂_ME ← Parameter sets accepted from "Best Samples" ABC-RS(x_obs, S', n_run, n_acc)
Standardise Θ̂_ME
for S' ∈ 2^S do
    Θ_acc ← Parameter sets accepted from "Best Samples"
      ABC-RS(x_obs, S', n_run, n_acc)
    Standardise Θ_acc
    MRSSE_S' ← MRSSE(Θ_acc, Θ̂_ME,i)
S* ← argmin_{S'∈2^S} MRSSE_S'
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, rather than choose from a set of hand-crafted ones. Their approach (**Algorithm 4.5**) uses a pilot run of an ABC-method to generate a naïve approximation of the parameter posterior which is used to generate summary statistics. The approximate posterior is used to generate a "training set" from which a regression model can be 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 however the identity function would be appropriate, although inefficient.

Algorithm 4.5 (Semi-Automatic ABC, 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[9], $\alpha \in \mathbb{R}, \beta \in \mathbb{R}^{|\theta|}$ 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

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

   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

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

the quadratic loss function L_2 is used meaning the problem is to find

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

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

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

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 $\mathbf{X}, \mathbf{Y}$[10] 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.

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

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[11]. Fearnhead & Prangle recommend the use of linear regression as it is straight-forward to implement and does not perform notably worse than the other approaches in general.

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

[11]

For their specific implementation of linear least-squares regression they treat each model parameter θ_i completely separately and allow for mappings $f(\cdot)$ of the response data. This means they are fitting $\rho = |\theta|$ different models

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

As ABC-methods only consider the distance between summary statistic values, the constant terms $\alpha^{(i)}$ can be neglect from our generate summary statistics. This means the summary statistic s_i for the i^{th} model parameter is defined as

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

The mapping $f(\cdot)$ is a parameter of this algorithm and should be used to encode likely relationships between observations and parameters, however 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, it 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 $\hat{X} \leftarrow \{X_{\hat{\theta}_1}, \ldots, X_{\hat{\theta}_m}\}$
5 $F \leftarrow f(\hat{X})$
6 $\tilde{F} \leftarrow F$ *with a preceding column of 1s*
7 **for** $i = 1, \ldots, \rho$ **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** $\{s_1, \ldots, s_\rho\}$

$\rho := |\theta|$, 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 the "Best Samples" version of ABC-Rejection Sampling is it has the simplest 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 time expert knowledge can be encoded is in the specification of map $f(\cdot)$.

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 avoids this issue as you only have to specify what acceptance rate you wish to achieve.

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[12] 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 Fearnhead & Prangle's semi-automatic approach 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 dangerous being overfitting. Overfitting occurs when a neural network models the training 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, especially when using the identity function as the only summary statistic (due to the sufficiency of the identity function). Alternatively, the use of less naïve statistic 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 very intricate investigation of the network.

[12] They also use a feedforward neural-network but these cannot model non-linear relationships unless they use a non-linear activation function.

4.3.6 Toy Example

4.4 Model Selection

Theorems which state when a model is misspecified that bayesian inference will put mass on the distributions "closest to the ground truth" rely on strong regularity conditions. [Grünwald and van Ommen, 2018]

Introduce learning rate (SafeBayes) [Grünwald and van Ommen, 2018]

5 ABC and Epidemic Events

6 Conclusion

6.1 Future Areas of Research

References

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

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

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

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.

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

Burr, T. and Skurikhin, A. (2013). Selecting summary statistics in approximate bayesian computation for calibrating stochastic models. *BioMed research international* **2013**, 210646.

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.

Darmois, G. (1935). Sur les lois de probabilité à estimation exhaustive. *Comptes Rendus de l'Académie des Sciences* , 1265–1266.

Dodge, Y., Institute, I. S. and Commenges, D. (2006). *The Oxford Dictionary of Statistical Terms*. Oxford University Press.

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**(1), 87–112.

Fearnhead, P. and Prangle, D. (2011). Constructing summary statistics for approximate bayesian computation: Semi-automatic abc .

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A* **222**, 309–368.

Grünwald, P. and van Ommen, T. (2018). Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it .

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2nd edition.

Hayashi, F. (2000). *Econometrics*. Princeton Univ. Press, Princeton, NJ [u.a.].

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.

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.

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.

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).

Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. *Proceedings of the Cambridge Philosophical Society* , 567–579.

Rao, C. R. (1945). *Information and accuracy attainable in the estimation of statistical parameters*. Bulletin of the Calcutta Mathematical Society, 81–91.

Roussas, G. (1998). *A Course in Mathematical Statistics*. Academic Press, 2nd edition, 263.

Ruli, E. (2018). On model selection with summary statistics.

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.

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* .