Bayesian Modelling of Epidemic Processes

D. Hutchinson

April 6, 2021

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

3 Approximate Bayesian Computation

What are “Simulation Methods”? (Possible now due to greater computational power and quality/quantity of data available) Two versions: allow for running models with stochastic components and testing variety of results; or, simulation for statistical inference and parameter estimation (vary parameters in simulation and compare results to sampled data).

Simulation methods are useful here as it is easier to simulate from a distribution than to calculate it.

ABC extends many Bayesian simulation methods to only require approximate matching of sampled and simulated data, rather than exact. (ABC allows for more simplified models?). Trade-offs?

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

ABC simulates from the likelihood, rather than explicitly determine the likelihood.

ABC has two uses: calibrating model, comparing models.

What is the computational effiency?

How to tune tolerance ε? Run tests showing performance with different tolerances (plots of posteriors when using different tolerances).

Curse of dimensionality (regression methods can counter this)

Accept-Reject-ABC

Perform poorly when prior and posterior and very different (especially when overlap is small).

Worth tracking “acceptance rate” to test algorithms performance. Don’t want too high or too low, time v quality.

Variations - Sample until N accepted (how to determine width of acceptance); Sample M times and keep best N (how to determine best); Sample M times, keep all and weight parameters by distance of simulated data from sampled data (weighted linear-regression).

Accept-reject algorithms for MLE (rather than Bayesian). (Use mode of posterior distribution as MLE or use simulation to approximate likelihood).

Using a uniform kernel can be interpreted as sampling from a model which has uniformly distributed additive error.

Independent samples

Acceptance-rate analogous to evidence.

User choices: summary stats; threshold; distance measure.

Acceptance-rate vs tolerance plot?

re

3.1 Importance Sampling

Use when have a better knowledge about parameter value distribution. (May be good for RONA as we are given good R rate estimates?)

Sample parameters θ from an “Importance distribution” $\xi(\theta)$ rather than a “prior” $\pi(\theta)$. Weight each accepted parameter θ as $\frac{\pi(\theta)}{\xi(\theta)}$.

Less variance between sampled and simulated summary statistic values.

Could use rejection algorithm to determine a good importance distribution.

Independent samples

https://bookdown.org/rdpeng/advstatcomp/importance-sampling.html

ABC-MCMC, -SMC & -PMC

MCMC, SMC are search processes

ABC-MCMC=[Marjoram *et al.*, 2003] (Markov Chain Monte Carlo).

ABC-SMC=[Sisson *et al.*, 2007] (Sequential Monte Carlo).

ABC-PMC=[Beaumont *et al.*, 2009] (Population Monte-Carlo).

Good approach for large data sets, and when prior & posterior are likely to be very different.

Sequences of dependent samples

Adaptive approach to ABC-SMC ([Moral *et al.*, 2012])

When does MCMC converge?

MCMC w/o summary stats converges to the true posterior $\mathbb{P}(\theta|D)$.

What is the burn-in period like? How good/important is mixing?

Advantage of ABC-MCMC over ABC-rejection? Fewer simulations required to get n accepted samples (for a given tolerance ε).

Acceptance-rate vs tolerance plot?

Stationary distribution of the MCMC is an estimate of the posterior?

How to choose perturbation kernels ([Filippi *et al.*, 2012]).

Model Choice

Regression Adjustment

Beaumont et al - Local Linear Regressions (LOCL)

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

Review

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

4 Summary Statistic Selection

In this chapter I motivate the research into summary statistic selection 4.1 and discuss features to consider when selecting summary statistics 4.2. I then describe five methods for summary statistic selection methods: three which use hand-crafted summary statistics 4.3.1-4.3.3; and two which automatically generate summary statistics 4.3.4-4.3.5; To close the section, a toy example of an SIR model to compare these methods 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.

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

Ideally, a summary statistic would compress the sampled data without any information loss (A property known as "sufficiency"). However, sufficient summary statistics are very rare in practice and we often have to trade-off information retention against dimensionality reduction.

For more complex models, with many parameters, it becomes difficult for a single summary statistic to accurately summarise the sampled data, thus, in practice it is common to apply a set of summary statistics $\{s_1, \dots, s_k\}$ to the same dataset with each targeting a different aspect of the model. 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 effect 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 the returned posteriors being the same as the priors supplied to the algorithm.

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 the more computationally efficient ABC methods (e.g. ABC-Rejection Sampling) will be sufficient to fit the model, given a good choice of summary statistics.

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 ABC 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 allow can be used to automate the process of determining these weights by specifying multiple versions of the same summary statistic but with each version having a different weight.

4.2 Properties of Summary Statistics

When evaluating a summary statistic for use in ABC there are several properties, both mathematical and practical, to consider. I first discuss practical properties to consider, and then the key mathematical property of sufficiency.

Practical Properties

The key reason for using summary statistics is for the computational efficiencies which result from their dimensionality reduction as this means more simulations can be processed in the same time-period. This naturally means summary statistics which result in greater dimensionality reduction are more 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. Some statistics, namely order statistics, are variable in the their time complexity for different data sets which will affect the reliability of models which utilise the. Integer overflow is a possible issue for some summary statistics, although it is often easy to avoid when actively been 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.

For real-world modelling problems, the interpretability of summary statistics used in the final model is a key factor in how useful this 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 than automatically generated statistics, as such generated statistics are rarely used in commercial problems[1]. In cases where it is chosen to used 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.

Sufficiency

Definition 4.1 (Sufficient Statistic)
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. [Dodge et al.*, 2006]*

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

More verbosely, a statistic is sufficient for a parameter(s) if it captures all the information which a sample of the model carries about said parameter(s). This means, 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 and but still captures all relevant information.

Theorem 4.1
Let $S_{1:k-1}(\cdot) := \{s_1(\cdot), \dots, s_{k-1}(\cdot)\}$ 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. Let $S_{1:k} := S_{1:k-1} \cup \{s_k\}$. Consider the posterior for model parameters θ given the summary statistics $S_{1:k}$

$$\begin{aligned}
\mathbb{P}(\theta|S_{1:k}) &= \frac{\mathbb{P}(\theta, S_{1:k})}{\mathbb{P}(S_{1:k})} \\
&= \frac{\mathbb{P}(\theta, S_{1:k})}{\mathbb{P}(S_{1:k})} \cdot \frac{\mathbb{P}(S_{1:k-1})}{\mathbb{P}(S_{1:k-1})} \\
&= \frac{\mathbb{P}(\theta, S_k|S_{1:k-1})}{\mathbb{P}(S_k|S_{1:k-1})} \\
&= \frac{\mathbb{P}(\theta, S_k|S_{1:k-1})}{\mathbb{P}(S_k|S_{1:k-1}, \theta)} \text{ as } S_{1:k-1} \text{ is sufficient} \\
&= \frac{\mathbb{P}(\theta, S_{1:k})}{\mathbb{P}(S_{1:k-1})} \cdot \frac{\mathbb{P}(\theta, S_{1:k-1})}{\mathbb{P}(\theta, S_{1:k})} \\
&= \frac{\mathbb{P}(\theta, S_{1:k-1})}{\mathbb{P}(S_{1:k-1})} \\
&= \mathbb{P}(\theta|S_{1:k-1})
\end{aligned}$$

[1] 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.

This shows that if $S_{1:k-1}$ is sufficient for θ, then the posterior for θ will be the same for any superset of $S_{1:k-1}$. This means that all supersets of a sufficient set of statistics is also sufficient. □

Theorem 4.1 shows 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.

Here is an example which proves that the sum of sampled values 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(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*

This example 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 4.2**), first recognised by Fisher in [Fisher, 1922], specifies a property which all sufficient statistics have. This property is used as basis for a more formulaic approach to finding sufficient statistics by seperating the terms of the conditional probability of a model given the summary statistic value in those which depend on the summary statistic and those which do not.

Theorem 4.2 (Fisher-Neyman Factorisation Criterion)
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 *<u>iff</u> 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.

[Fisher, 1922; Neyman, 1935]

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. □

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

The example 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 $s(\mathbf{X}) = \displaystyle\sum_{i=1}^{n} X_i$ 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 mean and identity function are both sufficient statistics, however the sample mean is a much more desirable statistic to use as it provides compression. This lack of uniqueness motivates the concept of minimal sufficiency.

Definition 4.2 (Minimally Sufficient Statistic)
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 $t(x) \mapsto s(x)$ [Dodge et al.*, 2006]*

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

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 the example below

Example 4.3 (Minimally Sufficient Statistic for IID Bernoulli Random Variables)
Let $X_1, \dots, X_n$ 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 $(1, 0, \dots, 0)$ and $(0, 1, \dots, 0)$. 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 4.3 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 propert can be used to identify minimally sufficient summary statistics, either by assisting in deduction or by checking a proposed statistic.

Theorem 4.3 (Condition for Minimal Sufficiency)
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.

[Balakrishnan, 2019]

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

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

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. □

Statistics carry information about sampled data, but in bayesian modelling most problems center around estimating parameter values. In some cases a sufficient statistic may be a good estimator of a model paramater, 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.

The Rao-Blackwell theorem (**Theorem 4.4**) 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 “easy” to derive a crude estimator and then this theorem can be applied 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.

Theorem 4.4 (Rao-Blackwell Theorem)
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. [Rao, 1945; Blackwell, 1947]

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 $v(X)$

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

□

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

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.

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.

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 maintain a large about of the relevant data from a sample. Moreover, the Pitman-Koopman-Darmois theorem **Theorem 4.5** shows that summary statistics which provide a high level of dimensionality reduction only exist for probability distributions from exponential families.

Theorem 4.5 (Pitman–Koopman–Darmois Theorem)
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. [Singh, 2015]

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

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] paper 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 sampled values is a sufficient statistics for the means of both geometric and Poisson models, 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 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 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 different model. This is the motivation behind most summary statistic selection methods.

4.3.1 Approximate Sufficient Subset

[Joyce and Marjoram, 2008]

4.3.2 Minimising Entropy

[Nunes and Balding, 2010]

4.3.3 Two-Step Minimum Entropy

[Nunes and Balding, 2010]

4.3.4 Semi-Automatic ABC

[Fearnhead and Prangle, 2011]

4.3.5 Non-Linear Projection

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

Balakrishnan, S. (2019). Lecture notes in 36-705: Intermediate statistics, lecture 12.

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.

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.

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 .

Filippi, S., Barnes, C., Cornebise, J. and Stumpf, M. P. H. (2012). On optimality of kernels for approximate bayesian computation using sequential monte carlo .

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 .

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.

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.

Moral, P., Doucet, A. and Jasra, A. (2012). An adaptive sequential monte carlo method for approximate bayesian computation **22**(5), 1009–1020.

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.

Singh, A. (2015). Lecture notes in 10-704: Information processing and learning, lecture 16.

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.

Wegmann, D. and Excoffier, L. (2010). Bayesian Inference of the Demographic History of Chimpanzees. *Molecular Biology and Evolution* **27**(6), 1425–1435.