Summary Statistic Selection

Dom Hutchinson

March 31, 2021

Contents

1 Summary Statistics

Definition 1.1 - *Summary Statistics* $s(\cdot)$
Summary Statistics are a projection of high-dimensional data to a lower dimensional space.

$$s(\cdot) : \mathbb{R}^m \to \mathbb{R}^n \quad \text{where } m > n$$

This is aimed to be done in such a way that as much information is retained as possible (ie the summary is as accurate as possible). The lower dimensional projection is important to make ABC computationally tractable.

The trade-off here is between computational requirements and data retention. As data is lost posterior accuracy and stability decreases.

Remark 1.1 - *Summary Statistics for ABC*
In ABC we want each *Summary Statistic* to map from all observations of the model to a lower dimensional space. As each observation can be multi-dimensional, this means *Summary Statistics* for ABC typically map from a matrix to a vector

$$s(\cdot) : \mathbb{R}^{n \times m} \to \mathbb{R}^p$$

where n is the number of observations, m is the dimension of each observation and p is the dimension being mapped to. Ideally $p \ll n \times m$. Moreover, $\sum p_i \ll n \times m$ where $\{p_1, \dots, p_k\}$

are the dimensions the k summary stats map to.

Proposition 1.1 - *Common Summary Statistics*
Typically summary statistics describe the following

- centre of the data (e.g. mean).
- spread of the data (e.g. variance).
- shape of the data (e.g. pearson's skew).
- dependence of different data fields (e.g correlation).
- start or end points.

Remark 1.2 - *Reduced Dimension*
Suppose we have N samples of each of M dimension, so all the data is represented by a matrix $\mathbf{X} \in \mathbb{R}^{N\times M}$[1]. If we were to just use the: mean, variance and pearson's skew to summarise each property, then the summarised data would only require a $\mathbb{R}^{3\times M}$ matrix[2] whose size is independent of the number of observations N and, significantly smaller than $\mathbb{R}^{N\times M}$.

The question is whether these three properties are sufficient to make valid/inciteful inferences from.

2 Sufficiency

Proposition 2.1 - *Approximate Sufficiency (AS)*
Approximate Sufficiency is the practice of finding the subset of summary statistics (from a larger set) which satisfy some optimality condition. This is done by identifying a large set of summary statistics S and then finding a subset $S' \subset S$ which is approximately as good as the superset S. There are several measures of sufficiency.

Proposition 2.2 - *Approach to AS*
A typical approach to AS is to do the following[3]

i). Choose a measure of sufficient $M(\cdot)$.

ii). Start with an empty set $S' = \emptyset$.

iii). Keep adding summary statistics $s \in S/S'$ to S' until $M(S')$ is no longer satisfied.

This approach has limitations since the final subset S' depends on the order in which summary statistics are added. Finding a way to order the elements of S would help this.

[1] ie N observations, each with M properties.
[2] 3 values per property
[3] This approach requires the underlying model to be known, as we specify the model we are trying to fit, this is ok."

2.1 Sufficiency

Remark 2.1 - *Distinguishing Models*
If two models have the same sufficient statistics it is impossible to distinguish between them.

Definition 2.1 - *Classical Sufficiency*
Let $X \sim f(\cdot;\theta)$[4] and $s(\cdot)$ be a summary statistic. $s(X)$ is a *sufficient statistic* for θ if

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

ie the conditional distribution for X, given the summary statistic $s(X)$, is independent of θ.

This can be read intuitively that $s(X)$ captures all the information the sample X contains about the parameter θ. (A lossless data-compression).

Example 2.1 - *Sufficient Statistic - Bernoulli Distribution*
Let $X \sim \text{Bern}(p)$ with $p \in [0,1]$ unknown and $\mathbf{x}$ be n independent samples of X. Note that

$$\begin{aligned} & f_X(x) &:=& \ p^x(1-p)^{1-x} \\ \implies & f_{\mathbf{X}}(\mathbf{x}) &=& \ \prod_{i=1}^n f_X(x_i) \text{ by independence of samples} \\ & &=& \ p^{\sum x_i}(1-p)^{n-\sum x_i} \end{aligned}$$

Consider the summary statistic $T(\mathbf{x}) := \sum x_i$. Note that $T(\mathbf{X})$ is only one dimensional (rather than n) and $T(\mathbf{X}) \sim \text{Binomial}(n,p)$. Thus

$$\begin{aligned} f_T(T(\mathbf{x})) &= \tbinom{T(\mathbf{x})}{n} p^{T(\mathbf{x})}(1-p)^{n-T(\mathbf{x})} \\ f_{\mathbf{X},T}(\mathbf{x},T(\mathbf{x})) &= p^{T(\mathbf{x})}(1-p)^{n-T(\mathbf{x})} \end{aligned}$$

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

$$\begin{aligned} f_{\mathbf{X}|T(\mathbf{X})}(\mathbf{x}|T(\mathbf{x})) &= \frac{f_{\mathbf{X},T(\mathbf{X})}(\mathbf{x})}{f_{T(\mathbf{X})}(\mathbf{x})} \\ &= \frac{p^{T(\mathbf{x})}(1-p)^{n-T(\mathbf{x})}}{\binom{T(\mathbf{x})}{n} p^{T(\mathbf{X})}(1-p)^{n-T(\mathbf{X})}} \\ &= \frac{1}{\binom{T(\mathbf{x})}{n}} \end{aligned}$$

The conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$ is independent of p, thus $T(\mathbf{X})$ is a sufficient statistic for p.

Example 2.2 - *Sufficient Statistic - Gaussian Distribution with Unknown Mean*
Let $X \sim \text{Normal}(\mu, \sigma_0^2)$ where $\mu \in \mathbb{R}$ is unknown and $\sigma_0^2 \in \mathbb{R}$ is known, and $\mathbf{x}$ be n independent observations of X. Note that

$$\begin{aligned} & f_X(x) &:=& \ \tfrac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\tfrac{1}{2\sigma_0^2}(x-\mu)^2\right\} \\ \implies & f_{\mathbf{X}}(\mathbf{x}) &=& \ \prod_{i=1}^n f_X(x_i) \text{ by independence of samples} \\ & &=& \ (2\pi\sigma_0^2)^{-n/2} \exp\left\{-\tfrac{1}{2\sigma_0^2}\sum(x_i-\mu)^2\right\} \end{aligned}$$

[4] X could be one or many observations.

Consider the following formulation of the distribution of $\mathbf{X}$ with an arbitrary term t introduced

$$\begin{aligned}
f_{\mathbf{X}}(\mathbf{x},t) &= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(x_i+t-t-\mu)^2\right\}\\
&= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum((x_i-t)-(\mu-t))^2\right\}\\
&= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum\left[(x_i-t)^2+(\mu-t)^2-2(\mu-t)(x_i-t)\right]\right\}\\
&= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(x_i-t)^2\right\}\cdot\exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(\mu-t)^2\right\}\cdot\exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum-2(\mu-t)(x_i-t)\right\}\\
&= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(x_i-t)^2\right\}\cdot\exp\left\{-\tfrac{n}{2\sigma_0^2}(\mu-t)^2\right\}\cdot\exp\left\{-\tfrac{-2(\mu-t)}{2\sigma_0^2}\textstyle\sum(x_i-t)\right\}\\
&= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(x_i-t)^2\right\}\cdot\exp\left\{-\tfrac{n}{2\sigma_0^2}(\mu-t)^2\right\}\cdot\exp\left\{-\tfrac{-2(\mu-t)}{2\sigma_0^2}\left[(\textstyle\sum x_i)-nt\right]\right\}
\end{aligned}$$

The third exponential disappears if $t := \frac{1}{n}x_i$ (the sample mean). Consider the summary statistic $T(\mathbf{X}) := \frac{1}{n}\sum X_i$ meaning $T(\mathbf{X}) \sim \text{Normal}(\mu, \frac{1}{n}\sigma_0^2)$ by the Central Limit Theorem.
The following are the marginal distribution for $T(\mathbf{X})$ and the joint distribution of $\mathbf{X}$ and $T(\mathbf{X})$.

$$\begin{aligned}
f_{T(\mathbf{X})}(T(\mathbf{x})) &= \sqrt{\tfrac{n}{2\pi\sigma_0^2}}e^{-\frac{n}{2\sigma_0^2}(\mu-T(\mathbf{x}))^2} = n^{1/2}(2\pi\sigma_0^2)^{-1/2}\exp\left\{-\tfrac{n}{2\sigma_0^2}(\mu-T(\mathbf{x}))^2\right\}\\
f_{\mathbf{X},T(\mathbf{X})}(\mathbf{x},T(\mathbf{x})) &= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(x_i-T(\mathbf{x}))^2\right\}\cdot\exp\left\{-\tfrac{n}{2\sigma_0^2}(\mu-T(\mathbf{x}))^2\right\}
\end{aligned}$$

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

$$\begin{aligned}
f_{\mathbf{X}|T(\mathbf{X})}(\mathbf{x}|T(\mathbf{x})) &= \frac{f_{\mathbf{X},T(\mathbf{X})}(\mathbf{x},T(\mathbf{x}))}{f_{T(\mathbf{X})}(T(\mathbf{x}))}\\
&= \frac{(2\pi\sigma_0^2)^{-n/2}\exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(x_i-T(\mathbf{x}))^2\right\}\cdot\exp\left\{-\tfrac{n}{2\sigma_0^2}(\mu-T(\mathbf{x}))^2\right\}}{n^{1/2}(2\pi\sigma_0^2)^{-1/2}\exp\left\{-\tfrac{n}{2\sigma_0^2}(\mu-T(\mathbf{x}))^2\right\}}\\
&= n^{-1/2}(2\pi\sigma_0^2)^{-n/2}\exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(x_i-T(\mathbf{x}))^2\right\}
\end{aligned}$$

The conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$ is independent of μ, thus $T(\mathbf{X})$ is a sufficient statistic for μ.

Proof 2.1 - *If $S_{1:k-1}$ are sufficient then $\mathbb{P}(\theta|S_{1:k}) = \mathbb{P}(\theta|S_{1:k-1})$ for all S_k.*
Let $S_{1:k-1}$ be a set of summary-statistics which are sufficient for parameters θ, and S_k be any other summary statistic. Then

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

2.2 Minimal Sufficiency

Definition 2.2 - *Minimal Sufficiency*
A sufficient statistic $s(X)$ is *Minimally Sufficient* if it can be represented as a function of any other sufficient statistic $t(X)$.

$$\exists\ f \text{ st } s(X) = f(t(X))$$

Example 2.3 - *Minimal Sufficient Statistics*
TODO

2.3 Fisher-Neyman Factorisation Theorem

Theorem 2.1 - *Fisher-Neyman Factorisation Theorem*
If $X \sim f(\cdot;\theta)$ then $s(\cdot)$ is sufficient for θ iff there exists non-negative functions $g(\cdot;\theta), h(\cdot)$ st

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

This shows that the data X only interacts with the parameter θ through the sufficient summary statistics $s(X)$.

Proof 2.2 - *Fisher-Neyman Factorisation Theorem*

$\longrightarrow$ Let s be a sufficient statistic and $h(x) = \mathbb{P}(X = x|s(X) = s(x))$ be a function which is independent of θ.

Let $g(t;\theta) = \mathbb{P}(S(x) = s(x))$. Then

$$\begin{aligned} f(x;\theta) &= \mathbb{P}(X = x|s(X) = s(x)) \cdot \mathbb{P}(s(X) = s(x)) \\ &= g(x)k(t(x);\theta) \end{aligned}$$

$\longleftarrow$ Suppose

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

Then

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

This final expression is independent of θ

□

Remark 2.2 - *Usefulness of Fisher-Neyman Factorisation Theorem*
In `Example 1.1` and `Example 1.2` we had to guess at a definition of $T(\mathbf{X})$ which produced a sufficient statistic. The *Fisher-Neyman Factorisation Theorem* removes a lot of that guesswork, in place of a more formulaic approach to finding sufficient statistics (as shown in

`Example 1.4` & `Example 1.5`).

Example 2.4 - *Sufficient Statistic - Bernoulli Distribution \w FNF Theorem*
Let $X \sim \text{Bern}(p)$ with $p \in [0,1]$ unknown and $\mathbf{x}$ be n independent samples of X. Note that

$$\begin{aligned} f_X(x) &:= p^x(1-p)^{1-x} \\ \implies f_{\mathbf{X}}(\mathbf{x}) &= \textstyle\prod_{i=1}^n f_X(x_i) \text{ by independence of samples} \\ &= p^{\sum x_i}(1-p)^{n-\sum x_i} \end{aligned}$$

Note that we can factorise $f_{\mathbf{X}}(\mathbf{x})$ as $f_{\mathbf{X}}(\mathbf{x}) = h(\mathbf{X})g(\sum X_i|p)$ where

$$\begin{aligned} h(\mathbf{X}) &:= 1 \\ g(\textstyle\sum X_i|p) &:= p^{\sum X_i}(1-p)^{n-\sum X_i} \\ \Leftrightarrow g(T(\mathbf{X})|p) &= p^{T(\mathbf{X})}(1-p)^{n-T(\mathbf{X})} \text{ where } T(\mathbf{X}) := \textstyle\sum X_i \end{aligned}$$

Notice that $h(\cdot)$ is independent of the unknown parameter p and that $g(\cdot|p)$ only interacts with p through the summary statistic $T(\mathbf{X})$. Thus by the *Fisher-Neyman Factorisation Theorem* $T(\mathbf{X}) := \sum X_i$ is a sufficient statistic for p.

Example 2.5 - *Sufficient Statistic - Gaussian Distribution \w FNF Theorem*
Let $X \sim \text{Normal}(\mu, \sigma_0^2)$ where $\mu \in \mathbb{R}$ is unknown and $\sigma_0^2 \in \mathbb{R}$ is known, and $\mathbf{x}$ be n independent observations of X. Note that

$$\begin{aligned} f_X(x) &:= \tfrac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\tfrac{1}{2\sigma_0^2}(x-\mu)^2\right\} \\ \implies f_{\mathbf{X}}(\mathbf{x}) &= \textstyle\prod_{i=1}^n f_X(x_i) \text{ by independence of samples} \\ &= (2\pi\sigma_0^2)^{-n/2} \exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(x_i-\mu)^2\right\} \end{aligned}$$

Consider the following formulation of the distribution of $\mathbf{X}$ with an arbitrary term t introduced

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x},t) &= (2\pi\sigma_0^2)^{-n/2} \exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(x_i+t-t-\mu)^2\right\} \\ &= (2\pi\sigma_0^2)^{-n/2} \exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum((x_i-t)-(\mu-t))^2\right\} \\ &= (2\pi\sigma_0^2)^{-n/2} \exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum\left[(x_i-t)^2+(\mu-t)^2-2(\mu-t)(x_i-t)\right]\right\} \\ &= (2\pi\sigma_0^2)^{-n/2} \exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(x_i-t)^2\right\} \cdot \exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(\mu-t)^2\right\} \cdot \exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum -2(\mu-t)(x_i-t)\right\} \\ &= (2\pi\sigma_0^2)^{-n/2} \exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(x_i-t)^2\right\} \cdot \exp\left\{-\tfrac{n}{2\sigma_0^2}(\mu-t)^2\right\} \cdot \exp\left\{-\tfrac{-2(\mu-t)}{2\sigma_0^2}\textstyle\sum(x_i-t)\right\} \\ &= (2\pi\sigma_0^2)^{-n/2} \exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(x_i-t)^2\right\} \cdot \exp\left\{-\tfrac{n}{2\sigma_0^2}(\mu-t)^2\right\} \cdot \exp\left\{-\tfrac{-2(\mu-t)}{2\sigma_0^2}\left[(\textstyle\sum x_i)-nt\right]\right\} \end{aligned}$$

The third exponential disappears if $t := \frac{1}{n}x_i$ (the sample mean). Define $T(\mathbf{X}) := \frac{1}{n}\sum X_i$, thus

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi\sigma_0^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma_0^2}\sum(x_i-T(\mathbf{x}))^2\right\} \cdot \exp\left\{-\frac{n}{2\sigma_0^2}(\mu-T(\mathbf{x}))^2\right\}$$

Note that we can factorise this expression $f_{\mathbf{X}}(\mathbf{x})$ as $f_{\mathbf{X}}(\mathbf{x}) = h(\mathbf{X})g(T(\mathbf{X})|\mu)$ where $T(\mathbf{X}) := \frac{1}{n}\sum X_i$ and

$$\begin{aligned} h(\mathbf{X}) &= (2\pi\sigma_0^2)^{-n/2} \exp\left\{-\tfrac{1}{2\sigma_0^2}\textstyle\sum(X_i-T(\mathbf{X}))^2\right\} \\ g(T(\mathbf{X})|\mu) &= \exp\left\{-\tfrac{n}{2\sigma_0^2}(\mu-T(\mathbf{X}))^2\right\} \end{aligned}$$

Notice that $h(\cdot)$ is independent of the unknown parameter μ and that $g(\cdot|\mu)$ only interacts with μ through the summary statistic $T(\mathbf{X})$. Thus by the *Fisher-Neyman Factorisation Theorem* $T(\mathbf{X}) := \frac{1}{n}\sum X_i$ is a sufficient statistic for μ.

2.4 Bayesian Sufficiency

Definition 2.3 - *Bayesian Sufficiency[1]*
In a Bayesian Setting, a summary statistic $s(X)$ of $X \sim f(\cdot;\theta)$ is sufficient if for (almost) all $x \in X$

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

ie the posterior for θ given the true model X is the same as the posterior for θ given the summary statistic $s(X)$ (for almost all $x \in X$).

2.5 ϵ-Approximate Sufficiency

Remark 2.3 - *Approximate Sufficiency*
Approximate sufficiency concerns sets of summary statistics, rather than a single statistic. We are looking for a set of statistics which are approximate sufficient, while no individual statistic in the set is.

This can be used when the distribution is not explicitly known (and can only be approximated)?

Remark 2.4 - *Motivation for using Approximate Sufficiency*
Suppose we have a set of summary statistics $\{s_1(\cdot), \dots, s_n(\cdot)\}$ for parameter θ and are consider adding a new statistic $s_{n+1}(\cdot)$. We only want to do this if$s_{n+1}(\cdot)$ offers a substantial amount of new information about θ.

$$\mathbb{P}(s_{1:n+1}(X)|\theta) \gg \mathbb{P}(s_{1:n}(X)|\theta)$$

Note that

$$\begin{aligned}
\mathbb{P}(s_{1:n+1}(X)|\theta) &= \mathbb{P}(s_{n+1}(X)|s_{1:n}(X),\theta)\mathbb{P}(s_{1:n}|\theta) \\
\implies \mathbb{P}(\theta|s_{1:n+1}(X)) &= \frac{\mathbb{P}(s_{n+1}(X)|s_{1:n}(X),\theta)\mathbb{P}(s_{1:n}(X)|\theta)\mathbb{P}(\theta)}{(s_{n+1}(X)|s_{1:n}(X))\mathbb{P}(s_{1:n}(X))} \text{ by Bayes} \\
&= \frac{\mathbb{P}(s_{1:n}(X)|\theta)\mathbb{P}(\theta)}{\mathbb{P}(s_{1:n}(X))} \text{ if } \mathbb{P}(s_{n+1}(X)|s_{1:n}(X)) = \mathbb{P}(s_{n+1}|s_{1:n}(X),\theta)^{[5]}
\end{aligned}$$

We don't want to add $s_{n+1}(\cdot)$ if $s_{1:n}(\cdot)$ are already sufficient for θ.

As it is unlikely for $s_{1:n}(\cdot)$ to be completely sufficient, especially in high-dimensional systems[6] so we don't add s_{n+1} if $s_{1:n}$ are *Approximately Sufficient.*

Remark 2.5 - *Motivation for Using Log-Likelihood*
Consider the following

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

This means $\ln\mathbb{P}(s_{n+1}(X)|s_{1:n}(X),\theta)$ is the only difference between $\ln\mathbb{P}(s_{1:n}|\theta)$ and $\ln\mathbb{P}(s_{1:n+1}|\theta)$. Thus, the smaller the value of $\ln\mathbb{P}(s_{n+1}(X)|s_{1:n}(X),\theta)$ the closer the two likelihoods are, implying $s_{1:n}$ and $s_{1:n+1}$ provide very similar amounts of information[7]

Definition 2.4 - *Score δ_{n+1}[2]*

[5] i.e. If $s_{1:n}(\cdot)$ are sufficient for θ.
[6] and as we are using empirical observation + stocastic noise.
[7] Meaning there is little to be gained by including $s_{n+1}(\cdot)$.

Let $\{T_1, \dots, T_n\}$ be a set of summary statistics for parameters θ and T_{n+1} be a statistic which we consider adding to the set.

The *Score* δ_{n+1} of the new statistic T_{n+1} relative to the set $\{T_1, \dots, T_n\}$ is defined as

$$\delta_{n+1} := \sup_\theta \ln \left[\mathbb{P}(T_{n+1}|T_1, \dots, T_n, \theta)\right] - \inf_\theta \ln \left[\mathbb{P}(T_{n+1}|T_1, \dots, T_n, \theta)\right]$$

Score δ_{n+1} is a measure of how much new information T_{n+1} introduces

Definition 2.5 - *ϵ-Approximate Sufficiency [2]*
Let $\{T_1, \dots, T_n\}$ be a set of summary statistics for parameter θ and T_{n+1} be a statistic which we consider adding to the set.

The set $\{T_1, \dots, T_n\}$ is *ϵ-Sufficient* to the new statistic T_{n+1}, if the score of T_{n+1} relative to $\{T_1, \dots, T_n\}$ is no-greater than some ϵ

$$\delta_{n+1} \leq \epsilon$$

If $T_{n+1} = \mathbf{X}$, the whole data set and $\epsilon = 0$ then this is the same as the definition for sufficiency.

Definition 2.6 - *Odds-Ratio $R_n(\theta)$*
Let $s_1(\cdot), \dots, s_n(\cdot)$ be a set of summary statistics and $\pi_0(\theta)$ be a prior for θ. Then the *Odds-Ratio* for s_n to $s_{1:n}$ is defined as

$$R_n(\theta) := \frac{\mathbb{P}(\theta|s_{1:n}(X))}{\mathbb{P}(\theta|s_{1:n-1}(X))}$$

We are able to estimate this quantity from simulation.

Theorem 2.2 -
Let $s_1(\cdot), \dots, s_n(\cdot)$ be a set of summary statistics, $\pi_0(\theta)$ be a prior for θ and δ_n be the score of $s_n(\cdot)$ relative to $s_{1:n-1}(\cdot)$. Then

$$e^{-\delta_n} \leq R_n(\theta) \leq e^{\delta_n}$$

Thus, the closer the score δ_n is to 0 the closer in value the two likelihoods $\mathbb{P}(\theta|s_{1:n}(X))$ and $\mathbb{P}(\theta|s_{1:n-1}(X))$ are (the more sufficient $s_{1:n-1}$ are for $s_{1:n}$).

Proof 2.3 - *Theorem 2.2*
Consider the numerator

$$\begin{aligned}
\mathbb{P}(\theta|S_{1:k}) &= \frac{\mathbb{P}(\theta, S_{1:k})}{\mathbb{P}(S_{1:k})} \\
&= \frac{\mathbb{P}(\theta, S_{1:k}|\theta)\pi(\theta)}{\int \mathbb{P}(S_{1:k}|\theta)\pi(\theta)d\theta} \\
&= \frac{\mathbb{P}(S_k|S_{1:k-1}, \theta)\mathbb{P}(S_{1:k-1}|\theta)\pi(\theta)}{\int \mathbb{P}(S_k|S_{1:k-1}, \theta)\mathbb{P}(S_{1:k-1}|\theta)\pi(\theta)d\theta} \\
&\leq \frac{\mathbb{P}(S_{1:k-1}|\theta)\pi(\theta)}{\int \mathbb{P}(S_{1:k-1}|\theta)\pi(\theta)d\theta} \cdot \frac{\sup_\theta \mathbb{P}(S_k|S_{1:k-1}, \theta)}{\inf_\theta \mathbb{P}(S_k|S_{1:k-1}, \theta)} \\
&= \frac{\mathbb{P}(S_{1:k-1}|\theta)\pi(\theta)}{\mathbb{P}(S_{1:k-1})} \cdot \exp\left\{\ln\left(\frac{\sup_\theta \mathbb{P}(S_k|S_{1:k-1}, \theta)}{\inf_\theta \mathbb{P}(S_k|S_{1:k-1}, \theta)}\right)\right\} \\
&= \underbrace{\mathbb{P}(\theta|S_{1:k-1})}_{\text{Bayes' Rule}} \cdot \exp\left\{\sup_\theta \ln\left(\mathbb{P}(S_k|S_{1:k-1}, \theta)\right) - \inf_\theta \ln\left(\mathbb{P}(S_k|S_{1:k-1}, \theta)\right)\right\} \\
&= \mathbb{P}(\theta|S_{1:k-1}) \cdot \exp\left\{\delta_k\right\} \\
\implies R_k(\theta) &\leq \exp\{\delta_k\}
\end{aligned}$$

A symmetric argument is made for the other bound.

Remark 2.6 - *Adding Statistics*
Consider a large set of statistics $T := \{T_1, \dots, T_n\}$. To find a sufficient subset T' of T the following process can be used

i). Define $T' = \emptyset$.

ii). Calculate the score of each

iii). Let $\delta_{\max} = \max_{t \in T} \text{Score}(t, T')$.

iv). Let $T_{\max} = \text{argmax}_{t \in T} \text{Score}(t, T')$.

v). If $(\delta_{\max} > \epsilon)$:

- $T' = T' \cup T_{\max}$.

vi). Repeat ii)-v) until no statistics have a score greater than ϵ.

This approach is deterministic wrt which statistics end up in the final T'. Another, stochastic, approach is to uniformly at random at one of the statistics with a score greater than ϵ to T' each iteration.

Proposition 2.3 - *Algorithm for Choosing Summary Statistics*
Here is an algorithm for finding a set of approximately-sufficient summary statistics S' from a larger set of summary statistics S.

- *Initialisation* - Define a set of summary statistics $S := \{s_1(\cdot), \dots, s_n(\cdot)\}$ and observe y_{obs} from the true model.
- Calculate observed values for each summary statistic

$$S_{obs} := \{s_1(y_{obs}), \dots, s_n(y_{obs})\}$$

3 Non-Linear Projection

4 Do we need Summary Statistics?

4.1 Minimum Distance ABC

References

[1] Wikipedia. Sufficient statistic — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Sufficient\%20statistic&oldid=995000742, 2021. [Online; accessed 03-January-2021].

[2] Paul Joyce and Paul Marjoram. Approximately sufficient statistics and bayesian computation. *Statistical applications in genetics and molecular biology*, 7(1), 2008.