Summary Statistic Selection

Dom Hutchinson

December 5, 2020

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.

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)

Remark 1.1 - *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] which is: fixed size regardless of N; and, significantly smaller than $\mathbb{R}^{N \times M}$.

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

Approximate Sufficiency

Proposition 1.2 - *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

[1] ie We sampled M different properties n times
[2] 3 values per property

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 1.3 - *Approach to AS*
A typical approach to AS is to do the following

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.

Sufficiency

Remark 1.2 - *Distinguishing Models*
If two models have the same sufficient statistics it is impossible to distinguish between them.

Definition 1.2 - *Classical Sufficiency*
Let $X \sim f(\cdot;\theta)$ and $T(X)$ be a summary statistic of X. $T(X)$ is a *sufficient statistic* for θ if

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

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

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

Example 1.1 - *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

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

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 1.2 - *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

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

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). 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.

Minimal Sufficiency

Definition 1.3 - *Minimal Sufficiency*
A sufficient statistic $T(X)$ is *Minimally Sufficient* if it can be represented as a function of any other sufficient statistic $S(X)$.

$$\exists f \text{ st } T(X) = f(S(X))$$

Example 1.3 - *Minimal Sufficient Statistics*
TODO

Fisher-Neyman Factorisation Theorem

Theorem 1.1 - *Fisher-Neyman Factorisation Theorem*
If $X \sim f(\cdot;\theta)$ then $T(X)$ is sufficient for θ iff there exists non-negative functions $g(\cdot;\theta), h(\cdot)$ st

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

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

Proof 1.1 - *Fisher-Neyman Factorisation Theorem*
TODO

Remark 1.3 - *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 1.4 - *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 1.5 - *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.

Bayesian Sufficiency

Definition 1.4 - *Bayesian Sufficiency*[3]
In a Bayesian Setting, a summary statistic $T(X)$ of $X \sim f(\cdot;\theta)$ is sufficient if for (almost) all $x \in X$

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

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

ϵ-Approximate Sufficiency

Remark 1.4 - *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 1.5 - *Motivation*

[3] https://en.wikipedia.org/wiki/Sufficient$_s tatistic\#Bayesian_s ufficiency$

Suppose we have a set of summary statistics $\{T_1, \dots, T_n\}$ for parameter θ, we only want to add a new statistic T_{n+1} to this set if it offers a substantial amount of new information about θ. We can consider the likelihoods of these sets of statistics

$$\mathbb{P}(T_1, \dots, T_n, T_{n+1}|\theta) \quad = \quad \mathbb{P}(T_{n+1}|T_1, \dots, T_n, \theta)\mathbb{P}(T_1, \dots, T_n|\theta)$$

Note that if $T_1, \dots, T_n$ where already sufficient then $\mathbb{P}(T_{n+1}|T_1, \dots, T_n, \theta)$ would be independent of θ (ie not contribute to inference about θ).

ϵ-Approximate Sufficiency defines when the set $\{T_1, \dots, T_n\}$ is to be consider sufficient and no new summary statistics need to added to the set.

Definition 1.5 - *Score δ_{n+1}*[4]
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 1.6 - *ϵ-Approximate Sufficiency* [4]
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

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

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

Remark 1.6 - *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.

[4] Joyce, P., & Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation *Statistical applications in genetics and molecular biology, 7(1).*

Non-Linear Projection

Do we need Summary Statistics?

Minimum Distance ABC