How to choose summary statistics for model selection and model checking.

Sarah Filippi

Imperial College London
Theoretical Systems Biology Group

13/02/2012
Model Selection: Which mountain is it a representation of?

- Uluru
- Kilimanjaro
- Mount Everest
Model selection vs model checking

- Model Selection: Which mountain is it a representation of?
  - Uluru
  - Kilimanjaro
  - Mount Everest

- Model Checking: Is it a representation of the Kilimanjaro?
Summary statistics for parameter inference

- Ideally, we should use a sufficient statistic to summarize the data $x^*$:

$$p(\theta|x^*) = p(\theta|S(x^*))$$

- When using an ABC method to approximate the posterior $p(\theta|x^*)$, the choice of the sufficient statistic is particularly important: a statistic with a small dimension is more efficient.

Information theoretical perspective

The idea of using a summary statistic instead of the whole data is to compress this information into a vector of minimum size. The information content may be measured by the mutual information. If $S$ is a sufficient statistic then

$$I(\Theta, X) = \int \int p(x, \theta) \log \frac{p(x, \theta)}{p(x)p(\theta)} dx d\theta = I(\Theta, S(X))$$
Summary statistics for parameter inference

- Ideally, we should use a sufficient statistic to summarize the data $x^*$:

$$p(\theta|x^*) = p(\theta|S(x^*))$$

- When using an ABC method to approximate the posterior $p(\theta|x^*)$, the choice of the sufficient statistic is particularly important: a statistic with a small dimension is more efficient.

Information theoretical perspective

The idea of using a summary statistic instead of the whole data is to compress this information into a vector of minimum size. The information content may be measured by the mutual information. If $S$ is a sufficient statistic then

$$I(\Theta, X) = \int \int p(x, \theta) \log \frac{p(x, \theta)}{p(x)p(\theta)} dx \, d\theta = I(\Theta, S(X))$$
What is the role of a summary statistic.

Two distinct perspectives:

• a summary statistic to compress a specific data \( x^* \) for the given model; ideally such that \( p(\theta|x^*) = p(\theta|S(x^*)) \)

  Joyce and Marjoram, SAGMB (2008); Nunes and Balding, SAGMB (2010)

• a summary statistic to compress the data for a given model (data-independent); ideally such that \( I(\Theta, X|S(X)) = 0 \)

What is the role of a summary statistic.

Two distinct perspectives:

- a summary statistic to compress a specific data $x^*$ for the given model; ideally such that $p(\theta|x^*) = p(\theta|S(x^*))$
  
  Joyce and Marjoram, SAGMB (2008); Nunes and Balding, SAGMB (2010)

- a summary statistic to compress the data for a given model (data-independent); ideally such that $I(\Theta, X|S(X)) = 0$
  

Link between the two perspectives:

$$I(\Theta, X|S(X)) = E_X \{KL [p(\theta|X); p(\theta|S(X))]\}$$
What is the role of a summary statistic.

Two distinct perspectives:

• a summary statistic to compress a specific data \( x^* \) for the given model; ideally such that \( p(\theta|x^*) = p(\theta|S(x^*)) \)
  Joyce and Marjoram, SAGMB (2008); Nunes and Balding, SAGMB (2010)

• a summary statistic to compress the data for a given model (data-independent); ideally such that \( I(\Theta, X|S(X)) = 0 \)

Link between the two perspectives:

\[
I(\Theta, X|S(X)) = E_X \{KL [p(\theta|X); p(\theta|S(X))]} \]

Our approach

Construct, from a set of candidate summary statistics, a set of minimal cardinality that describes the data \( x^* \) in a compact but lossless form, using the mutual information as a tool.
Selection of summary statistics

- Suppose we have a set of statistics $S = \{S_1, \cdots, S_w\}$
- Aim: determine the subset of $S$ with minimum cardinality which contains all the information provided by $S(x^*)$ about $\Theta$.
- If $S$ contains a sufficient statistic (or the data $x^*$ itself) then the constructed subset is a minimal sufficient statistic.
Selection of summary statistics

• Suppose we have a set of statistics $S = \{S_1, \ldots, S_w\}$

• Aim: determine the subset of $S$ with minimum cardinality which contains all the information provided by $S(x^*)$ about $\Theta$.

• If $S$ contains a sufficient statistic (or the data $x^*$ itself) then the constructed subset is a minimal sufficient statistic.

An impossible algorithm

• for all subsets $\mathcal{T} \subset S$, perform ABC to obtain estimates of $p_\epsilon(\theta|\mathcal{T}(x^*))$

• determine the set

$$Q = \{\mathcal{T} \subset S \text{ such that } KL [p_\epsilon(\theta|S(x^*)); p_\epsilon(\theta|\mathcal{T}(x^*))] = 0\}$$

• the desired subset is argmin$_{\mathcal{T} \in Q} |\mathcal{T}|$
An incremental algorithm

- Start with an informative statistic
  \[ Z \leftarrow \text{argmax}_{1 \leq k \leq w} \log \mathbb{E}_{\Theta} [p_{\epsilon}(\Theta|S_k(x^*))] \]

- Add step by step statistics which contains new information compared to the already selected statistics:
  add to \( Z \) \[ \text{argmax}_U KL [p_{\epsilon}(\Theta|Z(x^*), U(x^*)); p_{\epsilon}(\Theta|Z(x^*))] \]
An incremental algorithm

• Start with an informative statistic

\[ Z \leftarrow \arg\max_{1 \leq k \leq w} \log E_\Theta [p_\epsilon(\Theta|S_k(x^*))] \]

• Add step by step statistics which contains new information compared to the already selected statistics:

\[
\text{add to } Z \arg\max_{U} KL [p_\epsilon(\Theta|Z(x^*), U(x^*)); p_\epsilon(\Theta|Z(x^*))]
\]

Idea

Given a set of already selected statistics \( Z \), we aim to determine a statistic \( U \) which minimizes

\[
I(\Theta; S(X)|Z(X), U(X)) = I(\Theta; S(X)|Z(X)) - I(\Theta; Z(X), U(X)|Z(X))
\]

\[ \Rightarrow \text{select the statistic } U \text{ that maximises } I(\Theta; Z(X), U(X)|Z(X)). \]
An incremental algorithm

- Start with an informative statistic

\[ Z \leftarrow \text{argmax}_{1 \leq k \leq w} \log E_{\Theta} [p_\epsilon(\Theta|S_k(x^*))] \]

- Add step by step statistics which contains new information compared to the already selected statistics:

  add to \( Z \) \( \text{argmax}_\mathcal{U} \text{KL} [p_\epsilon(\Theta|Z(x^*), \mathcal{U}(x^*)); p_\epsilon(\Theta|Z(x^*))] \)

- **Stop the algorithm** as soon as the newly added statistic does not bring enough information i.e.

\[ \text{KL} [p_\epsilon(\Theta|Z(x^*), \mathcal{U}(x^*)); p_\epsilon(\Theta|Z(x^*))] \leq \delta \]

Barnes et al, Arxiv (2011)
In practise

- Estimation of $KL[p_\epsilon(\Theta|Z(x^*), \mathcal{U}(x^*)); p_\epsilon(\Theta|Z(x^*))]$ from weighted samples $(\theta_i, w_i)_{1 \leq i \leq N_s}$ and $(\theta'_i, w'_i)_{1 \leq i \leq N_s}$ by

$$\sum_{i=1}^{N_s} w_i \log \frac{w_i}{\bar{w}'_i}, \text{ where } \bar{w}'_i = \frac{\sum_{j=1}^{N_s} w'_j K_h(\theta_i; \theta'_j)}{\sum_{i,j=1}^{N_s} w'_j K_h(\theta_i; \theta'_j)}.$$

$K_h(\cdot; \mu)$ is the normal probability density with mean $\mu$ and variance $1/h$.

- $\delta$ reflects how small the estimated KL divergence between two similar probability distribution should be.

- A stochastic version of the algorithm may be used if the set of statistic is large; a test for order dependency is then required.
Summary statistics for model selection

- As pointed out recently, sufficiency within models is not enough to reliably perform model choice in the ABC framework.
  
  Robert et al, PNAS (2011)

- In particular it is not straightforward to determine a set of sufficient statistics for model selection even if sufficient statistics for parameter inference are available for each model.
Summary statistics for model selection

- As pointed out recently, sufficiency within models is not enough to reliably perform model choice in the ABC framework.
  
  Robert et al, PNAS (2011)

- In particular it is not straightforward to determine a set of sufficient statistics for model selection even if sufficient statistics for parameter inference are available for each model.

Information Theory perspective

Consider $q$ models; we require a statistic that is sufficient for the joint space $\{M, \{\Theta_i\}_{1 \leq i \leq q}\}$. For all statistics $S$,

$$I(M, \Theta_1, \ldots, \Theta_q; X|S) = I(M; X|\Theta_1, \ldots, \Theta_q, S) + \sum_{i=1}^{q} I(\Theta_i; X|S)$$

where $S = S(X)$.

Barnes et al, Arxiv (2011)
Summary statistics for model selection

Information Theory perspective

Consider $q$ models; we require a statistic that is sufficient for the joint space \( \{M, \{\Theta_i\}_{1 \leq i \leq q}\} \). For all statistics $S$,

\[
I(M, \Theta_1, \ldots, \Theta_q; X|S) = I(M; X|\Theta_1, \ldots, \Theta_q, S) + \sum_{i=1}^{q} I(\Theta_i; X|S)
\]

Method

- For each model $1 \leq m \leq q$, determine the set of statistics $S^{(m)}$ which minimizes $I(\Theta_i; X|S(X))$
- Add sequentially statistics to $\bigcup_{1 \leq m \leq q} S^{(m)}$ using the previously described algorithm on the joint space.
$y_1, \ldots, y_d \sim \mathcal{N}(\mu, \sigma_1^2)$ and $y_1, \ldots, y_d \sim \mathcal{N}(\mu, \sigma_2^2)$; $\sigma_1^2 \neq \sigma_2^2$

Statistics chosen for parameter inference

Additional statistics chosen for model selection
Examples: Population Genetics

Constant Population Size

[S1] Number of Segregating Sites; [S2] Number of Distinct Haplotypes; [S3] Haplotype Homozygosity; [S4] Average SNP Homozygosity; [S5] Number of occurrences of most common haplotype; [S6] Mean number of pair-wise differences between haplotypes; [S7] Number of Singleton Haplotypes; [S8] Number of Singleton SNPs; [S9] Linkage Disequilibrium.
[S1] Number of Segregating Sites; [S2] Number of Distinct Haplotypes; [S3] Haplotype Homozygosity; [S4] Average SNP Homozygosity; [S5] Number of occurrences of most common haplotype; [S6] Mean number of pair-wise differences between haplotypes; [S7] Number of Singleton Haplotypes; [S8] Number of Singleton SNPs; [S9] Linkage Disequilibrium.
Examples: Population Genetics

Constant Population Size

Exponential Population Growth

Two-Island Model with Migration

[S1] Number of Segregating Sites; [S2] Number of Distinct Haplotypes; [S3] Haplotype Homozygosity; [S4] Average SNP Homozygosity; [S5] Number of occurrences of most common haplotype; [S6] Mean number of pair-wise differences between haplotypes; [S7] Number of Singleton Haplotypes; [S8] Number of Singleton SNPs; [S9] Linkage Disequilibrium.

Summary Statistic Choice

The choice of summary statistics appears to depend subtly on the true data-generating model.
Examples: Random Walks

Classical Random Walk

Persistent Random Walk

Biased Random Walk

[S1] Mean square displacement; [S2] Mean x and y displacement; [S3] Mean square x and y displacement; [S4] Straightness index; [S5] Eigenvalues of gyration tensor.

Liepe et al, Integrative Biology (2012), In press

Choice of summary statistics Sarah Filippi
Model checking

• The appropriateness of a model is generally assessed based on the posterior predictive distribution,

\[ p(x|x^*) = \int f(x|\theta)p(\theta|x^*)d\theta, \]

where \( x^* \) is the observed data and \( x \) a hypothetical data set.

• If the model is reasonable then \( x \) should resemble the data \( x^* \) in some sense.

⇒ Which summary statistic should we use?
Summary statistics for model checking

- It should contain the aspects of the data that we wish to consider when assessing the appropriateness of a potential generative model.
- The information used for model checking should not rest on the same information that is used for parameter inference.
Summary statistics for model checking

- It should contain the aspects of the data that we wish to consider when assessing the appropriateness of a potential generative model.
- The information used for model checking should not rest on the same information that is used for parameter inference.
- Example: data = sample from $\mathcal{N}(0, 2)$; model = $\mathcal{N}(\mu, 1)$
Summary statistics for model checking

Method 1
Do parameter inference with a sufficient statistic and model checking with an ancillary statistic

- We developed an algorithm to select an ancillary statistic $A$ of maximum cardinality out of a pool of possible statistics:

$$A = \arg\max_{T \in Q} |T|$$

where

$$Q = \{ T \subset S \text{ such that } KL [p_\epsilon(\theta|T(x^*)); \pi(\theta)] = 0 \} .$$

- For a sequential algorithm, be aware that two ancillary statistics are not necessarily jointly ancillary.
Data: $y_1, \cdots, y_n \sim \mathcal{L}(0, 1/\sqrt{2})$  
Model: $\mathcal{N}(\mu, 1)$ 
Selected statistics: variance, range, 4th and 6th moments
Example: Population Genetics

Selected statistics for parameter inference

Selected statistics for model checking

[S1] Number of alleles; [S2] Haplotype Homozygosity; [S3] Number of occurrences of most common haplotype; [S4] Number of alleles in frequency one; [S5] A constant statistic always equal to 0.

Impossible to do model checking

There is no ancillary statistic for this model.
Limitation of Method 1: for many models, non-constant ancillary statistics does not exist.

Method 2
Do parameter inference with a sufficient statistic $S$ and model checking with $X|S$

- In Method 2, we only have to generate one and not two sets of statistics as we would have to for Method 1.
- If $S$ is sufficient then the conditional distribution of $X$ given $S$ is always independent of $\theta$. 
Example: Population Genetics

Aim: assess the constant population size model.

\[ y \sim \text{constant population size model} \]

\[ y \sim \text{expon. population growth model} \]
Take home messages

• Fundamental differences between model selection and model checking.

• Summary statistics are data compression tools. Sufficient statistics are lossless data compression.

• For parameter inference and model selection, we have to make sure that the loss of information does not affect our inference.

• For model checking, the summary statistic should contain information that is important to assess the model but not informative about the parameter.
Thanks for listening!

Acknowledgements:

• Chris Barnes  
  (University College London)
• Michael Stumpf
• Thomas Thorne
• Carsten Wiuf  
  (University of Copenhagen)

s.filippi@imperial.ac.uk
Theoretical Systems Biology group at Imperial College London
www.theosysbio.bio.ic.ac.uk