University of Bristol

Summary Statistic Selection for Approximate Bayesian Computation

Can machines do it?

Dom Hutchinson

06/05/2021

Approximate Bayesian Computation

ABC Motivation

Computational method for approximating posteriors for the parameters θ of models X where the likelihood $\mathbb{P}(X|\theta)$ is intractable.

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta)\mathbb{P}(\theta)}{\mathbb{P}(X)} \propto \mathbb{P}(X|\theta)\mathbb{P}(\theta)$$

ABC methods target sampling from the joint distribution $\pi_{ABC}(\theta, s|s_{obs})$.

$$\pi_{ABC}(\theta, s|s_{obs}) := K_\varepsilon(\|s - s_{obs}\|)\mathbb{P}(s|\theta)\pi_0(\theta)$$

Since

$$\begin{aligned}
\pi_{ABC}(\theta|s_{obs}) &= \int \pi_{ABC}(\theta, s|s_{obs})\,ds \\
\implies \quad \lim_{\varepsilon\to 0} \pi_{ABC}(\theta|x_{obs}) &= \lim_{\varepsilon\to 0} \int K_\varepsilon(\|x - x_{obs}\|)\mathbb{P}(x|\theta)\pi_0(\theta)\,dx \\
&= \int \delta_{x_{obs}}(x)\mathbb{P}(x|\theta)\,dx \cdot \pi_0(\theta) \\
&= \mathbb{P}(x_{obs}|\theta)\pi_0(\theta) \propto \mathbb{P}(\theta|x_{obs})
\end{aligned}$$

Approximate Bayesian Computation

General ABC Schema

Require: Observed values x_{obs}; Summary statistics $s(\cdot)$; Priors $\pi_0(\cdot)$; Theorised model $f(X|\cdot)$; Acceptance Kernel $K_\varepsilon(\cdot)$; Distance Measure $\|\cdot\|$.

1. Calculate summary statistic values $s_{obs} = s(x_{obs})$.
2. Until stopping condition reached:
   2.1 Sample a set of parameters $\tilde{\theta}$.
   2.2 Run the theorised model with sampled parameter $\tilde{x} = f_{\tilde{\theta}}(X|\tilde{\theta})$.
   2.3 Calculate summary statistic values $\tilde{s} = s(\tilde{x})$.
   2.4 Accepted parameters $\tilde{\theta}$ with probability $K_\varepsilon(\|\tilde{s} - s_{obs}\|)$.
3. Return all accepted parameter sets $\hat{\Theta}$.

ABC Schema

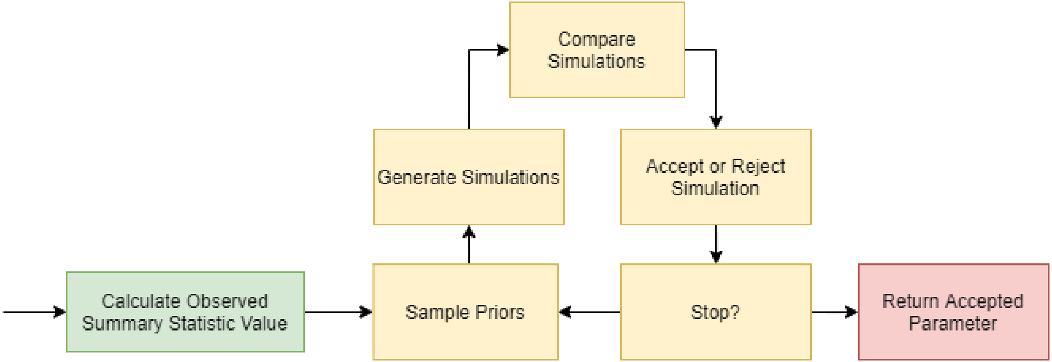


Figure: Flow-diagram for general ABC schema

ABC Algorithm Parameters

Standard ABC methods require the user to specify the following

- Theorised Model.
- Priors
- Summary Statistics.
- Acceptance Kernel.
- Distance Measure.
- Stopping Condition.

There are several common approaches to ABC which vary how parameters are sampled, how simulations are accepted and the stopping condition:

- Rejection Sampling ABC.
- Markov Chain Monte Carlo ABC (ABC-MCMC) [Marjoram et al., 2003].
- Sequential Monte Carlo ABC (ABC-SMC). [Sisson al., 2007]

[Beaumontet al., 2009] discuss the most adaptive version of ABC, a variation of ABC-SMC.

Summary Statistics

Summary Statistics

Large datasets take longer to process and suffer from the curse of dimensionality. For example, the SIR model from *Figure 2* covers only a short period but still generates 90 data-points.

Summary statistics s project data to lower dimensions whilst retain information.

$$\begin{aligned} s &: \mathbb{R}^m \to \mathbb{R}^n && \text{with } m > n \\ s &: \mathbb{R}^{m \times p} \to \mathbb{R}^n && \text{with } m \times p > n \end{aligned}$$

In general, each dimension of the output of a summary statistic is defined independently.

Summary statistics have traditionally been chosen using intuition and by the prevalence of the statistic in the literature.

Summary Statistics - Comments

A good summary statistic has the following properties:

- High levels of information extraction.
- High dimensionality reduction.
- Unbiased.
- Computationally efficient.
- Interpretable.

In general a good fit can be achieved, for simple models, with at most one summary statistic per parameter. If parameters are highly correlated or co-linear then less summary statistics are required.

Sufficient Summary Statistic

Sufficient summary statistics are those which can reduce dimensionality whilst still retain all the information contained in the full data set.

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

e.g. The sample mean is a sufficient statistic for a normal distribution with unknown mean, but known variance.

The identity function is a sufficient statistic for all models, but this is not very helpful for computational problems.

Sufficient Statistics in Practice

Identifying sufficient statistics is very difficult in practice.

Theorem (Fisher-Neyman Factorisation Criterion)

$s(\cdot)$ 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[1] and

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

This formulation shows that the distribution of the model X only depends on the parameter θ through the information extracted by the statistic s. A consequence of the sufficiency of s.

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

Summary Statistic Selection Methods

Joyce-Marjoram - Preliminaries

Since identifying sufficient statistics is difficult, [Joyce and Marjoram, 2008] proposes finding a set of statistics S' which are approximately sufficient to some super-set S.

They define the score metric measures how much more information a set of statistics extracts when it is extended by one statistic.

Score δ_k

The score of s_k relative to the set $s_{1:k-1} := \{s_1, \ldots, s_{k-1}\}$ is defined as

$$\delta_k := \sup_\theta \{\ln \mathbb{P}(s_k | s_{1:k-1})\} - \inf_\theta \{\ln \mathbb{P}(s_k | s_{1:k-1})\}$$

Joyce-Marjoram - Algorithm

Joyce-Marjoram Algorithm

require: Set of summary statistics S; Score threshold ε

$S' \leftarrow \emptyset$

while *true* **do**

    Calculate the score for each statistic in S wrt S'

    $\delta_{max} \leftarrow \max_{s \in S} \mathrm{Score}(s; S')$

    $s_{max} \leftarrow \mathrm{argmax}_{s \in S} \mathrm{Score}(s; S')$

    if $\delta_{max} > \varepsilon$ **then** $S' \leftarrow S' \cup \{s\}$;

    else **return** S';

However, the score metric is intractable. The approach proposed by Joyce & Marjoram compares the posteriors of the proposed sets and switches if the posteriors are notably different. This does not perform well with truly random summary statistics.

Minimising Entropy - Preliminaries

Entropy is a measure of information in a distribution, with lower values indicating more information.

Entropy

The entropy $H(X)$ of a probability distribution X is a measure of the information and uncertainty in distribution.

$$\text{Discrete} \quad H(X) \;:=\; -\sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \cdot \ln \mathbb{P}(X = x)$$

k^{th} Nearest Neighbour Estimator of Entropy

$$\hat{H} = \ln\left(\frac{\pi^{\rho/2}}{\Gamma\left(1 + \frac{\rho}{2}\right)}\right) - \frac{\Gamma'(k)}{\Gamma(k)} + \ln(n) + \frac{\rho}{n}\sum_{i=1}^{n} \ln D_{i,k}$$

where $n = |\Theta|$, ρ is the number of parameters, $D_{i,k}$ is the Euclidean distance between the i^{th}

Minimising Entropy - Algorithm

[Nunes and Balding, 2010] propose an approach to summary statistic selection which choose whichever set of statistics minimises entropy.

Minimising Entropy Summary Statistic Selection (ME)

require: Set of summary statistics S

for $S' \in 2^S$ **do**

  $\Theta \leftarrow$ Parameter sets accepted from ABC-Rejection Sampling using S'

  $\hat{H}_{S'} \leftarrow \hat{H}(\Theta)$

$S^*_{ME} \leftarrow \text{argmin}_{S' \in 2^S} \hat{H}_{S'}$

return S^*_{ME}

There are several ways to estimate entropy $\hat{H}$. [Nunes and Balding, 2010] recommend the k^{th}-Nearest Neighbour Estimator of Entropy with $k = 4$.

Two Step Minimising Entropy - Algorithm

Two Step ME Summary Statistic Selection

require: Observations from true model x_{obs}, Set of summary statistics S, Number of simulations to run n_{run}, Number of simulations to accept n_{obs}

$S_{ME} \leftarrow$ME(S)
$\hat{\Theta}_{ME} \leftarrow$Parameter sets accepted from “Best Samples” ABC-RS($x_{obs}, S_{ME}, n_{run}, n_{acc}$)
Standardise $\hat{\Theta}_{ME}$
for $S' \in 2^S$ **do**
    $\Theta_{acc} \leftarrow$Parameter sets accepted from “Best Samples” ABC-RS($x_{obs}, S', n_{run}, n_{acc}$)
    Standardise Θ_{acc}
    MRSSE$_{S'} \leftarrow$ MRSSE$(\Theta_{acc}, \hat{\Theta}_{ME,i})$
$S^* \leftarrow \text{argmin}_{S' \in 2^S} MRSSE_{S'}$
return S^*

Semi-Automatic ABC

[Fearnhead and Prangle, 2011] propose a method which generates its own summary statistics using linear regression.

Least-Squares Semi-Automatic ABC

```
f_θ ← Posterior from pilot run of an ABC-method using x_obs and S
Θ̂ ← m simulations from f_θ
X_θ̂ ← X(θ̂) for each θ̂ ∈ Θ̂; X̂ ← {X_θ̂1, ..., X_θ̂m}
F ← f(X̂); F̃ ← F with a preceding column of 1s
for i = 1, ..., ρ do
    A_i ← i^th element of each set in Θ̂
    (α^(i), β^(i)) ← (F̃^T F̃^-1) F̃^T A_i
    s_i(x) := β^(i) x
return {s_1, ..., s_ρ}
```

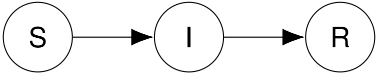
Alternatively, Lasso regression or Canonical correlation analysis can be used. Linear regression is straightforward and has closed form solutions.

SIR Model

SIR Model

The SIR model is a compartmental model which models the movements of individuals in a population between three compartments:

- **S**usceptible.
- **I**nfectious.
- **R**emoved.



A deterministic SIR model with constant population size N can be defined by the following ordinary differential equations.

$$\frac{dS}{dt} = -\frac{\beta}{N}S(t)I(t) \quad \frac{dI}{dt} = \frac{\beta}{N}S(t)I(t) - \gamma I(t) \quad \frac{dR}{dt} = \gamma I(t)$$

β mean infections generated by each infectious individual; γ probability of recovering.

Example Realisation of an SIR Model

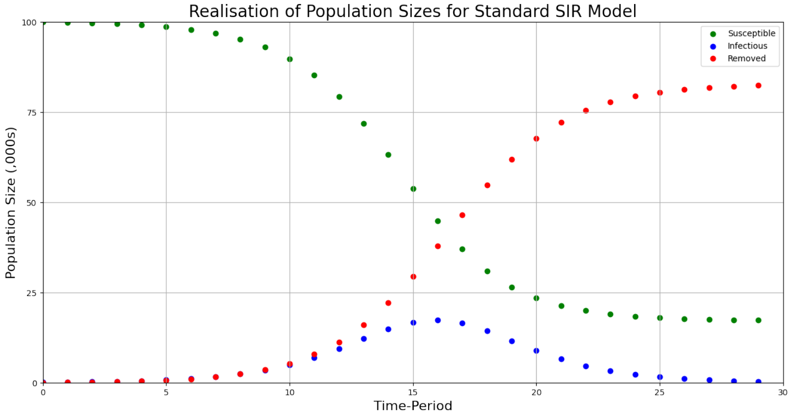


Figure: Realisation of a standard SIR model for a population of size $N = 100,000$ over 30 time-periods where $\beta = 1$ and $\gamma = 0.5$. ($R_0 = 2$)

ABC Methods Fitting an SIR Model

Algorithm	**LOO-CV Score**
Rejection Sampling	184,063
ABC-MCMC	90,713
ABC-SMC	19,300
ABC-SMC with adaptive perturbance & acceptance criteria.	13,160

Table: Leave-One-Out Cross-Validation Scores for different ABC Algorithms fitting to the SIR model in *Figure 2*. All using identity function as the summary statistic, performing $\sim 10,000$ simulations and rough tuning of acceptance rate.

Adaptive ABC-SMC fitting an SIR Model

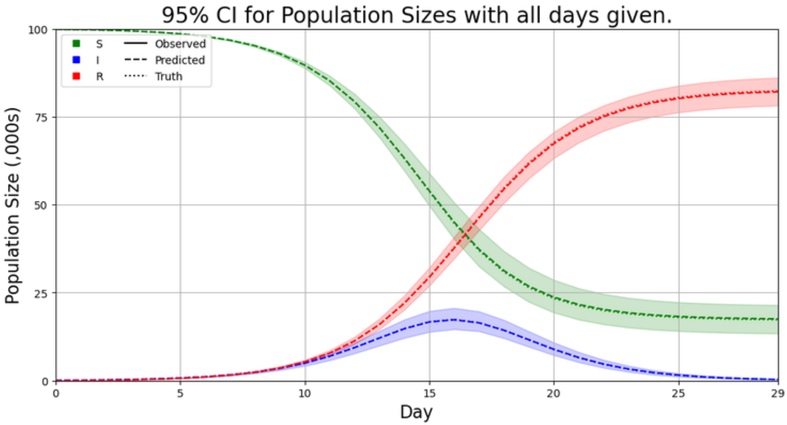


Figure: 95% confidence interval for adaptive ABC-SMC fitting to the SIR model in *Figure 2* using the identity function as the summary statistic. 95% CI for R_0 is [1.871,2.170].

Summary Statistic Methods and SIR Model

Proposed Summary Statistics

These statistics were manipulated to ensure they were on a similar scale, so as to be.

- Peak size of infectious population.
- Date of infectious population.
- Final Size of susceptible population.
- Final Size of infectious population.
- Final Size of removed population.
- Mean Size of susceptible population.
- Mean Size of infectious population.
- Mean Size of removed population.
- Maximum number of infections in a day.
- Maximum number of removals in a day.
- Net Weekly changes in susceptible population ($d = 4$).
- Net Weekly changes in infectious population.
- Net Weekly changes in removed population.
- Populations sizes on days 1,...,30 (as different statistics).
- Uniform random value in [12,20]
- $s(x) = 16$.

The random and constant statistics were never chosen by any algorithm.

Performance

Algorithm	Statistics	ABC-SMC MSE
Control	Identity Function	121,777
Joyce-Marjoram	[Final Susceptible Population]	101,730,336
Minimum Entropy	[Mean Infectious Population, Mean Removed Population]	1,131,712
2-Step ME	[Peak Infectious Population Size, Mean Infectious Population, Mean Removed Population]	228,150
Semi-Automatic ABC	N/A	643,255

Table: Mean Square Error when using Adaptive ABC-SMC with the recommended summary statistics from each algorithm

Visual Comparison

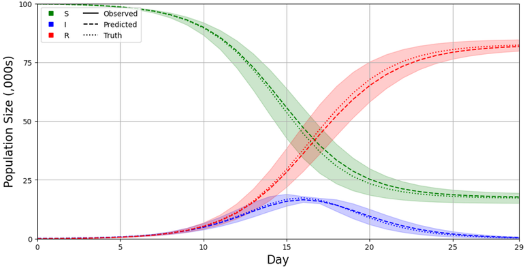


Figure: 95% CI when using summary statistics chosen by Two-Step Minimum Entropy and adaptive ABC-SMC. 95% CI for R_0 is [1.944,2.073].

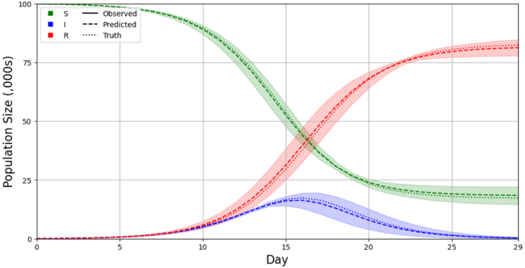


Figure: 95% CI when using summary statistics generated by semi-automatic ABC and adaptive ABC-SMC. 95% CI for R_0 is [1.847,2.127].

Projection

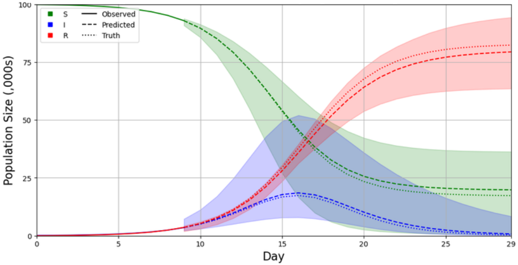


Figure: 95% CI when using summary statistics generated by semi-automatic ABC and adaptive ABC-SMC but with only the first 10 days of data. 95% CI for R_0 is [1.544,3.071].

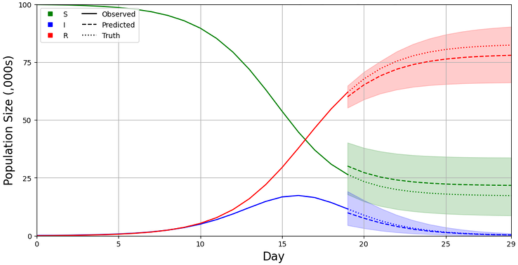


Figure: 95% CI when using summary statistics generated by semi-automatic ABC and adaptive ABC-SMC but with only the first 20 days of data. 95% CI for R_0 is [1.582,2.407].

Summary

There are effective methods which automate the process of choosing summary statistics.