Approximate Bayesian Computation Overview

Davi Sales Barreira

FGV - Escola de Matemática Aplicada, Rio de Janeiro, Brasil
davisbarreira@gmail.com

Abstract. Approximate Bayesian Computation (ABC) methods are known as likelihood-free techniques, thus are a useful approach in problems that the likelihood is intractable, e.g., likelihood not available in closed form, or likelihood too expensive to calculate. In this article, we present an overview of the method by replicating the paper Approximate Bayesian computational methods by Marin et al. (2012).

Keywords: Approximate Bayesian Computation · likelihood-free · Monte Carlo.

1 Introduction

1.1 Original ABC

The Approximate Bayesian Computation method was originally described by Rubin (1984) as a thought experiment to explain how to sample from a posterior distribution with a frequency interpretation. The method became proeminent due to the fact that it circumvents the need to calculate the likelihood function in order to obtain the posterior distribution. This can be a very useful feature in scenarios where the likelihood is intractable or too expensive to calculate. One example is in the case where one has latent variables, thus, the likelihood is expressed as:

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int \ell^*(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{u}) d\boldsymbol{u} \tag{1}$$

with $\boldsymbol{y}$ being the observed variable, $\boldsymbol{u}$ the latent variable and $\boldsymbol{\theta}$ is the parameter of interest.

Tavaré et al. (1997) introduced the ABC algorithm as a rejection techinique to obtain the posterior distribution without the explicit calculation of the likelihood. This original algorithm is given below.

Algorithm 1: Original ABC method

for *i=1 to N* **do**
  repeat
    Sample $\boldsymbol{\theta}' \sim \pi(\cdot)$
    Generate $\boldsymbol{z} \sim p(\cdot \mid \boldsymbol{\theta}')$
  until $\boldsymbol{y} = \boldsymbol{z}$;
end

The proof that the algorithm indeed results in an iid sample from the posterior is shown below. Let $\boldsymbol{y}$ be the observed, $\boldsymbol{\theta}$ the parameter of interest and $\boldsymbol{z}$ the generated samples.

$$f(\boldsymbol{\theta}_i) \propto \sum_{\boldsymbol{z} \in \mathbb{D}} \pi(\boldsymbol{\theta}_i) p(\boldsymbol{z} \mid \boldsymbol{\theta}_i) \mathbb{I}_{\boldsymbol{y}}(\boldsymbol{z}) = \pi(\boldsymbol{\theta}_i) p(\boldsymbol{y} \mid \boldsymbol{\theta}_i) \propto \pi(\boldsymbol{\theta}_i \mid \boldsymbol{y}) \tag{2}$$

The original ABC formulation only works for the case where $\boldsymbol{y}$ is discrete taking finite values, and therefore, an exact match is possible to be obtained in a finite number of simulations. Pritchard et al. (1999) then extended the method to a more general form considering an approximation instead of an exact match. This extended algorithm is shown below, where

- η is a function defining a statistic (e.g. the mean),
- ρ is a distance function,
- ϵ is an acceptance tolerance.

Algorithm 2: ABC method for discrete and continuous distributions

for *i=1 to N* **do**
  repeat
    Sample $\boldsymbol{\theta}' \sim \pi(\cdot)$
    Generate $\boldsymbol{z} \sim p(\cdot \mid \boldsymbol{\theta}')$
  until $\rho[\eta(\boldsymbol{y}), \eta(\boldsymbol{z})] \leq \epsilon$;
end

For this ABC algorithm, instead of the actual posterior, we get

$$\pi_\epsilon(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta}) p(\boldsymbol{z} \mid \boldsymbol{\theta}) \mathbb{I}_{A_{\epsilon,\boldsymbol{y}}}(\boldsymbol{z})}{\int_{A_{\epsilon,\boldsymbol{y}} \times \boldsymbol{\theta}} \pi(\boldsymbol{\theta}) p(\boldsymbol{z} \mid \boldsymbol{\theta}) d\boldsymbol{z} d\boldsymbol{\theta}} \tag{3}$$

where, $A_{\epsilon,\boldsymbol{y}} = \{\boldsymbol{z} \in \mathcal{D} \mid \rho[\eta(\boldsymbol{z}), \eta(\boldsymbol{y}) \leq \epsilon]\}$. Hence, for a tolerance (ϵ) "small enough", we expect a good approximation of the real posterior.

$$\pi_\epsilon(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int \pi_\epsilon(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{y}) d\boldsymbol{z} \approx \pi(\boldsymbol{\theta} \mid \boldsymbol{y}) \tag{4}$$

1.2 Moving Average

We will use the Moving Average model, also denoted as MA(q), for assessing the performance of the ABC methods. The MA(q) process is a stochastic process defined by:

$$y_k = u_k + \sum_{i=1}^{q} \theta_i u_{k-i} \tag{5}$$

where $(u_k)_{k \in \mathbb{Z}} \overset{iid}{\sim} N(0, 1)$. The true posterior distribution of MA(2) and MA(1) models can be numerically computed, since the likelihood function is indeed

avaliable. Therefore, the approximations obtained through ABC can be compared with the true posterior. The marginal posterior distributions are also obtained numerically.

For $q = 2$, imposing the standard identifiability condition we obtain the following conditions:

$$-2 < \theta_1 < 2, \qquad \theta_1 + \theta_2 > -1, \qquad \theta_1 - \theta_2 < 1. \tag{6}$$

hence, we use an uniform distribution over this triangular region as prior for $\boldsymbol{\theta}$.

We generate a synthetic sample of length 100 using $(\theta_1, \theta_2) = (0.6, 0.2)$. For $q = 2$, the true posterior has the following form:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta}), \qquad \boldsymbol{y} \mid \boldsymbol{\theta} \sim MVN(0, \Sigma) \tag{7}$$

$$\Sigma = \begin{bmatrix} 1+\theta_1^2+\theta_2^2 & \theta_1+\theta_2\theta_1 & \theta_2 & 0 & 0 & 0 & \dots & 0 \\ \theta_1+\theta_2\theta_1 & 1+\theta_1^2+\theta_2^2 & \theta_1+\theta_2\theta_1 & \theta_2 & 0 & 0 & \dots & 0 \\ \theta_2 & \theta_1+\theta_2\theta_1 & 1+\theta_1^2+\theta_2^2 & \theta_1+\theta_2\theta_1 & \theta_2 & 0 & \dots & 0 \\ 0 & \theta_2 & \theta_1+\theta_2\theta_1 & 1+\theta_1^2+\theta_2^2 & \theta_1+\theta_2\theta_1 & \theta_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \theta_2 & \theta_1+\theta_1\theta_2 & 1+\theta_1^2+\theta_2^2 \end{bmatrix}$$

For this model, applying the ABC algorithm consisted in the following steps:

- Sample $\boldsymbol{\theta}^*$ from the uniform triangular prior using rejection sampling;
- For each $k \in \{-1, 0, 1, ..., 100\}$, sample $u_k \overset{iid}{\sim} N(0, 1)$.
- For each $k \in \{1, 2, ..., 100\}$, calculate $z_k = u_k + \sum_{i=1}^{2} \theta_i^* u_{k-i}$.

Two distance metrics were initially compared. The raw distance between the series

$$\rho^2\{\boldsymbol{z}, \boldsymbol{y}\} = \sum_{k=1}^{n=100} (y_k - z_k)^2 \tag{8}$$

and the sum of the quadratic distances between the first $q = 2$ autocovariances.

$$\tau_j(\boldsymbol{x}) = \sum_{k=j+1}^{n=100} x_k x_{k-j}, \qquad \rho^2 = \sum_{j=0}^{q=2} (\tau_j(\boldsymbol{y}) - \tau_j(\boldsymbol{z}))^2 \tag{9}$$

Below we present the results of running ABC for the MA(2) process using the autocovariances distance.

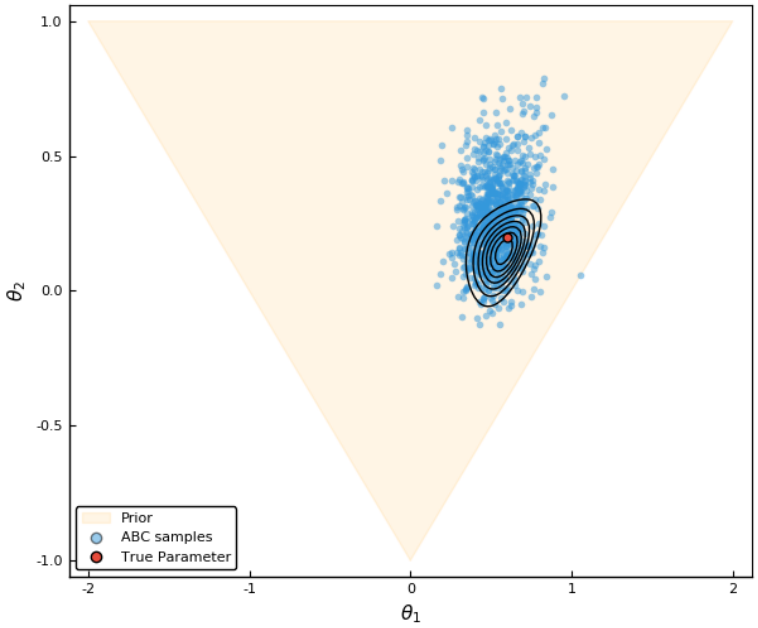

Fig. 1: Comparison between the true posterior (*line in black*), with the samples produced using the ABC . The number of simulations is $N = 10^6$, and the threshold ϵ corresponds to the quantile of accepting 0.1%. The ρ used was the distance of the autocovariances.

2 ABC Calibration

2.1 Summary Statistics (η)

As the number of observations grow, using the raw distance between each observation becomes too prohibitive, due to the rarity of actually obtaining samples close to each observation. The alternative is to try using summary statistics of low dimension. The ideal case is using sufficient statistics, which guarantee that the method indeed approximates the true posterior. The problem is that low-dimensional sufficient statistics are rarely available. Hence, choosing an appropriate low-dimensional statistic is paramount for obtaining good approximations with ABC (Marin et al., 2012).

Beaumont (2019) separates the approaches to address this problem into two categories: one is optimally choosing subsets of summary statistcs, and the other is projecting a set of summary statistics onto lower dimensional maps.

In the first category, Joyce and Marjoram (2008) introduced the concept of approximate sufficiency. The main idea is that given a set of summary statistics $s \subset S$, an approximately sufficient subset can be found by sequantially including those statistics into the ABC target. The method develops a score written as

$$\delta_k = sup_\theta\{\log f(s_k \mid s_1, ..., s_{k-1}, \theta)\} - inf_\theta\{\log f(s_k \mid s_1, ..., s_{k-1}, \theta)\} \quad (10)$$

and tests whether δ_k is less than a given tolerance. In the case this is true, the statistic is deemed approximately sufficient.

Marin et al. (2012) present some reservations regarding this method. They state that the construction of the statistics is not discussed in the paper by Joyce and Marjoram (2008). Secondly, the order in which the statistics are tested may alter the final subset. And finally, that the corrections proposed do not address the impact of correlation between the summary statistics.

In the second category, Fearnhead and Prangle (2010) propose a way of constructing appropriate summary statistics for ABC in a semiautomatic manner. Their method aims at minimizing the expected posterior loss

$$\mathbb{E}[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \mid \boldsymbol{y}] \implies \hat{\boldsymbol{\theta}} = \mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{y}] \tag{11}$$

hence, the optimal summary statistic is

$$s = \mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{y}] \tag{12}$$

Since $E[\boldsymbol{\theta} \mid \boldsymbol{y}]$ is unknown, it can instead be estimated by performing a linear regression on each component of $\boldsymbol{\theta}$. Therefore, the single optimal summary statistic is written as

$$s_{opt} = \beta^T f(s) \tag{13}$$

where β is the vector of the regression coefficients and $f(s)$ is the vector of summary statistics.

2.2 Tolerance threshold(ϵ)

The choice of ϵ is mostly driven by computational limitations. The lower the value of ϵ, the higher the number of simulations required. The standard practice (Beaumont et al., 2002) is to chose ϵ as a quantile of the simulated distances ρ, e.g., for 10^6 simulations, taking $\epsilon = 0.1\%$ corresponds to accepting 10^3 sampled $\boldsymbol{\theta}$'s. This implies that the choice of ϵ is just a proxy for the number of simulations to be performed.

Bibliography

Beaumont, M., Zhang, W., and Balding, D. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035. ID number: ISI:000180502300043.

Beaumont, M. A. (2019). Approximate bayesian computation. *Annual Review of Statistics and Its Application*, 6(1):379–403.

Fearnhead, P. and Prangle, D. (2010). Constructing summary statistics for approximate bayesian computation: Semi-automatic abc.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and bayesian computation. *Statistical applications in genetics and molecular biology*, 7:Article26.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, 12(4):1151–1172.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518.