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 · computational statistics.

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 $\boldsymbol{y} = \boldsymbol{z}$;
  Set $\boldsymbol{\theta}_i = \boldsymbol{\theta}'$
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 $\rho[\eta(\boldsymbol{y}), \eta(\boldsymbol{z})] \leq \epsilon$;
  Set $\boldsymbol{\theta}_i = \boldsymbol{\theta}'$
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.

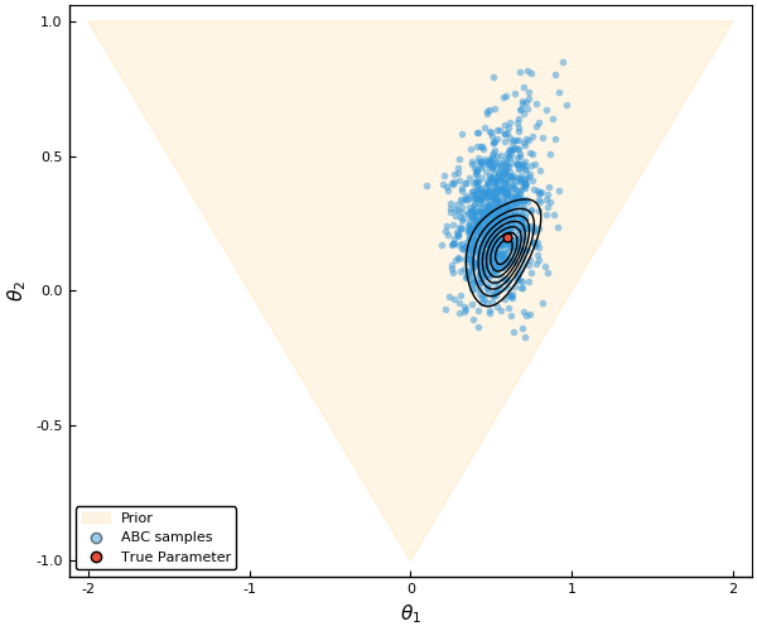

Fig. 1: Comparison between the true posterior (*line in black*), with the samples produced using the ABC . The number of simulations is $N = 10^6$, and the threshold ϵ corresponds to the quantile of accepting 0.1%. The ρ used was the distance of the autocovariances.

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 guarantees 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) criticize 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

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

hence, the optimal summary statistic is

$$s = \mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{y}] \quad (12)$$

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

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

where $\boldsymbol{\beta}$ is the vector of the regression coefficients and $\boldsymbol{f}(s)$ is the vector of summary statistics functions.

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 distance 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.

2.3 Calibration comparison in MA(2)

Using the MA(2) model, we run the ABC algorithm comparing different calibrations. As stated before, two different summary statistics are used, the raw distance and the autocovariances distance. Figure 2 makes it clear that the autocovariances distance perfom better that the raw distance, therefore, through the rest of this article we will only be using the autocovariances distance.

Figure 3 shows the improvement of the ABC approximation with the decrease of the tolerance comparing the marginal distributions of each parameter.

Regarding θ_1, the method seems to be converging to the real distribution, but for θ_2 the approximation doesn't seem to be improving much.

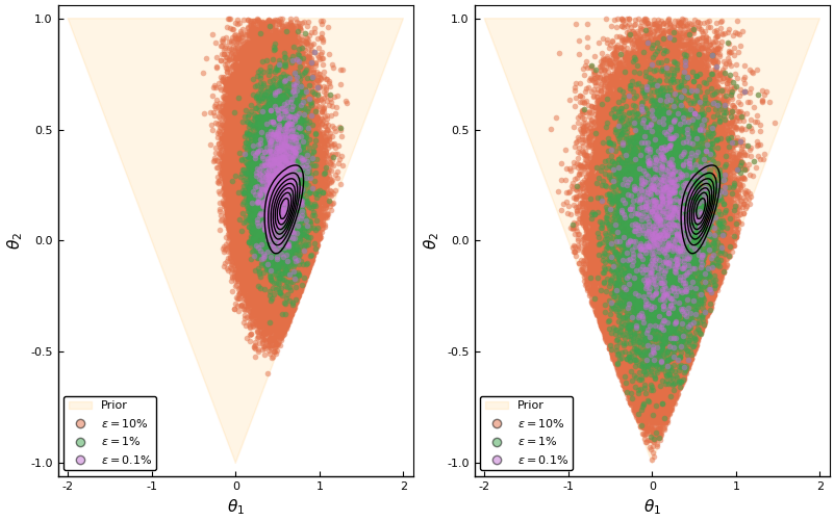

Fig. 2: Comparison of ABC method when using autocovariance distance (*left*) versus raw distance (*right*). The number of simulations is $N = 10^6$ and different thresholds ϵ are used.

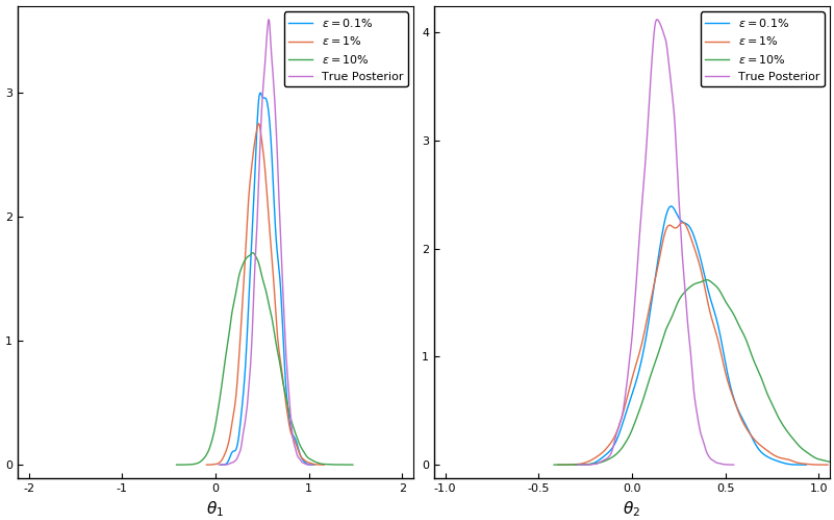

Fig. 3: Comparison of ABC samples with the true posterior marginal distribution for θ_1 (*left*) and θ_2 (*right*).

3 ABC Variations

Over the years, the "vanilla" ABC algorithm has been modified in order to get both better approximations for the posterior and lessen the need of using prohibitively small threshold values, thus decreasing the number of simulations necessary. In this section, some of these variations are presented.

3.1 MCMC-ABC

Using non-informative priors is usually very inefficient, because it leads to lots of rejections. To tackle this problem, Marjoram et al. (2003) came up with MCMC-ABC. The algorithm is presented below.

Algorithm 3: MCMC-ABC

Use Algorithm 2 to get $(\boldsymbol{\theta}^{(0)}, \boldsymbol{z}^{(0)})$ from the target $\pi_\epsilon(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{y})$.
for *i=1 to N* **do**
  repeat
    Sample $\boldsymbol{\theta'}$ from the Markov kernel $q(\cdot \mid \theta^{(i-1)})$
    Generate $\boldsymbol{z} \sim p(\cdot \mid \boldsymbol{\theta'})$
    Sample $u \sim U[0,1]$
    if $u \leq \frac{\pi(\boldsymbol{\theta'})q(\boldsymbol{\theta}^{(i-1)}|)}{\pi(\boldsymbol{\theta}^{(i-1)})q(\boldsymbol{\theta}^{(i-1)}|)}$ *and* $\rho\{\eta(\boldsymbol{z'}), \eta(\boldsymbol{y})\} \leq \epsilon$ **then**
      Set $(\boldsymbol{\theta}^{(i)}, \boldsymbol{z}^{(i)}) = (\boldsymbol{\theta'}, \boldsymbol{z'})$
    end
    else
      Set $(\boldsymbol{\theta}^{(i)}, \boldsymbol{z}^{(i)}) = (\boldsymbol{\theta}^{(i-1)}, \boldsymbol{z}^{(i-1)})$
    end
  until $\rho[\eta(\boldsymbol{y}), \eta(\boldsymbol{z})] \leq \epsilon$;
end

As can be seen, the MCMC-ABC maintains the likelihood-free feature of the original ABC method, but it also estimates $\pi_\epsilon(\boldsymbol{\theta} \mid \boldsymbol{y})$ instead of the true posterior distribution. The initialisation of the algorithm actually uses the "vanilla" ABC method, thus the burn-in of the first iterations can be avoided.

The MCMC-ABC method performs a bit better for our MA(2) example, as shown in Figure 4.

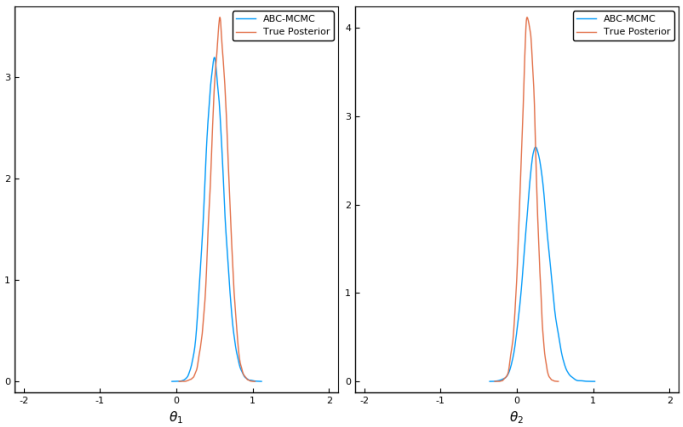

Fig. 4: Comparison of ABC-MCMC samples with the true posterior marginal distribution for θ_1 (*left*) and θ_2 (*right*) using $\epsilon = 0.1\%$.

3.2 Noisy ABC

Another variation of ABC is called Noisy ABC, that was proposed by Wilkinson (2013). The original ABC algorithm can be thought as a rejection algorithm using a uniform kernel ($\mathbb{I}_{A_{\epsilon,\boldsymbol{y}}(\boldsymbol{z})}$). The *Noisy* version generalizes this, allowing the use of different kernels, hence:

$$\pi_\epsilon(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta}) p(\boldsymbol{z} \mid \boldsymbol{\theta}) K_\epsilon(\boldsymbol{y} - \boldsymbol{z})}{\int \pi(\boldsymbol{\theta}) p(\boldsymbol{z} \mid \boldsymbol{\theta}) K_\epsilon(\boldsymbol{y} - \boldsymbol{z}) d\boldsymbol{z} d\boldsymbol{\theta}} \tag{14}$$

Now, instead of accepting if $\rho\{\eta(\boldsymbol{y}), \eta(\boldsymbol{z})\} \leq \epsilon$, we accept with probability $\frac{K_\epsilon(\boldsymbol{y}-\boldsymbol{z})}{max\{K_\epsilon(\boldsymbol{y}-\boldsymbol{z})\}}$.

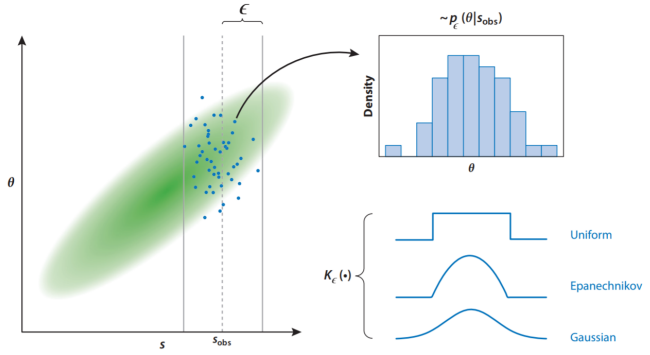

Fig. 5: Illustration of *Noisy* ABC rejection kernels, where s is the statistic from the ABC sampler and s_{obs} is the observed value from the data. Figure from Beaumont (2019).

3.3 Sequantial Techiniques

Sequential techniques are also used with ABC to enhance the efficiency of the algorithms. A popular method in this regard is the ABC-PMC (ABC population Monte Carlo) by Beaumont et al. (2009). It estimates the scale of the random walk step from the previous simulations and uses a sequence of tolerance thresholds ($\epsilon_1 \geq ... \geq \epsilon_T$) to approximate the distribution.

A recent work by Simola et al. (2019) proposed a method for adaptively selecting this sequence of tolerances in a way that improves the computationl efficiency and defines a stopping rule, thus assisting in automating the termination of the sampling procedure. The algorithm is presented below.

Algorithm 4: ABC-PMC

At iteration t=1,
for *i=1 to N* **do**
  repeat
    Sample $\boldsymbol{\theta}_i^{(1)} \sim \pi(\cdot)$
    Generate $\boldsymbol{z} \sim p(\cdot \mid \boldsymbol{\theta}_i^{(1)})$
  until $\rho[\eta(\boldsymbol{y}), \eta(\boldsymbol{z})] \leq \epsilon$;
  Set $w_i^{(1)} = 1/N$.
end
Set Σ_1 as twice the empirical variance of the $\boldsymbol{\theta}_i^{(1)}$'s
for *t=2 to T* **do**
  for *i=1 to N* **do**
    repeat
      Sample $\boldsymbol{\theta}_i^*$ from $\boldsymbol{\theta}_j^{(t-1)}$'s with probabilities $w_j^{(t-1)}$
      Generate $\boldsymbol{\theta}_i^{(t)} \sim N(\boldsymbol{\theta}_i^*, \Sigma_{(t-1)})$ and $\boldsymbol{z} \sim p(\cdot \mid \boldsymbol{\theta}_i^{(t)})$
    until $\rho[\eta(\boldsymbol{y}), \eta(\boldsymbol{z})] \leq \epsilon$;
    Set $w_i^{(t)} \propto \frac{\pi(\boldsymbol{\theta}_i^{(t)})}{\sum_{j=1}^{N} w_j^{(t-1)} \phi\{(\Sigma_{t-1})^{-1/2}(\boldsymbol{\theta}_i^{(t)} - \boldsymbol{\theta}_j^{(t-1)})\}}$.
  end
  Set Σ_t as twice the weighted variance of the $\boldsymbol{\theta}_i^{(t)}$'s
end

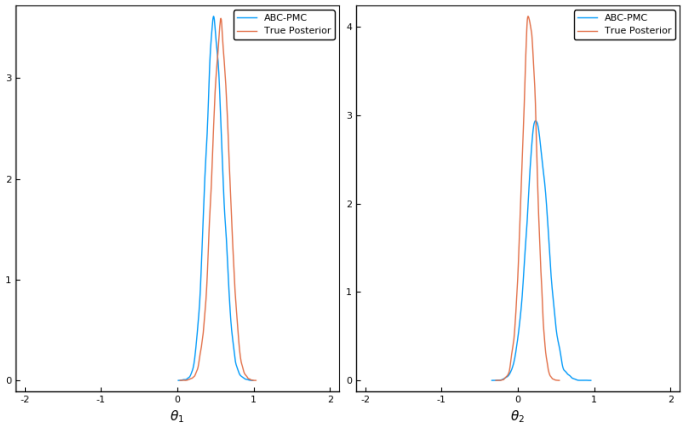

Fig. 6: Comparison of ABC-PMC samples with the true posterior marginal distribution for θ_1 (*left*) and θ_2 (*right*) using $\epsilon = 0.1\%$.

4 Post-processing of ABC

Instead of altering the ABC algorithm, methods have been developed that improve the estimated posterior by post-processing results of the ABC sampler. One of this methods is the *local linear regression* proposed by Beaumont et al. (2002). The idea is to use a weighted least squares regression of $\boldsymbol{\theta}$ on $(\eta(\boldsymbol{y}) - \eta(\boldsymbol{z}))$, with weights according to a chosen kernel.

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} - (\eta(\boldsymbol{y}) - \eta(\boldsymbol{z}))^T \hat{\beta} \tag{15}$$

Hence, this method adjusts the values of the sampled $\boldsymbol{\theta}$'s by projecting them in the axis where the error is equal to zero. Thus, the threshold of acceptance can be lessened without harming the posterior approximation.

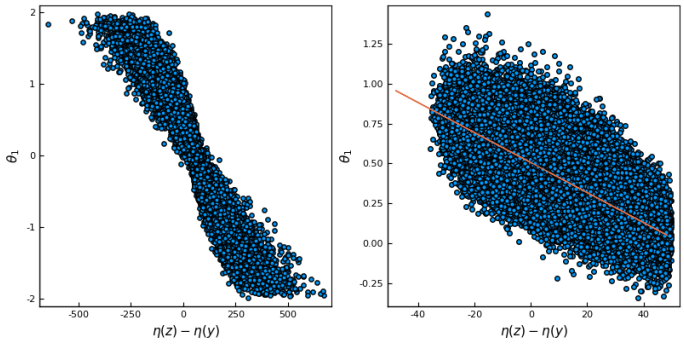

Fig. 7: Scatter plots of simulated θ_1 and $(\eta(\boldsymbol{y}) - \eta(\boldsymbol{z}))$ for autocovariance with $lag = 1$. On the *left* there are all the $N = 10^6$ simulations, while on the *right* only the accepted samples for $\epsilon = 10\%$ with the regression line.

Note that the linear regression is not done for every simulated value, but only those that have $\eta(\boldsymbol{y}) - \eta(\boldsymbol{z}) \leq \epsilon$. As shown in Figure 7, after a certain error, the scatter plot starts to present a non-linear behavior, which is why the correction by Beaumont et al. (2002) is only local. The results of this correction applied to the MA(2) model are shown below, and they are indeed as good, if not better, then when we used $\epsilon = 0.1\%$ in the "vanilla" algorithm.

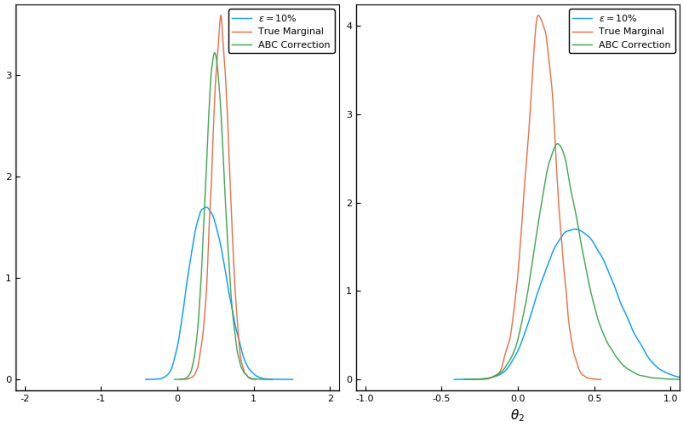

Fig. 8: Comparison of ABC samples corrected through local linear regression versus the true marginal posterior distribution for θ_1 (*left*) and θ_2 (*right*) using $\epsilon = 10\%$.

Blum and François (2010) proposed a nonlinear model with heteroskedasticity, instead of the linear regression. In this approach, the parameters are estimated by a neural network with one hidden layer. This model reduces event further the necessity of using low thresholds, thus accepting more simulated values.

5 Model Choice

Model choice is part of Bayesian analysis in which different models are compared and the end goal is to evaluate the probability a model generated the data compared to others. In addition to the parameters of each model, the inference also estimates a parameter $\mathcal{M}$ which corresponds to each specific model, hence, $\mathcal{M}$ takes values in $\{1, 2, ..., m\}$, where m is the total number of models being compared.

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.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). Adaptive approximate bayesian computation. *Biometrika*, 96(4):983–990.

Blum, M. G. B. and François, O. (2010). Non-linear regression models for approximate bayesian computation. *Statistics and Computing*, 20(1):63–73.

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.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

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.

Simola, U., Cisewski-Kehe, J., Gutmann, M. U., and Corander, J. (2019). Adaptive approximate bayesian computation tolerance selection.

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

Wilkinson, R. D. (2013). Approximate bayesian computation (abc) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, 12(2).