Approximate Bayesian Computation for Model Selection

Yuxi Jiang Natalia Garcia Martin Lorenzo Pacchiardi

October 18, 2019

Abstract

In this work we review generalization of the Approximate Bayesian Computation to the problem of model selection. In particular, the effect of the use of summary statistics on the resulting estimate of the Bayes factor is explained, and some techniques and insights on choosing in order to get a coherent approximation are presented. Also, we describe an improvement of the algorithm exploiting Sequential Monte Carlo, as well as an algorithm applying the Random Forests to get a reliable prediction in the ABC setting. Finally, some simulations on both synthetic and real data are performed.

1 Introduction

Approximate Bayesian Computation (ABC) techniques have been developed in the recent decades in order to perform Bayesian inference in problems for which a closed form of the likelihood does not exist or it is too expensive to compute. Specifically, such technique is able to obtain an approximation of the posterior probability for the parameters, provided that it is possible to simulate from both the likelihood and the prior.

The ABC framework was originally developed to solve problems in population genetics [1] but given its flexibility and robustness its use has been extended to areas such as ecology, conservation, molecular evolution and epidemiology [2]. The application of approximate methods in these areas is driven by the increased availability of molecular markers, the increased computer power and the recently developed DNA sequencing technologies among others, which give rise to large amounts of data, often high-dimensional and more complex models often involving intractable likelihood functions. The ABC approach is characterised by two main features: the use of summary statistics in order to represent the maximum amount of information in the simplest possible way [3] and the use of Monte Carlo simulations that avoid the need to use explicit likelihood functions.

2 Approximate Bayesian Computation

Consider a Bayesian model made up of:

- a parametric statistical model $(\mathcal{X}, f(\boldsymbol{x}|\boldsymbol{\theta}))$, where $\mathcal{X}$ is the observation space and $f(\boldsymbol{x}|\boldsymbol{\theta}) = l(\boldsymbol{\theta}|\boldsymbol{x})$ is the model likelihood,
- a prior distribution $\pi(\boldsymbol{\theta})$ on the parameter space $\boldsymbol{\Theta}$.

In Bayesian setting, we are interested in computing the posterior distribution having observed a dataset $\boldsymbol{y}$, i.e. $\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto l(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta})$. Of course, if the likelihood function is not available, direct computation of the posterior is not possible.

In the simple case of a finite or countable set $\mathcal{X}$, one can consider Algorithm 1:

that will produce an outcome $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N)$ that is distributed from the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

However, in realistic cases where $\mathcal{X}$ is a continuous space, the condition $\boldsymbol{y} = \boldsymbol{z}$ is too strict. The condition is therefore relaxed to $\rho\{\boldsymbol{z}, \boldsymbol{y}\} \leq \epsilon$, where ρ is some distance and ϵ is a chosen threshold.

Moreover, comparing the raw datasets can be expensive (if it is high dimensional, for instance); for this reason, some statistics $\eta(\cdot)$ are usually computed, and the distance between them used, in the following way: $\rho\{\eta(\boldsymbol{z}), \eta(\boldsymbol{y})\} \leq \epsilon$.

The resulting algorithm is therefore Algorithm 2:

Algorithm 1 Likelihood free rejection sampler

for $i = 1$ to N **do**
  repeat
    Generate $\boldsymbol{\theta}'$ from the prior distribution $\pi(\cdot)$
    Generate $\boldsymbol{z}$ from the likelihood $f(\cdot|\boldsymbol{\theta}')$
  until $\boldsymbol{z} = \boldsymbol{y}$
  Set $\boldsymbol{\theta}_i = \boldsymbol{\theta}'$
end for

Algorithm 2 ABC algorithm

for $i = 1$ to N **do**
  repeat
    Generate $\boldsymbol{\theta}'$ from the prior distribution $\pi(\cdot)$
    Generate $\boldsymbol{z}$ from the likelihood $f(\cdot|\boldsymbol{\theta}')$
  until $\rho\{\eta(\boldsymbol{z}), \eta(\boldsymbol{y})\} \leq \epsilon$
  Set $\boldsymbol{\theta}_i = \boldsymbol{\theta}'$
end for

The variables $(\boldsymbol{\theta}', \boldsymbol{z})$ corresponding to the accepted draws are distributed according to the joint distribution:

$$\pi_\epsilon(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta}) f(z|\boldsymbol{\theta}) \mathbb{I}_{A_{\epsilon,\boldsymbol{y}}}(z),$$

where $\mathbb{I}_B(\cdot)$ is the indicator function of set B and $A_{\epsilon,\boldsymbol{y}} = \{z \in \mathcal{X} | \rho\{\eta(\boldsymbol{z}), \eta(\boldsymbol{y})\} \leq \epsilon\}$

Using a sufficiently representative statistic η together with a small tolerance ϵ should produce a good approximation to the posterior distribution, i.e:

$$\pi_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}) = \int \pi_\epsilon(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}) dz \approx \pi(\boldsymbol{\theta}|\boldsymbol{y})$$

The choice of the summary statistics has a huge impact on the resulting approximation. In fact, when applying algorithm 2, the corresponding probability distribution is actually $\pi(\boldsymbol{\theta}|\eta(\boldsymbol{y}))$; indeed, in the case where $\eta(\cdot)$ is a sufficient statistic, the equality $\pi(\boldsymbol{\theta}|\eta(\boldsymbol{y})) = \pi(\boldsymbol{\theta}|\boldsymbol{y}))$ holds. In general, it is almost impossible to obtain a finite-dimensional sufficient statistics (an exception being the case of exponential families). Empirically, non-sufficient statistics are almost always used. Some results regarding the choice of statistic and the tolerance parameter have been obtained in the case of point estimation. At the same time, improved ABC algorithms have been extensively studied. As these topics are not the main focus of the present work, we refer the interested reader to the review performed in [4]. We will focus henceforth on applying ABC to the model selection.

In the same way, the choice of the distance and the threshold level ϵ will affect the performance; typically, Euclidean distance is used. Regarding the choice of ϵ, it is clear that a small value will produce a better approximation but will be more computational requiring. A possible approach is the one of selecting ϵ such that a given quantile of the distribution of distances (for instance, the smallest 1%) will be accepted.

3 ABC for model selection

Bayesian model selection makes use of Bayes Factors (BF) for comparing the evidence across a set of M models indexed by $\mathcal{M} = m$ ($m = 1, 2, \ldots, M$). Given two models $\mathcal{M} = 1$ and $\mathcal{M} = 2$ with likelihoods $f_1(\boldsymbol{y}|\boldsymbol{\theta_1})$ and $f_2(\boldsymbol{y}|\boldsymbol{\theta_2})$, the Bayes factor is given by the ratio of marginal likelihoods of the two models, which equals the posterior odds over the prior odds ratio:

$$B_{12} = \frac{\int_{\boldsymbol{\theta}_1} \pi_1(\boldsymbol{\theta_1}) f_1(\boldsymbol{y}|\boldsymbol{\theta_1}) d\boldsymbol{\theta_1}}{\int_{\boldsymbol{\theta}_2} \pi_2(\boldsymbol{\theta_1}) f_2(\boldsymbol{y}|\boldsymbol{\theta_2}) d\boldsymbol{\theta_2}} = \frac{P(\boldsymbol{y}|\mathcal{M}=1)}{P(\boldsymbol{y}|\mathcal{M}=2)} = \frac{P(\mathcal{M}=1|\boldsymbol{y})}{P(\mathcal{M}=2|\boldsymbol{y})} \frac{\pi(\mathcal{M}=2)}{\pi(\mathcal{M}=1)}.$$

This method naturally penalises model complexity, hence guarding against overfitting. However, an explicit version of the likelihood is often not available or computationally costly, suggesting the use of approximate Bayesian computation. The generalization of ABC to the model selection setting is quite straightforward. In particular, we can consider the inference to involve the model index $\mathcal{M}$ as well, and put a prior distribution $\pi(\mathcal{M} = m)$ on

that. The prior distribution on the parameters will also be dependent on the considered model: $\pi_m(\boldsymbol{\theta}_m)$, each of them being defined on a (potentially) different parameter space $\boldsymbol{\Theta}_m$. Within this setting, ABC is able to obtain an approximation to the posterior by following the same idea as before, described in algorithm 3 (note that $\boldsymbol{\eta}(\boldsymbol{z}) = (\eta_1(\boldsymbol{z}), \eta_2(\boldsymbol{z}), \ldots, \eta_M(\boldsymbol{z}))$ is the concatenation of the summary statistics used in all models).

Algorithm 3 ABC-MC

for $i = 1$ to N **do**
  repeat
    Generate m from the prior $\pi(\mathcal{M} = m)$
    Generate $\boldsymbol{\theta}_m$ from the prior $\pi_m(\boldsymbol{\theta}_m)$
    Generate $\boldsymbol{z}$ from the model $f_m(\boldsymbol{z}|\boldsymbol{\theta}_m)$
  until $\rho\{\boldsymbol{\eta}(\boldsymbol{z}), \boldsymbol{\eta}(\boldsymbol{y})\} \leq \epsilon$
  Set $m^{(i)} = m$ and $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}_m$
end for

The posterior probability will therefore be estimated as $\hat{\pi}(\mathcal{M} = m|\boldsymbol{y}) = N^{-1} \sum_{i=1}^{N} \mathbb{I}_{m^{(i)}=m}$. However, as it is pointed out in [5], model selection performed through ABC can be inconsistent with the results that would be obtained in the case where the true posterior were known. To see this, we can write the ABC approximation to the Bayes factor B_{12} above as

$$\widehat{B_{12}}(\boldsymbol{y}) = \frac{\pi(\mathcal{M} = 2) \sum_{i=1}^{N} \mathbb{I}_{m^{(i)}=1}}{\pi(\mathcal{M} = 1) \sum_{i=1}^{N} \mathbb{I}_{m^{(i)}=2}} = \frac{\pi(\mathcal{M} = 2) \sum_{t=1}^{T} \mathbb{I}_{m^t=1} \mathbb{I}_{\rho\{\boldsymbol{\eta}(\boldsymbol{z}^t), \boldsymbol{\eta}(\boldsymbol{y})\} \leq \epsilon}}{\pi(\mathcal{M} = 1) \sum_{t=1}^{T} \mathbb{I}_{m^t=2} \mathbb{I}_{\rho\{\boldsymbol{\eta}(\boldsymbol{z}^t), \boldsymbol{\eta}(\boldsymbol{y})\} \leq \epsilon}},$$

where T is the number of simulations that would be necessary for N acceptances using Algorithm 3. We will study the limit of this expression by letting T go to infinity. For simplification, we assume a uniform prior on the model index $\mathcal{M}$. The limit is given by

$$\begin{aligned} B_{12}^{\epsilon}(\boldsymbol{y}) &= \frac{\mathbb{P}[\mathcal{M} = 1, \rho\{\boldsymbol{\eta}(\boldsymbol{z}), \boldsymbol{\eta}(\boldsymbol{y})\} \leq \epsilon]}{\mathbb{P}[\mathcal{M} = 2, \rho\{\boldsymbol{\eta}(\boldsymbol{z}), \boldsymbol{\eta}(\boldsymbol{y})\} \leq \epsilon]} \\ &= \frac{\int\int \mathbb{I}_{\rho\{\boldsymbol{\eta}(\boldsymbol{z}), \boldsymbol{\eta}(\boldsymbol{y})\} \leq \epsilon} \, \pi_1(\boldsymbol{\theta_1}) f_1(\boldsymbol{z}|\boldsymbol{\theta_1}) \mathrm{d}\boldsymbol{z} \mathrm{d}\boldsymbol{\theta_1}}{\int\int \mathbb{I}_{\rho\{\boldsymbol{\eta}(\boldsymbol{z}), \boldsymbol{\eta}(\boldsymbol{y})\} \leq \epsilon} \, \pi_2(\boldsymbol{\theta_2}) f_2(\boldsymbol{z}|\boldsymbol{\theta_2}) \mathrm{d}\boldsymbol{z} \mathrm{d}\boldsymbol{\theta_2}} \\ &= \frac{\int\int \mathbb{I}_{\rho\{\boldsymbol{\eta}, \boldsymbol{\eta}(\boldsymbol{y})\} \leq \epsilon} \, \pi_1(\boldsymbol{\theta_1}) f_1^{\boldsymbol{\eta}}(\boldsymbol{\eta}|\boldsymbol{\theta_1}) \mathrm{d}\boldsymbol{\eta} \mathrm{d}\boldsymbol{\theta_1}}{\int\int \mathbb{I}_{\rho\{\boldsymbol{\eta}, \boldsymbol{\eta}(\boldsymbol{y})\} \leq \epsilon} \, \pi_2(\boldsymbol{\theta_2}) f_2^{\boldsymbol{\eta}}(\boldsymbol{\eta}|\boldsymbol{\theta_2}) \mathrm{d}\boldsymbol{\eta} \mathrm{d}\boldsymbol{\theta_2}}, \end{aligned}$$

where $f_1^{\boldsymbol{\eta}}(\boldsymbol{\eta}|\boldsymbol{\theta_1})$ and $f_2^{\boldsymbol{\eta}}(\boldsymbol{\eta}|\boldsymbol{\theta_2})$ represent the densities of $\boldsymbol{\eta}(\boldsymbol{z})$ when $\boldsymbol{z} \sim f_1(\boldsymbol{z}|\boldsymbol{\theta_1})$ and $\boldsymbol{z} \sim f_2(\boldsymbol{z}|\boldsymbol{\theta_2})$ respectively. When ϵ tends to zero, this converges to

$$B_{12}^{\boldsymbol{\eta}}(\boldsymbol{y}) = \frac{\int \pi_1(\boldsymbol{\theta_1}) f_1^{\boldsymbol{\eta}}(\boldsymbol{\eta(y)}|\boldsymbol{\theta_1}) \mathrm{d}\boldsymbol{\theta_1}}{\int \pi_2(\boldsymbol{\theta_2}) f_2^{\boldsymbol{\eta}}(\boldsymbol{\eta(y)}|\boldsymbol{\theta_2}) \mathrm{d}\boldsymbol{\theta_2}} = \frac{P(\boldsymbol{\eta}(\boldsymbol{y})|\mathcal{M} = 1)}{P(\boldsymbol{\eta}(\boldsymbol{y})|\mathcal{M} = 2)},$$

which is the Bayes factor based on $\boldsymbol{\eta(y)}$ alone. Therefore, there is a great loss of information in the limiting case as the information contained in $\boldsymbol{\eta(y)}$ will be less than that contained in $\boldsymbol{y}$.

This implies that the Bayes factor estimated with a summary statistic is in general different from the true one. Interestingly, this is also true when the statistic used for each model is sufficient under that model; even in this optimistic case, $\boldsymbol{\eta}(\cdot)$ is not in general sufficient for the joint parameter $(\mathcal{M}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$. To see this, one can apply the factorization theorem: $f_i(\boldsymbol{y}|\boldsymbol{\theta}_i) = g_i(\boldsymbol{y}) f_i^{\boldsymbol{\eta}}(\boldsymbol{\eta}(\boldsymbol{y})|\boldsymbol{\theta}_i)$ and inserting this in the definition of the Bayes factor yields:

$$B_{12} = \frac{g_1(\boldsymbol{y})}{g_2(\boldsymbol{y})} B_{12}^{\boldsymbol{\eta}}.$$

The ratio $g_1(\boldsymbol{y})/g_2(\boldsymbol{y})$ is equal to 1 in very few cases, one of them being the Gibbs Random Fields one, investigated in [6]. As noted heuristically in [4], using a large enough number of statistics still provides an acceptable level of approximation. In the next paragraphs, we will present more precise results.

3.1 Finding a summary statistic sufficient for model choice

As outlined in the previous paragraph, sufficient statistics under each of the separate models are not necessarily sufficient for model selection. More specifically, we say that a statistics is sufficient for model choice if the following equality holds:

$$B_{12}^{\boldsymbol{\eta}}(\boldsymbol{y}) = B_{12}(\boldsymbol{y}) \iff g_1(\boldsymbol{y}) = g_2(\boldsymbol{y}). \tag{1}$$

The authors of [7] describe a method to find a summary statistic $\boldsymbol{\eta}$ that is sufficient in the previous sense, provided that the you have sufficient statistics for the two models separately. Specifically, you can consider a model, which we will denote as $\mathcal{M} = 0$, under which all the considered models are nested. In the simple case of comparison between 2 models, this means the following: the set of parameters of the two nested models will be contained in the set of parameters of the more general one, meaning $\Theta_1, \Theta_2 \subset \Theta_0$. Typically, the nested models will be recovered when some components of the parameter vector $\boldsymbol{\theta}_0$ are set to constant values (see for instance Section 3.1.1). Then, any sufficient statistic for the model $\mathcal{M} = 0$ will be sufficient for performing model choice. To prove this, note that the previous definition of the models imply that $f_i(\boldsymbol{y}|\boldsymbol{\theta}) = f_0(\boldsymbol{y}|\boldsymbol{\theta})$, for $\boldsymbol{\theta} \in \Theta_i$, as well as the fact that $\boldsymbol{\theta} \notin \Theta_i \implies \pi_i(\boldsymbol{\theta}) = 0$. From these, the following follows (for $i = 1, 2, \ldots$):

$$\begin{aligned}
P(\boldsymbol{y}|\mathcal{M} = i) &= \int_{\Theta_i} f_i(\boldsymbol{y}|\boldsymbol{\theta})\pi_i(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_{\Theta_i} f_0(\boldsymbol{y}|\boldsymbol{\theta})\pi_i(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \int_{\Theta_i} f_0(\boldsymbol{y}|\boldsymbol{\eta}(\boldsymbol{y}), \boldsymbol{\theta}) f_0^{\boldsymbol{\eta}}(\boldsymbol{\eta}(\boldsymbol{y})|\boldsymbol{\theta})\pi_i(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= f_0(\boldsymbol{y}|\boldsymbol{\eta}(\boldsymbol{y})) \int_{\Theta_i} f_i^{\boldsymbol{\eta}}(\boldsymbol{\eta}(\boldsymbol{y})|\boldsymbol{\theta})\pi_i(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= f_0(\boldsymbol{y}|\boldsymbol{\eta}(\boldsymbol{y}))P(\boldsymbol{\eta}(\boldsymbol{y})|\mathcal{M} = i).
\end{aligned}$$

The above results therefore implies that:

$$\frac{P(\boldsymbol{\eta}(\boldsymbol{y})|\mathcal{M} = 1)}{P(\boldsymbol{\eta}(\boldsymbol{y})|\mathcal{M} = 2)} = \frac{P(\boldsymbol{y}|\mathcal{M} = 1)}{P(\boldsymbol{y}|\mathcal{M} = 2)},$$

so that the condition in (1) is satisfied. Again, it is important to note that these results are valid in the very specific case in which sufficient statistics for the different models are available.

3.1.1 Sufficient statistics for exponential families

For an exponential family, the likelihood can be written in the following form:

$$f_i(\boldsymbol{y}|\boldsymbol{\theta}_i) \propto \exp(\eta_i(\boldsymbol{y}) \cdot \boldsymbol{\theta}_i + \gamma_i(\boldsymbol{y})), \tag{2}$$

where it is explicit the dependence on the vector of sufficient statistics η_i; γ_i captures instead the intrinsic relation between the model and the data, which is not dependent on the parameters. When this term is different under different models, the collection of sufficient statistics for each model is not sufficient for model selection.

For an exponential family, it is quite straightforward to build an embedding model. In the simple case of $M = 2$, you can define $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \alpha_1, \alpha_2), \alpha_i \in \{0, 1\}$ and the likelihood as:

$$f_0(\boldsymbol{y}|\boldsymbol{\theta}_0) \propto \exp(\eta_1(\boldsymbol{y}) \cdot \boldsymbol{\theta}_1 + \eta_2(\boldsymbol{y}) \cdot \boldsymbol{\theta}_2 + \alpha_1\gamma_1(\boldsymbol{y}) + \alpha_2\gamma_2(\boldsymbol{y})), \tag{3}$$

and it is immediate to show that this reduces to the likelihood for model $\mathcal{M} = 1$ when $\boldsymbol{\theta}_2 = 0,\ \alpha_1 = 1,\ \alpha_2 = 0$, and reduces to the one for model $\mathcal{M} = 2$ when $\boldsymbol{\theta}_1 = 0,\ \alpha_1 = 0,\ \alpha_2 = 1$. It is also clear that the model defined through (3) is an exponential family for which the combined statistic $\boldsymbol{\eta} = (\eta_1, \eta_2, \gamma_1, \gamma_2)$ is sufficient, and our previous analysis implies therefore that $\boldsymbol{\eta}$ is sufficient for comparing the two considered models. Moreover, the Gibbs Random Field example can be seen as a specific case of this, where $\gamma_i = 0$.

Note that the described technique is useful to understand which summary statistics will be sufficient to perform model choice, but they do not imply a different simulation algorithm. Specifically, you still run Algorithm 3, after having determined the summary statistics by applying the analysis described in the present section.

3.2 Relevant statistics in the more general case

The previous analysis applies specifically to the case where sufficient statistics for the models under comparison are available and hold in a non-asymptotic regime, in the ideal case of $\epsilon = 0$, i.e. exact simulation.

A subsequent work [8] consider the more general case, and provides conditions for the Bayes factor computed using ABC to be convergent, i.e. to select the correct model asymptotically. These conditions amount to the

expectation of the summary statistic to differ asymptotically under the two models; the ideal sufficient statistic is therefore an ancillary one with different expectation under the two models (unfortunately, finding ancillary statistic is not easy in general).

The authors of the above paper also outline a methodology to identify a suitable statistic whilst the standard ABC algorithm is run; the method is the following:

- for each model m, run the standard ABC algorithm, and store, along with the parameters values, the value of the statistics η_m;
- among the collection $\boldsymbol{\eta}$, choose those statistics whose empirical average over the different models are significantly different.

Exploiting the chosen statistics to compute the Bayes factor should yield a good estimate of the true one. A more precise formulation as an hypothesis testing problem can be found in [8].

3.3 ABC-SMC (ABC Sequential Monte Carlo for Model Selection)

If the prior distribution used in ABC algorithm discussed above is very different from the posterior distribution, the algorithm would suffer from low acceptance rate. To improve the computational efficiency, Toni *et al.* proposed an ABC method based on sequential Monte Carlo (SMC)[9], that is derived from a sequential importance sampling (SIS) algorithm [10].

Using the idea of SIS, ABC-SMC approximates the posterior distribution sequentially by constructing intermediate distributions $\pi(\boldsymbol{\theta}|d(x^*, y) \leq \epsilon_t)$, $t = 0, \ldots, T-1$ that converges to the posterior distribution $\pi(\boldsymbol{\theta}|d(x^*, y) \leq \epsilon_T)$. Firstly, define a tolerance set $\{\epsilon_0, \ldots, \epsilon_T\}$ which satisfies $\epsilon_0 > \cdots > \epsilon_T \geq 0$; sample N particles from the prior distribution to obtain $\{\theta^{(1)}, \ldots, \theta^{(N)}\}$, call these accepted particles a "population" and calculate the corresponding weights for these particles. Then for each tolerance ϵ_i, successfully sample N particles, where each of the accepted particle in the new population is a perturbed sample from the previous population with its distance of simulated dataset within the tolerance level to the observed dataset. This process is repeated until tolerance ϵ_T, and the final population would be a sample that approximates the posterior distribution. The detailed derivation of ABC-SMC from SIS as well as the ABC-SMC algorithm can both be found in [9].

ABC-SMC can be adapted to a model selection algorithm. Let $m \in \{1, \ldots, M\}$ be the model parameter, $\theta(m) = (\theta(m)^{(1)}, \ldots, \theta(m)^{(k_m)})$, $m = 1, \ldots, M$, be the discrete model-specific parameters, where k_m is the number of parameters in model m. Let K_t be the perturbation kernel, which is chosen to be a random walk in the model selection framework proposed in [9]; the full algorithm is given as Algorithm 4, parameters are estimated for each model as the model selection algorithm is performed. Notice that if $T = 1$ and $M = 1$, then Algorithm 4 reduces to Algorithm 3.

The ABC-SMC algorithm for model selection would result in the greatest number of particles belonging to the model with the highest posterior density; for the models that only have a few particles it is recommended to estimate the parameters of these models independently, as the small number of particles could not provide good estimation to the posterior distribution of these model parameters. The algorithm provides approximations of the marginal posterior distribution of the model parameter $P(m|x)$ which can be used to directly calculate the Bayes factors, and the marginal posterior distributions of parameters $P(\theta_i|x, m)$. In the cases when there is no single model that is clearly the best, Bayesian model averaging (which requires $P(m|x)$) can be used to obtain a better inference than using a single model [11].

3.4 ABC-RF (ABC Model Choice via Random Forests)

ABC suffers from two major difficulties [12]. First, the number of simulations needs to be large in order to ensure reliability, which is often a difficulty for large datasets as is the case of genomics. Second, a calibration process involving the selection of summary statistics quantifying the difference between the observed and the simulated data needs to be performed. Pudlo *et al.* [12] propose the use of random forests (RF) for reliable ABC model choice by rephrasing model choice as a classification problem involving a first stage of prediction of the best fitting model via RF followed by the approximation of the posterior probability of the selected model through a secondary RF that regresses the selection error over the available summary statistics. RF allows for the inclusion of an arbitrary number of summary statistics and does not require a preliminary selection. Hence, we can include a large collection of summary statistics some of which may be potentially irrelevant or poorly informative [13].

Classification and Regression Trees (CART) refer to Decision Tree algorithms that can be used for classification or regression predictive modelling problems and are represented by binary trees where each node root refers to a

Algorithm 4 ABC-SMC algorithm for model selection

Initialize $\epsilon_0, \ldots, \epsilon_T$
for $t = 0$ to T **do**
  for $i = 1$ to N **do**
    repeat
      repeat
        Sample m^* from $\pi(m)$
        if $t = 0$ **then**
          Sample θ^{**} from $\pi(\theta(m^*))$
        else
          Sample θ^* from the previous population $\{\theta(m^*)_{t-1}\}$ with weights $w(m^*)_{t-1}$
          Perturb the particle θ^* to obtain $\theta^{**} \sim K_t(\theta|\theta^*)$
        end if
      until $\pi(\theta^{**}) \neq 0$
      Simulate a candidate dataset $x^* \sim f(x|\theta^{**}, m^*)$
    until $d(x^*, y) < \epsilon_t$
    Set $m_t^{(i)} = m^*$ and add θ^{**} to the population of particles $\{\theta(m^*)_t\}$, and calculate its weight as

$$w_t^{(i)} = \begin{cases} 1, & \text{if } t = 0 \\ \frac{\pi(\theta^{**})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\theta_{t-1}^{(j)}, \theta^{**})}, & \text{if } t > 0 \end{cases}$$

  end for
  Normalize the weights for every model parameter m
end for

Algorithm 5 Stage 1: ABC-RF

1. Generate a reference table made of the set of $(m,\ \boldsymbol{\eta}_m(\boldsymbol{z}))$ from the N simulations using $\pi(\mathcal{M} = m)$, $\pi_m(\boldsymbol{\theta}_m)$ and $f_m(\boldsymbol{z}|\boldsymbol{\theta}_m)$ (see Algorithm 3)
2. Construct N_{tree} randomised CART which predict m using $\boldsymbol{\eta}_m(\boldsymbol{z})$
   for $b = 1$ to N_{tree} **do**
     draw a bootsrap (sub-)sample of size N_{boot} from the reference table
     grow a randomised CART T_b
     for $n = 1$ to N_{nodes} **do**
       Select n of the predictors at random
       Determine the best split from among those predictors
     end for
   end for
   Predict new data by aggregating the predictions of the N_{tree} trees
3. Determine the predicted indexes for $\boldsymbol{\eta(y)}$ and the trees $\{T_b;\ b = 1 \ldots N_{\text{tree}}\}$
4. Determine $\hat{m}$ according to a majority vote among the predicted indexes

Algorithm 6 Stage 2: Estimating the posterior probability of the selected model

1. Use the trained RF (Algorithm 5) to predict model by $\hat{m}(\boldsymbol{\eta(z)})$ for each $(m,\ \boldsymbol{\eta}_m(\boldsymbol{z}))$ in the reference table and compute the out-of-bag classifier error $\mathbb{I}(\hat{m}(\boldsymbol{\eta}) \neq m)$
2. Use the reference table to build a RF regression function $\rho(\boldsymbol{\eta})$ regressing the model prediction error $\mathbb{I}(\hat{m}(\boldsymbol{\eta}) \neq m)$ on the summary statistics. $\rho(\boldsymbol{\eta})$ is an estimate of $\mathcal{P}\left[m \neq \hat{m}(\boldsymbol{\eta})|\boldsymbol{\eta}\right]$
3. Apply the RF to the actual observations summarised as $\boldsymbol{\eta}(\boldsymbol{y})$ and return $1 - \rho(\boldsymbol{\eta}(\boldsymbol{y}))$ as the estimate of $\mathcal{P}\left[m = \hat{m}(\boldsymbol{\eta(y)})|\boldsymbol{\eta}(\boldsymbol{y})\right]$

single covariate X_j. The binary rule consists on comparing this covariate to a bound t where the left-hand branch rising from the vertex defines $X_j < t$. This bound is chosen by minimising the entropy or Gini index. In order to predict Y given covariates X we follow a specific path along the tree by applying these binary rules. The outcome of the prediction is given by the value of the final leaf (terminal node) reached at the end of the path. The Random Forest (RF) algorithm was introduced by Breiman in 2001 [14] and consists on bagging (bootstrap aggregating) randomised CART [12].

Bootstrap Aggregation (or Bagging for short), is a simple ensemble method (that is, a technique which combines predictions from multiple algorithms in order to obtain accurate predictions) used to reduce the variance for high variance algorithms like CART, The use of a bootstrap algorithm is highly advantageous in the case of decision trees which are very sensitive to the training data [15]. However, while reducing the variance, bagging alone tends to increase the bias in the model as the existence of very strong predictors will lead to really similar trees. Random forests provide a solution to this issue by considering only a random subset of the features for splitting each node, resulting in decorrelated trees.

The algorithm suggested by Pudlo *et al.* is shown in Algorithm 5 and is implemented in the R package `abcrf`. First of all, some datasets are generated beforehand using $\pi(\mathcal{M} = m)$, $\pi_m(\boldsymbol{\theta}_m)$ and $f_m(\boldsymbol{z}|\boldsymbol{\theta}_m)$, and their model indices together with their summary statistics are stored in a reference table, on which the RF is trained. Once a model $\hat{m}$ is selected, we want to approximate the posterior $\pi(\hat{m}|\boldsymbol{\eta}(\boldsymbol{y}))$ by another RF to evaluate its performance. We do this by creating a Bayesian classifier which selects the model with the largest posterior probability by minimising the prior error rate, that is, the expected misclassification error over the hierarchical prior:

$$\sum_{\mathcal{M}} \pi(\mathcal{M} = m) \int \mathbf{1}\{\hat{m}(\boldsymbol{\eta}(\boldsymbol{z})) \neq m)\} f_m(\boldsymbol{z}|\boldsymbol{\theta}_m)\pi_m(\boldsymbol{\theta}_m)\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{\theta}_{\boldsymbol{m}}.$$

The estimate of the posterior probability $\mathcal{P}\left[m = \hat{m}(\boldsymbol{\eta}(\boldsymbol{y}))|\boldsymbol{\eta}(\boldsymbol{y})\right]$ is computed using Algorithm 6.

4 Experiments and results

4.1 Toy example for exponential family

We reproduced the toy example from [7]. Specifically, we consider two models in which observations are assumed to be independent and identically distributed, according to a Poisson(λ) in the first model, and a Geometric(μ) in the second one. Moreover, we put equal prior probability on the two models, place Exponential(1) prior on λ and Uniform(0,1) prior on μ; in this way, the model evidence can be computed analytically in both cases, as well as the Bayes Factor.

For n observations, the likelihoods are the following, under the two models:

$$f_1(\boldsymbol{y}) \propto \exp\left(\sum_{j=1}^{n} y_j\theta_1 - \sum_{j=1}^{n} \log y_j!\right), \qquad f_2(\boldsymbol{y}) \propto \exp\left(\sum_{j=1}^{n} y_j\theta_2\right),$$

where $\theta_1 = \log\lambda$ and $\theta_2 = \log(1-\mu)$. Note that in our parametrisation of the Geometric likelihood, $\boldsymbol{y}$ represents the number of failures, therefore each $y_j = \{0, 1, \ldots\}$. By the analysis carried out in Section 3.1.1, it is clear that $\boldsymbol{\eta}(\boldsymbol{y}) = (s,t) = \left(\sum_j y_j, \sum_j \log y_j!\right)$ is a sufficient statistic for model selection. Note that s is a sufficient statistic for each model, but not for model selection.

The experiments are performed as in [7]: exploiting rejection sampling, we generated 1000 datasets of size $n = 100$ from a Poisson(0.5) distribution, in a way such that the quantity $\frac{P(\boldsymbol{y}|\mathcal{M}=1)}{P(\boldsymbol{y}|\mathcal{M}=2)+P(\boldsymbol{y}|\mathcal{M}=1)}$ is uniformly distributed in $[0.01, 0.99]$; by doing that, testing is performed in different scenarios. For each of these datasets, Algorithm 3 is run, by first using both statistics and then only s. In practice, the procedure to run the algorithm in an efficient way is the following:

1. Fix T and draw $T/2$ trial datasets according to each of the Bayesian models, and store their statistics. These will constitute the reference table (each of them was denoted by $\boldsymbol{x}$ in the above tractation). We used $T = 30,000$ in our experiments.

2. For each of the Poisson(0.5) datasets (the $\boldsymbol{y}$), use the above reference table to estimate the Bayes Factor. Specifically, the distance of each of these datasets with respect to each of the trial one is computed, and only the ones whose distance is smaller than a given threshold are kept to estimate the BF.

Contrarily to what the original paper [7] states, the use of a fixed threshold did not prove to be reliable. A better approach seemed to be the one of using the part of trial datasets constituting the smallest quantile of the distance distribution to estimate the BF. In this way, the same number of trial datasets are used for each dataset, meaning that the threshold is chosen automatically.

The choice of the quantile is crucial as well. Indeed, the smaller quantile we take, the better the approximation to the true Bayes Factor; but at the same time, taking a very small quantile means considering a very small number of trial simulations, so that the estimate has a large variance. To understand this, we considered a fixed dataset Poisson(0.5) dataset, with $n = 100$, and estimated the BF using 50 different reference tables with size $N = 30,000$, for different choice of quantile. The results are shown in Figure 1. Basing on this experiment, we chose to use 0.25% as quantile in the subsequent analysis, meaning that we retain only 75 trial datasets for estimating the BF.

In Figure 2, the analytical and approximate BF are represented, for both choices of summary statistics. You can see that by using only s the estimate that you get is consistently wrong. Moreover, the BF estimate with both statistics is more precise when the true BF has values close to 1. As noted in [7], this is not a problem, as having very large or very small values implies strong evidence for one of the two models, thus a larger uncertainty does not prevent you to draw conclusions in this regime.

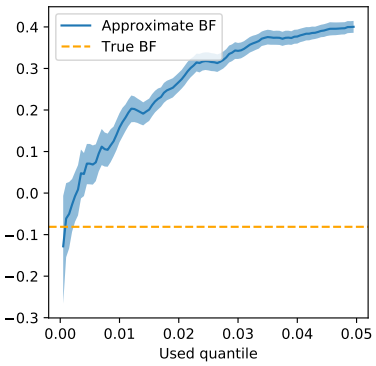

Figure 1: Approximate BF as a function of the used quantile, with 95% confidence bands estimated over 50 different reference tables.

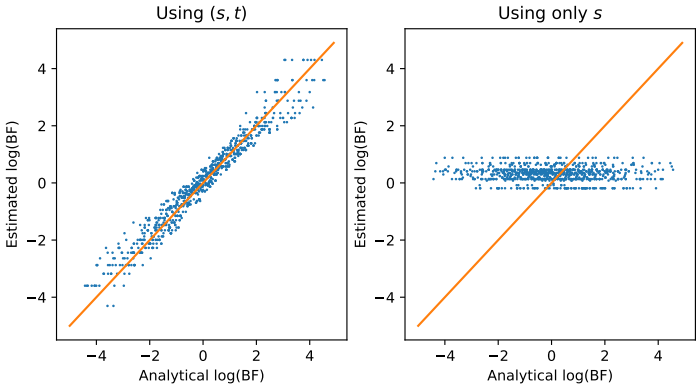

Figure 2: Comparison between analytical and estimated BF using a quantile 0.25%, using both statistics (left) and only s, that is sufficient for each model but not for model selection (right).

4.2 ABC-SMC for SIR model selection

We will reproduce the example in [9] which illustrates the use of the ABS-SMC model selection algorithm. Three different SIR models are compared using a simulated dataset, and the algorithm is implemented using the ABC-SysBio package [16].

SIR models can be used to describe the epidemiology of infectious disease; a simple SIR model separates a population into three types: susceptible (S), infected (I) and recovered (R) individuals. Using the notation $\dot{x} = \frac{dx}{dt}$ to represent the time derivative, let the birth rate of an individual be α, death rate be d, infection rate be γ, and recovery rate be ν. Model 1 assumes that every individual can be infected only once.

$$\dot{S} = \alpha - \gamma SI - dS$$
$$\dot{I} = \gamma SI - \nu I - dI$$
$$\dot{R} = \nu I - dR$$

Model 2 includes an additional latent phase of infection, L, individual in this phase is infected, but does not yet have the ability to infect others. The transition rate between latent state and infected state is assumed to be δ.

$$\dot{S} = \alpha - \gamma SI - dS$$
$$\dot{L} = \gamma SI - \delta L - dL$$
$$\dot{I} = \delta L - \nu I - dI$$
$$\dot{R} = \nu I - dR$$

Model 3 removes the assumption that recovered individual cannot be infected from Model 1 and assumes that the rate a recovered individual becomes susceptible again is e.

$$\dot{S} = \alpha - \gamma SI - dS + eR$$
$$\dot{I} = \gamma SI - \nu I - dI$$
$$\dot{R} = \nu I - dR - eR$$

For the implementation, 1000 particles are used for each population, and Gaussian noise with standard deviation of 0.2 is added to the dataset generated by Model 1. The distributions of the three models at each iteration are shown in Figure 3. After 11 iterations, Model 1 is selected by 989 particles.

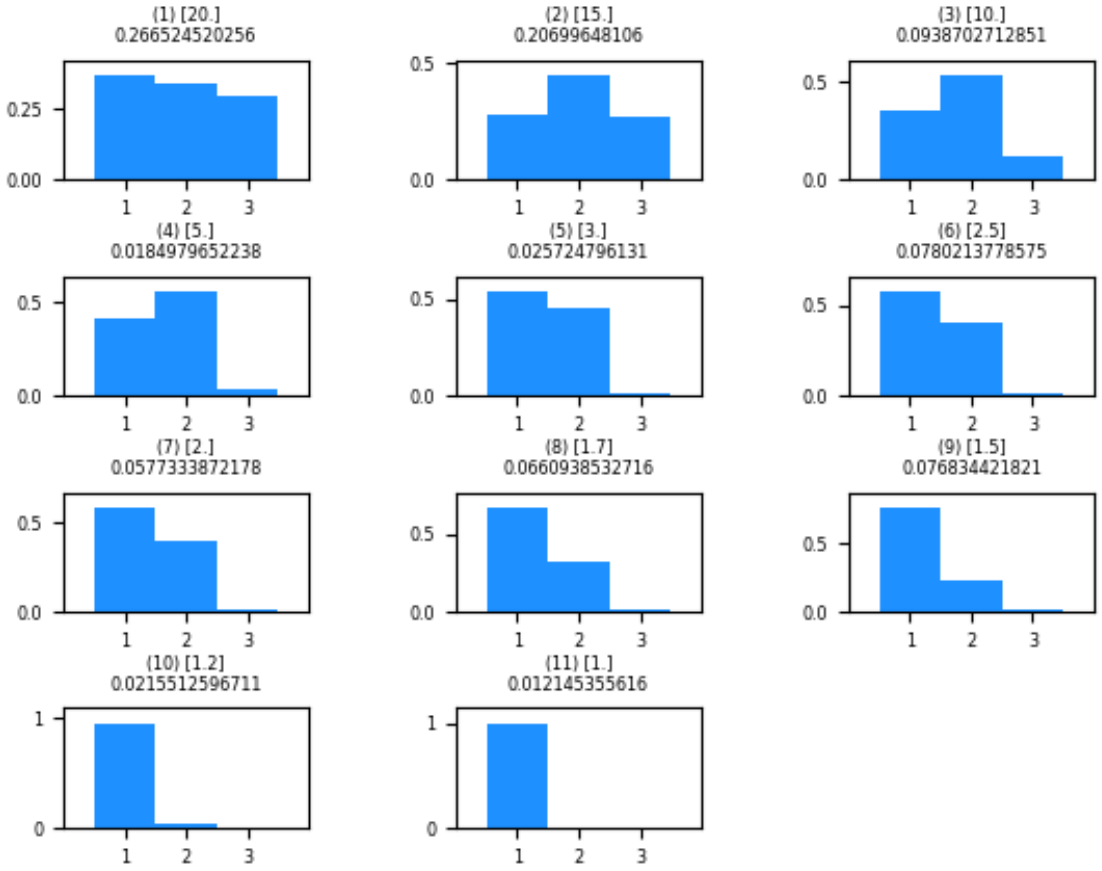

Figure 3: Posterior distributions of the three SIR models with (i) indicating the i^{th} population, $[\epsilon_t]$ indicating the tolerance level and acceptance rate is given below the two.

4.3 ABC-RF for SNP data

We will try to replicate the results of the paper by Pudlo *et al.* [12] where they analyse a SNP dataset containing a reference table of 10,000 simulations on which to perform ABC model choice. A pseudo-observed dataset is available to use as the observed data $\boldsymbol{y}$. The dataset includes information on 1,000 autosomal SNP (single nucleotide polymorphisms) markers. A set of 48 summary statistics has been obtained using the software DIYABC (Do It Yourself ABC) and contains single populations statistics, such as the proportion of loci of a certain type or the variance of gene diversity across loci, as well as two population and three population statistics, such as population distances. Two LDA (linear discriminant analysis) axes are also suggested to be included as additional statistics.

We perform model choice across three models of evolution and calculate the prior error rates to compare the performance of the standard versus the random forest (RF) ABC.

Model 3 is chosen as the most appropriate model by all the algorithms (ABC-MC, ABC-RF with LDA and ABC-RF with no LDA). ABC-RF with no LDA (DIYABC summaries only) gave a (out-of-bag) prior error rate of 20.01% while ABC-RF including LD1 and LD2 gave a prior error rate of 22.30%. For the standard ABC (k-nn) with DIYABC and $k = 5$ we obtained a prior error rate of 29.25%.

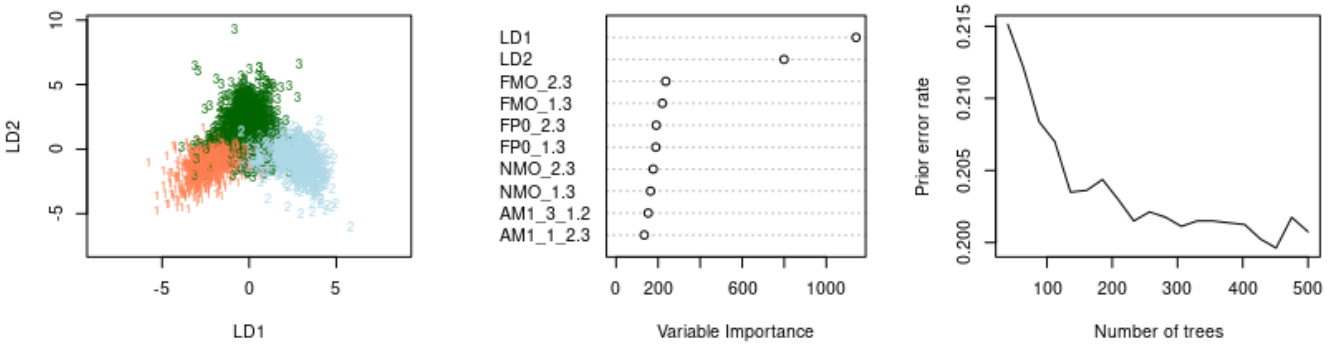

Figure 4: (Left) LDA projection of reference table. (Middle) Top 10 statistics. (Right) Prior error rate versus number of trees for the ABC-RF model.

5 Discussion

We reviewed the framework of ABC for model selection, highlighting the main issues, that are connected to the use of summary statistics and to high computational burden. In particular, we have shown that the use of summary statistics may lead to non-coherent estimates of the Bayes factor, as first pointed out in [5]; few techniques and results have been obtained guiding the choice of summary statistics in order to get coherent estimates, as presented in [7, 8]. An algorithm to improve on the computational aspect of the problem has also been described and tested, together with a random-forest based one, that partly overcomes both problems, by allowing to use a large number of sufficient statistics whose importance in the estimation is automatically selected using decision trees (that are also computationally efficient).

References

[1] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman, "Population growth of human y chromosomes: a study of y chromosome microsatellites.," *Molecular biology and evolution*, vol. 16, no. 12, pp. 1791–1798, 1999.

[2] J. Lopes and M. Beaumont, "ABC: a useful Bayesian tool for the analysis of population data," *Infection, Genetics and Evolution*, vol. 10, no. 6, pp. 825–832, 2010.

[3] K. Csilléry, M. G. Blum, O. E. Gaggiotti, and O. François, "Approximate Bayesian computation (ABC) in practice," *Trends in ecology & evolution*, vol. 25, no. 7, pp. 410–418, 2010.

[4] J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder, "Approximate bayesian computational methods," *Statistics and Computing*, vol. 22, no. 6, pp. 1167–1180, 2012.

[5] C. P. Robert, J.-M. Cornuet, J.-M. Marin, and N. S. Pillai, "Lack of confidence in approximate bayesian computation model choice," *Proceedings of the National Academy of Sciences*, 2011.

[6] A. Grelaud, C. P. Robert, J.-M. Marin, F. Rodolphe, J.-F. Taly, *et al.*, "ABC likelihood-free methods for model choice in gibbs random fields," *Bayesian Analysis*, vol. 4, no. 2, pp. 317–335, 2009.

[7] X. Didelot, R. G. Everitt, A. M. Johansen, D. J. Lawson, *et al.*, "Likelihood-free estimation of model evidence," *Bayesian analysis*, vol. 6, no. 1, pp. 49–76, 2011.

[8] J.-M. Marin, N. S. Pillai, C. P. Robert, and J. Rousseau, "Relevant statistics for bayesian model choice," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 5, pp. 833–859, 2014.

[9] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf, "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," *Journal of the Royal Society Interface*, vol. 6, no. 31, pp. 187–202, 2008.

[10] P. Del Moral, A. Doucet, and A. Jasra, "Sequential monte carlo samplers," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 3, pp. 411–436, 2006.

[11] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: a tutorial," *Statistical science*, pp. 382–401, 1999.

[12] P. Pudlo, J.-M. Marin, A. Estoup, J.-M. Cornuet, M. Gautier, and C. P. Robert, "Reliable ABC model choice via random forests," *Bioinformatics*, vol. 32, no. 6, pp. 859–866, 2015.

[13] L. Raynal, J.-M. Marin, P. Pudlo, M. Ribatet, C. P. Robert, and A. Estoup, "ABC random forests for Bayesian parameter inference," *arXiv preprint arXiv:1605.05537*, 2016.

[14] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] J. Brownlee, "Bagging and Random Forest Ensemble Algorithms for Machine Learning." `https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/`, 2018. [Online; accessed 03-Feb-2019].

[16] J. Liepe, C. Barnes, and E. Clue, "ABC-SysBio: Approximate Bayesian Computation in Python with GPU support," 2011. [Online; accessed 05-Feb-2019].