Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

Genetics and population analysis

ABC random forests for Bayesian parameter inference

Louis Raynal [1,*], Jean-Michel Marin [1,2,*], Pierre Pudlo [3], Mathieu Ribatet [1], Christian P. Robert [4,5] and Arnaud Estoup [2,6]

[1]IMAG, Univ Montpellier, CNRS, Montpellier, France, [2]IBC, Univ Montpellier, CNRS, Montpellier, France, [3]Institut de Mathématiques de Marseille, Aix-Marseille Université, France, [4]Université Paris Dauphine, PSL Research University, Paris, France, [5]Department of Statistics, University of Warwick, Coventry, UK, [6]CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier, Montpellier, France

[*]To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Approximate Bayesian computation (ABC) has grown into a standard methodology that manages Bayesian inference for models associated with intractable likelihood functions. Most ABC implementations require the preliminary selection of a vector of informative statistics summarizing raw data. Furthermore, in almost all existing implementations, the tolerance level that separates acceptance from rejection of simulated parameter values needs to be calibrated.

Results: We propose to conduct likelihood-free Bayesian inferences about parameters with no prior selection of the relevant components of the summary statistics and bypassing the derivation of the associated tolerance level. The approach relies on the random forest methodology of Breiman (2001) applied in a (non parametric) regression setting. We advocate the derivation of a new random forest for each component of the parameter vector of interest. When compared with earlier ABC solutions, this method offers significant gains in terms of robustness to the choice of the summary statistics, does not depend on any type of tolerance level, and is a good trade-off in term of quality of point estimator precision and credible interval estimations for a given computing time. We illustrate the performance of our methodological proposal and compare it with earlier ABC methods on a Normal toy example and a population genetics example dealing with human population evolution.

Availability and implementation: All methods designed here have been incorporated in the R package `abcrf` (version 1.7.1) available on CRAN.

Contacts: louis.raynal@umontpellier.fr, jean-michel.marin@umontpellier.fr

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

As statistical models and data structures get increasingly complex, managing the likelihood function becomes a more and more frequent issue. We now face many realistic fully parametric situations where the likelihood function cannot be computed in a reasonable time or simply is unavailable. As a result, while the corresponding parametric model is well-defined, with unknown parameter θ, standard solutions based on the density function $f(y \mid \theta)$ like Bayesian or maximum likelihood analyses are prohibitive to implement. To bypass this hurdle, the last decades witnessed different inferential strategies, among which composite likelihoods (Lindsay, 1988; Varin et al., 2011), indirect inference (Gourieroux et al., 1993) and likelihood-free methods such as approximate Bayesian computation (ABC, Beaumont et al., 2002; Csilléry et al., 2010; Marin et al., 2012), became popular options. We focus here on improving the latter solution.

Since their introduction in population genetics (Tavaré et al., 1997; Pritchard et al., 1999; Beaumont et al., 2002), ABC methods have been used in an ever increasing range of applications, corresponding to different types of complex models in diverse scientific fields (see, e.g., Beaumont, 2008; Toni et al., 2009; Beaumont, 2010; Csilléry et al., 2010; Theunert et al., 2012; Chan et al., 2014; Arenas et al., 2015; Sisson et al., 2018).

Posterior distributions are the cornerstone of any Bayesian analysis as they constitute both a sufficient summary of the data and a means to deliver all aspects of inference, from point estimators to predictions and uncertainty quantification. However, it is rather common that practitioners and users of Bayesian inference are not directly interested in the posterior distribution *per se*, but rather in some summary aspects, like posterior mean, posterior variance or posterior quantiles, since these are easier to interpret and report. With this motivation, we consider a version of ABC focussing on the approximation of unidimensional transforms of interest like the above, instead of resorting to the classical ABC approach that aims at approximating the entire posterior distribution and then handling it as in regular Bayesian inference. The approach we study here is based on random forests (RF, Breiman, 2001), which produces non-parametric regressions on an arbitrary set of potential regressors. We recall that the calibration side of RF (i.e. the choice of the RF parameters: typically the number of trees and the number of summary statistics sampled at each node) was successfully exploited in Pudlo et al. (2016) for conducting ABC model choice.

After exposing the ABC and RF principles, we explain how to fuse both methodologies towards Bayesian inference about parameters of interest. We then illustrate the performance of our proposal and compare it with earlier ABC methods on a Normal toy example and a population genetics example dealing with human population evolution.

2 Methods

Let $\{f(y\mid\theta)\colon y\in\mathscr{Y},\theta\in\Theta\}$, $\mathscr{Y}\subseteq\mathbb{R}^n$, $\Theta\subseteq\mathbb{R}^p$, $p,n\geq 1$ be a parametric statistical model and $\pi(\theta)$ be a prior distribution on the parameter θ. Given an observation (or sample) y issued from this model, Bayesian parameter inference is based on the posterior distribution $\pi(\theta\mid y)\propto\pi(\theta)f(y|\theta)$. The computational difficulty addressed by ABC techniques is that a numerical evaluation of the density (a.k.a., likelihood) $f(y\mid\theta)$ is impossible or at least very costly, hence preventing the derivation of the posterior $\pi(\theta\mid y)$, even by techniques like MCMC (Marin and Robert, 2014).

2.1 ABC for parameter inference

The principle at the core of ABC is to approximate traditional Bayesian inference from a given dataset by simulations from the prior distribution. The simulated values are accepted or rejected according to the degree of proximity between the observed dataset y and a simulated one $y(\theta)$ thanks to a (usually normalized Euclidean) distance d. ABC relies on the operational assumptions that, while the likelihood is intractable, observations can be generated from the statistical model $f(\cdot\mid\theta)$ under consideration for a given value of the parameter θ.

The ABC resolution of the intractability issue with the likelihood is to produce a so-called *reference table*, recording a large number of datasets simulated from the prior predictive distribution, with density $f(y\mid\theta)\times\pi(\theta)$, and then extracting those that bring the simulations close enough to the actual sample. In most ABC implementations, for both computational and statistical efficiency reasons, the simulated $y^{(t)}$'s $(t=1,\dots,N)$ are summarized through a dimension-reduction function $\eta\colon\mathscr{Y}\to\mathbb{R}^k$, often called a vector of k summary statistics. While the outcome of the ABC algorithm is then an approximation to the posterior distribution of θ given $\eta(y)$, rather than given the entire data y (Marin et al., 2012), arguments are to be found in the literature supporting the (ideal) choice of a summary statistic η of the same dimension as the parameter (Fearnhead and Prangle, 2012; Li and Fearnhead, 2015; Frazier et al., 2017). Algorithm 1 details how the *reference table* is constructed. The *reference table* will latter be used as a training dataset for the different RF methods explained below.

In practice, a *reference table* of size N is simulated, distances $\left(d(\eta(y),\eta(y^{(t)}))\right)_{t=1,\dots,N}$ are computed and then given a tolerance proportion $0<p_\varepsilon\leq 1$, pairs $(\theta^{(t)},\eta(y^{(t)}))$ within the p_ε range of lowest distances are selected. The parameter sample thus derived is deemed to approximate the posterior distribution $\pi(\theta\mid\eta(y))$.

Algorithm 1: Generation of a *reference table* from the prior predictive distribution $\pi(\theta)f(y\mid\theta)$

```
for $t \leftarrow 1$ to $N$ do
    Simulate $\theta^{(t)} \sim \pi(\theta)$;
    Simulate $y^{(t)} = (y_1^{(t)},\dots,y_n^{(t)}) \sim f\left(y \mid \theta^{(t)}\right)$;
    Compute $\eta(y^{(t)}) = \{\eta_1(y^{(t)}),\dots,\eta_k(y^{(t)})\}$;
end
```

The method is asymptotically consistent in the sense that the true parameter behind the data can be exhibited when both the sample size and the number of simulations grow to infinity and the tolerance decreases to zero (Frazier et al., 2017). However, it suffers from two major drawbacks. First, to ensure a sufficient degree of reliability, the number N of simulations must be quite large, even if some new sequential ABC scheme provides interesting improvements in that respect (Prangle, 2017; Klinger and Hasenauer, 2017; Klinger et al., 2018). Hence, it may prove difficult to apply ABC on large or complex datasets since producing data may prove extremely costly. Second, the calibration of the ABC algorithm (i.e. a tolerance level indicating the separation of accepted from rejected simulated parameter values) is a critical step and impacts the resulting approximation (Marin et al., 2012; Blum et al., 2013). Since the justification of the method is doubly asymptotic, it is delicate if not impossible (Li and Fearnhead, 2015; Frazier et al., 2017) to optimally tune ABC for finite sample sizes. A third feature of major importance in this algorithm is that it requires selecting a vector of summary statistics that captures enough information from the observed and simulated data. For most problems, using the raw data to compare datasets is indeed impossible due to their high dimension. Fearnhead and Prangle (2012) give a natural interpretation of the vector of summary statistics as an estimator of θ, but this puts a clear restriction on the dimension and nature of the components of $\eta(y)$.

It is worth noting that the original rejection ABC method, which can be interpreted as a K-nearest neighbour method, has been recurrently improved by linear or non-linear regression strategies, mentioned in literature as adjusted local linear (Beaumont et al., 2002), ridge regression (Blum et al., 2013), and by methods based on adjusted neural networks (Blum and François, 2010). Instead of ridge regularization within local linear adjustment, Saulnier et al. (2017) propose to use a lasso regularization in order to select among the summary statistics. The obtained results are promising excepted when the summary statistics are highly correlated. In such cases, Saulnier et al. (2017) suggest to use random forests. We will only consider ridge regularization in the present work. Finally, additional methods have been developed to exploit the information of already accepted parameter values, by sampling according to a sequential Monte-Carlo based simulation approach (ABC-SMC, Sisson et al., 2007, 2009; Del Moral et al., 2012; Klinger and Hasenauer, 2017; Klinger et al., 2018), or from an importance sampling perspective (ABC-PMC, Beaumont et al., 2009; Prangle, 2017). Such methods make use of a sequence of simulated datasets and include Markov transition kernels as well as importance weights for accepted datasets.

2.2 Random forest methodology

The random forest methodology (RF) of Breiman (2001) is pivotal in our proposal. We use Breiman's RF in a regression setting where a response variable $Y\in\mathbb{R}$ is explained by a vector of covariates $X=(X^{(1)},\dots,X^{(k)})$. A collection of N datasets, made of responses and associated covariates, is used to train a RF.

A given regression RF of size B is composed of B regression trees. A tree is a structure made of binary nodes, which are iteratively built from top to bottom until a stopping rule is satisfied. There are two types of nodes in such trees, the internal and terminal nodes, the latter being also called leaves. At an internal node, a binary rule of the form $X^{(j)} \leq s$ versus $X^{(j)} > s$ compares a covariate $X^{(j)}$ with a bound s. The result of the test divides the predictor space and the training dataset depending on this splitting rule into two parts in two new different nodes. When constructing the tree based on a training sample, the covariate index j and the splitting bound s are determined towards minimising a L^2-loss criterion. The same covariate may be used multiple times for the choice of j at different levels of the tree construction. Splitting events stop when all the observations of the training dataset in a given node have the same covariate values, in which case the node becomes a leaf. Moreover, when a node has less than N_{min} observations, the node also becomes a leaf, typically $N_{\text{min}} = 5$ in the regression framework. Once the tree construction is complete, a value of the response variable is allocated to each tree leaf, corresponding to the average of the response values of the present datasets. For a given and an observed dataset that corresponds to a new covariate X, predicting the associated value of Y implies following the path of the binary rules. The outcome of the prediction is the allocated value of the leaf where this dataset ends after following this path.

The RF method consists in aggregating (or bagging) randomized regression trees. A large number of trees are trained on bootstrap samples of the training dataset and furthermore a subset of n_{try} covariates among the k available covariates are randomly considered at each split. The predicted value of a regression RF is determined by averaging the B predictions over its B tree components.

2.3 ABC parameter inference using random forest

2.3.1 Motivations and main principles

The particular choice of RF as a (non-parametric) estimation method in a regression setting is justified by the robustness of both random forests and quantile methods to "noise", that is, to the presence of irrelevant predictors, even when the proportion of such covariates amongst the entire set of proposed predictors is substantial (Marin et al., 2018). By comparison, the method of K-nearest neighbour classifiers lacks such characteristics (Biau et al., 2015). In the setting of building an ABC algorithm without preliminary selection of some summary statistics, our conjecture is that RF allows for the inclusion of an arbitrary and potentially large number of summary statistics in the derivation of the forest and therefore that it does not require the usual preliminary selection of summary statistics. When implementing this approach, we hence bypass the selection of summary statistics and include a large collection of summary statistics, some or many of which being potentially poorly informative if not irrelevant for the regression. For earlier considerations on the selection of summary statistics, see Joyce and Marjoram (2008); Nunes and Balding (2010); Jung and Marjoram (2011); Fearnhead and Prangle (2012) and the review paper of Blum et al. (2013) where different dimension reduction techniques are compared.

A regression RF produces an expected predicted value for an arbitrary transform of θ, conditional on an observed dataset. This prediction is the output of a piece-wise constant function of the summary statistics. RF aggregates trees, partitions the feature space (here the space of summary statistics) in a way tuned to the forecast of a scalar output, i.e., a one dimensional functional of the parameter. This partition and prediction are done without requiring the definition of a particular distance on the feature space and is hence not dependant of any type of tolerance level. From an ABC perspective, each tree of a RF provides a partition of the covariate space, in our case the k-dimensional space of summary statistics, adapted for the forecasting of the response variable, corresponding to a scalar transformation $h(\theta)$ of the parameter θ. In the following subsection we present how to compute quantities of interest in a context of parameter inference, thanks to the calculation of weights.

2.3.2 Calculation of weights and approximation of the posterior expectation

Assume we have now grown a RF made of B trees that predicts $\tau = h(\theta) \in \mathbb{R}$ using the summarized observed dataset $\eta(y)$ and the training sample $(\eta(y^{(t)}), \tau^{(t)})_{t=1,\ldots,N}$, where $\tau^{(t)} = h(\theta^{(t)})$. In the examples below, we will consider the case where h is the projection on a given coordinate of θ. To sum up, we are training a RF using simulated datasets from the *reference table*, where the covariates are the summary statistics and the response variable is a unidimensional parameter of interest. Each of these B trees produces a partition of the space of summary statistics, with a constant prediction of the expected value of τ on each set of the partition. More precisely, given b-th tree in the forest, let us denote $n_b^{(t)}$ the number of times the pair $(\eta(y^{(t)}), \tau^{(t)})$ is repeated in the bootstrap sample that is used for building the b-th tree. Note that $n_b^{(t)}$ is equal to zero when the pair does not belong to the bootstrap sample. These pairs form the so-called out-of-bag sample of the b-th tree. Now, let $L_b(\eta(y))$ denote the leaf reached after following the path of binary choices given by the tree, which depends on the value of $\eta(y)$. The number of items of the bootstrap sample that fall in that leaf is

$$\left|L_b(\eta(y))\right| = \sum_{t=1}^{N} n_b^{(t)} \mathbf{1}\left\{\eta(y^{(t)}) \in L_b(\eta(y))\right\},$$

where $\mathbf{1}$ denotes the indicator function, and the mean value of τ of that leaf of the b-th tree is

$$\frac{1}{\left|L_b(\eta(y))\right|} \sum_{t=1}^{N} n_b^{(t)} \mathbf{1}\left\{\eta(y^{(t)}) \in L_b(\eta(y))\right\} \tau^{(t)}.$$

Averaging these B predictions of τ leads to an approximation of the posterior expected value of τ, also denoted mean value of τ, which can be written as follows:

$$\widetilde{\mathbb{E}}(\tau \big| \eta(y)) = \frac{1}{B} \sum_{t=1}^{N} \sum_{b=1}^{B} \frac{1}{\left|L_b(\eta(y))\right|} n_b^{(t)} \mathbf{1}\left\{\eta(y^{(t)}) \in L_b(\eta(y))\right\} \tau^{(t)}.$$

As exhibited by Meinshausen (2006), the above can be seen as a weighted average of τ along the whole training sample of size N made by the *reference table*. In fact, the weight of the t-th pair $(\eta(y^{(t)}), \tau^{(t)})$ given $\eta(y)$ is

$$w_t(\eta(y)) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{\left|L_b(\eta(y))\right|} n_b^{(t)} \mathbf{1}\left\{\eta(y^{(t)}) \in L_b(\eta(y))\right\}.$$

2.3.3 Approximation of the posterior quantile and variance

The weights $w_t(\eta(y))$ provide an approximation of the posterior cumulative distribution function (cdf) of τ given $\eta(y)$ as

$$\widetilde{F}(\tau \mid \eta(y)) = \sum_{t=1}^{N} w_t(\eta(y)) \mathbf{1}\{\tau^{(t)} < \tau\}.$$

Posterior quantiles, and hence credible intervals, are then derived by inverting this empirical cdf, that is by plugging $\tilde{F}$ in the regular quantile definition

$$\widetilde{\mathbb{Q}}_\alpha\{\tau \mid \eta(y)\} = \inf\left\{\tau : \widetilde{F}(\tau \mid \eta(y)) \geq \alpha\right\}.$$

This derivation of a quantile approximation is implemented in the R package `quantregForest` and the consistency of $\tilde{F}$ is established in Meinshausen (2006).

An approximation of $\mathrm{Var}(\tau \mid y)$ can be derived in a natural way from $\widetilde{F}$, leading to

$$\widehat{\mathrm{Var}}(\tau \mid \eta(y)) = \sum_{t=1}^{N} w_t(\eta(y)) \left(\tau^{(t)} - \sum_{u=1}^{N} w_u(\eta(y))\tau^{(u)} \right)^2 .$$

2.3.4 Alternative variance approximation

Regarding the specific case of the posterior variance of τ, we propose a slightly more involved albeit manageable version of a variance estimate. Recall that, in any given tree b, some entries from the *reference table* are not included since each tree relies on a bootstrap sample of the training dataset. The out-of-bag simulations, i.e. unused in a bootstrap sample, can be exploited toward returning an approximation of $\mathbb{E}\{\tau \mid \eta(y^{(t)})\}$, denoted $\hat{\tau}^{(t)}_{\mathrm{oob}}$. Indeed, given a vector of summary statistics $\eta(y^{(t)})$ of the training dataset, passing this vector down the ensemble of trees where it has not been used and mean averaging the associated predictions provide such an approximation. Since

$$\mathrm{Var}(\tau \mid \eta(y)) = \mathbb{E}\left([\tau - \mathbb{E}\{\tau \mid \eta(y)\}]^2 \mid \eta(y) \right) ,$$

we advocate applying the original RF weights $w_t(\eta(y))$ to the out-of-bag square residuals $(\tau^{(t)} - \hat{\tau}^{(t)}_{\mathrm{oob}})^2$, which results in the alternative approximation

$$\widetilde{\mathrm{Var}}(\tau \mid \eta(y)) = \sum_{t=1}^{N} w_t\{\eta(y)\}(\tau^{(t)} - \hat{\tau}^{(t)}_{\mathrm{oob}})^2 .$$

Under the same hypotheses as Meinshausen (2006), this estimator converges when $N \to \infty$. Indeed, $\hat{\tau}^{(t)}_{\mathrm{oob}}$ and $\sum_{t=1}^{N} w_t(\eta(y))\tau^{(t)}$ tends to the same posterior expectation. Hence, the two variance estimators above mentioned are equivalent. A comparison between different variance estimators is detailed in the section 2 of Supplementary Information. Owing to the results of this comparative study, we choose to use the above alternative variance estimator when presenting the results from two examples.

As a final remark, it is worth stressing that the approximation of the posterior covariance between a pair of parameters can be achieved thanks to a total of three RFs. The details of that statistical extension are presented in the Section 3 of Supplementary Information.

2.3.5 A new R package for conducting parameter inferences using ABC-RF

When several parameters are jointly of interest, our recommended global strategy consists in constructing one independent RF for each parameter of interest and estimate from each RF several summary measurements of the posterior distribution (i.e. posterior expectation, quantiles and variance) of each parameter. However, if one is interested in estimating the posterior covariance between pair of parameters, an additional RF is required. Our R library `abcrf` was initially developed for Bayesian model choice using ABC-RF as in Pudlo et al. (2016). The version 1.7.1 of `abcrf` includes all the methods proposed in this paper to estimate posterior expectations, quantiles, variances (and covariances) of parameter(s). `abcrf` version 1.7.1 is available on CRAN. We provide in the Section 4 of Supplementary Information, a commented R code that will allow non expert users to run random forest inferences about parameters using the `abcrf` package version 1.7.1.

3 Results

We illustrate the performances of our ABC-RF method for Bayesian parameter inference on a Normal toy example and on a realistic population genetics example. In the first case and only in that case, approximations of posterior quantities can be compared with their exact counterpart. This example is detailed in Section 1 of Supplementary Information. For both examples, we further compare the performances of our methodology with those of earlier ABC methods based on solely rejection, adjusted local linear (Beaumont et al., 2002), ridge regression (Blum et al., 2013), adjusted neural networks (Blum and François, 2010), and adaptive PMC (ABC-PMC, Beaumont et al., 2009; Prangle, 2017). Moreover, we carried out additional comparisons between ABC-RF, adaptive ABC-PMC (Beaumont et al., 2009; Prangle, 2017), ABC-SMC (Del Moral et al., 2012) and adaptive ABC-SMC (Klinger and Hasenauer, 2017) methods for various tuning parameters (see Section 1 of Supplementary Information). Due to excessive computational heaviness and in agreement with the content of the results obtained on the Normal toy example, we did not extended the later comparisons to the population genetics example. Normalized mean absolute errors (NMAE) are used to measure performance on test datasets, the normalization being done by dividing the absolute error by the true value of the target. A normalized version offers the advantage of being hardly impacted when only a few observations get poorly predicted.

For both illustrations, RFs were trained based on the functions of the R package `ranger` (Wright and Ziegler, 2017) with forests made of $B = 500$ trees, with $n_{\mathrm{try}} = k/3$ selected covariates (i.e. summary statistics) for split-point selection at each node, and with a minimum node size equals to 5 (Breiman, 2001, and see Section 3.2, Practical recommendations regarding the implementation of the ABC-RF algorithm). The other ABC methods in the comparison were based on the same *reference tables*, calling the corresponding functions in the R package `abc` (Csilléry et al., 2012, 2015) with its default parameters. ABC with neural network adjustment require the specification of the number of layers composing the neural network: we opted for the default number of layers in the R package `abc`, namely 10. A correction for heteroscedasticity is applied by default when considering regression adjustment approaches. Note that regression corrections are univariate for local linear and ridge regression as well as for RF, whereas neural network - by construction - performs multivariate corrections.

The Normal toy example detailed in Section 1 of Supplementary Information has two parameter of interest θ_1 and θ_2. We observed good overall performances concerning estimation of posterior expectations and quite acceptable for posterior variances (Figure S1). Quantile estimation are good for θ_1 if less accurate for θ_2 (Figure S2). See also Figure S3 for a direct comparison of the true posterior density distribution function of θ_1 in the Normal model with a sample of 40 ABC-RF approximations of the posterior density (using RF weights), based on 40 independent reference tables and for two different test datasets. Table S1 and Figure S4 shows that ABC-RF provides lower NMAE than all other ABC methods. More specifically, we found that the ABC-RF clearly outperforms all adaptive and sequential methods (and designs) considered, and that, in contrast to other methods, ABC-RF was only weakly affected by the presence of a large number of noise variables (see Table S2 and S3 in Section 1 of Supplementary Information).

3.1 Human population genetics example

We illustrate our methodological findings with the study of a population genetics dataset including 50 000 single nucleotide polymorphic (SNP) markers genotyped in four human population samples (The 1000 Genomes Project Consortium, 2012; see details in Pudlo et al., 2016). The four populations include Yoruba (Africa; YRI), Han (East Asia; CHB), British (Europe; GBR) and American individuals of African ancestry (North America; ASW). The considered evolutionary model is represented in Figure 1. It includes a single out-of-Africa event with a secondarily split into one European and one East Asian population lineage and a recent genetic admixture of Afro-Americans with their African ancestors and with Europeans. The model was robustly chosen as most appropriate among a

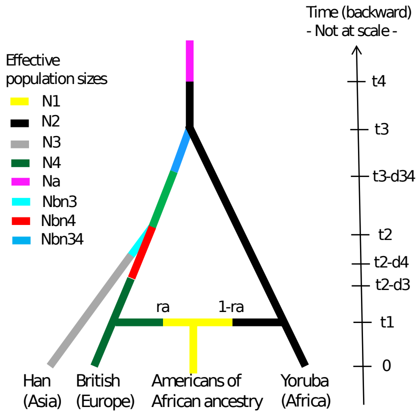

Fig. 1. Evolutionary model of four human populations considered for Bayesian parameter inference using ABC-RF. The prior distributions of the demographic and historical parameters used to simulate SNP datasets are as followed: Uniform[100; 10000] for the split times t2 and t3 (in number of generations), Uniform[1; 30] for the admixture time t1, Uniform[0.05; 0.95] for the admixture rate ra (proportion of genes with a non-African origin), Uniform[1000; 100000] for the stable effective population sizes N1, N2, N3, N4 and N34 (in number of diploid individuals), Uniform[5; 500] for the bottleneck effective population sizes Nbn3, Nbn4, and Nbn34, Uniform[5; 500] for the bottleneck durations d3, d4, and d34, Uniform[100; 10000] for both the ancestral effective population size Na and t4 the time of change to Na. Conditions on time events were t4>t3>t2. See Pudlo et al. (2016) for details. Regarding the genetic model, we simulated biallelic polymorphic SNP datasets using the algorithm proposed by Hudson (2002) (cf "-s 1" option in the program ms associated to Hudson (2002)). This coalescent-based algorithm provides the simulation efficiency and speed necessary in the context of ABC, where large numbers of simulated datasets including numerous (statistically independent) SNP loci have to be generated (see Supplementary Appendix S1 of Cornuet et al. (2014) for additional comments on Hudson's algorithm).

set of eight evolutionary models, when compared using ABC-RF for model choice in Pudlo et al. (2016).

We here focused our investigations on two parameters of interest in this model: (i) the admixture rate ra (i.e. the proportion of genes with a non-African origin) that describes the genetic admixture between individual of British and African ancestry in Afro-Americans individuals; and (ii) the ratio N2/Na between the ancestral effective population size Na and African N2 (in number of diploid individuals), roughly describing the increase of African population size in the past. Considering ratios of effective population sizes allows preventing identifiability issues of the model.

We used the software DIYABC v.2.0 (Cornuet et al., 2008, 2014) to generate a *reference table* of size 200000, with $N = 199000$ datasets being used as training dataset and $N_{pred} = 1000$ remaining as test datasets. RFs are built in the same way as for our Normal example and make use of the $k = 112$ summary statistics provided for SNP markers by DIYABC, (see Pudlo et al., 2016, and the Section 5 of Supplementary Information).

Due to the complexity of this model, the exact calculation of any posterior quantity of interest is infeasible. To bypass this difficulty we compute NMAE using simulated parameters from the test table, rather than targeted posterior expectations ; in this case the normalization is performed by dividing by simulated parameter values. Here, 95% credible intervals (CI) are deduced from posterior quantile estimate of order 2.5% and 97.5%. Performances are measured via mean range and coverage, with coverage corresponding to the percentage of rightly bounded parameters. For example a 95% CI should provide coverage equal to 95% of the test table.

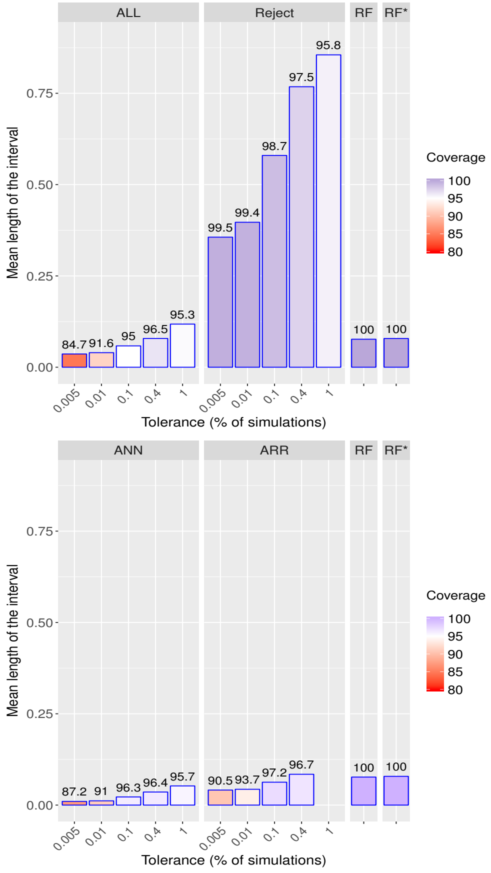

Fig. 2. Range and coverage comparison of approximate 95% credible intervals on the admixture parameter ra (Figure 1) obtained with ABC-RF (RF) and with earlier ABC methods : rejection (Reject), adjusted local linear (ALL) or ridge regression (ARR) or neural network (ANN) with various tolerance levels for Reject, ALL, ARR and ANN. Coverages values are specified by bar colors and superimposed values. Heights indicate CI mean lengths. Results for ALL, Reject and RF are presented in the upper figure whereas those for ANN, ARR and RF are in the lower figure. See Figure S7 for a similar representation of results for the parameter N2/Na. RF* refers to results obtained using ABC-RF when adding 20 additional independent noise variables generated from a uniform $\mathscr{U}_{[0,1]}$ distribution. RF refers to results without noise variables.

Figure 2, Figure S7 and Table S5 illustrate the quality of the ABC-RF method when compared with ABC with either rejection, local linear, ridge or neural network adjustment (with logit transforms of the parameters for non rejection methods) using different tolerance levels (i.e., with tolerance proportion ranging from 0.005 to 1). We recall that considering the ABC rejection method with a tolerance equals to 1 is equivalent to work with the prior. Note that, due to memory allocation issues when using ABC method with adjusted ridge regression and a tolerance level of 1 on large *reference table*, we did not manage to recover results in this specific case.

Interesting methodological features can be observed in association with this example. ABC with rejection performs poorly in terms of NMAE and provides conservative and hence wide CIs (i.e., with coverage higher than the formal level). For ABC with adjustment, the lower the tolerance the lower the error (Table S5). The CI quality however highly suffers from low tolerance, with underestimated coverage (Figures 2 and S7). The smaller the tolerance value, the narrower the CI. Results for the ABC method with adjusted ridge regression seems however to be unstable for the parameter N2/Na depending on the considered level of tolerance. The ABC method using neural network and a tolerance level of 0.005 provides the lowest NMAE for both parameters of interest. The corresponding coverages are however underestimated, equal to 87.2% for ra and 81.6% for N2/Na, when 95% is expected (lower part of Figure 2 and Figure S7). Note that results with this method can be very time consuming to obtain when the tolerance level and the number of layers are large. The ABC-RF method provides an appealing trade-off between parameter estimation quality (ABC-RF is the method with the second lowest NMAE values in Table S5) and slightly conservative CIs (Figures 2 and S7). Similar results and methodological features were observed when focusing on the 90% CI (results not shown). It is also worth stressing that not any calibration of any kind of a tolerance level parameter are needed with ABC-RF, which is an important plus for this method. On the opposite, earlier ABC methods require calibration to optimize their use, such calibration being time consuming when different levels of tolerance are used.

For the observed dataset used in this study, posterior expectations and quantiles of the parameters of interest ra and N2/Na are reported in Tables S6 and S7. Expectation and CI values substantially vary for both parameters, depending on the method used. The impact of the tolerance levels is noteworthy for both the rejection and local linear adjustment ABC methods. The posterior expectation of ra obtained using ABC-RF was equal to 0.221 with a relatively narrow associated 95% CI of $[0.112, 0.287]$. The latter estimation lays well within previous estimates of the mean proportion of genes of European ancestry within African American individuals, which typically ranged from 0.070 to 0.270 − with most estimates around 0.200 −, depending on individual exclusions, the population samples and sets of genetic markers considered, as well as the evolutionary models assumed and inferential methods used (reviewed in Bryc et al., 2015). Interestingly, a recent genomic analysis using a conditional random field parametrized by random forests trained on reference panels (Maples et al., 2013) and 500 000 SNPs provided a similar expectation value of ra for the same African American population ASW (i.e. ra $= 0.213$), with a somewhat smaller 95% CI (i.e. $[0.195, 0.232]$), probably due to the ten times larger number of SNPs in their dataset (Baharian et al., 2016).

The posterior expectation of N2/Na obtained using ABC-RF was equal to 4.508 with a narrow associated 95% CI of $[3.831, 5.424]$. Such values point to the occurrence of the substantial ancestral demographic and geographic expansion that is widely illustrated in previous Human population genetics studies, including African populations (e.g. Henn et al., 2012). Although our modeling setting assumes a naïve abrupt change in effective population sizes in the ancestral African population, the equivalent of N2/Na values inferred from different methods and modeling settings fit rather well with our own posterior expectations and quantiles for this parameter (e.g. Schiffels and Durbin, 2014).

In contrast to earlier ABC methods, the RF approach is deemed to be mostly insensitive to the presence of covariates whose the distributions does not depend on the parameter values (i.e. ancillary covariates) (e.g. Breiman, 2001; Marin et al., 2012). To illustrate this feature, we have added 20 additional independent noise variables generated from a uniform $\mathscr{U}_{[0,1]}$ distribution (results designated by RF*) in the *reference table* generated for the present Human population genetics example. We found that the presence of such noise covariates do not impact the results in terms of NMAE, coverage and only slightly on parameter estimation for the observed dataset (Tables S6, S7 and S8, and Figures 2 and S7). For the rest of the article, no noise variables were used.

3.2 Practical recommendations regarding the implementation of the ABC-RF algorithm

We mainly consider in this section two important practical issues, namely the choice of the number of simulations (N) in the *reference table* and of the number of trees (B) in the random forest. For sake of simplicity and concision, we focus our recommendations on the above human population genetics example (subsection 3.1). We stress here that, although not generic, our recommendations fit well with other examples of complex model settings that we have analysed so far (results not shown). We also stress that for simpler model settings substantially smaller N and B values were sufficient to obtain good results. Finally, we provide practical comments about the main sources of variabilities in inferences typical of the ABC-RF methodology.

Reference table size − We consider a reference table made of $N = 199\,000$ simulated datasets. However, Table S9 shows a negligible decrease of NMAE when using $N = 100\,000$ to $N = 199\,000$ datasets. Table S10 also exhibits small variations between predictions on the observed dataset, especially for $N \geq 7500$. The level of variation thus seems to be compatible with the random variability of the RF themselves. Altogether, using a *reference table* including 100 000 datasets seems to be a reasonable default choice. It is worth stressing that the out-of-bag mean squared error can be easily retrieved without requiring the simulation of a (small size) secondary test table. It provides a good indicator of the quality of the RF at a low computational cost (Tables S9 and S11).

Number of trees − A forest including 500 trees is a default choice when building RFs, as this provides a good trade-off between computation efficiency and statistical precision (Breiman, 2001; Pudlo et al., 2016). To evaluate whether or not this number is sufficient, we recommend to compute the out-of-bag mean squared error depending on the number of trees in the forest for a given *reference table*. If 500 trees is a satisfactory calibration, one should observe a stabilization of the error around this value. Figure S8 illustrates this representation on the human population genetics example and points to a negligible decrease of the error after 500 trees. This graphical representation is produced via our R package `abcrf`.

Minimum node size (maximum leaf size) − We recall that splitting events during a tree construction stop when a node has less than $N_{\min}$ observations, in that case, the node becomes a leaf. Note that the higher $N_{\min}$ the quicker RF treatments. In all RF treatments presented here, we used the default size $N_{\min} = 5$. Table S11 illustrates the influence of $N_{\min}$ on the human population genetics example and highlights a negligible decrease of the error for $N_{\min}$ lower than 5.

Finally, we see no reason to change the number of summary statistics sampled at each split n_{try} within a tree, which is traditionally chosen as $k/3$ for regression when k is the total number of predictors (Breiman, 2001).

Variability in the ABC-RF methodology − The ABC-RF methodology is associated with different sources of variabilities the user should be aware of. Using a simulated *reference table* is the main source, RF being the second. Indeed, predicting quantities of interest for the same test dataset with two different *reference tables* of equal size N will result in slightly different estimates. This variation has been previously highlighted in Figure S3 dealing with the analysis of the Normal toy example. We recall RF are composed of trees trained on bootstrap samples, each one considering n_{try} covariates randomly selected amongst the k available at each split. This random aspects of RF results in variability. In practice, a good user habit should be to run ABC-RF more than once on different training datasets to ensure that the previously mentioned variabilities are negligible. If this variability is significant, we recommend considering a *reference table* of higher size.

4 Discussion

This paper introduces a novel approach to parameter estimation in likelihood-free problems, relying on the machine-learning tool of regression RF to automate the inclusion of summary statistics in ABC algorithms. Our simulation experiments demonstrate several advantages of our methodological proposal compared with earlier ABC methods.

While using the same *reference table* and test dataset for all compared methods, our RF approach appears to be more accurate than previous ABC solutions. Approximations of expectations are quite accurate, while posterior variances are only slightly overestimated, which is an improvement compared with other approaches that typically underestimate these posterior variances. The performances for covariance approximation are quite encouraging as well, although the method is still incomplete and need further developments on this particular point (more details are given in Section 3 of Supplementary Information). We found that quantile estimations depend on the corresponding probability and we believe this must be related to the approximation error of the posterior cumulative function $F(x \mid \eta(y))$. More specifically, we observed that upper quantiles may be overestimated, whereas lower quantiles may be underestimated (Figure S2), indicating fatter tails in the approximation. Hence, credible intervals produced by the RF solution may be larger than the exact ones. However from a risk assessment point of view, this overestimation aspect clearly presents less drawbacks than underestimation of credible intervals. Altogether, owing to the various models and datasets we analysed, we argue that ABC-RF provides a good trade-off in terms of quality between parameter estimation of point estimators (e.g. expectation, median or variance) and credible interval coverage. A comparison of computing times is given in Section 9 of the Supplementary Information.

Throughout our experiments, we found that, contrary to earlier ABC methods, the RF approach is mostly insensitive to the presence of covariates whose the distributions does not depend on the parameter values (ancillary covariates). Therefore, we argue that the RF method can deal with a very large number of summary statistics, bypassing any form of pre-selection of those summaries. Interestingly, the property of ABC-RF to extract and adaptively weight information carried by each of the numerous summary statistics proposed as explanatory variables can be represented by graphs, showing the relative contribution of summary statistics in ABC-RF estimation for each studied parameter (see Section 8 of the Supplementary Information for details).

As an alternative, Papamakarios and Murray (2016) propose to approximate the whole posterior distribution by using Mixture Density Networks (MDN, Bishop, 1994). The MDN strategy consists in using Gaussian mixture models with parameters calibrated thanks to neural networks. The strategy of Papamakarios and Murray (2016) is to iteratively learn an efficient proposal prior (approximating the posterior distribution), then to use this proposal to train the posterior, both steps making use of MDN. This strategy can be easily applied when the prior is uniform or Gaussian, but other prior choices can involve difficulties. This is because in such cases, it might be difficult to simulate from the corresponding proposal. The approximation accuracy of the posterior as a Gaussian mixture model depends of the number of components and the number of hidden layers of the networks. Those two parameters require calibration. Finally, by using MDN, one loses the contribution of summary statistics provided by RF and thus some useful interpretation elements. Despite these remarks, this promising method remains of interest and is worth mentioning.

The RF method focuses on unidimensional parameter inference. Multi-objective random forest (Kocev et al., 2007) could be a solution to deal with multidimensional parameter using RF. However, our attempts based on the later methodology were so far unfruitful (results not shown). An alternative approach could be based on using the RF strategies to approximate some conditionals distributions and then recover the joined posterior using either a Gibbs sampler (based on approximated full conditionals) or Russian rule decompositions to which a product of embedded full conditionals is associated. We are presently comparing the two strategies on simulated datasets.

In population genetics, which historically corresponds to the field of introduction of ABC methods, next generation sequencing technologies result in large genome-wide datasets that can be quite informative about the demographic history of the genotyped populations. Several recently developed inferential methods relying on the observed site frequency spectrum appear particularly well suited to accurately characterizing the complex evolutionary history of invasive populations (Gutenkunst et al., 2009; Excoffier et al., 2013). Because of the reduced computational resources demanded by ABC-RF and the above-mentioned properties of the method, we believe that ABC-RF can efficiently contribute to the analysis of massive SNP datasets, including both model choice (Pudlo et al., 2016) and Bayesian inference about parameters of interest. More generally, the method should appeal to all scientific fields in which large datasets and complex models are analysed using simulation-based methods such as ABC (e.g. Beaumont, 2010; Sisson et al., 2018).

Acknowledgements

The use of random forests was suggested to J.-M.M. and C.P.R. by Bin Yu during a visit at CREST, Paris. We are very grateful to her for this conversation and also to Nicolai Meinshausen for some very helpful exchanges on the quantile regression forest method. We thank Dennis Prangle and Michael Blum for their very helpful suggestions, helping us to improve the quality of that paper. We thank Oliver Stegle and the three anonymous reviewers for the time invested in reviewing our paper and for their useful comments. We also thank Dennis Prangle and Michael Blum, both members of PCI Evol Biol (Peer Community in Evolutionary Biology; https://evolbiol.peercommunityin.org/), for their insightful comments on an earlier version of the manuscript, and Michael Blum for recommendation of a preprint of the manuscript (recommendation text as well as reviews and reply of authors are available at https://evolbiol.peercommunityin.org/public/rec?articleId=90).

Funding

A.E. acknowledges financial support by the National Research Fund ANR (France) through the European Union program ERA-Net BiodivERsA (project EXOTIC) and the project ANR-16-CE02-0015-01 (SWING), and the INRA scientific department SPE (AAP-SPE 2016). C.P.R. is supported by an IUF 2016–2021 senior grant.

References

Arenas, M. et al. (2015) CodABC: a computational framework to coestimate recombination, substitution, and molecular adaptation rates by approximate Bayesian computation. *Molecular Biology and Evolution*, 32(4), 1109–1112.

Baharian, S. et al. (2016) The great migration and African-American genomic diversity. *PLOS Genetics*, 12(5), e1006059.

Beaumont, M. A. et al. (2002) Approximate Bayesian computation in population genetics. *Genetics*, 162(4), 2025–2035.

Beaumont, M. A. (2008) Joint determination of topology, divergence time and immigration in population trees. In Simulations, Genetics and Human Prehistory, pages 134–154. Eds. Matsumura S., Forster P. and Renfrew C. *Cambridge: (McDonald Institute Monographs), McDonald Institute for Archaeological Research*, 2008.

Beaumont, M. A. et al. (2009) Adaptive approximate Bayesian computation. *Biometrika*, 96(4), 983-990.

Beaumont, M. A. (2010) Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41, 379–406.

Biau, G. (2012) Analysis of a random forest model. *Journal of Machine Learning Research*, 13, 1063–1095.

Biau, G. et al. (2015) New insights into approximate Bayesian computation. *Annales de l'Institut Henri Poincaré B, Probability and Statistics*, 51(1), 376–403.

Bishop, C. M. (1994) Mixture density networks. Tech. Rep. NCRG/94/004, Neural Computing Research Group, Aston University.

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

Blum, M. G. B. et al. (2013) A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28 (2), 189–208.

Breiman, L. (2001) Random forests. *Machine Learning*, 45 (1), 5–32.

Bryc, K. et al. (2015) The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics*, 96(1), 37–53.

Chan, Y. L. et al. (2014) Detecting concerted demographic response across community assemblages using hierarchical approximate Bayesian computation. *Molecular Biology and Evolution*, 31 (9), 2501–2515.

Csilléry, K. et al. (2010) Approximate Bayesian computation (ABC) in practice. *Trends in Ecology and Evolution*, 25(7), 410–418.

Csilléry, K. et al. (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, 3(3), 475–479.

Csilléry, K. et al. (2015) abc: Tools for approximate Bayesian computation (ABC). R package version 2.1.

Cornuet, J.-M. et al. (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, 24(23), 2713–2719.

Cornuet, J.-M. et al. (2014) DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, 30, (8), 1187–1189.

Del Moral, P. et al. (2012) An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22, 1009–1020.

Excoffier, L. et al. (2013) Robust demographic inference from genomic and SNP data. *PLOS Genetics*, 9(10), e1003905.

Fearnhead, P. and Prangle, D. (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74 (3), 419–474.

Frazier, D. T. et al. (2017) Asymptotic properties of approximate Bayesian computation. *arXiv*, 1609.06903.

Gourieroux, C. et al. (1993) Indirect inference. *Journal of Applied Econometrics*, 8, 85–118.

Gutenkunst, R. N. et al. (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLOS Genetics*, 5(10), e1000695.

Henn, B. M. et al. (2012) The great human expansion. *Proc. Natl. Acad. Sci. USA*, 109(44), 17758–17764.

Hudson, R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2), 337–338.

Joyce, P. and Marjoram, P. (2008) Approximately sufficient statistics and Bayesian computation. *Statistical Application in Genetics and Molecular Biology*, 7 (1), Article 26.

Jung, H. and Marjoram, P. (2011) Choice of summary statistics weights in approximate Bayesian computation. *Statistical Application in Genetics and Molecular Biology*, 10 (1), Article 45.

Klinger, E. and Hasenauer, J. (2017). A scheme for adaptive selection of population sizes in approximate Bayesian computation - sequential Monte Carlo. *Computational Methods in Systems Biology: 15th International Conference*, 128–144.

Klinger, E., Rickert, D. and Hasenauer, J. (2018). pyABC: distributed, likelihood-free inference. *Bioinformatics*, to appear.

Kocev, D. et al. (2007) Ensembles of multi-objective decision trees. In Machine Learning: ECML 2007. Lecture Notes in Computer Science, vol 4701. pages 624–631. Eds. Kok J. N., Koronacki J., Mantaras R. L., Matwin S., Mladenič D. and Skowron A. *Springer, Berlin, Heidelberg*, 2007.

Li, W. and Fearnhead, P. (2015) On the asymptotic efficiency of ABC estimators. *arXiv*, 1506.03481.

Lindsay, B. (1988) Composite likelihood methods. *Contemporary Mathematics*, 80, 221–239.

Maples, B. K. et al. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2), 278–288.

Marin, J.-M. et al. (2012) Approximate Bayesian computational methods. *Statistics and Computing*, 22 (6), 1167–1180.

Marin, J.-M. and Robert, C. P. (2014) *Bayesian Essentials with R*. Springer.

Marin, J.-M. et al. (2018) Likelihood-free model choice. In Handbook of Approximate Bayesian Computation. Eds. Sisson S. A., Fan Y. and Beaumont M. A. *Chapman and Hall/CRC Press*.

Meinshausen, N. (2006) Quantile regression forests. *Journal of Machine Learning Research*, 7, 983–999.

Nunes, M. A. and Balding, D. J. (2010) On optimal selection of summary statistics for approximate Bayesian computation. *Statistical Application in Genetics and Molecular Biology*, 9 (1), Article 34.

Papamakarios, G. and Murray, I. (2016) Fast ε-free inference of simulation models with Bayesian conditional density estimation. In Advances in Neural Information Processing Systems 29, pages 1028–1036. Eds. Lee D. D., Sugiyama M., Luxburg U. V., Guyon I. and Garnett R. *Curran Associates, Inc.*, 2016.

Prangle, D. (2017) Adapting the ABC distance function. *Bayesian Analysis*, 12(1), 289–309.

Pritchard, J. K. et al. (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12), 1791–1798.

Pudlo, P. et al. (2016) Reliable ABC model choice via random forests. *Bioinformatics*, 32(6), 859–866.

Saulnier, E. et al. (2017) Inferring epidemiological parameters from phylogenies using regression-ABC: a comparative study. *PLOS Computational Biology*, 13(3), e1005416.

Schiffels, S. and Durbin, R. (2014) Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46, 919–925.

Sisson, S. A. et al. (2007) Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci.*, 104, 1760–1765.

Sisson, S. A. et al. (2009) Correction: sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci.*, 106, 1760.

Sisson, S. A. et al. (2018) Handbook of Approximate Bayesian Computation. *Chapman & Hall/CRC Press.*

Tavaré, S. et al. (1997) Inferring coalescence times from DNA sequence data. *Genetics*, 145(2), 505–518.

The 1000 genomes project consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491, 56–65.

Theunert, C. et al. (2012) Inferring the history of population size change from genome-wide SNP Data. *Molecular Biology and Evolution*, 29 (12), 3653–3667.

Toni, T. et al. (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6 (31), 187–202.

Varin, C. et al. (2011) An overview of composite likelihood methods. *Statistica Sinica*, 21(1), 5–42.

Wright, M. N. and Ziegler, A. (2017) ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17.