Bioinformatics, 32(6), 2016, 859–866
doi: 10.1093/bioinformatics/btv684
Advance Access Publication Date: 20 November 2015
Original Paper

OXFORD

Genetics and population analysis

Reliable ABC model choice via random forests

Pierre Pudlo[1,2,†], Jean-Michel Marin[1,2,*,†], Arnaud Estoup[2,3], Jean-Marie Cornuet[3], Mathieu Gautier[2,3] and Christian P. Robert[4,5]

[1]Université de Montpellier, IMAG, Montpellier, [2]Institut de Biologie Computationnelle (IBC), Montpellier, [3]CBGP, INRA, Montpellier, [4]Université Paris Dauphine, CEREMADE, Paris, France and [5]University of Warwick, Coventry, UK

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Inanc Birol

Received on May 29, 2015; revised on September 2, 2015; accepted on September 30, 2015

Abstract

Motivation: Approximate Bayesian computation (ABC) methods provide an elaborate approach to Bayesian inference on complex models, including model choice. Both theoretical arguments and simulation experiments indicate, however, that model posterior probabilities may be poorly evaluated by standard ABC techniques.

Results: We propose a novel approach based on a machine learning tool named random forests (RF) to conduct selection among the highly complex models covered by ABC algorithms. We thus modify the way Bayesian model selection is both understood and operated, in that we rephrase the inferential goal as a classification problem, first predicting the model that best fits the data with RF and postponing the approximation of the posterior probability of the selected model for a second stage also relying on RF. Compared with earlier implementations of ABC model choice, the ABC RF approach offers several potential improvements: (i) it often has a larger discriminative power among the competing models, (ii) it is more robust against the number and choice of statistics summarizing the data, (iii) the computing effort is drastically reduced (with a gain in computation efficiency of at least 50) and (iv) it includes an approximation of the posterior probability of the selected model. The call to RF will undoubtedly extend the range of size of datasets and complexity of models that ABC can handle. We illustrate the power of this novel methodology by analyzing controlled experiments as well as genuine population genetics datasets.

Availability and implementation: The proposed methodology is implemented in the R package abcrf available on the CRAN.

Contact: jean-michel.marin@umontpellier.fr

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

1 Introduction

Approximate Bayesian computation (ABC) represents an elaborate statistical approach to model-based inference in a Bayesian setting in which model likelihoods are difficult to calculate (due to the complexity of the models considered).

Since its introduction in population genetics (Beaumont *et al.*, 2002; Pritchard *et al.*, 1999; Tavaré *et al.*, 1997), the method has found an ever increasing range of applications covering diverse types of complex models in various scientific fields (see, e.g. Arenas *et al.*, 2015; Beaumont, 2008, 2010; Chan *et al.*, 2014; Csillèry *et al.*, 2010; Theunert *et al.*, 2012; Toni *et al.*, 2009). The principle of ABC is to conduct Bayesian inference on a dataset through comparisons with numerous simulated datasets. However, it suffers from two major difficulties. First, to ensure reliability of the method, the number of simulations is large; hence, it proves difficult to apply ABC for large datasets (e.g. in population genomics where tens to

hundred thousand markers are commonly genotyped). Second, calibration has always been a critical step in ABC implementation (Blum *et al.*, 2013; Marin *et al.*, 2012). More specifically, the major feature in this calibration process involves selecting a vector of summary statistics that quantifies the difference between the observed data and the simulated data. The construction of this vector is therefore paramount and examples abound about poor performances of ABC model choice algorithms related with specific choices of those statistics (Didelot *et al.*, 2011; Marin *et al.*, 2014; Robert *et al.*, 2011), even though there also are instances of successful implementations.

We advocate a drastic modification in the way ABC model selection is conducted: we propose both to step away from selecting the most probable model from estimated posterior probabilities and to reconsider the very problem of constructing efficient summary statistics. First, given an arbitrary pool of available statistics, we now completely bypass selecting among those. This new perspective directly proceeds from machine learning methodology. Second, we postpone the approximation of model posterior probabilities to a second stage, as we deem the standard numerical ABC approximations of such probabilities fundamentally untrustworthy. We instead advocate selecting the posterior most probable model by constructing a (machine learning) classifier from simulations from the prior predictive distribution (or other distributions in more advanced versions of ABC), known as the ABC *reference table*. The statistical technique of random forests (RF) (Breiman, 2001) represents a trustworthy machine learning tool well adapted to complex settings as is typical for ABC treatments. Once the classifier is constructed and applied to the actual data, an approximation of the posterior probability of the resulting model can be produced through a secondary RF that regresses the selection error over the available summary statistics. We show here how RF improves upon existing classification methods in significantly reducing both the classification error and the computational expense. After presenting theoretical arguments, we illustrate the power of the ABC-RF methodology by analyzing controlled experiments as well as genuine population genetics datasets.

2 Materials and methods

Bayesian model choice (Berger, 1985; Robert, 2001) compares the fit of M models to an observed dataset $\mathbf{x}^0$. It relies on a hierarchical modelling, setting first prior probabilities $\pi(m)$ on model indices $m \in \{1, \ldots, M\}$ and then prior distributions $\pi(\theta|m)$ on the parameter θ of each model, characterized by a likelihood function $f(\mathbf{x}|m, \theta)$. Inferences and decisions are based on the posterior probabilities of each model $\pi(m|\mathbf{x}^0)$.

2.1 ABC algorithms for model choice

While we cannot cover in much detail the principles of ABC, let us recall here that ABC was introduced in Tavaré *et al.* (1997) and Pritchard *et al.* (1999) for solving intractable likelihood issues in population genetics. The reader is referred to, e.g. Beaumont (2008, 2010), Toni *et al.* (2009), Csillèry *et al.* (2010) and Marin *et al.* (2012) for thorough reviews on this approximation method. The fundamental principle at work in ABC is that the value of the intractable likelihood function $f(\mathbf{x}^0|\theta)$ at the observed data $\mathbf{x}^0$ and for a current parameter θ can be evaluated by the proximity between $\mathbf{x}^0$ and pseudo-data $\mathbf{x}(\theta)$ simulated from $f(\mathbf{x}|\theta)$. In discrete settings, the indicator $\mathbb{I}(\mathbf{x}(\theta) = \mathbf{x}^0)$ is an unbiased estimator of $f(\mathbf{x}^0|\theta)$ (Rubin, 1984). For realistic settings, the equality constraint is replaced with a tolerance region $\mathbb{I}(d(\mathbf{x}(\theta), \mathbf{x}^0) \leq \epsilon)$, where $d(\mathbf{x}^0, \mathbf{x})$ is a measure of divergence between the two vectors and $\epsilon > 0$ is a tolerance value. The implementation of this principle is straightforward: the ABC algorithm produces a large number of pairs $(\theta, \mathbf{x})$ from the prior predictive, a collection called the *reference table*, and extracts from the table the pairs $(\theta, \mathbf{x})$ for which $d(\mathbf{x}(\theta), \mathbf{x}^0) \leq \epsilon$.

To approximate posterior probabilities of competing models, ABC methods (Grelaud *et al.*, 2009) compare observed data with a massive collection of pseudo-data, generated from the prior predictive distribution in the most standard versions of ABC; the comparison proceeds via a normalized Euclidean distance on a vector of statistics $S(\mathbf{x})$ computed for both observed and simulated data. Standard ABC estimates posterior probabilities $\pi(m|\mathbf{x}^0)$ at stage (B) of Algorithm 1 below as the frequencies of those models within the k nearest-to-$\mathbf{x}^0$ simulations, proximity being defined by the distance between $S(\mathbf{x}^0)$ and the simulated $S(\mathbf{x})$'s.

Selecting a model means choosing the model with the highest frequency in the sample of size k produced by ABC, such frequencies being approximations to posterior probabilities of models. We stress that this solution means resorting to a k-nearest neighbor (k-nn) estimate of those probabilities, for a set of simulations drawn at stage (A), whose records constitute the so-called *reference table*, see Biau *et al.* (2015) or Stoehr *et al.* (2015).

Algorithm 1. ABC model choice algorithm

(A) Generate a reference table including N_{ref} simulations $(m, S(\mathbf{x}))$ from $\pi(m)\pi(\theta|m)f(\mathbf{x}|m, \theta)$

(B) Learn from this set to infer about m at $\mathbf{s}^0 = S(\mathbf{x}^0)$

Selecting a set of summary statistics $S(x)$ that are informative for model choice is an important issue. The ABC approximation to the posterior probabilities $\pi(m|\mathbf{x}^0)$ will eventually produce a right ordering of the fit of competing models to the observed data and thus select the right model for a specific class of statistics on large datasets (Marin *et al.*, 2014). This most recent theoretical ABC model choice results indeed shows that some statistics produce nonsensical decisions and that there exist sufficient conditions for statistics to produce consistent model prediction, albeit at the cost of an information loss due to summaries that may be substantial. The toy example comparing MA(1) and MA(2) models in Supplementary Informations and Figure 1 clearly exhibits this potential loss in using only the first two autocorrelations as summary statistics. Barnes *et al.* (2012) developed an interesting methodology to select the summary statistics but with the requirement to aggregate estimation and model pseudo-sufficient statistics for all models. This induces a deeply inefficient dimension inflation and can be very time consuming.

It may seem tempting to collect the largest possible number of summary statistics to capture more information from the data. This brings $\pi(m|S(\mathbf{x}^0))$ closer to $\pi(m|\mathbf{x}^0)$ but increases the dimension of $S(\mathbf{x})$. ABC algorithms, like k-nn and other local methods suffer from the curse of dimensionality [see e.g. Section 2.5 in Hastie *et al.* (2009)] so that the estimate of $\pi(m|S(\mathbf{x}^0))$ based on the simulations is poor when the dimension of $S(\mathbf{x})$ is too large. Selecting summary statistics correctly and sparsely is therefore paramount, as shown by the literature in the recent years. [See Blum *et al.* (2013) surveying ABC parameter estimation.] For ABC model choice, two main projection techniques have been considered so far. First, Prangle *et al.* (2014) show that the Bayes factor itself is an acceptable summary (of dimension one) when comparing two models, but its practical evaluation via a pilot ABC simulation induces a poor approximation

of model evidences (Didelot *et al.*, 2011; Robert *et al.*, 2011). The recourse to a regression layer like linear discriminant analysis (LDA, Estoup *et al.*, 2012) is discussed below and in Supplementary Section S1. Other projection techniques have been proposed in the context of parameter estimation: see, e.g. Fearnhead and Prangle (2012); Aeschbacher *et al.* (2012).

Given the fundamental difficulty in producing reliable tools for model choice based on summary statistics (Robert *et al.*, 2011), we now propose to switch to a different approach based on an adapted classification method. We recall in the next section the most important features of the RF algorithm.

2.2 RF methodology

The classification and regression trees (CART) algorithm at the core of the RF scheme produces a binary tree that sets allocation rules for entries as labels of the internal nodes and classification or predictions of Y as values of the tips (terminal nodes). At a given internal node, the binary rule compares a selected covariate X_j with a bound t, with a left-hand branch rising from that vertex defined by $X_j < t$. Predicting the value of Y given the covariate X implies following a path from the tree root that is driven by applying these binary rules. The outcome of the prediction is the value found at the final leaf reached at the end of the path: majority rule for classification and average for regression. To find the best split and the best variable at each node of the tree, we minimize a criterium: for classification, the Gini index and, for regression, the L^2-loss. In the randomized version of the CART algorithm (see Supplementary Algorithm S1), only a random subset of covariates of size n_{try} is considered at each node of the tree.

The RF algorithm (Breiman, 2001) consists in bagging (which stands for bootstrap aggregating) randomized CART. It produces N_{tree} randomized CART trained on samples or sub-samples of size N_{boot} produced by bootstrapping the original training database. Each tree provides a classification or a regression rule that returns a class or a prediction. Then, for classification we use the majority vote across all trees in the forest, and, for regression, the response values are averaged.

Three tuning parameters need be calibrated: the number N_{tree} of trees in the forest, the number n_{try} of covariates that are sampled at a given node of the randomized CART and the size N_{boot} of the bootstrap sub-sample. This point will be discussed in Section 3.4.

For classification, a very useful indicator is the *out-of-bag* error (Hastie *et al.*, 2009, Chapter 15). Without any recourse to a test set, it gives some idea on how good is your RF classifier. For each element of the training set, we can define the out-of-bag classifier: the aggregation of votes over the trees not constructed using this element. The out-of-bag error is the error rate of the out-of-bag classifier on the training set. The out-of-bag error estimate is as accurate as using a test set of the same size as the training set.

2.3 ABC model choice via RF

The above-mentioned difficulties in ABC model choice drives us to a paradigm shift in the practice of model choice, namely to rely on a classification algorithm for model selection, rather than a poorly estimated vector of $\pi(m|S(\mathbf{x}^0))$ probabilities. As shown in the example described in Section 3.1, the standard ABC approximations to posterior probabilities can significantly differ from the true $\pi(m|\mathbf{x}^0)$. Indeed, our version of stage (B) in Algorithm 1 relies on a RF classifier whose goal is to predict the suited model $\hat{m}(s)$ at each possible value s of the summary statistics $S(\mathbf{x})$. The RF is trained on the simulations produced by stage (A) of Algorithm 1, which constitute the reference table. Once the model is selected as $m^\star$, we opt to approximate $\pi(m^*|S(\mathbf{x}^0))$ by another RF, obtained from regressing the probability of error on the (same) covariates, as explained below.

A practical way to evaluate the performance of an ABC model choice algorithm (test a given set of summary statistics and a given classifier) is to check whether it provides a better answer than others. The aim is to come near the so-called *Bayesian classifier*, which, for the observed $\mathbf{x}$, selects the model having the largest posterior probability $\pi(m|\mathbf{x})$. It is well known that the Bayesian classifier minimizes the 0–1 integrated loss or error (Devroye *et al.*, 1996). In the ABC framework, we call the integrated loss (or risk) the *prior error rate*, since it provides an indication of the global quality of a given classifier $\hat{m}$ on the entire space weighted by the prior. This rate is the expected value of the misclassification error over the hierarchical prior

$$\sum_m \pi(m) \int \mathbf{1}\{\hat{m}(S(\mathbf{y})) \neq m\} f(\mathbf{y}|\theta, m)\pi(\theta|m)\mathrm{d}\mathbf{y}\mathrm{d}\theta\,.$$

It can be evaluated from simulations $(\theta, m, S(\mathbf{y}))$ drawn as in stage (A) of Algorithm 1, independently of the reference table (Stoehr *et al.*, 2015), or with the out-of-bag error in RF that, as explained above, requires no further simulation. Both classifiers and sets of summary statistics can be compared via this error scale: the pair that minimizes the prior error rate achieves the best approximation of the ideal Bayesian classifier. In that sense, it stands closest to the decision we would take were we able to compute the true $\pi(m|\mathbf{x})$.

We seek a classifier in stage (B) of Algorithm 1 that can handle an arbitrary number of statistics and extract the maximal information from the reference table obtained at stage (A). As introduced above, RF classifiers (Breiman, 2001) are perfectly suited for that purpose. The way we build both a RF classifier given a collection of statistical models and an associated RF regression function for predicting the allocation error is to start from a simulated ABC *reference table* made of a set of simulation records made of model indices and summary statistics for the associated simulated data. This table then serves as training database for a RF that forecasts model index based on the summary statistics. The resulting algorithm, presented in Algorithm 2 and called ABC-RF, is implemented in the R package `abcrf` associated with this article.

Algorithm 2: ABC-RF

(**A**) Generate a reference table including N_{ref} simulation $(m, S(\mathbf{x}))$ from $\pi(m)\pi(\theta|m)f(\mathbf{x}|m,\theta)$

(**B**) Construct N_{tree} randomized CART which predict m using $S(\mathbf{x})$

for $b = 1$ **to** N_{tree} **do**

draw a bootstrap (sub-)sample of size N_{boot} from the reference table

grow a randomized CART T_b (Supplementary Algorithm S1)

end for

(**C**) Determine the predicted indexes for $S(\mathbf{x}^0)$ and the trees $\{T_b; b = 1, \ldots, N_{\text{tree}}\}$

(**D**) $S(\mathbf{x}^0)$ according to a majority vote among the predicted indexes

The justification for choosing RF to conduct an ABC model selection is that, both formally (Biau, 2012; Scornet *et al.*, 2015) and experimentally (Hastie *et al.*, 2009, Chapter 5), RF classification was shown to be mostly insensitive both to strong correlations

between predictors (here the summary statistics) and to the presence of noisy variables, even in relatively large numbers, a characteristic that k-nn classifiers lack.

This type of robustness justifies adopting a RF strategy to learn from an ABC reference table for Bayesian model selection. Within an arbitrary (and arbitrarily large) collection of summary statistics, some may exhibit strong correlations and others may be uninformative about the model index, with no terminal consequences on the RF performances. For model selection, RF thus competes with both local classifiers commonly implemented within ABC: It provides a more non-parametric modelling than local logistic regression (Beaumont, 2008), which is implemented in the DIYABC software (Cornuet *et al.*, 2014) but is extremely costly—see the method of Estoup *et al.* (2012) to reduce the dimension using linear discriminant projection before resorting to local logistic regression. This software also includes a standard k-nn selection procedure [i.e. the so-called direct approach in Cornuet *et al.* (2008)] which suffers from the curse of dimensionality and thus forces selection among statistics.

2.4 Approximating the posterior probability of the selected model

The outcome of RF computation applied to a given target dataset is a classification vote for each model which represents the number of times a model is selected in a forest of n trees. The model with the highest classification vote corresponds to the model best suited to the target dataset. It is worth stressing here that there is no direct connection between the frequencies of the model allocations of the data among the tree classifiers (i.e. the classification vote) and the posterior probabilities of the competing models. Machine learning classifiers hence miss a distinct advantage of posterior probabilities, namely that the latter evaluate a confidence degree in the selected model. An alternative to those probabilities is the prior error rate. Aside from its use to select the best classifier and set of summary statistics, this indicator remains, however, poorly relevant since the only point of importance in the data space is the observed dataset $S(\mathbf{x}^0)$.

A first step addressing this issue is to obtain error rates conditional on the data as in Stoehr *et al.* (2015). However, the statistical methodology considered therein suffers from the curse of dimensionality and we here consider a different approach to precisely estimate this error. We recall (Robert, 2001) that the posterior probability of a model is the natural Bayesian uncertainty quantification since it is the complement of the posterior error associated with the loss $\mathbb{I}(\hat{m}(S(\mathbf{x}^0)) \neq m)$. While the proposal of Stoehr *et al.* (2015) for estimating the conditional error rate induced a classifier given $S = S(\mathbf{x}^0)$

$$\mathbb{P}(\hat{m}(S(Y)) \neq m | S(Y) = S(\mathbf{x}^0)), \tag{1}$$

involves non-parametric kernel regression, we suggest to rely instead on a RF regression to undertake this estimation. The curse of dimensionality is then felt much less acutely, given that RF can accommodate large dimensional summary statistics. Furthermore, the inclusion of many summary statistics does not induce a reduced efficiency in the RF predictors, while practically compensating for insufficiency.

Before describing in more details the implementation of this concept, we stress that the perspective of Stoehr *et al.* (2015) leads to effective estimates of the posterior probability that the selected model is the true model, thus providing us with a non-parametric estimation of this quantity. Indeed, the posterior expectation (1) satisfies

$$\begin{aligned}\mathbb{E}[\mathbb{I}(\hat{m}(\mathbf{s}^0) \neq m)|S(\mathbf{x}^0)] &= \sum_{i=1}^{k} \mathbb{E}[\mathbb{I}(\hat{m}(S(\mathbf{x}^0)) \neq m = i)|S(\mathbf{x}^0)] \\ &= \sum_{i=1}^{k} \mathbb{P}[m = i|S(\mathbf{x}^0)] \times \mathbb{I}(\hat{m}(S(\mathbf{x}^0)) \neq i) \\ &= \mathbb{P}[m \neq \hat{m}(S(\mathbf{x}^0))|S(\mathbf{x}^0)] \\ &= 1 - \mathbb{P}[m = \hat{m}(S(\mathbf{x}^0))|S(\mathbf{x}^0)]\,.\end{aligned}$$

It therefore provides the complement of the posterior probability that the true model is the selected model.

To produce our estimate of the posterior probability $\mathbb{P}[m = \hat{m}(S(\mathbf{x}^0))|S(\mathbf{x}^0)]$, we proceed as follows:

1. We compute the value of $\mathbb{I}(\hat{m}(s) \neq m)$ for the trained RF $\hat{m}$ and for all terms in the ABC reference table; to avoid overfitting, we use the out-of-bag classifiers;
2. We train a RF regression estimating the variate $\mathbb{I}(\hat{m}(s) \neq m)$ as a function of the same set of summary statistics, based on the same reference table. This second RF can be represented as a function $\varrho(s)$ that constitutes a machine learning estimate of $\mathbb{P}[m \neq \hat{m}(s)|s]$;
3. We apply this RF function to the actual observations summarized as $S(\mathbf{x}^0)$ and return $1 - \varrho(S(\mathbf{x}^0))$ as our estimate of $\mathbb{P}[m = \hat{m}(S(\mathbf{x}^0))|S(\mathbf{x}^0)]$.

This corresponds to the representation of Algorithm 3 which is implemented in the R package `abcrf` associated with this paper.

Algorithm 3: Estimating the posterior probability of the selected model

(a) Use the RF produced by Algorithm 2 to compute the out-of-bag classifiers of all terms in the reference table and deduce the associated binary model prediction error
(b) Use the reference table to build a RF regression function $\varrho(s)$ regressing the model prediction error on the summary statistics
(c) Return the value of $1 - \varrho(S(\mathbf{x}^0))$ as the RF regression estimate of $\mathbb{P}[m = \hat{m}(S(\mathbf{x}^0))|S(\mathbf{x}^0)]$

3 Results: illustrations of the ABC-RF methodology

To illustrate the power of the ABC-RF methodology, we now report several controlled experiments as well as two genuine population genetic examples.

3.1 Insights from controlled experiments

The Supplementary Information details controlled experiments on a toy problem, comparing MA(1) and MA(2) time-series models, and two controlled synthetic examples from population genetics, based on single-nucleotide polymorphism (SNP) and microsatellite data. The toy example is particularly revealing with regard to the discrepancy between the posterior probability of a model and the version conditioning on the summary statistics $S(\mathbf{x}^0)$. Figure 1 shows how far from the diagonal are realizations of the pairs $(\pi(m|\mathbf{x}^0), \pi(m|S(\mathbf{x}^0)))$, even though the autocorrelation statistic is quite informative (Marin *et al.*, 2012). Note in particular the vertical accumulation of points near $\mathbb{P}(m = 2|\mathbf{x}^0) = 1$. Supplementary Table S1 demonstrates the further gap in predictive power for the full Bayes solution with a true error rate of 12% versus the best solution (RF) based on the summaries barely achieving a 16% error rate.

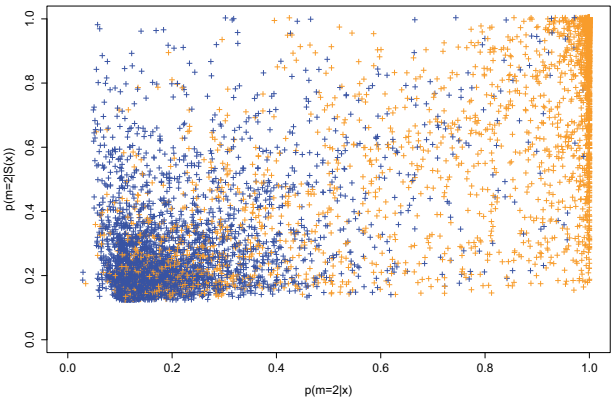

Fig. 1. Illustration of the discrepancy between posterior probabilities based on the whole data and based on a summary. The aim is to choose between two nested time series models, namely moving averages of order 1 and 2 [denoted MA(1) and MA(2), respectively; see Supplementary Information for more details]. Each point of the plot gives two posterior probabilities of MA(2) for a dataset simulated either from the MA(1) (*blue*) or MA(2) model (*orange*), based on the whole data (*x*-axis) and on only the first two autocorrelations (*y*-axis)

For both controlled genetics experiments in the Supplementary Information, the computation of the true posterior probabilities of the three models is impossible. The predictive performances of the competing classifiers can nonetheless be compared on a test sample. Results, summarized in Supplementary Tables S2 and S3 in the Supplementary Information, legitimize the use of RF, as this method achieves the most efficient classification in all genetic experiments. Note that that the prior error rate of any classifier is always bounded from below by the error rate associated with the (ideal) Bayesian classifier. Therefore, a mere gain of a few percents may well constitute an important improvement when the prior error rate is low. As an aside, we also stress that, since the prior error rate is an expectation over the entire sampling space, the reported gain may exhibit much better performances over some areas of this space.

Supplementary Figure S2 displays differences between the true posterior probability of the model selected by Algorithm 2 and its approximation with Algorithm 3. Moreover, we found that the values of the votes provided by Algorithm 2 is only useful to assess the model that best fits the data but that any conclusion regarding level of confidence necessitates the computation of the posterior probability of the selected model provided by Algorithm 3.

3.2 Microsatellite dataset: retracing the invasion routes of the Harlequin ladybird

The original challenge was to conduct inference about the introduction pathway of the invasive Harlequin ladybird (*Harmonia axyridis*) for the first recorded outbreak of this species in eastern North America. The dataset, first analyzed in Lombaert *et al.* (2011) and Estoup *et al.* (2012) via ABC, includes samples from three natural and two biocontrol populations genotyped at 18 microsatellite markers. The model selection requires the formalization and comparison of 10 complex competing scenarios corresponding to various possible routes of introduction [see Supplementary Information for details and analysis 1 in Lombaert *et al.* (2011)]. We now compare our results from the ABC-RF algorithm with other classification methods for three sizes of the reference table and with the original solutions by Lombaert *et al.* (2011) and Estoup *et al.* (2012). We included all summary statistics computed by the DIYABC software for microsatellite markers (Cornuet *et al.*, 2014), namely 130 statistics, complemented by the nine LDA axes as additional summary statistics (see Supplementary Section S4).

In this example, discriminating among models based on the observation of summary statistics is difficult. The overlapping groups of Supplementary Figure S8 reflect that difficulty, the source of which is the relatively low information carried by the 18 autosomal microsatellite loci considered here. Prior error rates of learning methods on the whole reference table are given in Table 1. As expected in such a high dimension settings (Hastie *et al.*, 2009, Section 2.5), *k*-nn classifiers behind the standard ABC methods are all defeated by RF for the three sizes of the reference table, even when *k*-nn is trained on the much smaller set of covariates composed of the nine LDA axes. The classifier and set of summary statistics showing the lowest prior error rate is RF trained on the 130 summaries and the nine LDA axes.

Supplementary Figure S9 shows that RFs are able to automatically determine the (most) relevant statistics for model comparison, including in particular some crude estimates of admixture rate defined in Choisy *et al.* (2004), some of them not selected by the experts in Lombaert *et al.* (2011). We stress here that the level of information of the summary statistics displayed in Supplementary Figure S9 is relevant for model choice but not for parameter estimation issues. In other words, the set of best summaries found with ABC-RF should not be considered as an optimal set for further parameter estimations under a given model with standard ABC techniques (Beaumont *et al.*, 2002).

The evolutionary scenario selected by our RF strategy agrees with the earlier conclusion of Lombaert *et al.* (2011), based on approximations of posterior probabilities with local logistic regression solely on the LDA axes, i.e. the same scenario displays the highest ABC posterior probability and the largest number of selection among the decisions taken by the aggregated trees of RF. Using Algorithm 3, we got an estimate of the posterior probability of the selected scenario equal to 0.4624. This estimate is significantly lower than the one of about 0.6 given in Lombaert *et al.* (2011) based on a local logistic regression method. This new value is more credible because it is based on all the summary statistics and, on a method adapted to such an high dimensional context and less sensitive to calibration issues. Moreover, this small posterior probability corresponds better to the intuition of the experimenters and indicates that new experiments are necessary to give a more reliable answer (e.g. the genotyping of a larger number of loci).

3.3 SNP dataset: inference about human population history

Because the ABC-RF algorithm performs well with a substantially lower number of simulations compared to standard ABC methods, it is expected to be of particular interest for the statistical processing of massive SNP datasets, whose production is on the increase in the field of population genetics. We analyze here a dataset including 50 000 SNP markers genotyped in four Human populations (The 1000 Genomes Project Consortium, 2012). The four populations include Yoruba (Africa), Han (East Asia), British (Europe) and American individuals of African ancestry, respectively. Our intention is not to bring new insights into Human population history, which has been and is still studied in greater details in research using genetic data but to illustrate the potential of ABC-RF in this context. We compared six scenarios (i.e. models) of evolution of the four Human populations which differ from each other by one ancient and one recent historical events: (i) a single out-of-Africa

colonization event giving an ancestral out-of-Africa population which secondarily split into one European and one East Asian population lineages, versus two independent out-of-Africa colonization events, one giving the European lineage and the other one giving the East Asian lineage; (ii) the possibility of a recent genetic admixture of Americans of African origin with their African ancestors and individuals of European or East Asia origins. The SNP dataset and the compared scenarios are further detailed in the Supplementary Information. We used all the summary statistics provided by DIYABC for SNP markers (Cornuet *et al.*, 2014), namely 112 statistics in this setting complemented by the five LDA axes as additional statistics.

To discriminate between the six scenarios of Supplementary Figure S12, RF and other classifiers have been trained on three refer ence tables of different sizes. The estimated prior error rates are reported in Table 2. Unlike the previous example, the information carried here by the 50 000 SNP markers is much higher, because it induces better separated simulations on the LDA axes (Fig. 2) and much lower prior error rates (Table 2). RF using both the initial summaries and the LDA axes provides the best results.

The ABC-RF algorithm selects Scenario 2 as the predicted scenario on the Human dataset, an answer which is not visually obvious on the LDA projections of Figure 2 in which Scenario 2 corresponds to the blue color. But considering previous population genetics studies in the field, it is not surprising that this scenario, which includes a single out-of-Africa colonization event giving an ancestral out-of-Africa population with a secondarily split into one European and one East Asian population lineage and a recent genetic admixture of Americans of African origin with their African ancestors and European individuals, was selected. Using Algorithm 3, we got an estimate of the posterior probability of scenario 2 equal to 0.998, corresponding to a high level of confidence in choosing scenario 2.

Computation time is a particularly important issue in the present example. Simulating the 10 000 SNP datasets used to train the classification methods requires 7 h on a computer with 32 processors (Intel Xeon(R) CPU 2 GHz). In that context, it is worth stressing that RF trained on the DIYABC summaries and the LDA axes of a 10 000 reference table has a smaller prior error rate than all other classifiers, even when they are trained on a 50 000 reference table. In practice, standard ABC treatments for model choice are based on reference tables of substantially larger sizes [i.e. 10^5 to 10^6 simulations per scenario (Bertorelle *et al.*, 2010; Estoup *et al.*, 2012)]. For the above setting in which six scenarios are compared, standard ABC treatments would hence request a minimum computation time of 17 days (using the same computation resources). According to the comparative tests that we carried out on various example datasets, we found that RF globally allowed a minimum computation speed gain around a factor of 50 in comparison to standard ABC treatments: see also Supplementary Section S4 for other considerations regarding computation speed gain.

Table 1. Harlequin ladybird data: estimated prior error rates for various classification methods and sizes of the reference table

Classification method trained on	Prior error rates (%)		
	$N_{ref} = 10\,000$	$N_{ref} = 20\,000$	$N_{ref} = 50\,000$
LDA	39.91	39.30	39.04
Standard ABC (*k*-nn) on DIYABC summaries	57.46	53.76	51.03
Standard ABC (*k*-nn) on LDA axes	39.18	38.46	37.91
Local logistic regression on LDA axes	41.04	37.08	36.05
RF on DIYABC summaries	40.18	38.94	37.63
RF on DIYABC summaries and LDA axes	36.86	35.62	34.44

Note. Performances of classifiers used in stage (B) of Algorithm 1. A set of 10 000 prior simulations was used to calibrate the number of neighbors *k* in both standard ABC and local logistic regression. Prior error rates are estimated as average misclassification errors on an independent set of 10 000 prior simulations, constant over methods and sizes of the reference tables. N_{ref} corresponds to the number of simulations included in the reference table.

Table 2. Human SNP data: estimated prior error rates for classification methods and three sizes of reference table

Classification method trained on	Prior error rates (%)		
	$N_{ref} = 10\,000$	$N_{ref} = 20\,000$	$N_{ref} = 50\,000$
LDA	9.91	9.97	10.03
Standard ABC (*k*-nn) using DYIABC summaries	23.18	20.55	17.76
Standard ABC (*k*-nn) using only LDA axes	6.29	5.76	5.70
Local logistic regression on LDA axes	6.85	6.42	6.07
RF using DYIABC initial summaries	8.84	7.32	6.34
RF using both DIYABC summaries and LDA axes	5.01	4.66	4.18

Note. Same comments as in Table 1.

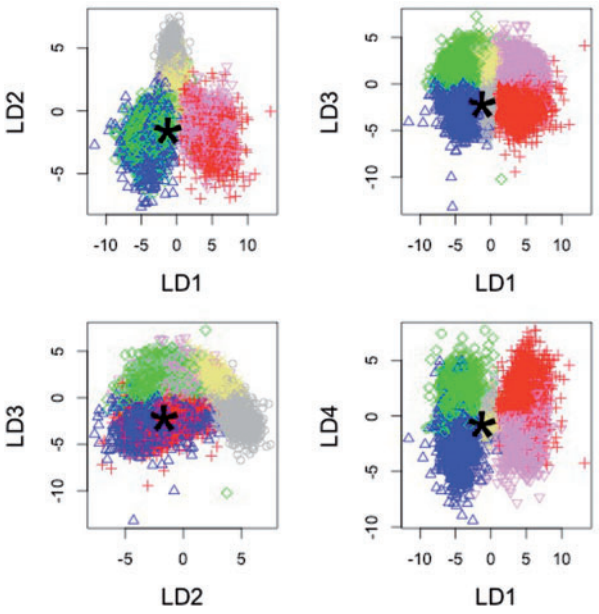

Fig. 2. Human SNP data: projection of the reference table on the first four LDA axes. Colors correspond to model indices. The location of the additional datasets is indicated by a *large black star*

3.4 Practical recommendations regarding the implementation of the algorithms

We develop here several points, formalized as questions, which should help users seeking to apply our methodology on their dataset for statistical model choice.

Are my models and/or associated priors compatible with the observed dataset?

This question is of prime interest and applies to any type of ABC treatment, including both standard ABC treatments and treatments based on ABC RF. Basically, if none of the proposed model - prior combinations produces some simulated datasets in a reasonable vicinity of the observed dataset, it is a signal of incompatibility and we consider it is then useless to attempt model choice inference. In such situations, we strongly advise reformulating the compared models and/or the associated prior distributions to achieve some compatibility in the above sense. We propose here a visual way to address this issue, namely through the simultaneous projection of the simulated reference table datasets and of the observed dataset on the first LDA axes. Such a graphical assessment can be achieved using the R package `abcrf` associated with this paper. In the LDA projection, the observed dataset need be located reasonably within the clouds of simulated datasets (see Fig. 2 as an illustration). Note that visual representations of a similar type (although based on PCA) as well as computation for each summary statistics and for each model of the probabilities of the observed values in the prior distributions have been proposed by Cornuet *et al.* (2010) and are already automatically provided by the DIYABC software.

Did I simulate enough datasets for my reference table?

A rule of thumb is to simulate between 5000 and 10 000 datasets per model among those compared. For instance, in the example dealing with Human population history (Section 3.3), we have simulated a total of 50 000 datasets from six models (i.e. about 8300 datasets per model). To evaluate whether or not this number is sufficient for RF analysis, we recommend to compute global prior error rates from both the entire reference table and a subset of the reference table (for instance from a subset of 40 000 simulated datasets if the reference table includes a total of 50 000 simulated datasets). If the prior error rate value obtained from the subset of the reference table is similar, or only lightly higher, than the value obtained from the entire reference table, one can consider that the reference table contains enough simulated datasets. If a substantial difference is observed between both values, then we recommend an increase in the number of datasets in the reference table. For instance, in the Human population history example, we obtained prior error rate values of 4.22% and 4.18% when computed from a subset of 40 000 simulated datasets and the entire 50 000 datasets of the reference table, respectively. In this case, the benefit of producing more simulated dataset in the reference table seems negligible.

Did my forest grow enough trees?

According to our experience, a forest made of 500 trees usually constitutes an interesting trade-off between computation efficiency and statistical precision (Breiman, 2001). To evaluate whether or not this number is sufficient, we recommend to plot the estimated values of the prior error rate and/or the posterior probability of the best model as a function of the number of trees in the forest. The shapes of the curves provide a visual diagnostic of whether such key quantities stabilize when the number of trees tends to 500. We provide illustrations of such visual representations in the case of the example dealing with Human population history in Figure 3. Such a graphical assessment can be achieved using the R package abcrf associated with this paper

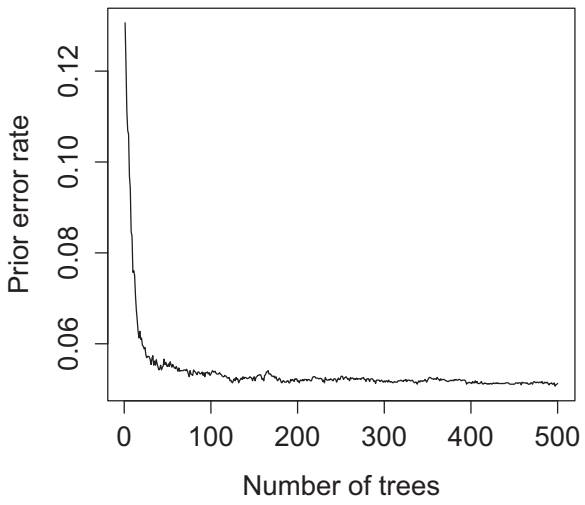

Fig. 3. Human SNP data: evolution of the ABC-RF prior error rate when $N_{\text{ref}} = 50\,000$ with respect to the number of trees in the forest

How do I set N_{boot} and n_{try} for classification and regression?

For a reference table with up to 100 000 datasets and 250 summary statistics, we recommend keeping the entire reference table, that is, $N_{\text{boot}} = N$ when building the trees. For larger reference tables, the value of N_{boot} can be calibrated against the prior error rate, starting with a value of $N_{\text{boot}} = 50\,000$ and doubling it until the estimated prior error rate is stabilized. For the number n_{try} of summary statistics sampled at each of the nodes, we see no reason to modify the default number of covariates n_{try} which is chosen as $\sqrt{d}$ for classification and $d/3$ for regression when d is the total number of predictors (Breiman, 2001). Finally, when the number of summary statistics is lower than 15, one might reduce N_{boot} to $N/10$.

4 Discussion

This article is purposely focused on selecting a statistical model, which can be rephrased as a classification problem trained on ABC simulations. We defend here the paradigm shift of assessing the best fitting model via a RF classification and in evaluating our confidence in the selected model by a secondary RF procedure, resulting in a different approach to precisely estimate the posterior probability of the selected model. We further provide a calibrating principle for this approach, in that the prior error rate provides a rational way to select the classifier and the set of summary statistics which leads to results closer to a true Bayesian analysis.

Compared with past ABC implementations, ABC-RF offers improvements at least at four levels: (i) on all experiments we studied, it has a lower prior error rate; (ii) it is robust to the size and choice of summary statistics, as RF can handle many superfluous statistics with no impact on the performance rates (which mostly depend on the intrinsic dimension of the classification problem (Biau, 2012; Scornet *et al.*, 2015), a characteristic confirmed by our results); (iii) the computing effort is considerably reduced as RF requires a much smaller reference table compared with alternatives (i.e. a few thousands versus hundred thousands to billions of simulations) and (iv) the method is associated with an embedded and error-free

evaluation which assesses the reliability of ABC-RF analysis. As a consequence, ABC-RF allows for a more robust handling of the degree of uncertainty in the choice between models, possibly in contrast with earlier and over-optimistic assessments.

Because of a massive gain in computing and simulation efforts, ABC-RF will extend the range and complexity of datasets (e.g. number of markers in population genetics) and models handled by ABC. In particular, we believe that ABC-RF will be of considerable interest for the statistical processing of massive SNP datasets whose production rapidly increases within the field of population genetics for both model and non-model organisms. Once a given model has been chosen and confidence evaluated by ABC-RF, it becomes possible to estimate parameter distribution under this (single) model using standard ABC techniques (Beaumont *et al.*, 2002) or alternative methods such as those proposed by Excoffier *et al.* (2013).

Acknowledgements

The authors are grateful to the referees for their supportive and constructive comments throughout the editorial process. 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 grateful to Gérard Biau for his help about the asymptotics of random forests. Some parts of the research were conducted at BIRS, Banff, Canada, and the authors (P.P. and C.P.R.) took advantage of this congenial research environment.

Funding

This research was partly supported by the ERA-Net BiodivERsA2013-48 (EXOTIC), with the national funders FRB, ANR, 25 MEDDE, BELSPO, PT-DLR and DFG, part of the 2012–2013 BiodivERsA call for research proposals. This work was also supported by the Labex NUMEV.

Conflict of Interest: none declared.

References

Aeschbacher,S. *et al.* (2012) A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics*, **192**, 1027–1047.

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

Barnes,C. *et al.* (2012) Considerate approaches to constructing summary statistics for ABC model selection. *Stat. Comput.*, **22**, 1181–1197.

Beaumont,M. (2008) Joint determination of topology, divergence time and immigration in population trees. In: Matsumura,S. *et al.* (eds.) *Simulations, Genetics and Human Prehistory*. McDonald Institute Monographs, McDonald Institute for Archaeological Research, Cambridge, pp. 134–154.

Beaumont,M. (2010) Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.*, **41**, 379–406.

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

Berger,J. (1985) *Statistical Decision Theory and Bayesian Analysis*, second edition. Springer-Verlag, New York.

Bertorelle,G. *et al.* (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol. Ecol.*, **19**, 2609–2625.

Biau,G. (2012) Analysis of a random forest model. *J. Machine Learn. Res.*, **13**, 1063–1095.

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

Blum,M. *et al.* (2013) A comparative review of dimension reduction methods in approximate Bayesian computation. *Stat. Sci.*, **28**, 189–208.

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

Chan,Y. *et al.* (2014) Detecting concerted demographic response across community assemblages using hierarchical approximate Bayesian computation. *Mol. Biol. Evol.*, **31**, 2501–2515.

Choisy,M. *et al.* (2004) Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Mol. Ecol.*, **13**, 955–968.

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

Cornuet,J.-M. *et al.* (2010) Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics*, **11**.

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**, 1187–1189.

Csillèry,K. *et al.* (2010) Approximate Bayesian computation (ABC) in practice. *Trends Ecol. Evol.*, **25**, 410–418.

Devroye,L. *et al.* (1996) *A Probabilistic Theory of Pattern Recognition,* volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York.

Didelot,X. *et al.* (2011) Likelihood-free estimation of model evidence. *Bayesian Anal.*, **6**, 48–76.

Estoup,A. *et al.* (2012) Estimation of demo-genetic model probabilities with approximate Bayesian computation using linear discriminant analysis on summary statistics. *Mol. Ecol. Resour.*, **12**, 846–855.

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

Fearnhead,P. and Prangle,D. (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **74**, 419–474.

Grelaud,A. *et al.* (2009) Likelihood-free methods for model choice in Gibbs random fields. *Bayesian Anal.*, **3**, 427–442.

Hastie,T. *et al.* (2009) *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, second edition. Springer-Verlag, New York.

Lombaert,E. *et al.* (2011) Inferring the origin of populations introduced from a genetically structured native range by approximate Bayesian computation: case study of the invasive ladybird Harmonia axyridis. *Mol. Ecol.*, **20**, 4654–4670.

Marin,J. *et al.* (2012) Approximate Bayesian computational methods. *Stat. Comput.*, **22**, 1167–1180.

Marin,J. *et al.* (2014) Relevant statistics for Bayesian model choice. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **76**, 833–859.

Prangle,D. *et al.* (2014) Semi-automatic selection of summary statistics for ABC model choice. *Stat. Appl. Genet. Mol. Biol.*, **13**, 67–82.

Pritchard,J. *et al.* (1999) Population growth of human Y chrom osomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.*, **16**, 1791–1798.

Robert,C. (2001) *The Bayesian Choice, second edition*. Springer-Verlag, New York.

Robert,C. *et al.* (2011) Lack of confidence in ABC model choice. *Proc. Natl Acad. Sci. USA*, **108**, 15112–15117.

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

Scornet,E. *et al.* (2015) Consistency of random forests. *Ann. Stat.*, **43**, 1716–1741.

Stoehr,J. *et al.* (2015) Adaptive ABC model choice and geometric summary statistics for hidden Gibbs random fields. *Stat. Comput.*, **25**, 129–141.

Tavaré,S. *et al.* (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 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. *Mol. Biol. Evol.*, **29**, 3653–3667.

Toni,T. *et al.* (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, **6**, 187–202.