Peer Community Journal

Section: Mathematical & Computational Biology

Research article

Published
2026-05-07

Cite as
François Rousset, Raphaël Leblois, Arnaud Estoup and Jean-Michel Marin (2026) *A new iterative framework for simulation-based population genetic inference with improved coverage properties of confidence intervals*, Peer Community Journal, 6: e43.

Correspondence
francois.rousset@umontpellier.fr

Peer-review
Peer reviewed and recommended by PCI Mathematical & Computational Biology, https://doi.org/10.24072/pci.mcb.100426

A new iterative framework for simulation-based population genetic inference with improved coverage properties of confidence intervals

François Rousset [1], Raphaël Leblois [2], Arnaud Estoup [2], and Jean-Michel Marin [3]

Volume **6** (2026), article e43

https://doi.org/10.24072/pcjournal.721

Abstract

Simulation-based methods such as approximate Bayesian computation (ABC) are widely used to infer the evolutionary history of populations from molecular genetic data. We describe and evaluate a new iterative method of statistical inference about model parameters, which revisits the idea of inferring a likelihood surface using simulation when the likelihood function cannot be evaluated. It is based on combining the random forest machine learning method, and multivariate Gaussian mixture (MGM) models, in an effective inference workflow, here used to fit models with up to 15 variable parameters. In addition to the traditional assessment of precision in terms of bias and mean square error, we also evaluate the coverage of confidence intervals. The method is compared with approximate Bayesian computation using random forests (ABC-RF), a non-iterative method sharing some technical features with the proposed approach, across scenarios of historical demographic inference from population genetic data. It is also compared to another iterative method, sequential neural likelihood estimation (SNLE). These comparisons highlight the importance of an iterative workflow for exploring the parameter space efficiently. For equivalent simulation effort of the data-generating process, the new summary-likelihood method provides intervals whose coverage is better controlled than the marginal coverage of intervals provided by ABC with random forests, and than generally reported for ABC methods. The iterative workflow can also yield greater improvements in estimator precision when larger datasets are used.

[1]ISEM, Univ Montpellier, CNRS, IRD, 34095 Montpellier, France, [2]CBGP, INRAE, CIRAD, IRD, Institut Agro, Univ Montpellier, 34980 Montferrier-sur-Lez, France, [3]IMAG, Univ Montpellier, 34095 Montpellier, France

Peer Community Journal is a member of the Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/

e-ISSN 2804-3871

Introduction

The likelihood function is a classical component of efficient statistical inference methodologies. In instances where the likelihood function cannot be computed in a reasonable time frame, alternative statistical methods can be employed. These include moment-matching techniques when the analytical results for moments of a response variable are known, or approaches based on the simulation of the putative data-generating process. Among the latter simulation-based methods, approximate Bayesian computation (ABC) has been particularly developed, notably for applications in population genetics (Beaumont, 2010; Beaumont et al., 2002; Bertorelle et al., 2010; Tavaré et al., 1997), but also in diverse other fields of biology (Schälte and Hasenauer, 2020) and beyond (e.g., Akeret et al., 2015; Sisson et al., 2019). In molecular evolutionary studies, ABC has been widely used to infer the past history of migration, founding events, invasion routes and introgressions among populations (e.g., Fraimout et al., 2017), but also selection pressures (e.g., Nakagome et al., 2015) or genomic rearrangement rates (e.g., Moshe et al., 2022).

The idea of estimating the likelihood function by simulation is less developed, although it goes back at least to Diggle and Gratton (1984). Some recent machine-learning methods, such as sequential neural likelihood estimation (SNLE, Papamakarios et al., 2019), incorporate this step, although they perform inference via the posterior distribution rather than the likelihood function. Diggle and Gratton's more strictly likelihood-based approach has had limited follow-up in the form of practically implemented likelihood methods for more complex models, particularly when the data are represented by many descriptive statistics (i.e., "summary statistics"). The "synthetic-likelihood" method (Wood, 2010) may be seen as derived from Diggle and Gratton (1984), except that it assumes the summary statistics to have a multivariate normal distribution for each parameter value. An extension of Diggle and Gratton's approach not making this assumption has been described by Rousset et al. (2017). As this is a likelihood inference using the information retained in the summary statistics, we refer to this method as "summary-likelihood (SL) inference". Summary likelihood is not full-data likelihood but is still a form of likelihood, which one can evaluate if the full data have been thrown away and only some summary statistics have been retained. When the number of summary statistics is large, inferring the likelihood surface becomes more tractable by applying dimension-reduction techniques. Various machine learning methods can be employed for this purpose; here, we adopt the random forest approach (Breiman, 2001; Geurts et al., 2006). This method is also used in ABC-RF (Collin et al., 2021; Pudlo et al., 2016; Raynal et al., 2019).

The aforementioned methods all address the issue of inferring the parameters θ of a hypothetical data-generating process, utilizing simulations of this process for varying values of the vector θ. In Rousset et al. (2017) as in Diggle and Gratton (1984) and Wood (2010), values of θ are drawn and for each draw, the distribution of summary statistics is estimated using a moderately large number of simulations of the data-generating process. In this study, we present a more efficient summary-likelihood inference workflow that requires significantly fewer simulations to provide accurate inferences. This iterative workflow facilitates the inference of more parameter-rich models, even when simulating observations is computationally expensive. The new workflow employs a single simulated observation for each drawn value of θ. This approach aligns with the general methodology of ABC methods, where each drawn θ and the corresponding summary statistics for each simulated observation form one line of a table commonly known as the reference table. The new iterative method has already been used to fit a simple two-parameter model of evolution of experimental populations by Laugier et al. (2025).

The idea of estimating the likelihood function for summary statistics from an ABC reference table already appears in the ABC literature (e.g., Fan et al., 2013; Papamakarios et al., 2019; Rubio and Johansen, 2013), but this has not yet led to accessible and validated software implementations of summary-likelihood inference. Indeed, a potentially substantial drawback of this idea is that inferring the likelihood surface in high dimensions is a complex task. For inference of likelihoods or joint full posterior distributions, the importance of iterative (or "sequential") methods of construction of the reference table has been repeatedly stressed (e.g. Blum and François, 2010; Cranmer et al., 2019; Del Moral et al., 2006; Lueckmann et al., 2021; Toni et

al., 2009). Modern variants employ neural networks for density estimation (Blum and François, 2010; Greenberg et al., 2019; Lueckmann et al., 2017; Papamakarios and Murray, 2016; Papamakarios et al., 2019). However, in practice, non-iterative ABC methods (e.g., Beaumont et al., 2002; Blum and François, 2010; Pudlo et al., 2016) remain widely used. Such methods can, in principle, enable amortized inference, whereby a single reference table (or even a single trained neural network) can be used to analyze many datasets. This can yield substantial computational savings when data are repeatedly generated under the same design (e.g., Zammit-Mangion et al., 2025). However, in many real-world applications of the non-iterative ABC methods described above, datasets differ enough in design or characteristics that such amortization is not feasible.

ABC methods are employed for the purpose of inferring posterior distributions for parameters, given a prior distribution for θ. Subsequently, credible intervals can be derived from the posterior distribution. As an alternative approach, likelihood-ratio based confidence intervals can be computed when a likelihood surface is inferred. The intervals returned by the different methods can be compared in terms of coverage, which is the probability that the interval contains a data-generating parameter value. It is anticipated that credibility intervals will provide correct coverage on average across the prior distribution of the parameters, while confidence intervals are constructed to ensure correct coverage for any possible value of the parameters: for different perspectives on these two concepts (credibility and confidence intervals) see for example Neyman (1977) or Casella and Berger (2002). However, it appears that a significant number of ABC methods are unable to effectively control the prior-averaged coverage of credibility intervals. For instance, Raynal et al. (2019) evaluated coverage of credibility intervals for ABC-RF as well as for basic rejection ABC and its elaborations using adjusted local regression (Beaumont et al., 2002), ridge regression (Blum et al., 2013), or neural networks (Blum and François, 2010). They found that ABC-RF intervals were conservative, with 100% coverage for 95% nominal probability. Additionally, they found that the coverage of other ABC methods was contingent upon a required rejection threshold, which is challenging to accurately set a priori. Consequently, they also generated anti-conservative intervals for certain threshold values. As observed by Hermans et al. (2022), the issue of coverage control has been seldom addressed in earlier machine learning literature. The deep learning-based methods they examined were all found to produce anti-conservative intervals, indicating a clear need for improvement in the coverage of intervals in simulation-based inference methods. Consistent with this, coverage of credibility intervals has received increasing attention in recent work (e.g., Frazier et al., 2024), with some studies explicitly distinguishing the coverage properties of Bayesian credible intervals from those of frequentist confidence intervals (Dalmasso et al., 2024).

In this paper, we present an evaluation of the performance of our new summary-likelihood workflow using simulation scenarios of inference of demographic history by analysis of population genetic data. A toy simulation scenario completes the tests of the method. Compared with the non-iterative ABC-RF method, our results underscore the value of an iterative workflow for improving inference when accurate exploration of the parameter space is critical. We show that the new workflow allows more uniform control of the coverage of intervals than previously reported for ABC methods. The main deviations from ideal control appear due to lack of information about parameters in the data, a common feature of demo-genetic inferences, and of other fields of application of simulation-based inference (e.g., Auger-Méthé et al., 2021; Daly et al., 2018; Fan et al., 2019). We also provide some comparison with the SNLE workflow implemented in the `sbi` package (Tejero-Cantero et al., 2020). This iterative workflow uses highly efficient neural network methods to infer likelihood surfaces and posterior densities, and provides credibility intervals. For problems of higher dimension than those considered here (up to 15 parameters), this method appears faster. However, our simulations show that, in the population-genetic scenarios examined here, the coverage of the intervals provided by SNLE is not always well controlled.

Material and Methods

The inference workflow

Starting from a reference table built from a limited number of simulations, an initial estimation of the summary-likelihood surface is derived. Then, new parameter points are sampled with greater probability in regions of high inferred likelihood, with the objective of more accurately inferring the likelihood surface in such regions. These sampling and likelihood-surface estimation steps are repeated iteratively to augment the reference table and to obtain progressively more precise inferences of the likelihood surface.

In this Section we provide a first description of these steps, introducing terminology and notation. For clarity, the simulation results included in the reference table are called "samples" to distinguish them from the "data" to be analyzed, whether the latter are real data or simulated ones. In the same way, we distinguish the data-generating parameters values (denoted $\boldsymbol{\theta}^{\dagger}$), which are not information used by the inference workflow, from the sample-generating parameters, which are essential information included in the reference table. Following well-established notation that we will repeatedly use below, probability densities will be denoted as P, with indices denoting the random variables whose values are the arguments of the function. In particular, $\mathrm{P}_{Y;\boldsymbol{\Theta}}$ denotes the density of Y values (where these values may represent the data, or some simulated sample) as function of parameter values $\boldsymbol{\Theta}$, and $\mathrm{P}_{Y,\boldsymbol{\Theta}}$ denotes the joint density of Y and $\boldsymbol{\Theta}$, under the assumption that $\boldsymbol{\Theta}$ is sampled from some distribution. $\mathrm{P}_{Y;\boldsymbol{\Theta}}(S;\boldsymbol{\theta})$ is thus the probability (or probability density) of a sample S for a given parameter vector $\boldsymbol{\theta}$.

Inferring the likelihood from a reference table. The likelihood $L(\boldsymbol{\theta};\mathcal{D})$ of $\boldsymbol{\theta}$ given observed data $\mathcal{D}$ is generally defined, up to a constant factor, as $\mathrm{P}_{Y;\boldsymbol{\Theta}}(\mathcal{D};\boldsymbol{\theta})$, viewed as a function of the parameters for fixed data. Given a distribution $i_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ for the parameters, the joint distribution of samples S and parameters $\boldsymbol{\theta}$ can be written $\mathrm{P}_{Y,\boldsymbol{\Theta}}(S,\boldsymbol{\theta}) = i_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\mathrm{P}_{Y;\boldsymbol{\Theta}}(S;\boldsymbol{\theta})$, and the likelihood can be written as

$$L(\boldsymbol{\theta};\mathcal{D}) = \mathrm{P}_{Y,\boldsymbol{\Theta}}(\mathcal{D},\boldsymbol{\theta})/i_{\boldsymbol{\Theta}}(\boldsymbol{\theta}). \tag{1}$$

Accordingly, to estimate the likelihood function from the reference table, one can first estimate a joint density $\mathrm{P}_{Y,\boldsymbol{\Theta}}$ of samples and parameters, from their realized joint distribution in the reference table. One can then divide the value in $Y = \mathcal{D}$ of this estimated joint density function by an estimate of the marginal parameter density function $i_{\boldsymbol{\Theta}}$, deduced from the joint density, to obtain the likelihood function. Since the term "instrumental distribution" is commonly used to refer to a probability distribution used in an algorithm for the estimation of a target quantity, such as a likelihood or posterior distribution, we thus view the $\boldsymbol{\Theta}$ vectors in the reference table as samples from an instrumental distribution which is estimated conjointly with the distribution of summary statistics in the reference table.

As in Papamakarios et al. (2019), the iterative proposal mechanism for new parameter points does not asymptotically bias likelihood learning. In the present workflow, new parameters points are sampled in each iteration in order to preferentially fill the region of parameter space with a high likelihood (as detailed below, Section "Refinements of likelihood surface inference through iterations"). The joint and marginal densities, and the likelihood, are re-estimated in each iteration. After a few iterations, the inferred $i_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is usually quite different from the distribution of parameters used to construct the initial reference table.

From raw statistics to projected statistics. Here, as in most applications of ABC, the information provided to the inference method is typically a vector of statistics $\mathbf{u}(\mathcal{D})$, summarizing higher-dimensional observed data $\mathcal{D}$ which is the information available to the analyst. For example, the largest genetic datasets considered in our simulation study involve genetic information at 10,000 genetic markers from 120 individuals, which is summarized by 130 statistics; but in some applications $\mathbf{u}(\mathcal{D})$ may just be a vectorial representation of the full $\mathcal{D}$. The simulated samples for each drawn $\boldsymbol{\theta}$ must then also be expressed as a vector of summary statistics $\mathbf{u}(S)$. As with ABC, it is up to the user to provide statistics that are informative for a given inference problem.

In our SL method, the potentially large number of *raw* summary statistics describing the data or the simulated samples in the reference table (up to 130 summary statistics in our simulations), is typically reduced by a *projection step* to a smaller number of *projected* summary statistics $\mathbf{p}[\mathbf{u}(D)]$, in order to reduce the dimension of the joint distribution, of parameters and of statistics, which is estimated. "Projection" here refers to the idea of non-linear projection to a lower-dimensional space. We obtain the projected summary statistics by performing a non-parametric regression of each parameter θ_j on the raw summary statistics. This regression is performed using a variant of the random-forest method, combining features of its original version (Breiman, 2001) and of "extremely-randomized trees" (Geurts et al., 2006). Various other projection methods could be used in SL, provided they provide predictions that avoid overfitting. The random-forest method was retained for the same reasons as in ABC-RF: it is fast, efficient, does not require ad-hoc adjustment of control settings for each application, and the out-of-bag predictions can be used to easily avoid overfitting (e.g., Breiman, 2001, p.11; Hastie et al., 2009). It thus fulfills the need for a method which can be applied automatically in any inference problem (Raynal et al., 2019). This choice also facilitates comparison with ABC-RF, as differences between the performance of ABC-RF and summary likelihood cannot be attributed to the choice of widely different projection methods.

In our use of random forests, a joint vector $\mathbf{p}[\mathbf{u}(S)]$ of projected summary statistics is defined, for each simulated sample S, as the vector of predictions $\hat{\theta}_j(S)$ of each element θ_j of the $\boldsymbol{\theta}$ parameter vector by its random-forest regression. This reduces the number of statistics to the number of parameters, which is the minimum required for identifiability of the parameters. Additional statistics can be retained insofar as computations do not become impractically slow, but have not been considered here. Fearnhead and Prangle (2012) already advocated summarising a sample by the posterior expectations of each parameter given that sample, which can be approximated in practice using random-forest regression. Here, however, the random-forest regression targets posterior expectations under the instrumental distribution induced by the iterative algorithm, rather than under any fixed, pre-specified prior.

An initial reference table is thus constructed as follows. Parameter vectors are drawn from the initial instrumental distribution $i_{\boldsymbol{\Theta}}^{(0)}(\boldsymbol{\theta})$ for parameters, and one sample S (i.e., one realization of the biological and sampling process) is drawn for each parameter vector. Raw summary statistics $\mathbf{u}(S)$ are computed, and projected statistics $\mathbf{p}[\mathbf{u}(S)]$ are deduced from them for each such sample.

Estimating the summary-likelihood function. The likelihood function for projected statistics is written

$$L(\theta; \mathbf{p}[\mathbf{u}(\mathcal{D})]) = \mathrm{P}_{T,\boldsymbol{\Theta}}(\mathbf{p}[\mathbf{u}(\mathcal{D})], \boldsymbol{\theta}) / i_{\boldsymbol{\Theta}}(\boldsymbol{\theta}), \tag{2}$$

in terms of the joint density $\mathrm{P}_{T,\boldsymbol{\Theta}}$ of projected statistics and parameters.

To model this joint density from the reference table, we use by default a Multivariate Gaussian Mixture (MGM) model. MGMs have previously been used to infer posterior distributions of parameters (Bonassi et al., 2011), and a related approach, Gaussian Locally Linear Mapping (GLLiM, Deleforge et al., 2014), has also been used in simulation-based inference (Häggström et al., 2024). We also considered Masked Autoregressive Flows (MAFs, Papamakarios et al., 2017), a deep-learning approach for estimating unconditional or conditional probability densities. In our first attempts to use MAFs, they were trained directly on the final reference table. Because this is substantially slower than fitting MGMs, MAFs did not appear suitable as a systematic replacement for MGMs in our workflow. In practice, MAFs are more efficiently trained sequentially, updating the inferred density at each iteration by warm-starting from the previous iteration; this is the strategy used in SNLE, whose performance will be compared to summary-likelihood for a set of selected simulation scenarios.

Once a joint density estimate $\hat{\mathrm{P}}_{T,\boldsymbol{\Theta}}$ is obtained using MGM models, the likelihood for any $\boldsymbol{\theta}$ is estimated by

$$\hat{L}(\boldsymbol{\theta}; \mathbf{p}[\mathbf{u}(\mathcal{D})]) = \hat{\mathrm{P}}_{T,\boldsymbol{\Theta}}(\mathbf{p}[\mathbf{u}(\mathcal{D})], \boldsymbol{\theta}) / \hat{\imath}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}), \tag{3}$$

where, when the joint density estimation uses MGM models, the estimate $\hat{\imath}(\boldsymbol{\theta})$ of the instrumental density is easily deduced from the joint density estimate by marginalization of the latter

density over the projected statistics. One may think that such estimation is not needed, at least in the first iteration when the instrumental distribution typically has some simple known form. However, in subsequent iterations such knowledge is not available because the instrumental density (as automatically generated by rules discussed later) is implicit and complex.

When Masked Autoregressive Flows are used, the likelihood function can be directly estimated by training a conditional MAF (Papamakarios et al., 2017, Section 3.4; Papamakarios et al., 2019), in order to learn the conditional density of samples given any $\boldsymbol{\theta}$ values, on the joint distribution of parameters and statistics in the reference table.

Maximum summary-likelihood estimates (summary-MLEs or MSLEs, denoted $\hat{\boldsymbol{\theta}}$) are deduced from the inferred likelihood surface, by numerical maximization with respect to $\boldsymbol{\theta}$ of the estimated likelihood as given by eq. 3. Summary-likelihood ratio tests (summary-LRTs) are also deduced from the likelihood surface. In particular, when testing a value θ_i of a given element i of $\boldsymbol{\theta}$, we compute the constrained maximum of the estimated summary-likelihood surface given θ_i (yielding the constrained maximum summary-likelihood estimates $\hat{\boldsymbol{\theta}}_{\theta_i}$), and compare it to the global summary-likelihood maximum. Thus, we use the profile likelihood (e.g., Davison, 2003, Section 4.5.2) for all likelihood-ratio tests, with test statistic

$$W = 2\left(\log\left[L\left(\hat{\boldsymbol{\theta}}; \mathbf{p}[\mathbf{u}(\mathcal{D})]\right)\right] - \log\left[L\left(\hat{\boldsymbol{\theta}}_{\theta_i}; \mathbf{p}[\mathbf{u}(\mathcal{D})]\right)\right]\right) \tag{4}$$

and p-value given by the tail probability $\mathrm{P}(X > W)$ for a χ^2-distributed variable with number of degrees of freedom equal to the number of fixed parameters in $\hat{\boldsymbol{\theta}}_{\theta_i}$, i.e. 1 in the present applications of the test.

Refinements of likelihood surface inference through iterations. In each iteration, new parameter values are drawn, simulations of the process are performed and added to the reference table, and the above steps of the workflow are repeated on the incremented reference table. The sampling of parameters should still permit further exploration to prevent entrapment in the initial high-likelihood region, which may be disjoint from the final one (as will be illustrated by the results on the 7-parameter human admixture scenario). The implemented sampling procedure further allows exploration of the parameter space beyond the previously sampled ranges, and users can specify "absolute" bounds that should not be exceeded during such exploration. The software also allows users to specify not only ranges for each parameter, but also arbitrary additional constraints on any combination of parameters (this is used in some of our simulations).

The rules for drawing n_{i+1} new parameter vectors after iteration i are detailed in the Supplementary Information (Section S.1.1). They are defined to preferentially fill the region of parameters with a given minimum likelihood ratio $l_{\min}$ relative to the current summary-MLEs. In practice, $l_{\min}$ was set to the 95% threshold for two-dimensional confidence regions. The sampling also allows exploration beyond the boundaries of this "top" region. The sampling step within the "top" aims to fill it uniformly rather than in proportion to the likelihood ratio.

The iterative workflow allows the projections to be re-computed. However, random-forest computations on large tables take time, so some elaborations have been implemented to shorten them. First, for large reference tables, they use a subset of the reference table, mostly defined from points identified as belonging to the top of the likelihood surface (see Supplementary Material for details). Second, projections are not re-computed when more than 90% of the selected points were already used in the previous computation of the projections. The easiest way to accelerate inferences in parameter-rich models without substantially compromising performance may be to reduce this threshold so that projections are re-computed less often.

Implementation. All the above-described steps of summary-likelihood inference have been implemented in an automated workflow in the `Infusion` R package (Rousset, 2025), which calls the `ranger` R package (Wright and Ziegler, 2017) for random-forest methods, and the `Rmixmod` R package (Lebret et al., 2015) for MGM modeling.

Control of inference workflow

Our workflow potentially depends on many control parameters, but we set default values for all of them, which were used for all simulations in this paper, unless mentioned otherwise. These values, as described here and further detailed in Supplementary Section S.1, are automatically selected by our implemented procedures unless users specifically request different values.

In particular, for fitting n_p parameters to a given dataset, a final number of $1000(3n_p - 1)$ sample simulations are run. Supplementary Table S.1 illustrates computations times for single datasets for different simulation designs, and Supplementary Section S.1.6 provides further details on how the number of samples added at each iteration is controlled. Smaller references tables appear to lead to degraded performance (one example being given in Supplementary Table S.15) and are therefore not recommended. Larger tables may require substantial increases in computation times, which may be a small concern when only one dataset is analyzed, but would be unpractical when inferences are performed on series of hundreds of simulated datasets.

While the final size of the reference tables was predefined in our simulation study, alternative and more adaptive stopping rules could be used, based in particular on two criteria implemented in the R package. The first criterion is the estimated precision of the log profile likelihood ratio statistics at the bounds of the inferred confidence intervals. A bootstrap procedure is implemented to evaluate root mean square errors (RMSEs) of prediction of these log-ratio statistics. It is thus possible to request termination of iterations when these RMSEs, averaged over the different interval bounds, or when all such RMSEs, are below a certain threshold. A second criterion is the comparison of the distribution of samples simulated in two ways: samples from the simulator of the process being inferred, versus samples from the inferred distribution. These distributions can be compared as in a classifier two-sample test (Lopez-Paz and Oquab, 2017), by training a classifier to assign samples to either distribution: the better the inferred distribution, the lower the classification performance. This has previously been used to compare different methods of inference of posterior distributions (Lueckmann et al., 2021), but can be used here to compare the distributions of samples given specific values of the inferred parameters, such as their MSLEs.

SNLE inference. To perform SNLE inference, we used the Python package `sbi` version 0.25.0, with default controls for training the MAF neural density estimator. In particular, the flow consisted of five autoregressive transformations, each parameterized by a Masked Autoencoder for Distribution Estimation (MADE)-type neural network (Germain et al., 2015) with hidden layers of size 50 and two blocks. Dropout and batch normalization were disabled. The autoregressive networks used ReLU activation functions. Between successive transforms, random permutations were applied to increase flexibility of the flow.
Posterior samples were obtained using the default Markov chain Monte Carlo (MCMC) sampler implemented in `sbi`. This corresponds to slice sampling with multiple parallel chains. The default configuration uses 20 chains, with 100 warmup steps per chain and thinning factor equal to one. Final reference tables had the same size as in our default summary-inference workflow, and we ran ten rounds of the sequential procedure with an equal number of simulated samples per round (Lueckmann et al., 2021, Appendix Section A.4).

Design of simulation study

Our simulation study is mainly based on two scenarios of demographic history of populations which have been considered in previous developments of ABC with random forests. However, we also consider 15-parameter toy examples where the number of raw summary statistics is the number of parameters, so that projections are not needed and inferences are relatively fast.

Toy examples. In the toy examples, we estimate the covariance parameters of a 5-dimensional normal distribution. The data-generating values, and the parametrization of this model are detailed in Supplementary Section S.3.1. The summary statistics are defined as the 15 distinct elements of the observed covariance matrix of 50 draws from a multivariate normal distribution of dimension 5 with the given covariance matrix.

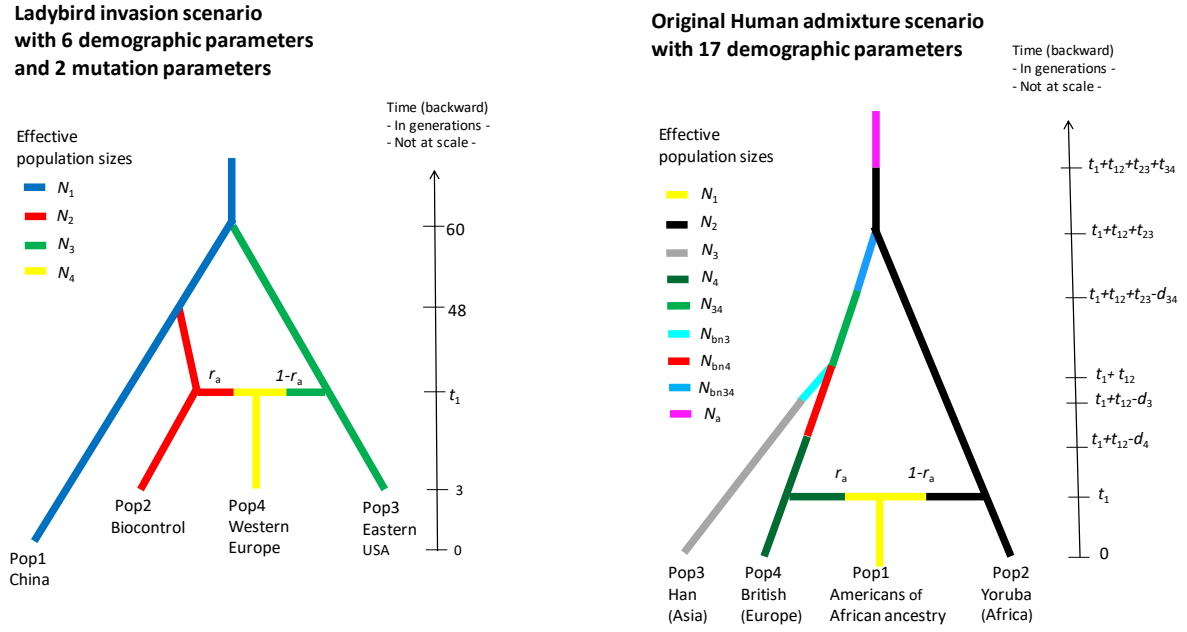

Figure 1 – The two scenarios of historical demography. Left: ladybird invasion scenario; right: Human admixture scenario. See Text for description of parameters.

Origin of invasive ladybird populations. This scenario (Figure 1, left) was already taken as an example for ABC-RF analyses by Pudlo et al. (2016). It is motivated by the invasion of the ladybird beetle *Harmonia axyridis* in Europe, and by a dataset of genotypes at 18 microsatellite loci, from a total of 126 individuals from four populations (Lombaert et al., 2011). The fitted model has 8 parameters: effective population sizes N_1 to N_4, admixture time t_1, admixture proportion r_a, and two mutation parameters $\bar{\mu}$ and $\bar{p}$. We estimated the composite parameters $\log(N_i\bar{\mu})$ instead of the N_is. Further details on model, data, and sampled parameter ranges are given in Supplementary Section S.4.1.

Human admixture scenario. For this second scenario (Figure 1, right), the model and data were also already considered in previous papers on ABC-RF (Collin et al., 2021; Raynal et al., 2019). It is a scenario of admixture between populations of European and African ancestry in America. More precisely, after an ancient demographic change at time t_4 in the ancestral African population, an out-of-Africa colonization event occurs at time t_3 that gives an ancestral out-of-Africa population which secondarily splits into one European population and one East Asian population at time t_2. Finally, a genetic admixture event occurs between populations of European and African ancestry in America at the time t_1, with a proportion r_a of European ancestry. Additional parameters N_{bn3}, N_{bn4}, N_{bn34} describe the population sizes during bottleneck events associated to colonization events along the population tree, and d_3, d_4, and d_{34} describe the duration of these bottlenecks. As detailed and justified in Supplementary Section S.5.1.1, we re-parametrized the model in terms of $t_{12} = t_2 - t_1$, $t_{23} = t_3 - t_2$, $t_{34} = t_4 - t_3$.

We explored different versions of this model, with 7 or 13 estimated parameters, by assuming that the values of other parameters were fixed and known. A real dataset of 5000 SNP markers (extended to 10000 in some simulations), genotyped in four human populations (The 1000 Genomes Project Consortium, 2012), defines the sampling design of the simulated data. The four populations include Yoruba (Africa), Han (East Asia), British (Europe) and American individuals of African ancestry.

Few parameters may be practically identifiable in the Human admixture scenario. For this reason, Raynal et al. (2019) and Collin et al. (2021) reported estimation performance for only a few parameters, with good performance only for the admixture rate r_a and for some composite parameters defined as ratios of population size to time interval parameters. However, such

composite parameters complicate the interpretation of the results. For example, the parameter space is no longer defined only by box constraints (i.e., by the range of each parameter), and marginal prior distributions are no longer uniform. This implies that low relative RMSEs, which are defined relative to the marginal range of each parameter, are not necessarily indicative that the parameter is practically identifiable.

For these reasons, we considered composite parameters only in one of two considered 13-parameter variants of the Human admixture scenario. More specifically, in this variant we considered the following composite parameters, sometimes referred to as "bottleneck intensity" parameters in the colonization/invasion literature: $b_3 = d_3/N_{\mathrm{bn3}}$, $b_4 = d_4/N_{\mathrm{bn4}}$, and $b_{34} = d_{34}/N_{\mathrm{bn34}}$. Estimating additional population size parameters specific to the branches of the tree that are affected by bottlenecks is difficult. In this simulation variant, we therefore do not try to estimate two such parameters, N_3 and N_4. Supplementary Table S.19 presents results for another 13-parameter variant without any composite parameters and with estimated N_3 and N_4.

Further details on model, data, and sampled parameter ranges are given in Supplementary Section S.5.1.1.

Assessment of confidence and credible intervals. The concept of confidence interval is based on the control of coverage whatever the data-generating parameter values $\boldsymbol{\theta}^{\dagger}$, rather than only on average over a prior distribution, so we assess coverage for given $\boldsymbol{\theta}^{\dagger}$ values, with only a few exceptions.

For fixed prior distributions, credibility intervals may become asymptotically equivalent to confidence intervals for "large samples", i.e., when the information contained in the parameters increases indefinitely with sample size (Lehmann and Casella, 1998, Section 6.8). Credibility intervals and confidence intervals may then be seen as asymptotic approximations to each other. But this asymptotic equivalence may fail when (i) coverage is evaluated conditionally for a parameter at the boundary of the prior distribution and (ii) for some definition of posterior intervals. In particular, intervals based on central quantiles of the posterior distribution ("central posterior intervals") will never cover the $\boldsymbol{\theta}^{\dagger}$ value when it is at a boundary.

This boundary effect should not be overlooked: (i) it may be widely ignored by practitioners (reports of performance focusing on good marginal coverage encouraging such ignorance), and (ii) it is a real concern in practice insofar as software (including ABC-RF) often report central posterior intervals, which leads to the exclusion of parameter values at range bounds from reported intervals, even when there is actually not enough information in the data to actually reject such values. For the following comparisons, we implemented the computation of highest posterior density (HPD) intervals from the ABC-RF output, which are expected to have better coverage at the boundaries. For unimodal posterior distributions, these are also the shortest credible intervals.

In principle, assessment of performance of confidence intervals calls for estimation of coverage for many different $\boldsymbol{\theta}^{\dagger}$ values. However, this would be unpractical here due to the high-dimensional parameter space and the computational cost of the simulations. Supplementary Section S.2 details the rules used to select several data-generating parameter values $\boldsymbol{\theta}^{\dagger}$; and Supplementary Table S.5 and Section S.5.1.2 show the values used for the ladybird invasion scenario and the Human admixture scenario, respectively.

Performance summaries

We evaluated the bias and root-mean-square error (RMSE) of the following point estimators: the summary-MLEs, and the mean and median of the posterior distributions for ABC-RF. Each Figure and Table for given data-generating parameter values is based on running the inference workflow on 400 simulated datasets (but 200 or 1000 datasets for some of the Supplementary Tables and Figures).

Parameter transformations were applied, mainly in order to homogenize the RMSEs of the estimators of transformed parameters in the population genetic scenarios. In practice this means that population size and mutation rate parameters were log-transformed, but for simplicity we also applied the following automatic transformation rule to other parameters based on their

explored ranges: if upper bound is ≥ 500, log(1+.) or log(.) transformation is used, depending on whether the lower bound is zero or not. All logarithms are base 10 logarithms. ABC-RF inferences used uniform priors on the transformed scales.

The RMSE summaries may not lead to clear conclusions. Which of ML and posterior estimates have lower RMSE depends on the location of data-generating values $\boldsymbol{\theta}^\dagger$ relative to the prior distribution used in a Bayesian inference (e.g., Casella and Berger, 2002, p. 333), so we do not expect RMSE criteria to systematically favor one class of estimators over the other when examined conditionally given $\boldsymbol{\theta}^\dagger$ values. Moreover, an efficient posterior estimate is expected to have lower RMSE on average when $\boldsymbol{\theta}^\dagger$ values are chosen randomly in their prior distributions. Yet, systematic departures from such theoretical expectations can occur if one of the methods does not locate well its point estimates, as a result of poorly inferring the likelihood function or the posterior distribution.

Bias and RMSE of estimators will be reported on a scale relative to the width of the explored ranges : for the ith element θ_i of the parameter vector, we use transformed parameter values ϑ_i as described above, and evaluate the following scaled means:

$$\text{relative bias} = \frac{\text{mean}(\hat{\boldsymbol{\vartheta}}_i) - \vartheta_i^\dagger}{\vartheta_{\text{U}i} - \vartheta_{\text{L}i}}, \tag{5}$$

$$\text{relative RMSE} = \frac{\text{RMSE}(\hat{\boldsymbol{\vartheta}}_i)}{\vartheta_{\text{U}i} - \vartheta_{\text{L}i}} \tag{6}$$

where $\hat{\boldsymbol{\vartheta}}_i$ is the vector of estimates of ϑ_i over the simulated datasets; and $\vartheta_i^\dagger$, $\vartheta_{\text{L}i}$ and $\vartheta_{\text{U}i}$ are the corresponding data-generating value, lower, and upper bound of explored range of the transformed ith parameter, respectively. With relative variance defined as $\text{Var}(\hat{\vartheta}_i)/(\vartheta_{\text{U}i} - \vartheta_{\text{L}i})^2$, the standard decomposition of mean-square error as variance plus squared bias holds for these relative values.

This is useful in particular to compare the RMSE of a summary-likelihood estimator to that of a non-identifiable parameter whose estimator would be uniformly distributed, and would then have a relative RMSE equal to $1/\sqrt{12} \approx 0.289$ if the data-generating value were the mid-range of the parameter (and higher otherwise). Another useful diagnostic pattern for poorly identifiable parameters is the relative variance of summary-MLE versus posterior estimates: the posterior mean estimator should have low variance, as it should approach the prior mean for all datasets. Thus, although an efficient posterior estimate is expected to have slightly lower RMSE on average over a prior distribution, a markedly lower ratio of the variance of posterior estimates to the variance of summary-MLEs instead suggests that the parameter is not well estimated by ABC-RF.

The actual conditional coverage of nominal 95% intervals will be reported for both the confidence and credibility intervals. For each element θ_i of the parameter vector, a more informative summary, not depending on a conventional level such as 95%, is the actual distribution of p-values of the test whose null hypothesis is that $\theta_i = \theta_i^\dagger$. This distribution should be uniform in ideal conditions, and will be reported in Figures. We used the profile likelihood function for such tests, as previously described.

We also evaluate the performance of intervals provided by parametric bootstrap simulations, where bootstrap samples are drawn from the inferred distribution of projected summary statistics as function of the parameters (see Supplementary Section S.1.7 for details about this procedure). Various definitions exist for bootstrap-based confidence intervals (e.g., Davison and Hinkley, 1997, Chapter 5). We thus compare the coverage of intervals defined by inverting profile-LRTs whose p-values are read from the χ^2 distribution, to two bootstrap intervals. One interval is defined by inverting the profile-LRTs whose p-values are read from the bootstrap distribution of the likelihood-ratio statistic. The second bootstrap interval uses a Bartlett correction (Bartlett, 1937): it is defined by inverting p-values read from the χ^2 distribution, for the profile likelihood ratio statistic corrected by the estimated mean of this statistic in the bootstrap samples. The definitions of the three confidence intervals are summarized in Table 1 and their coverage values are

Table 1 – Tests that yield p-values and profile LRTs used to define confidence intervals. W^* is the log profile likelihood ratio statistic (eq. 4) evaluated on a parametric bootstrap sample S^* from the fitted model. Realized coverage of implied confidence intervals at nominal level C is the frequency of the event $p > 1 - C$ over simulated data sets.

Variable	p-value definition
profLR	$p = \Pr(X > W)$ where $X \sim \chi^2_1$ and W is the log profile likelihood ratio statistic (eq. 4).
bootLR	p is the frequency of the event $W^* > W$ over different bootstrap values W^*.
BcorCI	$p = \Pr(X > W/\bar{W}^*)$ where $X \sim \chi^2_1$ and $\bar{W}^*$ is the mean value of W^* over bootstrap samples.

reported in later Tables as "profLR", "bootLR" and "BcorCI" respectively. We also considered percentile bootstrap intervals, but these appear to have less predictable coverage over the different simulations conditions.

We report the conditional coverage of two types of credibility intervals: the central posterior intervals computed by the `abcrf` R package (Marin et al., 2025), and highest posterior density (HPD) intervals also deduced from the ABC-RF output.

It is not always reasonable to require exact control of the coverage or distribution of p-values. If there is no information about a parameter in the data (i.e., if likelihood is flat with respect to it), the conclusion of the analysis should be that there is no information. This lack of information is represented by unbounded intervals with 100% coverage. On the other hand, weak information will result in intervals with more than the nominal coverage and in distributions of summary-LRTs that show a deficit of low p-values relative to the uniform distribution. These patterns readily occur in our simulations of demo-genetic scenarios.

Results

15-parameter toy model

In this model, we consider the estimation of the covariance matrix of a 5-dimensional multivariate normal distribution. It has two variants, one where all 15 parameters are identifiable, the other where only 5 variances are identifiable. As this example mainly serves as a first check of the inference workflow in a case with a relatively high number of parameters and where we know which parameters are identifiable and which are not, no variation of data-generating parameters nor any comparison with ABC are presented.

Performance summaries are detailed in Supplementary Table S.2 and distributions of p-values of summary-LRTs are shown in Figure 2. In each case 400 datasets were simulated and reference tables of 44000 samples were constructed independently for each dataset. Coverage is approximately controlled for identifiable parameters, mean coverage over the twenty rows of Table S.2 being 0.954 for the profile-likelihood ratio intervals and 0.96 for the Bartlett-corrected bootstrap intervals. There is expectedly a strong deficiency of low p-values for unidentifiable ones. Ideally, in the latter case, the distribution of p-values should be a step function in $p = 1$, and the observed non-stepwise distributions are the result of the fluctuations of the estimated surface compared to the exact one, these fluctuations being too small to result in low p-values. Overall, when parameters are not identifiable, the confidence intervals still have at least nominal coverage.

Ladybird invasion scenario (8 parameters)

The simulation results reported in Table 2 show that there is little information about admixture rate r_a. Posterior estimators for this parameter appear to have low bias only because the data-generating value $r_a^\dagger$ of r_a is close to the mean value of the prior distribution. This also explains why these posterior estimators have lower RMSE (conditional on $r_a^\dagger$) than the summary-MLE. There is also weak information about $\log(N_4\bar{\mu})$, t_1, and $\bar{p}$. Coverage of the likelihood-ratio

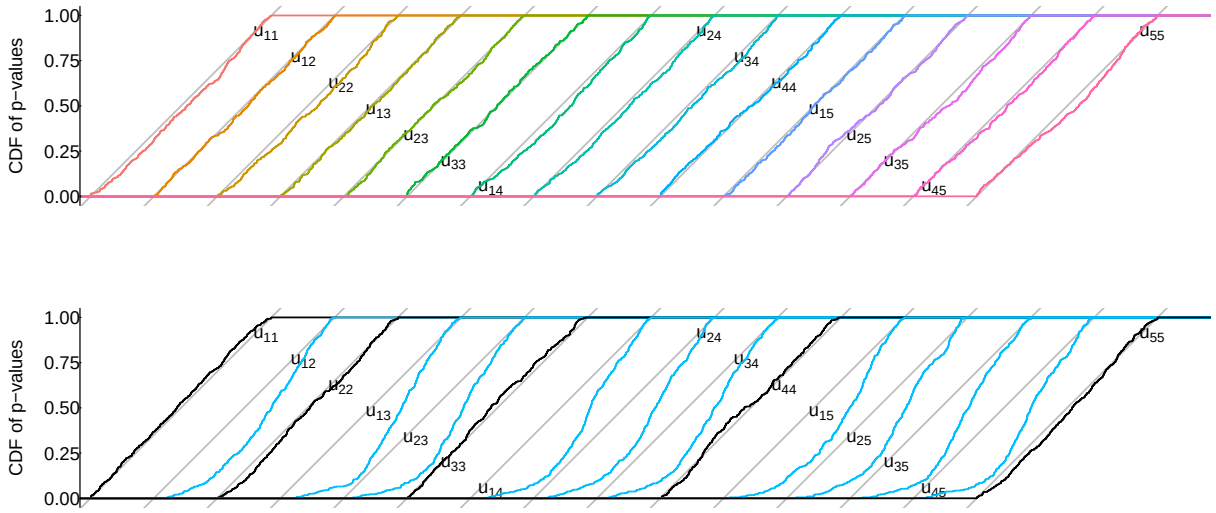

Figure 2 – Distributions of p-values of summary-LRTs in multivariate-normal toy example. The cumulative distributions of p-values are shown for the tests of each parameter of the covariance matrix (i.e., each element u_{ij} of its Cholesky factor, eq. S.2 in Supplementary Material). The grey diagonal lines represent uniform densities on [0,1], shifted in the x axis for visibility. Top: fully-identifiable model. Bottom: partially-identifiable model, with distributions for identifiable and non-identifiable parameters shown in black and blue, respectively. The distributions for non-identifiable parameters show a marked deficiency of low p-values (the cumulative distribution being under the diagonal), as expected since the inferred likelihood surface should be flat with respect to such parameters.

intervals is too high, 98% on average, for these four parameters, but also for the four other ones (97,05% on average, Table 2, "profLR" column). These excesses largely disappear in the bootstrap-corrected versions of the summary-LRTs, even for the parameters with low information (Table 2, "bootLR" and "BcorCI" columns, and Supplementary Figure S.1, bottom), with the Bartlett-corrected intervals having 95.4% coverage on average for the four better-estimated parameters and 95.5% for the four other ones.

The ABC-RF estimators of the different parameters often have higher bias, but there is no clear trend for RMSEs, consistently with the fact that the relative magnitude of the RMSE of ML vs. posterior estimates depends on where the data-generating values lie within the prior support. While the dependence of relative RMSE performance on location of the data-generating parameters in the prior distribution may explain some of the heterogeneity in apparent performance of the two methods, additional simulations show that more obscure specificities of the ABC-RF method contribute to this heterogeneity. Notably, the ABC-RF estimates for $\log(N_4\bar{\mu})$ are biased away from the prior mean (-0.5), their correlation with summary-ML estimates is low (0.215), and they have a much lower variance than the latter. Together, these results are not expected from general theory for posterior estimators, given the uniform priors and the fact that summary-MLE estimates do not exhibit a similar bias. Varying the $\log(N_4\bar{\mu})$ value, with other parameters being fixed, shows that the ABC-RF $\log(N_4\bar{\mu})$ estimator is biased away from the prior mean over much of the prior range (Supplementary Figure S.2), whereas a bias toward the prior mean would be more commonly expected (e.g., Casella and Berger, 2002, Section 7.3.4). Yet, it has lower averaged RMSE (Supplementary Table S.6) than the summary-ML estimator, as expected when comparing the prior-averaged performance of ML and posterior-mean estimators.

In Section S.4.2 of the Supplementary Information, we assessed performance of inferences for alternative data-generating parameter values derived from a preliminary fit of the actual data. These simulations exhibit even less information about parameters t_1 and r_a, and repeat the striking patterns of Figure S.2 for estimation of $\log(N_4\bar{\mu})$.

Table 2 – Performance summaries for the ladybird invasion scenario. Bias and root-mean-square error (RMSE) are reported for the maximum-summary likelihood estimates (MSLE), and the posterior mean (postEv) and posterior median (postMed) for ABC-RF. Coverage of confidence intervals with nominal 95% level is reported for summary-likelihood inference (profLR) as implied by the distribution of p-values shown in Figure S.1, for two forms of bootstrap correction described in the Text (profLR and BcorCI), for the central intervals provided by ABC-RF (postCI), and for highest posterior density intervals (HPD CI). Bold font is used to emphasize for each parameter the bias value minimal in absolute value, the minimum RMSE, and the coverage closest to 0.95.

	Relative bias			Relative RMSE			coverage				
parameter	MSLE	postEv	postMed	MSLE	postEv	postMed	profLR	bootLR	BcorCI	postCI	HPD CI
$\log(N_1\bar{\mu})$	**0.0008**	0.010	0.008	**0.035**	0.036	0.036	0.970	0.965	0.967	**0.960**	0.997
$\log(N_2\bar{\mu})$	**-0.0004**	-0.078	-0.064	**0.092**	0.148	0.150	0.975	**0.950**	0.945	0.980	0.975
$\log(N_3\bar{\mu})$	**0.009**	-0.040	-0.024	**0.096**	0.129	0.134	0.965	**0.952**	0.960	0.990	0.990
$\log(N_4\bar{\mu})$	0.044	**0.026**	0.056	0.264	**0.050**	0.073	0.992	**0.952**	0.960	1.000	1.000
t_1	**0.049**	0.128	0.081	0.281	0.149	**0.131**	0.987	0.940	**0.947**	1.000	1.000
r_a	0.017	0.013	**0.008**	0.171	**0.103**	0.121	0.972	0.960	**0.957**	1.000	1.000
$\log(\bar{\mu})$	**-0.012**	-0.100	-0.078	**0.064**	0.129	0.111	0.972	**0.952**	0.945	0.985	0.982
$\bar{p}$	**0.006**	0.023	0.018	0.171	**0.143**	0.146	0.967	**0.952**	**0.947**	0.987	0.985

https://doi.org/10.24072/pcjournal.721

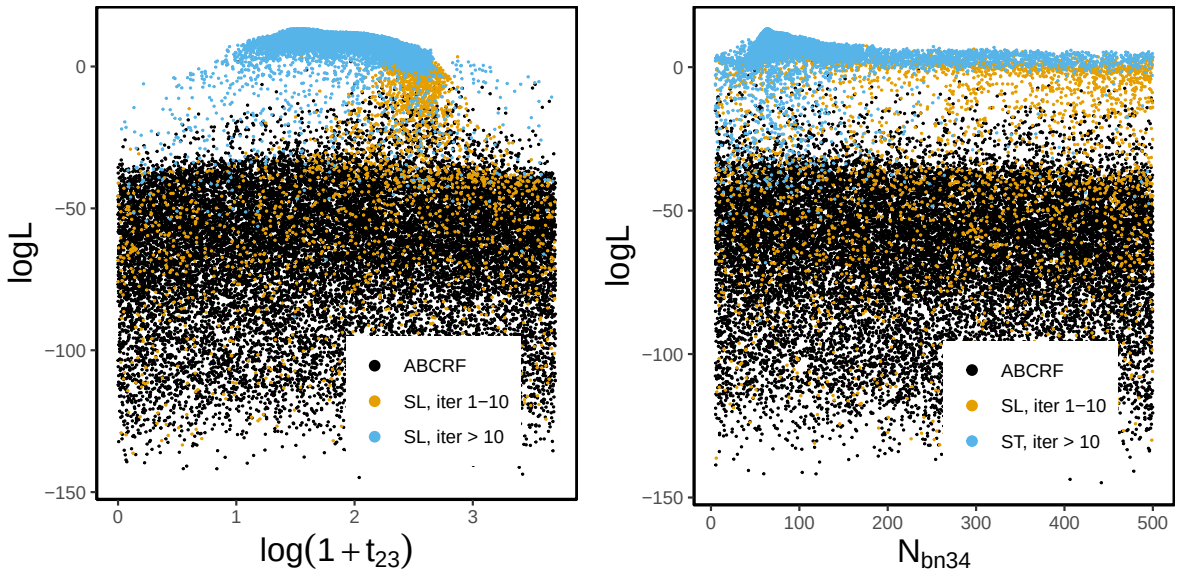

Figure 3 – log-likelihoods of points $\boldsymbol{\theta}$ from ABC-RF and summary likelihood reference tables. x-axis values are those of $\log(1 + t_{23})$ (left) or N_{bn34} (right) from each $\boldsymbol{\theta}$, for points from the ABC-RF reference table (in black) and for points from the summary inference reference table (in orange for points from the first 10 iterations, and in blue for points from later iterations). y-axis values are log-likelihood values according to the likelihood surface inferred in the final iteration of the summary likelihood workflow.

Overall, this means that ABC-RF estimators can be strongly biased toward a value quite distinct from the prior mean, and their variance around this value may be small. ABC-RF estimates may then exhibit much higher or much lower RMSE than summary-MLEs, depending on the data-generating parameter values $\boldsymbol{\theta}^{\dagger}$ chosen, and this pattern cannot be simply interpreted in terms of location of $\boldsymbol{\theta}^{\dagger}$ relative to the prior mean.

7-parameter human admixture scenario

This inference scenario is obtained by fixing the following ten parameters from the full 17-parameter scenario schematized in the right part of Figure 1 : $\log(N_1) = 4.1$, $\log(N_3) = 4.3$, $\log(N_4) = 3.5$, $r_{\text{a}} = 0.2$, $d_3 = 42$, $N_{\text{bn3}} = 160$, $d_4 = 9$, $N_{\text{bn4}} = 98$, $\log(N_{34}) = 3.1$, $d_{34} = 24$.

Performance summaries are reported in Table 3 and Supplementary Figure S.4. The summary-likelihood method generally exhibits lower bias and RMSE than ABC-RF (Table 3). ABC-RF estimates can be strongly biased. In particular, we observe for parameter t_{23} a bias away from the prior mean, similarly to the pattern previously discussed for $\log(N_4)$ in the ladybird invasion scenario. In a transformed log scale, the data-generating value was $\log(1 + t_{23}) = 1.5$, the mean summary-MLE was 1.66, the realized prior mean in the ABC-RF reference table was 1.855 and the mean ABC-RF estimate was 2.491. Examination of the log-likelihood values $\ell(\boldsymbol{\theta})$ for parameters values $\boldsymbol{\theta}$ from both the ABC-RF reference table and from the summary-likelihood reference table for a given dataset suggests an explanation, which we illustrate in Figure 3, left panel (similar results have been obtained for other simulated datasets). All $\ell(\boldsymbol{\theta})$ are here the values of the log-likelihood function estimated by the summary-likelihood inference for the dataset. The points from the ABC-RF reference table form a cloud with a likelihood maximum near $x = 2.5$. This readily explains the mean posterior estimate near 2.5. In this case, the uniform prior sampling used by the ABC-RF method hence appears to miss the more relevant parameter region. One consequence is that the conditional coverage of posterior intervals for t_{23} is quite low (15.25%). On the other hand, the iterative exploration of the parameter space by the summary-likelihood workflow allows higher likelihood values to be attained for lower t_{23} values.

Table 3 – Performance summaries for the 7-parameter human admixture scenario. Bias and root-mean-square error (RMSE) are reported for the maximum-summary likelihood estimates (MSLE), and the posterior mean (postEv) and posterior median (postMed) for ABC-RF. Coverage of confidence intervals with nominal 95% level is reported for summary-likelihood inference (profLR), for two forms of bootstrap correction described in the Text (bootLR and BcorCI), for the central intervals provided by ABC-RF (postCI), and for highest posterior density intervals (HPD CI). Bold font is used to emphasize for each parameter the bias value minimal in absolute value, the minimum RMSE, and the coverage closest to 0.95.

	Relative bias			Relative RMSE			coverage				
parameter	MSLE	postEv	postMed	MSLE	postEv	postMed	profLR	bootLR	BcorCI	postCI	HPD CI
$\log(N_2)$	**0.007**	0.100	0.105	**0.047**	0.103	0.108	0.955	**0.947**	**0.947**	**0.952**	0.980
t_1	**0.101**	0.304	0.293	0.317	0.313	**0.309**	0.985	0.967	**0.965**	1.000	0.997
$\log(t_{12})$	**-0.004**	-0.033	-0.029	**0.046**	0.056	0.057	0.960	**0.942**	0.927	1.000	1.000
$\log(1+t_{23})$	**0.039**	0.245	0.275	**0.069**	0.246	0.276	0.972	0.970	**0.950**	0.152	0.195
N_{bn34}	**0.007**	0.359	0.326	**0.019**	0.362	0.334	0.992	0.980	0.982	0.885	**0.955**
$\log(1+t_{34})$	0.006	-0.037	**-0.004**	0.028	0.046	**0.022**	**0.947**	0.937	0.940	1.000	1.000
$\log(N_{\mathrm{a}})$	**0.004**	0.080	0.103	**0.052**	0.085	0.108	0.962	0.945	**0.947**	1.000	1.000

https://doi.org/10.24072/pcjournal.721

For both t_1 and N_{bn34}, estimates provided by ABC-RF are largely driven by the mean of the prior distribution, which is more expected in the case of a parameter with less apparent information about it (t_1) than with more information about it (N_{bn34}). The latter case might be explained as a less striking effect of the failure of the ABC-RF analysis to explore the narrow region of parameter space with high likelihood (Figure 3, right panel). Two consequences of this failure, combined with attraction of posterior estimates towards the prior mean, are that the conditional coverage of "central" posterior intervals for N_{bn34} is distinctively low (88.5%), but that HPD intervals are less affected.

Deficiencies of low p-values are observed for LRTs, without bootstrap correction, for admixture time t_1, bottleneck size N_{bn34}, and time t_{23}, with corresponding higher than nominal coverage of the intervals. For t_1 there is little statistical information in the data: likelihood profiles are rather flat, and many estimates are at the lower bound of the parameter space. For N_{bn34} there is more information, as shown in particular by the low RMSE of summary-MLEs. The coverage of the Bartlett-corrected intervals for the other parameters is 94.6% on average.

To mitigate the effects of a low level of information in the data for some parameters, we performed additional simulations with datasets of 10,000 SNP rather than 5,000 as previously assumed. The detailed results reported in Supplementary Table S.11 show reductions of RMSE of summary-MLEs by a factor between 1.36 and 1.58 for the different parameters except N_{bn34}, which exhibits a reduction by 2.02. By contrast, the RMSE of the posterior mean estimator is reduced by a lower factor (between 0.93 and 1.28). Hence, the summary-likelihood results exhibit reductions in RMSE roughly as expected from doubling the information in the data (i.e. a reduction of RMSE by a $\sqrt{2}$ factor), but ABC-RF does not. This could reflect another benefit of a more efficient exploration of the region of interest of the parameter space by the iterative workflow. Indeed, when larger datasets are considered, the parameter region with high likelihood ratios becomes narrower and is thus expected to become more difficult to appraise by non-iterative methods. We thus expect iterative methods to provide larger gains in precision relative to non-iterative ones for larger datasets.

We have repeated the simulations as in Table 3 except that the N_{bn34} value was set to either the lower or the upper bound of its explored range (this parameter was chosen because it is the best estimated among the seven parameters, in terms of relative RMSE). The results, detailed in Supplementary Table S.12, show that, as expected, HPD intervals perform much better than central intervals in this case, where they perform similarly to profile-likelihood based intervals.

13-parameter Human admixture scenario

Results for the 13-parameter simulation scenario are presented in Table 4 and Supplementary Figure S.6, with composite bottleneck intensity parameters defined as ratios of bottleneck duration to bottleneck population size: $b_3 = d_3/N_{\mathrm{bn3}}$, $b_4 = d_4/N_{\mathrm{bn4}}$, and $b_{34} = d_{34}/N_{\mathrm{bn34}}$. Summary-MLEs have lower absolute bias than posterior estimates, while for RMSEs the pattern is more heterogeneous, and can be interpreted as resulting from a combination of the effects previously considered to explain discrepancies between the two methods. A more detailed analysis for each parameter is presented in Supplementary Section S.5.3.2. In particular, it suggests that poor exploration of parameter space again affects estimation of $\log(1+t_{23})$, but also of $\log(t_{12})$, $\log(N_{\mathrm{bn3}})$, $\log(N_{\mathrm{bn4}})$, and b_{34}. The admixture rate r_{a} and the composite bottleneck intensity parameters b_3 and b_4 appear well estimated by both methods, and population size N_2 may also be relatively well estimated by SL.

Other parameters are poorly estimated by both methods. For $\log(N_1)$ and N_{bn34} in particular, the position of the prior mean relative to the data-generating $\theta^{\dagger}$ value, rather than efficient use of information, explains the low RMSE of the ABC-RF estimates. In Supplementary Section S.17, we show that summary-likelihood estimates are not improved by running further iterations of the workflow.

Intervals generally have higher than nominal coverage, which is not unexpected given the limited information present in the data about most parameters, and given additional boundary effects such as those detailed for b_3 in Supplementary Section S.5.3.5. The coverage approaches

Table 4 – Performance summaries for the 13-parameter human admixture scenario. Bias and root-mean-square error (RMSE) are reported for the maximum-summary likelihood estimates (MSLE), and the posterior mean (postEv) and posterior median (postMed) for ABC-RF. Coverage of confidence intervals with nominal 95% level is reported for summary-likelihood inference (profLR), for two forms of bootstrap correction described in the Text (bootLR and BcorCI), for the central intervals provided by ABC-RF (postCI), and for highest posterior density intervals (HPD CI). Bold font is used to emphasize for each parameter the bias value minimal in absolute value, the minimum RMSE, and the coverage closest to 0.95.

	Relative bias			Relative RMSE			coverage				
parameter	MSLE	postEv	postMed	MSLE	postEv	postMed	profLR	bootLR	BcorCI	postCI	HPD CI
$\log(N_1)$	**-0.121**	-0.296	-0.279	0.363	0.302	**0.290**	0.967	**0.947**	0.955	1.000	1.000
$\log(N_2)$	**-0.006**	0.027	0.023	0.033	0.034	**0.031**	**0.962**	0.935	**0.937**	1.000	1.000
t_1	**0.076**	0.252	0.195	0.248	0.261	**0.220**	0.980	0.932	**0.940**	1.000	1.000
r_a	**-0.001**	-0.002	-0.002	**0.015**	0.017	0.019	**0.955**	0.925	0.930	0.997	1.000
$\log(t_{12})$	**-0.020**	-0.055	-0.061	**0.056**	0.066	0.077	0.997	**0.992**	0.995	1.000	1.000
d_3/N_{bn3}	**0.008**	0.030	0.017	0.033	0.039	**0.027**	0.990	**0.967**	**0.967**	1.000	1.000
$\log(N_{\mathrm{bn3}})$	**-0.075**	-0.133	-0.146	0.164	**0.138**	0.152	1.000	**0.972**	0.982	1.000	1.000
d_4/N_{bn4}	**0.001**	0.008	-0.002	0.025	**0.024**	0.026	0.985	**0.962**	0.972	1.000	1.000
$\log(N_{\mathrm{bn4}})$	**-0.021**	-0.069	-0.086	0.185	**0.081**	0.109	0.992	**0.965**	0.975	1.000	1.000
$\log(1+t_{23})$	**0.022**	0.204	0.234	**0.106**	0.206	0.236	0.985	**0.967**	0.970	1.000	1.000
d_{34}/N_{bn34}	**-0.033**	-0.109	-0.108	**0.055**	0.111	0.110	0.982	**0.972**	0.982	1.000	1.000
$\log(N_{\mathrm{bn34}})$	**0.002**	0.017	-0.035	0.121	**0.029**	0.045	0.985	**0.960**	0.967	1.000	1.000
$\log(N_\mathrm{a})$	**-0.009**	-0.037	-0.018	0.042	0.045	**0.038**	0.965	**0.940**	0.935	1.000	1.000

https://doi.org/10.24072/pcjournal.721

the nominal 95% for the best estimated parameters, and the bootstrap corrections partially correct the coverage for the other parameters. These results may be seen as evidence that the summary-likelihood method is able to provide confidence intervals for clearly identifiable parameters, from limited simulation effort in such a 13-parameter scenario.

Finally, we compared the results of summary-likelihood and ABC-RF inferences from the real SNP dataset used to design our simulation-based study (Supplementary Table S.16). It is generally not easy to make sense of such comparisons unless additional empirical information, not included in the data, is available about the values to be inferred. We found that the estimates ≈ 0.2 of the admixture rate r_a (the proportion of genes of European ancestry within African American individuals) reported by both methods are consistent with previous studies (as discussed by Collin et al., 2021). The most notable pattern was that for the four parameters identified by the simulation study as being estimable with some precision, r_a, b_3, b_4 and N_2, the confidence intervals of our summary-likelihood method lay within the credible intervals of the ABC-RF method, and were two to five times narrower than the latter intervals. The largest ratio is observed for N_2, which was found in the simulation to be much better estimated by summary-likelihood than by ABC-RF.

Comparison to SNLE

We aimed to compare the performance of the summary likelihood approach with SNLE across the four simulation scenarios. However, we did not include SNLE for the 13-parameter admixture example because the parameter space is subject to non-rectangular constraints (Eq. S.3), which the `sbi` implementation does not readily accommodate. Addressing these constraints within the `sbi` framework would require at least a non-trivial reparameterization of the parameters. In the remaining scenarios, point-estimate performance (bias and RMSE) was broadly similar between the two methods, in particular when contrasted to ABC-RF. In contrast, interval estimation yielded more heterogeneous results (Table 5). In the 15-parameter toy example, SNLE posterior intervals achieved good coverage, comparable to summary-likelihood, but were consistently slightly wider (width ratio 1.02–1.16 across parameters; mean 1.07). By contrast, in the 7-parameter admixture scenario, SNLE intervals were mostly too narrow and exhibited poorly calibrated coverage. Results for the ladybird invasion scenario were intermediate: coverage was, on average, improved relative to uncorrected profile likelihood-ratio intervals, but showed greater variability across parameters (0.912–1).

Discussion

In this study, we present and assess the performance of an automated workflow for summary-likelihood, a method of simulation-based inference based on inferring a likelihood surface for summaries of the data. The efficiency of the method is contingent upon its iterative procedure of exploration of the parameter space. In a 15-parameter toy example, the coverage of confidence intervals provided by summary-likelihood is nearly optimal. However, in cases where some parameters are not practically identifiable, the results are inherently more complex. In these cases, the intervals derived from the likelihood profiles for parameters with low information content are, as expected, too large. Nevertheless, bootstrap procedures appear to be a viable solution for obtaining better coverage. Our simulations suggest that the "bootLR" bootstrap intervals may be systematically used to improve the coverage of intervals. These intervals are defined by the profile-likelihood ratio threshold value determined as the q-quantile of the bootstrap distribution of the likelihood ratio, where q is the intended coverage. The Bartlett-corrected intervals often performed similarly to these intervals. We also considered bootstrap-corrected intervals based on the distribution of parameter estimates, namely the "basic" and "percentile" intervals (Davison and Hinkley, 1997). However, their coverage (only reported in the electronic Supplementary Material) was more variable across simulation conditions, so we cannot recommend them as a general-purpose option.

Although we could not perform an extensive simulation study of the sensitiveness of the performance of our workflow to its control parameters, we performed simulation of the effect

Table 5 – Compared performance of the summary likelihood and SNLE approaches. Bias and root-mean-square error (RMSE) are reported for the maximum-summary likelihood estimates (MSLE), and the posterior mean (postEv) for SNLE. Coverage of intervals with nominal 95% level is reported for summary-likelihood inference (profLR) for the central posterior intervals provided by SNLE (postCI). Boldface indicates, for each parameter, the smallest absolute bias, the minimum RMSE, and the coverage closest to 0.95.

	Summary likelihood					SNLE		
parameter	bias	RMSE	profLR	bootLR	BcorCI	bias	RMSE	postCI
15-parameter toy example								
u_{11}	**0.003**	**0.048**	0.970	0.970	**0.967**	0.027	0.056	0.925
u_{12}	**-0.003**	**0.071**	**0.945**	0.962	0.960	-0.021	0.075	0.930
u_{22}	**-0.001**	**0.063**	0.967	0.980	0.970	0.026	0.073	**0.960**
u_{13}	**0.006**	**0.063**	0.962	0.975	0.965	0.010	0.069	**0.947**
u_{23}	**-0.001**	**0.062**	0.962	0.982	0.970	-0.003	0.066	**0.952**
u_{33}	**-0.008**	**0.069**	0.922	0.942	0.940	0.020	0.076	**0.945**
u_{14}	**0.002**	**0.063**	0.947	0.972	0.970	0.018	0.069	**0.950**
u_{24}	**0.0009**	**0.058**	**0.952**	0.960	0.955	-0.011	0.063	0.940
u_{34}	**-0.003**	**0.050**	0.955	0.965	**0.952**	0.008	0.053	0.955
u_{44}	-0.014	**0.050**	0.935	**0.952**	0.945	**-0.003**	0.051	0.957
u_{15}	**0.006**	0.060	0.970	0.977	0.972	0.020	**0.057**	**0.950**
u_{25}	**0.0009**	**0.052**	**0.952**	0.967	0.960	0.002	0.055	**0.947**
u_{35}	-0.001	**0.050**	**0.952**	0.965	0.960	**-0.0003**	0.052	0.955
u_{45}	0.006	**0.046**	**0.947**	0.960	0.942	**0.001**	0.048	**0.947**
u_{55}	-0.018	0.050	0.932	**0.945**	**0.945**	**-0.009**	**0.049**	**0.955**
Ladybird invasion scenario								
$\log(N_1\bar{\mu})$	0.0008	**0.035**	0.970	**0.965**	0.967	**-0.00008**	0.036	0.912
$\log(N_2\bar{\mu})$	**-0.0004**	0.092	0.975	**0.950**	0.945	-0.009	**0.081**	0.930
$\log(N_3\bar{\mu})$	0.009	0.096	0.962	**0.952**	0.960	**0.002**	**0.083**	0.927
$\log(N_4\bar{\mu})$	**0.044**	0.264	0.992	**0.952**	0.960	0.077	**0.124**	0.977
t_1	**0.049**	0.281	0.987	0.940	**0.947**	0.105	**0.174**	1.000
r_a	**0.017**	0.171	0.972	0.960	**0.957**	0.017	**0.146**	0.962
$\log(\bar{\mu})$	**-0.012**	0.064	0.972	**0.952**	0.945	-0.023	**0.063**	0.942
$\bar{p}$	**0.006**	0.171	0.967	**0.952**	**0.947**	0.016	**0.151**	**0.947**
7-parameter Human admixture scenario								
$\log(N_2)$	0.007	0.047	0.955	**0.947**	**0.947**	**0.002**	**0.043**	0.845
t_1	**0.101**	0.317	0.982	0.967	0.965	0.211	**0.270**	**0.937**
$\log(t_{12})$	**-0.004**	0.046	0.960	**0.942**	0.927	-0.009	**0.042**	0.820
$\log(1+t_{23})$	0.039	0.069	0.972	0.970	**0.950**	**0.032**	**0.057**	0.775
$N_{\mathrm{bn}34}$	**0.007**	**0.019**	0.997	0.980	0.982	0.009	0.028	**0.977**
$\log(1+t_{34})$	0.006	0.028	**0.947**	0.937	0.940	**0.005**	**0.026**	0.890
$\log(N_\mathrm{a})$	0.004	0.052	0.960	0.945	**0.947**	**-0.0001**	**0.046**	0.880

of increasing of decreasing the number of Gaussian components in MGMs (Supplementary Section S.3.2.2). We also verified that increasing the size of the reference table from 38000 to 50000 in the 13-parameter admixture model yields no clear benefit (Supplementary Section S.17). Conversely, we found that decreasing this size from 20000 to 6000 in the 7-parameter model led to degraded performance (Supplementary Table S.15).

Beyond documenting the general performance of the summary-likelihood method, our simulations provide a comparison with the ABC-RF method. Raynal et al. (2019) compared their ABC-RF method to some iterative (or "sequential") ABC workflows (ABC-PMC, Beaumont et al., 2009; Prangle, 2017; SMC-ABC, Del Moral et al., 2012). For a toy example, they found ABC-RF

to better infer marginal posterior distributions, without the computational overhead of the iterative methods. However, they did not extend their comparisons to their population genetics example. By contrast, our results highlight the benefits of an iterative workflow for the exploration of parameter space, and are thus consistent with the premises of sequential ABC methods.

There are further differences between the summary-likelihood and ABC-RF methods, beyond the nature of the intervals sought, and whether the workflow is iterative or not. In particular, ABC-RF is currently constrained to infer marginal posterior distributions for each parameter separately, which makes it more difficult to identify sets of parameters that can be estimated only in combination with each other. By contrast, the likelihood surface inherently retains information about the joint effect of parameters on the likelihood of the data. It is also possible to extend ABC-RF for the inference of multivariate parameters, using distributional random forests (Dinh et al., 2025).

The comparisons with ABC-RF are based on previously considered simulation scenarios. Our findings indicate that ABC-RF is capable of producing estimates with favorable bias and RMSE characteristics, comparable to those of summary-likelihood. However, we also identified instances where its estimation performance diverged significantly, which can be attributed, at least in part, to the imperfect exploration of parameter space by non-iterative methods. Further evidence of this imperfect exploration can be observed in the markedly smaller reduction in the root mean square error (RMSE) of ABC-RF estimates relative to SL estimates when the amount of data is increased, as was investigated in the 7-parameter human admixture scenario.

Moreover, we confirmed the previous observation that ABC-RF often produces 95% credibility intervals with 100% coverage. This is true even for the most easily estimated parameter. For instance, Raynal et al. (2019, Figure 2) already found that 95% credibility intervals for the admixture rate r_a in the Human admixture scenario had 100% coverage. Comparison of credibility and confidence intervals shows that in such cases the credibility intervals consistently extend in both directions beyond the confidence intervals.

A further issue is that ABC-RF estimates may be significantly biased. This phenomenon can be attributed to the fact that sampling from pre-specified priors, a common practice in non-sequential ABC methods, is inadequate for fully exploring the parameter space in certain models. Consequently, the credibility interval coverage provided by ABC-RF may be considerably diminished in such instances.

We found that replacing reference tables, each constructed for a different simulated dataset, with a single reference table of the same size leads to poor inference, even if it is obtained by subsampling the reference tables constructed for each different simulated dataset (Supplementary Section S.5.2.4). While this result has little impact on the analysis of real datasets, each requiring a new reference table matching the details of the sampling design of the dataset, it prevents a drastic reduction of computation time in our simulation studies, which would be possible if a single reference table could be used for accurate inference from all simulated datasets.

Random-forest regression has been used in this work to reduce the dimension of the summary statistics. The implementation of our workflow allows other reduction methods to be used. However, we used the Random-forest regression approach here for reasons previously discussed (Collin et al., 2021; Pudlo et al., 2016; Raynal et al., 2019) and because it provides a convenient baseline for comparing other components of the inference workflow. This approach is fast, easily automated, and although it is not necessarily the most efficient, the additional effort needed to develop more efficient summaries on a case-by-case basis should be considered when designing applications of simulation-based inference. For example, Quelin et al. (2025) compared random forests, gradient boosting, and neural network methods. Neural networks achieved a 5.8% reduction in RMSE compared to random forests (average value other the 10 cases in their Tables 1 and 2), for 10 to 100 fold increases in training times (their supplementary Figure S1). Moreover, a preliminary study was necessary to optimize their alternative methods separately for each inferred parameter.

In our simulations, the size of the reference table was determined only by the number of parameters. Our implementation allows alternative and more adaptive stopping criterion to be

used, based on the estimated precision of the profile likelihood ratios at the bounds of the inferred confidence intervals. A bootstrap procedure is implemented in our R package to evaluate mean square errors of prediction of these ratios. It is thus possible to request termination of iterations when the average RMSE for the different bounds, or when all RMSEs, are below a certain threshold. However, such a termination condition is necessary rather than sufficient, because likelihood surfaces may be inferred with high precision but low accuracy, in particular in cases where large reference tables are built from few iterations, whereas many iterations would be needed to identify narrow parameter regions with high likelihood.

The fixed reference-table sizes in our simulations were deliberately kept small enough to allow performance evaluation over hundreds of samples, but larger tables may be required to achieve accurate inference in many applications. In particular, the linear scaling with the number of parameters used here may become increasingly optimistic as dimensionality grows. Further, even with few parameters, much larger numbers of simulations may be required when the distribution of the data is heavy-tailed, as illustrated by the g-and-k distribution (Frazier et al., 2024).

Further iterative ABC methods have been developed, notably sequential neural likelihood estimation (SNLE, Papamakarios et al., 2019) and related workflows based on deep-learning methods that learn probability distributions and related functions (e.g., Sharrock et al., 2024 and references therein). In our comparison of summary-likelihood with SNLE, we found that the relative performance of the two methods strongly depended on the simulation scenario. SNLE showed excellent performance in one case but more variable interval coverage in the other two, particularly in the 7-parameter admixture scenario. As currently implemented, the main limitation of summary-likelihood is its rapidly increasing runtime with the number of parameters, whereas SNLE's iterative likelihood learning becomes comparatively more efficient as dimensionality grows. This leaves open the possibility that a workflow based on iterative MAF training to learn the likelihood surface, as in SNLE, could ultimately deliver consistently better-calibrated intervals at more moderate computational costs.

Conclusion

In this study, we have implemented and evaluated summary-likelihood inference, an iterative simulation-based method for approximate likelihood inference. In addition to documenting the method's favorable performance, including in regard to confidence interval inference, we have conducted comparisons with the ABC-RF method. These comparisons demonstrate that constructing a reference table of simulations based on pre-specified priors may result in the omission of the most pertinent parameter regions. In contrast, an iterative method of exploring the parameter space may prove more effective in identifying these regions. Using summary-likelihood inference, we found that the full inference for any given dataset required a moderate number of simulations, of the order of 3000 times the number of estimated parameters, showing that approximate likelihood inference is a practically achievable objective even when intensive simulation of genomic datasets is required.

While iterative ABC methods have the potential to outperform non-iterative simulation-based approaches such as standard rejection ABC or ABC-RF, they are less commonly used by non-experts. This may be due to factors such as simplicity, robustness, computational requirements and familiarity with non-iterative approaches. However, as the field advances and tools become more user-friendly, there should be a shift towards iterative methods, particularly for studies involving high-dimensional parameter spaces and where computational efficiency is critical. The SNLE iterative method, which uses MAF training to learn the likelihood surface, already enables fast and efficient inference, but the intervals it provides do not always appear to be well-calibrated. Our method can yield better intervals within a reasonable timeframe for models with a moderate number of parameters, but our results leave open the possibility that improved use of iterative MAF training may also provide better intervals at even lower computational costs.

Acknowledgments

We thank A. Courtiol, D. Prangle, P. Druilhet and an anonymous reviewer for comments on the manuscript, and the Genotoul bioinformatics platform Toulouse Occitanie (https://doi.org/10.15454/1.5572369328961167E12) for providing computing resources. Preprint version 5 of this article has been peer-reviewed and recommended by PCI Math Comp Biol (https://doi.org/10.24072/pci.mcb.100426; Druilhet, 2026).

Fundings

This work has been supported by funds from the Occitanie Regional Council's program "Key challenge BiodivOc" managed by the University of Montpellier (DevOCGen project), and has benefited from state aid managed by the project AgroStat (reference: ANR-23-EXMA-0002) of the Maths-VivES France 2030 program handled by the French Agence Nationale de la Recherche.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. The authors declare the following non-financial conflicts of interest: three authors are recommenders for one or several thematic PCIs, and one is co-representative for PCI Evol Biol.

Data, script, code, and supplementary information availability

Code, data, scripts, and supplementary information are available as a Zenodo record (https://doi.org/10.5281/zenodo.19615138; Rousset et al., 2026). This record includes executables for the version of the `diyabc` simulator used in this study, whose sources are available at https://github.com/diyabc/diyabc/releases/tag/v1.1.36.

References

Akeret J, Refregier A, Amara A, Seehars S, Hasner C (2015). *Approximate Bayesian computation for forward modeling in cosmology*. *Journal of Cosmology and Astroparticle Physics* **2015**, 043. https://doi.org/10.1088/1475-7516/2015/08/043.

Auger-Méthé M, Newman K, Cole D, Empacher F, Gryba R, King AA, Leos-Barajas V, Mills Flemming J, Nielsen A, Petris G, Thomas L (2021). *A guide to state–space modeling of ecological time series*. *Ecological Monographs* **91**, e01470. https://doi.org/https://doi.org/10.1002/ecm.1470.

Bartlett MS (1937). *Properties of sufficiency and statistical tests*. *Proceedings of the Royal Society (London) A* **160**, 268–282. https://doi.org/10.1098/rspa.1937.0109.

Beaumont MA (2010). *Approximate Bayesian Computation in evolution and ecology*. *Annual Review of Ecology, Evolution and Systematics* **41**, 379–406. https://doi.org/10.1093/genetics/162.4.2025.

Beaumont MA, Cornuet JM, Marin JM, Robert CP (2009). *Adaptive approximate Bayesian computation*. *Biometrika* **96**, 983–990. https://doi.org/10.1093/biomet/asp052.

Beaumont MA, Zhang W, Balding DJ (2002). *Approximate Bayesian computation in population genetics*. *Genetics* **162**, 2025–2035. https://doi.org/10.1093/genetics/162.4.2025.

Bertorelle G, Benazzo A, Mona S (2010). *ABC as a flexible framework to estimate demography over space and time: some cons, many pros*. *Molecular Ecology* **19**, 2609–2625. https://doi.org/10.1111/j.1365-294X.2010.04690.x.

Blum MGB, François O (2010). *Non-linear regression models for approximate Bayesian computation*. *Statistics and Computing* **20**, 63–73. https://doi.org/10.1007/s11222-009-9116-0.

Blum MGB, Nunes MA, Prangle D, Sisson SA (2013). *A comparative review of dimension reduction methods in approximate Bayesian computation*. *Statistical Science* **28**, 189–208. https://doi.org/10.1214/12-STS406.

Bonassi FV, You L, West M (2011). *Bayesian learning from marginal data in bionetwork models*. *Statistical Applications in Genetics and Molecular Biology* **10**. https://doi.org/10.2202/1544-6115.1684.

Breiman L (2001). *Statistical modeling: the two cultures (with Discussion)*. *Statist. Sci.* **16**, 199–231. https://doi.org/10.1214/ss/1009213726.

Casella G, Berger RL (2002). *Statistical inference*. Pacific Grove, CA: Duxbury.

Collin FD, Durif G, Raynal L, Lombaert E, Gautier M, Vitalis R, Marin JM, Estoup A (2021). *Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest*. *Molecular Ecology Resources* **21**, 2598–2613. https://doi.org/10.1111/1755-0998.13413.

Cranmer K, Brehmer J, Louppe G (2019). *The frontier of simulation-based inference*. *Proceedings of the National Academy of Sciences* **117**, 30055–30062. https://doi.org/10.1073/pnas.1912789117.

Dalmasso N, Masserano L, Zhao D, Izbicki R, Lee AB (2024). *Likelihood-free frequentist inference: bridging classical statistics and machine learning for reliable simulator-based inference*. *Electronic Journal of Statistics* **18**, 5045–5090. https://doi.org/10.1214/24-ejs2307.

Daly AC, Gavaghan D, Cooper J, Tavener S (2018). *Inference-based assessment of parameter identifiability in nonlinear biological models*. *Journal of The Royal Society Interface* **15**, 20180318. https://doi.org/10.1098/rsif.2018.0318.

Davison AC (2003). *Statistical models*. Cambridge, UK: Cambridge Univ. Press. https://doi.org/10.1017/CBO9780511815850.

Davison AC, Hinkley DV (1997). *Bootstrap methods and their applications*. Cambridge Univ. Press. https://doi.org/10.1017/CBO9780511802843.

Del Moral P, Doucet A, Jasra A (2006). *Sequential Monte Carlo samplers*. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **68**, 411–436. https://doi.org/10.1111/j.1467-9868.2006.00553.x.

Del Moral P, Doucet A, Jasra A (2012). *An adaptive sequential Monte Carlo method for approximate Bayesian computation*. *Statistics and Computing* **22**, 1009–1020. https://doi.org/10.1007/s11222-011-9271-y.

Deleforge A, Forbes F, Horaud R (2014). *High-dimensional regression with gaussian mixtures and partially-latent response variables*. *Statistics and Computing* **25**, 893–911. https://doi.org/10.1007/s11222-014-9461-5.

Diggle PJ, Gratton RJ (1984). *Monte Carlo methods of inference for implicit statistical models*. *Journal of the Royal Statistical Society B* **46**, 193–227. https://doi.org/10.1111/j.2517-6161.1984.tb01290.x.

Dinh KN, Liu C, Xiang Z, Liu Z, Tavaré S (2025). *Approximate Bayesian Computation sequential Monte Carlo via random forests*. *Statistics and Computing* **35**. https://doi.org/10.1007/s11222-025-10748-x.

Druilhet P (2026). *Computationally efficient iterative summary-likelihood inference for calibrated uncertainty in population genetics*. *Peer Community in Mathematical and Computational Biology*, 100426. https://doi.org/10.24072/pci.mcb.100426.

Fan Y, Meikle SR, Angelis GI, Sitek A (2019). *ABC in nuclear imaging*. In: *Handbook of Approximate Bayesian Computation*. Ed. by S. A. Sisson, Y. Fan, and M. A. Beaumont. Boca Raton, Florida: CRC Press, pp. 623–647. https://doi.org/10.1201/9781315117195-21.

Fan Y, Nott DJ, Sisson SA (2013). *Approximate Bayesian computation via regression density estimation*. *Stat* **2**, 34–48. https://doi.org/https://doi.org/10.1002/sta4.15.

Fearnhead P, 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**, 419–474. https://doi.org/10.1111/j.1467-9868.2011.01010.x.

Fraimout A, Debat V, Fellous S, Hufbauer RA, Foucaud J, Pudlo P, Marin JM, Price DK, Cattel J, Chen X, Deprá M, François Duyck P, Guedot C, Kenis M, Kimura MT, Loeb G, Loiseau A, Martinez-Sañudo I, Pascual M, Polihronakis Richmond M, et al. (2017). *Deciphering the routes*

of invasion of Drosophila suzukii *by means of ABC random forest*. *Molecular Biology and Evolution* **34**, 980–996. https://doi.org/10.1093/molbev/msx050.

Frazier DT, Kelly R, Drovandi C, Warne DJ (2024). *The statistical accuracy of neural posterior and likelihood estimation*. https://doi.org/10.48550/arXiv.2411.12068. arXiv: 2411.12068.

Germain M, Gregor K, Murray I, Larochelle H (2015). *MADE: Masked Autoencoder for Distribution Estimation*. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 881–889. URL: https://proceedings.mlr.press/v37/germain15.html.

Geurts P, Ernst D, Wehenkel L (2006). *Extremely randomized trees*. *Machine Learning* **36**, 3–42. https://doi.org/10.1007/s10994-006-6226-1.

Greenberg D, Nonnenmacher M, Macke J (2019). *Automatic posterior transformation for likelihood-free inference*. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2404–2414. URL: https://proceedings.mlr.press/v97/greenberg19a.html.

Häggström H, Rodrigues PLC, Oudoumanessah G, Forbes F, Picchini U (2024). *Fast, accurate and lightweight sequential simulation-based inference using Gaussian locally linear mappings*. https://doi.org/10.48550/arXiv.2403.07454. arXiv: 2403.07454.

Hastie T, Tibshirani R, Friedman J (2009). *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer. https://doi.org/10.1007/978-0-387-84858-7.

Hermans J, Delaunoy A, Rozet F, Wehenkel A, Begy V, Louppe G (2022). *A trust crisis in simulation-based inference? Your posterior approximations can be unfaithful*. https://doi.org/10.48550/arXiv.2110.06581. arXiv: 2110.06581.

Laugier F, Béthune K, Plumel F, Froissard C, Donnay JM, Chenin T, Rousset F, David P (2025). *Cytoplasmic male sterility declines in the presence of resistant nuclear backgrounds*. *The American Naturalist* **206**, 16–30. https://doi.org/10.1086/735820.

Lebret R, Iovleff S, Langrognet F, Biernacki C, Celeux G, Govaert G (2015). *Rmixmod: The R package of the model-based unsupervised, supervised, and semi-supervised classification Mixmod library*. *Journal of Statistical Software* **67**, 1–29. https://doi.org/10.18637/jss.v067.i06.

Lehmann EL, Casella G (1998). *Theory of point estimation*. New York: Springer-Verlag.

Lombaert E, Guillemaud T, Thomas CE, Lawson Handley LJ, Li J, Wang S, Pang H, Goryacheva I, Zakharov IA, Jousselin E, Poland RL, Migeon A, Van Lenteren J, De Clercq P, Berkvens N, Jones W, Estoup A (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. *Molecular Ecology* **20**, 4654–4670. https://doi.org/10.1111/j.1365-294X.2011.05322.x.

Lopez-Paz D, Oquab M (2017). *Revisiting classifier two-sample tests*. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SJkXfE5xx.

Lueckmann JM, Boelts J, Greenberg DS, Gonçalves PJ, Macke JH (2021). *Benchmarking simulation-based inference*. In: *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research, pp. 343–351. URL: https://proceedings.mlr.press/v130/lueckmann21a.html.

Lueckmann JM, Goncalves PJ, Bassetto G, Öcal K, Nonnenmacher M, Macke JH (2017). *Flexible statistical inference for mechanistic models of neural dynamics*. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 1289–1299. URL: http://papers.nips.cc/paper/by-source-2017-855.

Marin JM, Raynal L, Pudlo P, Robert CP, Estoup A (2025). *abcrf: Approximate Bayesian Computation via Random Forests. R package version 2.0*. https://doi.org/10.32614/CRAN.package.abcrf.

Moshe A, Wygoda E, Ecker N, Loewenthal G, Avram O, Israeli O, Hazkani-Covo E, Pe'er I, Pupko T (2022). *An approximate bayesian computation approach for modeling genome rearrangements*. *Molecular Biology and Evolution* **39**, msac231. https://doi.org/10.1093/molbev/msac231.

Nakagome S, Alkorta-Aranburu G, Amato R, Howie B, Peter BM, Hudson RR, Di Rienzo A (2015). *Estimating the ages of selection signals from different epochs in human history*. *Molecular Biology and Evolution* **33**, 657–669. https://doi.org/10.1093/molbev/msv256.

Neyman J (1977). *Frequentist probability and frequentist statistics*. *Synthese* **36**, 97–131. https://doi.org/10.1007/BF00485695.

Papamakarios G, Murray I (2016). *Fast ϵ-free inference of simulation models with bayesian conditional density estimation*. In: *Advances in Neural Information Processing Systems 29*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., pp. 1028–1036. URL: http://papers.nips.cc/paper/by-source-2016-598.

Papamakarios G, Pavlakou T, Murray I (2017). *Masked autoregressive flow for density estimation*. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 2335–2344. URL: http://papers.nips.cc/paper/by-source-2017-1368.

Papamakarios G, Sterratt D, Murray I (2019). *Sequential neural likelihood: fast likelihood-free inference with autoregressive flows*. In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. Naha, Okinawa, Japan: PMLR, pp. 837–848. URL: https://proceedings.mlr.press/v89/papamakarios19a.html.

Prangle D (2017). *Adapting the ABC distance function*. *Bayesian Analysis* **12**, 289–309. https://doi.org/10.1214/16-BA1002.

Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP (2016). *Reliable ABC model choice via random forests*. *Bioinformatics* **32**, 859–866. https://doi.org/10.1093/bioinformatics/btv684.

Quelin A, Austerlitz F, Jay F (2025). *Assessing simulation-based supervised machine learning for demographic parameter inference from genomic data*. *Heredity* **134**, 417–426. https://doi.org/10.1038/s41437-025-00773-x.

Raynal L, Marin JM, Pudlo P, Ribatet M, Robert CP, Estoup A (2019). *ABC random forests for Bayesian parameter inference*. *Bioinformatics* **35**, 1720–1728. https://doi.org/10.1093/bioinformatics/bty867.

Rousset F (2025). *Infusion: inference using simulation. R package version 2.3.0*. https://doi.org/10.32614/cran.package.infusion.

Rousset F, Gouy A, Martinez-Almoyna C, Courtiol A (2017). *The summary-likelihood method and its implementation in the Infusion package*. *Molecular Ecology Research* **17**, 110–119. https://doi.org/10.1111/1755-0998.12627.

Rousset F, Leblois R, Estoup A, Marin JM (2026). *Data, scripts, code, and supplementary information for "A new iterative framework for simulation-based population genetic inference with improved coverage properties of confidence intervals"*. Zenodo. https://doi.org/10.5281/ZENODO.19615138.

Rubio FJ, Johansen AM (2013). *A simple approach to maximum intractable likelihood estimation*. *Electron. J. Stat.* **7**, 1632–1654. https://doi.org/10.1214/13-EJS819.

Schälte Y, Hasenauer J (2020). *Efficient exact inference for dynamical systems with noisy measurements using sequential approximate Bayesian computation*. *Bioinformatics* **36**, i551–i559. https://doi.org/10.1093/bioinformatics/btaa397.

Sharrock L, Simons J, Liu S, Beaumont M (2024). *Sequential neural score estimation: likelihood-free inference with conditional score based diffusion models*. In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 44565–44602. URL: https://proceedings.mlr.press/v235/sharrock24a.html.

Sisson S. A., Y. Fan, and M. A. Beaumont, eds. (2019). *Handbook of Approximate Bayesian Computation*. Boca Raton, Florida: CRC Press. https://doi.org/10.1201/9781315117195.

Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997). *Inferring coalescence times from DNA sequence data*. *Genetics* **145**, 505–518. https://doi.org/10.1093/genetics/145.2.505.

Tejero-Cantero A, Boelts J, Deistler M, Lueckmann JM, Durkan C, Gonçalves PJ, Greenberg DS, Macke JH (2020). *sbi: A toolkit for simulation-based inference*. *Journal of Open Source Software* **5**, 2505. https://doi.org/10.21105/joss.02505.

The 1000 Genomes Project Consortium (2012). *An integrated map of genetic variation from 1,092 human genomes*. *Nature* **491**, 56–65. URL: http://dx.doi.org/10.1038/nature11632.

Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MP (2009). *Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems*. *Journal of The Royal Society Interface* **6**, 187–202. https://doi.org/10.1098/rsif.2008.0172.

Wood SN (2010). *Statistical inference for noisy nonlinear ecological dynamic systems*. *Nature* **466**, 1102–1104. https://doi.org/10.1038/nature09319.

Wright MN, Ziegler A (2017). *ranger: A fast implementation of random forests for high dimensional data in C++ and R*. *Journal of Statistical Software* **77**, 1–17. https://doi.org/10.18637/jss.v077.i01.

Zammit-Mangion A, Sainsbury-Dale M, Huser R (2025). *Neural methods for amortized inference*. *Annual Review of Statistics and Its Application* **12**, 311–335. https://doi.org/10.1146/annurev-statistics-112723-034123.