Resampling: An improvement of importance sampling in varying population size models

C. Merle a,b,c,*, R. Leblois b,c, F. Rousset c,d, P. Pudlo a,c,e

a Institut Montpelliérain Alexander Grothendieck UMR CNRS 5149, Université de Montpellier, Place Eugène Bataillon, 34095 Montpellier, France
b Centre de Biologie pour la Gestion des Populations, Inra, CBGP (UMR INRA / IRD / Cirad / Montpellier SupAgro), Campus International de Baillarguet, 34988 Montferrier sur Lez, France
c Institut de Biologie Computationnelle, Université de Montpellier, 95 rue de la Galéra, 34095 Montpellier, France
d Institut des Sciences de l’Evolution, Université de Montpellier, CNRS, IRD, EPHE, Place Eugène Bataillon, 34392 Montpellier, France
e Institut de Mathématiques de Marseille, I2M UMR 7373, Aix-Marseille Université, CNRS, Centrale Marseille, 39, rue F. Joliot Curie, 13453 Marseille, France

A R T I C L E   I N F O
Article history:
Received 31 January 2016
Available online 3 October 2016

Keywords:
Importance sampling
Resampling
Markov process
Population genetics
Demographic inference
Coalescent

A B S T R A C T
Sequential importance sampling algorithms have been defined to estimate likelihoods in models of ancestral population processes. However, these algorithms are based on features of the models with constant population size, and become inefficient when the population size varies in time, making likelihood-based inferences difficult in many demographic situations. In this work, we modify a previous sequential importance sampling algorithm to improve the efficiency of the likelihood estimation. Our procedure is still based on features of the model with constant size, but uses a resampling technique with a new resampling probability distribution depending on the pairwise composite likelihood. We tested our algorithm, called sequential importance sampling with resampling (SISR) on simulated data sets under different demographic cases. In most cases, we divided the computational cost by two for the same accuracy of inference, in some cases even by one hundred. This study provides the first assessment of the impact of such resampling techniques on parameter inference using sequential importance sampling, and extends the range of situations where likelihood inferences can be easily performed.

1. Introduction

Under genetic neutrality, the distribution of the genetic polymorphism in a sample of individuals depends on the evolution of the population size through unobserved stochastic processes. Typically, these stochastic processes describe the evolution of the alleles at a given locus of the individuals from a population backward to their Most Recent Common Ancestor (MRCA). When the population size is constant and finite, Wright-Fisher models describe this evolution and the coalescent theory approximates these models when the population size is large. In this context, the history (genealogy with mutations) of the observed sample is a latent process. One of the major challenges to conduct a parametric inference analysis with these models is computing the likelihood of the data at any point \( \phi \) of the parametric space. Indeed, the likelihood at \( \phi \) is the integral of the probabilities of each possible realization of the latent process. In population genetics, the likelihood of an observed sample is an integral over the distribution of ancestral histories that may have led to this sample. In this work, we consider a class of Monte Carlo methods based on Sequential Importance Sampling (SIS) which provides an estimate of the integral. In this scheme, the importance sampling distribution proposes paths of the process among those who contribute the most to the sum defining the likelihood.

For models of panmictic population with constant size, Griffiths and Tavaré (1994b) described an algorithm wherein a proposal distribution suggests histories of the sample by stepwise reduction of the data set, either by coalescence of two identical genes or by removal of a mutation on a single gene lineage. Stephens and Donnelly (2000, Theorem 1) characterized the optimal proposal distribution for a large class of time homogeneous models, but not for varying population size models. However, in most cases (even in time-homogeneous models) the optimal distribution cannot be practically computed and has to be approximated. De Iorio and Griffiths (2004a) developed a method for
constructing such approximations for any model where the mutation process can be described as a Markov chain on gene types and De Iorio and Griffiths (2004b) extended this to subdivided population models. These methods have been further elaborated for stepwise mutation models in a subdivided population by De Iorio et al. (2005). The latent process, namely the history of the data, of models with past changes in population size exhibits inhomogeneity in time. Thus the previous theoretical arguments which derive efficient proposal distributions are no longer applicable in this context. Nevertheless, as shown by Leblois et al. (2014), we can adapt an importance distribution from the importance distributions of models with constant size populations. But their simulation tests demonstrated some limits of the algorithm. Most importantly, under demographic scenarios with strong changes, they face large computation times due to the large variance of the Monte Carlo estimate of the likelihood.

Our aim in this work is to improve the accuracy of the likelihood estimation for a given computational cost. One direction could be to derive a new importance sampling proposal distribution like Hoboth et al. (2008) did for the infinite site model. In this paper we chose another direction which consists of resampling among the paths proposed by the importance distribution. Under a demographic scenario of constant population size, the likelihood is a sum of products of probabilities of each event whatever their time because of the time homogeneity of the latent process. By contrast, under a varying population size model, these probabilities depend on the times of occurrence of the events. Thus we have to integrate over the possible times of occurrences. It follows that the whole integral (namely the likelihood of the data) we estimate with importance sampling in the general demographic case is an integral over a space of much larger dimension. Because the efficiency of importance sampling decreases with the dimension of the integral, even if the importance distribution adapted by Leblois et al. (2014) were the most efficient distribution among a certain class of distributions, we expect more variance of the likelihood estimate.

We took the opportunity of the paper to present rigorously the stochastic model for a rather general demographic scenario in Section 2. The time inhomogeneous latent process is part of the folklore in the neutral population genetic literature, but has never been written down explicitly. In particular, Eq. (6) at the end of Section 2.2 shows that the integral defining likelihood is of much larger dimension than in the constant demographic case. Indeed, when the population size varies over time, the integrals over the random times cannot be removed, as explained in Section 2.4, after a presentation of the sequential importance sampling (SIS) algorithm.

The major contribution of the paper is the addition and the calibration of the resampling procedure of Section 3 in the SIS algorithm, based on Liu et al. (2001) and Liu (2008). The novelty is mostly in the choice of the resampling distribution that we propose in Section 3.3, which depends on both the current weight of the latent path and the pairwise composite likelihood (PCL) of the current state of the latent process. Section 4 presents numerical results on the likelihood estimates of simulated data sets. These results highlight the benefit due to the proposed resampling distributions in the likelihood estimates. We then plug the likelihood estimates in an inference method presented in Section 5.1. The remainder of Section 5 highlights how our proposals improve the estimate of the parameters, and the likelihood surface around the maximum likelihood estimate used to compute confidence intervals (CIs). We can thus confirm that the gain due to resampling also benefits to the demographic parameter estimation. We end Section 5 with a discussion on cases where the data do not hold much information regarding the parameter of interest, leading to flat likelihood surfaces. Finally we show the relevance of our methods by presenting numerical results on a bat data set where strong evidence for population contraction had been already provided by Storoz and Beaumont (2002). All computations for this work were performed using an updated version of the Migraine software, available at http://kimura.univ-montp2.fr/~rousset/Migraine.htm (Rousset and Leblois, 2007, 2012; Leblois et al., 2014).

## 2. The stochastic model and its likelihood

To illustrate our method we consider genetic data from individuals of a single population sampled at time \( t = 0 \). Let \( N(t) \) be the population size, expressed in number of gene copies in the whole paper, \( t \) generations away from the sampling time \( t = 0 \). We assume that \( N(t) \) is a parametric function of \( t \), see Section 4.1 for examples. In this Section we focus only on data from a given locus.

### 2.1. Stochastic model

Kingman (1982)’s coalescent process is the usual model to describe ancestral relationships between gene copies of the sample under neutrality in a population of constant, but relatively large size. We superimpose a mutation model on the coalescent process to describe gene modifications along lineages. Since the evolution is neutral, the coalescent is independent of the mutation process. To describe the resulting process, we introduce a random vector \( H_t \), indexed by the set of possible types of genes (possible alleles) \( E \). If \( A \in E \), the component \( H_t(A) \) counts the number of genes of type \( A \) at time \( t \) (i.e., \( t \) generations away, backward in time, from the sampling time) in the genealogy of the sample. The likelihood of the genetic data is given by the distribution of \( H_t \), which cannot be written as an explicit function of the parameter of interest, denoted \( \phi \).

### 2.2. Markovian description of the evolution

Actually, we only have at our disposal the following description of the process \( H_t \) forward in time.

Let \( e_A \) denote the vector indexed by \( E \) whose components are all equal to 0, except the \( A \)-component which is equal to one. On one hand, the probability of a new lineage of type \( A \) in the genealogy at time \( t - \delta \), knowing that \( h \) is the value of the process at time \( t \), is

\[
\mathbb{P}(H_{t-\delta} = h + e_A | H_t = h) = \frac{(h + 1) \phi(A)}{2N(t)} + o(\delta) \tag{1}
\]

where \( |h| = \sum_{A \in E} h(A) \) is the total number of lineages at time \( t \) in the genealogy and \( o(\delta) \) a quantity that is negligible in front of \( \delta \) when \( \delta \to 0 \). On the other hand the probability of a mutation of a gene of type \( B \) at time \( t \) to a gene of type \( A \) at time \( t - \delta \), knowing that \( h \) is the value of the process at time \( t \), is

\[
\mathbb{P}(H_{t-\delta} = h + e_A - e_B | H_t = h) = \mu \cdot h(B) p_{BA} \delta + o(\delta) \tag{2}
\]

where \( p_{BA} \) is the mutation probability from allele \( B \) to allele \( A \) forward in time, \( \mu \) is the mutation rate per generation per lineage, and \( o(\delta) \) is a quantity that is negligible in front of \( \delta \) when \( \delta \to 0 \). Additionally, when there is only one lineage in the genealogy, the distribution of the gene type (the allele) is supposed to be the stationary distribution \( \psi(\cdot) \) of the transition matrix \( p_{B,A}; B, A \in E \):

\[
\mathbb{P}(H_t = e_A | |H_t| = 1) = \psi(A). \tag{3}
\]

Hence, forward in time (i.e., when \( t \) decreases), the coalescent based model \( H_t \) is a pure jump, continuous time Markov process taking values in the set of integer vectors indexed by \( E \), namely \( h^E \). But the process is time inhomogeneous because the coalescence
rate of Eq. (1) depends on the current date \( t \) through the function \( N(t) \). Note that Eqs. (1) and (2) can both be written
\[
P\left(H_{t-\delta} = h' | H_t = h\right) = \Lambda_t(h' | h) \delta + o(\delta)
\]
for any \( h \neq h' \) in \( \mathbb{N}^2 \) where
\[
\Lambda_t(h' | h) = \begin{cases} 
\left((|h| + 1)A_h(\alpha) / (2N(t))\right) & \text{if } h' = h + e_{\alpha} \\n\mu h(\beta) p_{\beta,\alpha} & \text{if } h' = h + e_{\alpha} - e_{\beta} \\n0 & \text{otherwise}
\end{cases}
\]
defines the intensity matrix of the Markov process.

We can give a more explicit description of the process since it is a pure jump process, or, in other words, \( H_t \) is a piecewise constant function of \( t \). We denote \( X_0 = H_0 \) as the first value of the process at time \( T_0 = 0 \). The process \( H_t \) remains constant until (random) time \( T_1 \), where it takes another value \( X_1 = H_{T_1} \). After \( T_1 \), the process \( H_t \) stays equal to \( X_1 \) until time \( T_2 \) where it jumps to a another value \( X_2 \), and so on.

Fig. S1 in the supplementary material represents a possible path of the process \( H \). We refer to such paths as (possible) histories of the sample (composed of three genes of type \( a \) and one gene type \( b \) in Fig. S1 in the supplementary material). Note that many genealogies correspond to a possible history since, at time \( T_1 \), we have chosen a genealogy with a coalescence between the two left-handmost genes, but we could have chosen another genealogy joining any pair of genes of type \( a \) at time \( T_1 \). Likewise at time \( T_2 \) we could have chosen any of the three possible cases. Actually each genealogy leading to the same path of the process \( H \) has the same probability because of the exchangeability of the lineages carrying the same allele.

We set \( \Delta_i = T_{i+1} - T_i \) which is usually named the holding time at value \( X_i \) for any integer number \( i \). The distribution of the process is given by the density
\[
P\left(X_i = h \right) \text{ for all } i = 0, \ldots, n \text{ and } \Delta_i \in (\delta_i; \delta_i + d\delta_i)
\]
for all \( i = 0, \ldots, n \) and \( \Delta_i \in (\delta_i; \delta_i + d\delta_i) \) with
\[
\lambda_x(h) = \lambda_x(h^0 | h^1)
\]
and
\[
P_t(h^0 | h^1) = \lambda_x(h^0 | h^1) / \lambda_x(h).
\]
The term \( \lambda_x(h) \) is interpreted as the (infinitesimal) jump rate of the process \( H \) at time \( t \), knowing that the process takes value \( h \) at time \( t \), namely
\[
P(H_{t-\delta} \neq H_t = h) = \lambda_x(h) \delta + o(\delta)
\]
when \( \delta > 0 \) and \( o(\delta) \) is negligible in front of \( \delta \) when \( \delta \to 0 \). With Eq. (3), it can be computed as
\[
\lambda_x(h) = \frac{|h| A(h) + \mu |h|}{2N(t)} = \frac{|h|}{2N(t)} \left((|h| + 1) + \theta(t)\right),
\]
where \( \theta(t) = 2\mu N(t) \) is the usual composite mutation rate parameter of population genetics. On the other hand, \( P_t(h^0 | h^1) \) is interpreted as a transition matrix (forward in time), and gives the probability that the new value (forward in time) of \( H_t \) is \( h' \) for \( t' < t \) knowing that the process jumps at time \( t' \) and that \( H_{t'} = h \).

Set \( \tau = \inf\{t > 0: |H_t| = 1\} \), which is the age of the most recent common ancestor (MRCA) of the sample. Since \( H \) is a pure jump process, \( \tau = T_0 \) where \( \sigma = \inf\{n > 0: |X_n| = 1\} \). And note that the distribution of \( H_\tau = X_{\sigma} \) is known explicitly as
\[
P(H_\tau = e_\alpha) = \psi(A).
\]
To recover the likelihood of the observed genetic data, that is to say the distribution of \( H_\tau \) from this knowledge and the (forward in time) transition mechanism of Eq. (4), we have to integrate over all possible histories from the MRCA, that is to say all possible values \( n \) of \( \sigma, t_1, \ldots, t_n \) in \( \mathbb{N}^2 \) and \( \delta_0, \ldots, \delta_{n-1} \geq 0 \). Hence
\[
P(H_0 = h_0) = \sum_{n=1}^{\infty} \int \cdots \int \sum_{t_1, \ldots, t_n} \sum_A \psi(A) 1 \left[ h_n = e_{\alpha}, |h_{n-1}| > 1 \right] 
\]
\[
\times \prod_{i=1}^{n} P_{t_i}(h_{t_i-1} | h_{t_i}) \times \prod_{i=0}^{n-1} \lambda_{t_i + \delta_i} (h_{t_i}) \times \exp \left( -\int_{0}^{\delta_i} \lambda_{t_i + \delta_i} (h_{t_i}) du \right) d\delta_i.
\]
\[
(6)
\]
The challenge to conduct a likelihood based inference in population size varying models is in computing the multidimensional integral of Eq. (6), which cannot be computed formally. Note that, to alleviate notations, we have dropped the dependency in the parameter of interest \( \phi \) of the quantities arising in Eq. (6) but both \( P_t(h^0 | h^1) \) and \( \lambda_x(h) \) are functions of \( \phi \), and even sometimes the stationary distribution \( \psi(A) \).

2.3. Evaluating the likelihood with importance sampling

The trick to compute the likelihood defined in Eq. (6) is to rely on another transition matrix \( Q_t(h^0 | h^1) \) backward in time and set
\[
W_n(h_{0:n}; t_{0:n}) = \prod_{i=1}^{n} P_{t_i}(h_{t_i-1} | h_{t_i}) \prod_{i=1}^{n} Q_{t_i}(h_{t_i} | h_{t_i-1}).
\]
Then Eq. (6) leads to
\[
P(H_\tau = h_\sigma) = \sum_{n=1}^{\infty} \int \cdots \int \sum_{t_1, \ldots, t_n} \sum_A \psi(A) 1 \left[ h_n = e_{\alpha}, |h_{n-1}| > 1 \right] 
\]
\[
\times W_n(h_{0:n}; t_{0:n}) \times \prod_{i=1}^{n} Q_{t_i}(h_{t_i} | h_{t_i-1}) \times \prod_{i=0}^{n-1} \lambda_{t_i + \delta_i} (h_{t_i}) \times \exp \left( -\int_{0}^{\delta_i} \lambda_{t_i + \delta_i} (h_{t_i}) du \right) d\delta_i.
\]
\[
(8)
\]
where \( t_i = \delta_0 + \cdots + \delta_i \) are implicit functions of the \( \delta_i \)'s. The right hand side of Eq. (8) can be interpreted as the expected value of
\[
W_\sigma = \sum_n \sum_A \psi(A) 1 \left[ h_n = e_{\alpha}, |h_{n-1}| > 1 \right] \times W_n(h_{0:n}; t_{0:n})
\]
when \( h_0, h_1, \ldots \) and \( t_0, t_1, \ldots \) are the realizations of the embedded chain and the jump times of an inhomogeneous Markov process \( H \) whose intensity matrix (backward in time) is given by \( \tilde{\Lambda}_t(h^0 | h^1) = \lambda_x(h) Q_t(h^0 | h^1) \) and which starts from \( \tilde{H}_0 = h_0 \), the observed data. In other words,
\[
P(H_{\tau} = e_\sigma) = \mathbb{E}(W_\sigma).
\]
Interpreting the likelihood of the data as an expected value over another distribution is the first step toward an importance sampling estimate of the likelihood. Indeed, the Monte Carlo estimation of the likelihood is the empirical average of \( W_0 \) computed on simulated replicates \( \tilde{H}^j \) (for \( j = 1, \ldots, n_{\text{rep}} \)) of the process \( H \):

\[
\hat{p}(H_0 = h_0) = \frac{1}{n_{\text{rep}}} \sum_{j=1}^{n_{\text{rep}}} W_0^{(j)}.
\]  

(9)

2.4. Practical aspects and efficiency

To sum up the above, the idea of importance sampling is to interpret the likelihood as an expected value over a sampling distribution of histories (the distribution of the process \( H \)) which may differ from the distribution of the genuine model described in Section 2.2. We can choose this importance distribution freely, as long as any path from the data \( h_0 \) to the MRCA with positive density under the distribution of the latent process \( H \) has also a positive density under the distribution of the importance process \( H \). But this choice has a major impact on the efficiency of the approximation in Eq. (9) since the variance of \( W_0 \) depends on the distribution of \( H \).

When the population size \( N(t) \) is fixed to \( N_0 \) for all \( t \), we have at our disposal an efficient sampling distribution from the literature (Stephens and Donnelly, 2000; De Iorio and Griffiths, 2004a; De Iorio et al., 2005) which represents a major improvement over the first proposal of Griffiths and Tavaré (1994a). In this simple demographic scenario, the importance sampling estimate is the most efficient under a parent independent mutation model (which means that \( p_{Bi} \) does not depend on \( B \)) and Eq. (8) provides an exact evaluation of the likelihood for \( n_{\text{rep}} = 1 \) replicate. In other words, the variance of \( W_0 \) under the efficient importance sampling distribution is zero. For other mutation models and a constant population size, the variance of \( W_0 \) is no longer zero, but the number \( n_{\text{rep}} \) of replicates required to get a sharp estimation of the likelihood with Eq. (9) is much smaller when relying on the efficient importance distribution rather than on the proposal of Griffiths and Tavaré (1994a). In the scenario where the population size varies over time, we resort to the proposal of Leblois et al. (2014) to define the transition matrix of \( H \). That is to say that for any value of \( t > 0 \), \( Q_t(h, h') \) is the transition matrix of the efficient importance distribution as if the population size were constant over time and fixed to the current value of \( N(t) \). But the efficiency of this importance distribution depends on the variation of \( N(t) \), as explained in Leblois et al. (2014).

Nevertheless, the choice of importance distribution leads to a process \( H \) which is defined explicitly as an inhomogeneous Markov process, backward in time, starting from the observed data \( h_0 \). The computation of a simulated path of \( H \), as well as of \( W_0 \) is performed with the sequential method of Algorithm 1.

Note that, since \( W_0 \) is a product along the path of the process \( H \), its value is computed sequentially at step 2.(b) of Algorithm 1. The update depends only on the current time \( t \), the value of \( H \) just before time \( t \), namely \( h \), and the new value \( h' \). Hence, Algorithm 1 does not have to keep track of the whole path of \( H \). Moreover, drawing the holding time \( \delta \) can be done either with a rejection algorithm, see Appendix B in the Supplementary Materials, or by inverting the cumulative distribution function when possible, e.g., see Griffiths and Tavaré (1994b). Finally, the above algorithm is run \( n_{\text{rep}} \) times to approximate the likelihood by the average in Eq. (9).

Unfortunately, the decrease of efficiency of the importance sampling scheme from the constant population size scenario to the general case (where \( N(t) \) varies over time) can be drastic, see Leblois et al. (2014). To understand the major difference, we recall that, when the population size \( N(t) \) is constant, neither \( \Lambda_t(h|h) \) nor \( P_t(h|h') \) depend on \( t \). In this simpler case we choose an importance distribution characterized by a \( Q_t(h|h') \) which does not depend on \( t \), and \( W_0 \) does not depend on the random jump times \( T_1, T_2, \ldots \) of the process \( H \). But in the varying size scenario, these random times do contribute to the variance of \( W_0 \).

Fig. 1 follows \( w \), computed at line 2.(b) of Algorithm 1, over successive coalescence for one hundred replicates of the SIS algorithm. It shows first that the variance of the final SIS weights is quite large, second that replicates that lead to the highest final \( W_0 \) also tend to have high \( w \) throughout the sequence of coalescence events. Moreover, Fig. 1 shows that, as the number of coalescence events undergone increases, the range of values of \( w \) increases exponentially. This exponential increase is evidence that importance sampling becomes inefficient in spaces of high dimension, and that each random jump times \( T_1, T_2, \ldots \) has a multiplicative contribution to the overall range of \( W_0 \).

The final approximation of the likelihood is the average of these one hundred replicates of \( W_0 \), see Eq. (9). The difference between the values of the final SIS weights \( W_0 \) is so huge that most of them are negligible compared to the top ten highest ones and do not contribute to the average in Eq. (9). If we look at Fig. 1, considerations of their trajectories further back in time show that these weights do not rise back much. Hence, we can reduce the computing effort and increase the accuracy of the method by ignoring the histories which give low values of \( w \) after some times through the sequence of coalescence events and by replacing them by histories with higher values of SIS weight \( w \). The above idea can be seen as a correction of the proposal distribution, and the estimate will remain unbiased if the SIS weights are accordingly corrected. In order to implement it, we rely on the resampling procedure explained in the next Section.

3. Resampling

We derive a resampling procedure from that of Liu (2008 Section 4.1.2) that consists of pruning away the partial histories associated with very small weights and reusing those associated with high weights instead of restarting from scratch.

3.1. Sequential importance sampling with resampling: the algorithm

The sequential importance sampling with resampling (SISR) is as follows. We initiate \( n_{\text{rep}} \) independent runs of Algorithm 1, hence \( n_{\text{rep}} \) draws of \( W_0 \) from the importance distribution. But we stop the repeat-until loop at step 2 of Algorithm 1 before hitting the MRCA. Indeed each run of Algorithm 1 is stopped at checkpoints (stopping times of the Markov process \( H_t \)); at these checkpoints we evaluate the quality of the collection of \( n_{\text{rep}} \) partial histories and, if necessary, we resample these histories. To that aim, we propose a new resampling distribution detailed in Section 3.2.
(a) Sorting after 1 coalescence event. (b) Sorting after 10 coalescence events. (c) Sorting after 30 coalescence events. (d) Sorting after 50 coalescence events. (e) Sorting after 70 coalescence events. (f) Sorting after 98 coalescence events.

Fig. 1. Evolution of the current partial SIS weights as a function of the number of coalescence events undergone by the sample, for 100 histories drawn according to the importance distribution. Ateach step, the current weights are normalized by the sum of the current weights of all the histories and represented on a logarithmic scale. In red: the 10 histories corresponding to the 10 highest partial SIS weights computed after (a) 1, (b) 10, (c) 30, (d) 50, (e) 70 and (f) 98 coalescence events.

The checkpoints at which we test whether to resample correspond to a given number $k$ of events (coalescences and mutations) undergone after the previous checkpoint or to a given number $k$ of coalescence events undergone. Once a checkpoint is reached, we evaluate the quality of the collection of partial histories to assess whether it is necessary to resample. The quality test is based on the Effective Sample Size (ESS) relative to the collection of partial histories, namely

$$\text{ESS} = \frac{\left( \sum_{j=1}^{n_H} w^{(j)} \right)^2}{\sum_{j=1}^{n_H} (w^{(j)})^2}$$

where $w^{(j)}$ is the current value of $w$ of the $j$th partial history at the checkpoint. The largest value of ESS is $n_H$ and occurs when $w^{(1)} = \ldots = w^{(n_H)}$; the ESS decreases when the range of values of $w^{(j)}$ expands. When a single weight, $w^{(1)}$, say, is much larger than the other ones, the ESS is approximately equal to 1 since both numerator and denominator of Eq. (10) are approximately equal to $(w^{(1)})^2$. Actually, the ESS assesses how the random holding times and events of the partial histories between two successive checkpoints contribute to the variance of the estimate in Eq. (9). Thus we resample the collection of partial histories whenever the ESS falls below a threshold value, for instance ESS/10, where ESS is the value of the ESS after the last resampling.

The checkpoints are frequent when $k = 1$, but computing the ESS, as well as resampling the whole collection is time consuming, so that higher values of $k$ might be more pragmatic.

The SISR is presented in Algorithm 2, where $w^{(j)}$, $t^{(j)}$, and $h^{(j)}$, for $j = 1, \ldots, n_H$, are arrays which keep tracks of the values of current weight $w$, time $t$ and state $h$ of each run of Algorithm 1. Moreover, we define $n_{\text{control}}$ which stores the number of undergone events (either the total number of undergone events, or the number of undergone coalescences) in-between checkpoints.

### 3.2 The resampling procedure

Assume that SISR has reached a checkpoint and that the ESS is low enough to resample. At this time, we create a new collection of $n_H$ simulated histories by drawing at random in the previous collection of histories according to a multinomial distribution $M(v, n_H)$ where $v = (v^{(1)}, \ldots, v^{(n_H)})$ is a resampling probability distribution on the collection of $n_H$ partial histories, see Section 3.3 for examples of such distributions.
Resampling is equivalent to applying a second importance sampling algorithm within the SIS. Indeed,

$$\sum_{j=1}^{n_H} w^{(j)} = \sum_{j=1}^{n_H} \frac{w^{(j)}}{v_j}$$

(11)

and the right hand side can be interpreted as the expected value of $w^{(j)}/v_j$ over the distribution of a random $j$ drawn from $v$. Thus, in order not to bias the procedure, the new weight associated to the $j$th history is $w^{(j)}/v_j$ whenever this history appears in the resampled collection of histories. Algorithm 3 summarizes the procedure, with the notations of Algorithm 2.

Algorithm 3 Resampling procedure

1. For $j = 1, \ldots, n_H$,
   a. Draw $r'$ from distribution $v$
   b. Set $h^{(0)} = h^{(r')}$ and $\mu = r^{(r')}$
   c. Set $w^{(0)} = w^{(r')}/v_j$

EndFor

2. Replace the old collection $\{(h^{(0)}_i, r^{(0)}, w^{(0)}_i)\}_{i=1}^{n_H}$ with the new collection $\{(h^{(0)}_i, r^{(0)}, w^{(0)}_i)\}_{i=1}^{n_H}$

3.3. The resampling distribution

Resampling introduces a new possible cause of variance, since we replace the sum of Eq. (11) with a Monte Carlo estimate. The resampling probability distribution $v = (v_1, \ldots, v_{n_H})$ on the collection of histories could be any distribution. To achieve efficiency we should pick a distribution $v$ that reflects the future trend of the partial histories, more precisely that sets relatively high probabilities on the partial histories that will correspond to the highest $w$. If the resampling distribution brings helpful information about histories, then the mean square error (MSE) of Eq. (9) should decrease. On the contrary if the resampling probability distribution brings useless information or no information, then the resampling just adds noise and the MSE increases. For example a uniform $v$ distribution that clearly does not bring any information about the histories introduces only an additional variance in the estimation of the likelihood.

The resampling algorithm could face two difficulties: first it does not always choose the replicates with the highest $W_r$; second, regardless of the optimal replicate, it is not perfectly predicted from an intermediary $w$. However, as explained at the end of Section 2.4, the current value of $w$ helps predicting the contribution of the final value $W_r$ to the average in Eq. (9). One might thus be tempted to resample the $j$th partial history with a probability proportional to $w^{(j)}$, i.e. $v_j \propto w^{(j)}$. But this resampling distribution $v$ might not well choose the optimal $W_r$ because of the spread of the $w^{(j)}$’s. Thus Liu et al. (2001) proposed to rely on $v_j \propto \left[w^{(j)}\right]^{\alpha}$ for some $\alpha \in [0, 1]$, with an arbitrary preference on $\alpha = 1/2$. Another possible predictor of the contribution of the $j$th partial history to the average in Eq. (9) is its current state $h^{(j)}$. Note that the probability of $h^{(j)}$ is the expected value of $W_r/v_j$. Of course the probability of $h^{(j)}$ is intractable, but we might replace it by an easily computed pseudo-likelihood. We propose here to rely on the pairwise composite likelihood $L_2(h^{(j)})$ defined in the Appendix A in the Supplementary Materials since we can compute it very easily. But such pseudo-likelihoods are much more contrasted than the true one and thus should be tempered by some exponent $\beta \ll 1$. Hence we advocate the following resampling distribution

$$v^{(j)} \propto \left(w^{(j)}\right)^{\alpha} \left(L_2(h^{(j)})\right)^{\beta}, \quad \alpha, \beta \in [0, 1], \beta \ll 1.$$  

(12)

The tuning parameters $\alpha$ and $\beta$ are used to balance the effect of the information provided by the SIS weight and by the composite likelihood. Section 4 provides numerical examples showing the efficiency of the above resampling distribution for a large range of values of the tuning parameters $\alpha$ and $\beta$.

4. Improvements on the likelihood estimate: numerical results

4.1. The simulated demographic model

We use the model described in Sections 2.1 and 2.2 to analyze microsatellite markers under a Stepwise Mutational Model (SMM). The set $E$ of allele types corresponding to microsatellite markers is the set $\mathbb{N}$ of non negative integers. When a mutation occurs under a SMM, the size of the allele is either increased by 1 or decreased by 1 with the same probability (Kimura and Ohta, 1978).

As in Leblois et al. (2014) we consider a single isolated population whose size has undergone past changes. We denote $N(t)$ the population size expressed as the number of gene copies, $t$ generations away from the sampling time $t = 0$. The haploid population size at sampling time is $N(0) = N$. Then, going backward in time, the population size changes according to a deterministic function until reaching an ancestral population size $N_{anc}$ at time $t = T$. Then, $N(t) = N_{anc}$ for all $t \geq T$. To illustrate our method we consider an exponentially contracting population size, represented in Fig. 2 in the supplementary material, which can be written

$$N(t) = \begin{cases} N \times \left(N_{anc}/N\right)^{\mu t} & \text{if } t \in [0; T], \\ N_{anc} & \text{if } t \geq T. \end{cases}$$

To ensure identifiability the demographic parameters are scaled as $\theta = 2\mu N, D = D/2N$ and $\theta_{anc} = 2\mu N_{anc},$ where $\mu$ is the mutation rate per locus per generation. The parameter space of the model is thus the set of vectors $\phi = (\theta, D, \theta_{anc})$. Additionally we set $\theta(\tau) = 2\mu N(\tau)$, and $N_{ratio} = \theta/\theta_{anc} = N/N_{anc}$, the latter being useful to characterize the strength of the contraction.

All the data sets simulated in this work where produced with the IBDSim software available at http://www1.montpellier.inra.fr/BCG/IBDSIM/index.html (Leblois et al., 2009). Monomorphic loci were not excluded a priori but were very unlikely given the chosen values of population sizes and mutation rates.

4.2. Reduction of the MSE between the true value of the likelihood and its estimate

Numerical results of Section 4 are presented on data sets simulated under the demographic scenario where $\phi = (\theta, D, \theta_{anc}) = (0.4, 0.25, 400)$ which models a recent and strong past contraction in population size. This scenario is the most representative of our conclusions although we studied more moderate and/or older contractions. To assess the efficiency of estimates of the likelihood at a given point of the parameter space, we compared these estimates with a reference value $L$ that is a sharp approximation of the true likelihood. We have computed the reference value $L$ by estimating the likelihood with SIS with a huge collection of $n_H = 20,000,000$ independent histories.

To evaluate the variability of the likelihood estimate returned by a given algorithm, with given values of its tuning parameters, we first plot the empirical distribution of 100 runs of the same algorithm with a boxplot (Figs. 2–3). We also measured the mean square error of 100 estimates $L$ around the reference value $L$, and averaged this MSE over 100 simulated data sets to remove the dependency on the simulated data set (Fig. 4). Note that this last method computes an empirical mean square error that approximates the MSE of the likelihood estimate.
To compare the efficiency of both SIS and SISR algorithms with the same computational effort of simulating a collection of $M = 2000$ histories, we considered the ratio of the empirical mean square error of SISR over that of SIS. The MSE ratio brings out the reduction of the mean square error in likelihood estimate due to the resampling technique.

For Figs. 2–3 discussed below, we note that all the SISR calibrations lead to a smaller variance in likelihood estimation than SIS and that the average is much closer to the reference value. It means that resampling improves the accuracy of the likelihood estimate and that even if the resampling procedure is not much calibrated, the likelihood inference is already better than without resampling. Our method is thus not too sensitive to the tuning of the resampling parameters $\alpha$, $\beta$. This is an important feature of our proposal: it will ease its use and facilitate its dissemination.

Type and frequency of checkpoints

As said in Section 3.1, the checkpoints might be defined either in term of a fixed number of events (coalescences and mutations) undergone on each partial history or in term of a fixed number of coalescences events. Note that the total number of gene copies in the current state $h$ is determined by the size of the observed data $B_0$ and the number of coalescence events undergone from time $t = 0$ to state $h$. Hence the second definition of a checkpoint proposes to compare partial histories that lead to current states $h$ with the same number of lineages. Thus Liu (2008) advocated the second type of checkpoints. Fig. 2 shows that resampling among histories with the same number of lineages (that is by the number of coalescence events) is slightly more efficient than the alternative condition. Indeed, when relying on the alternative checkpoints, the departure of the likelihood estimate from the reference value is sometimes much larger (see the number of points outside the whiskers in Fig. 2). This large departure can be particularly misleading when the true likelihood surface is almost flat. In this situation, it can bring out a false pick in the estimated likelihood surface which shift the position of the maximum likelihood estimate (see the end of Section 5.4). In general, for a given number of events, histories with fewer lineages (more coalescences) often correspond to a current weight which is lower than the others but which might have a high final weight.

In Fig. 3(a) and (c), we propose to resample every $k = 8, 6, 2$ or 1 coalescence events, which corresponds respectively to resampling 12, 16, 49 or 98 times during Algorithm 3 since the simulated data set is composed of 100 genes. We observe that the variation in likelihood estimation decreases with the frequency of the checkpoints, indicating that we should propose a resampling step as often as possible. Figs. 4(a) and (b) leads to the same conclusion regarding the best calibration of $k$: MSE ratio comparing SISR with $k = 1$ to the SIS estimate (represented by a *) are often the lowest MSE ratios when compared to other MSE ratios with the same tuning parameters $\alpha$ and $\beta$. Both Figs. 4(a) and (b) also indicate that the MSE ratio always decreases with the frequency of checkpoints when the resampling distribution depends on the pairwise composite likelihood ($\beta = 0.01$).

Calibration of the powers $\alpha$ and $\beta$ of the resampling distribution

The efficiency of the resampling distribution of Eq. (12) greatly differs between different values of the resampling parameters $\alpha$ and $\beta$. For this reason, they must be adequately chosen to balance the effect of the information provided by the SIS weights and by the composite likelihood. Note that, when relying on SISR, inaccurate likelihood estimates can be obtained only with very poor choices of $\alpha$ and $\beta$ but here we present only the choices of $\alpha$ and $\beta$ that provide an improvement of the likelihood estimation. First of all, we notice that the MSE Ratio is substantially below one for all sixty $(\alpha, \beta)$ couples considered (see Figs. 4(a) and (b)) meaning again that resampling improves the inference of the likelihood in terms of MSE, even when the resampling parameters are not much calibrated. Then we observe that when the resampling distribution only depends on the SIS weights ($\beta = 0$), the MSE Ratio globally decreases when $\alpha$ increases (see Fig. 4(b)). When resampling is performed after each coalescence, as recommended above, we find that any $\alpha$ between 0.5 and 1 is a reasonably good choice (Fig. 4(c)). This is further supported by Fig. 3(b) which represents the boxplots obtained with $k = 1, \alpha = 0.4, 0.5, 0.6, 0.7$ and 1 for three different values $\beta = 0, 0.001$ and 0.01. Indeed the variation in likelihood estimates corresponding to $\alpha = 0.4$ is larger than for other values of $\alpha$ whereas the variation in likelihood estimates corresponding to $\alpha = 0.7$ is slightly reduced compared to other $\alpha$ values. In the following we therefore set $\alpha = 0.7$, which is higher than the arbitrary calibration $\alpha = 0.5$ proposed by Liu et al. (2001). Likewise with $k = 1$, Fig. 4(c) represents the MSE ratios obtained with $\beta = 0$ and 0.01, and different values of $\alpha$. First we note that the ten values of MSE ratios are between 0.05 and 0.20, which means that the resampling reduces the MSE by a factor 5 to 20. Second the horizontal dotted blue line indicates that the lowest MSE ratio obtained with $\beta = 0$ is higher than the MSE ratio obtained with five out of seven values of $\alpha$ when $\beta = 0.01$, that is when using the composite likelihood. Moreover, this choice of $\beta$ is supported by Fig. 3(d) which sets $\alpha = 0.7$ and shows that the span of the distribution of the likelihood estimate is smaller when $\beta = 0.01$ (and $k = 1$) than with other values of $\beta$.

In conclusion we can choose $k = 1, \alpha = 0.7$ and $\beta = 0.01$ although other choices of tuning parameters of the SISR also lead to an improvement of the likelihood estimation. The MSE ratio for the likelihood estimates using SIS vs. SISR is lower than 60% for most choices of the resampling parameters, and below 10% when the procedure proposes to resample after every coalescence event and when using the PCL, which allows a strong improvement of the likelihood estimation in a parameter point.
Fig. 3. Boxplots of 100 estimates of the likelihood in a given parameter point with different inference algorithms. The red box corresponds to SIS estimates. The red horizontal line represents the reference value and the cross on each box represents the mean of the 100 estimates of this box. Each block of three to five boxes with similar colors (pink, orange, green, blue and mauve) corresponds to SISR with a fixed value: (a) of $\alpha$ for four different values of $k$ when $\beta = 0.01$ is fixed for all the SISR estimations, (b) of $\beta$ for five different values of $\alpha$ when $k = 1$ is fixed for all the SISR estimations, (c) of $\beta$ for four different values of $k$ when $\alpha = 0.7$ is fixed for all the SISR estimations, and (d) of $k$ for three different values of $\beta$ when $\alpha = 0.7$ is fixed for all the SISR estimations. See main text for details about $k$, $\alpha$ and $\beta$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Improvements in the likelihood based inference of demographic parameters

The final aim of inference from a dataset is not to compute the likelihood at a given point in parameter space. Rather, it is to provide an estimate of the parameter $\phi = (\theta, D, \theta_{anc})$ and confidence intervals (CI) around each coordinate of $\phi$, i.e., marginal CIs. Thus, we will quantify the impact of our procedures on the performance of such inferences.

5.1. Inference method

In the numerical results below, we conduct a maximum likelihood analysis of the data. An estimate of the parameter $\phi$ is given by the maximum likelihood estimate (MLE) $\hat{\phi}(x)$ of a multilocus data set $x$. The multilocus likelihood is the product of the likelihoods for each locus $x_\ell, \ell = 1, \ldots, d$:

$$L_d(\phi, x) = \prod_{\ell=1}^{d} L(\phi, x_{\ell}),$$

where $d$ is the number of loci in the sample, and each term of the product can be estimated with a SIS or a SISR algorithm. The biological assumption that permits the above writing is that the loci are distant enough in the genome to have independent past histories.

Then we derived marginal CIs on each coordinate, from likelihood-ratio test based on the profile likelihood (see Davison, 2003, for details). To obtain a numerical value of the marginal CIs, we rely on asymptotic theory which states that under $H_0$ profile log likelihood-ratio is approximately $\chi^2$-distributed (e.g., Severini,
2000), here with a degree of freedom equal to 1 when the size of the data set is large, i.e., when there is enough information on the parameter in the data.

5.2. Inference algorithm and its evaluation

As in Rouset and Leblois (2007, 2012) and Leblois et al. (2014) we can conduct the inference process as follows. We first define a set of parameter points through a stratified random sample within a range of the parameter space provided by the user. Then, at each parameter point, the multilocus likelihood is the product of the likelihoods for each locus, which are estimated through the SIS or SISR algorithm. The likelihoods inferred at the different parameter points through a stratified random sample within a range of the parameters provide the user. Then, at each parameter point, the multilocus likelihood is the product of the likelihoods for each locus, which are estimated through the SIS or SISR algorithm. The likelihoods inferred at the different parameter points are then smoothed by a Kriging scheme. After a first analysis of the smoothed likelihood surface, the algorithm can be repeated a second time to increase the density of parameter points in the neighborhood of the first MLE. The Kriging step removes part of the estimation error of likelihood in any given parameter point by assuming that the likelihood is a smooth function of the parameter. In this Section, we thus conducted numerical experiments to show that the gain of SISR over SIS in accuracy of likelihood estimates is retained through the Kriging step.

A convenient way to evaluate numerically the whole inference procedure, and in particular the coverage of the marginal CIs, is to check that distribution of p-value of the likelihood-ratio test of $H_0 : \phi_1 = \phi_1^*$ against $H_1 : \phi_1 \neq \phi_1^*$ is uniform on the interval [0; 1] when the data set $x$ is simulated from the model with $\phi_1 = \phi_1^*$. To this end we represent the empirical cumulative distribution function (ECDF) of the p-value on many simulated data sets $x$, which should be closed to the 1:1 diagonal. Deviation from a uniform distribution can occur: either because the likelihood is poorly estimated or because the exact profile log likelihood ratios do not follow the asymptotic $\chi^2$ distribution. We perform a Kolmogorov–Smirnov test to assess the uniform distribution of the sample of p-values.

We have also computed other measures of the performance of the inference method of $\phi_1$ on many simulated data sets $x_i$ from $\phi_1 = \phi_1^*$, namely

- the mean relative bias of the MLE, computed as
  $\frac{(\text{observed bias on } \phi_1^*)}{\phi_1^*} = \frac{1}{\phi_1^*} \left( \text{mean}_i (\hat{\phi}_1(x_i) - \phi_1^*) \right).

- and the relative root mean square error (relative RMSE) of the MLE, computed as
  $\sqrt{\text{MSE on } \phi_1^*} = \frac{1}{\phi_1^*} \sqrt{\text{mean}_i \left( \hat{\phi}_1(x_i) - \phi_1^* \right)^2}$.

5.3. Numerical experiment cases and previous results

We performed numerical experiments on four different demographic scenarios, all modeled according to the exponential contraction of Section 4.1, which are as follows.

(i) $\phi = (\theta, D, \theta_{anc}) = (0.4, 1.25, 40)$, which is the baseline scenario of Leblois et al. (2014): the population size has undergone a contraction of strength $N_{ratio} = 0.01$ and $D = T/2N = 1.25$, where $N$ is the size of the population at time $t = 0$.

(ii) $\phi = (\theta, D, \theta_{anc}) = (0.4, 1.25, 400)$ which differs from the baseline scenario in the strength $N_{ratio} = 0.001$ of the contraction,

(iii) $\phi = (\theta, D, \theta_{anc}) = (0.4, 0.25, 40) \quad \text{which differs from the baseline scenario in the speed of the contraction,}$

(iv) $\phi = (\theta, D, \theta_{anc}) = (0.4, 0.25, 400)$ which represents the strongest and recentest contraction of the four scenarios. Numerical results on the estimation of the likelihood in this last demographic scenario have already been presented in Section 4 above.

The baseline scenario is a case where the inference procedure performs well when the likelihood of each locus at each point of the parameter space is estimated with the sequential importance sampling on $n_H = 2000$ histories, but $n_H = 100$ gives satisfactory
results. A more careful look at the results shows that the profile likelihoods exhibit clear peaks around the MLE for all parameters. This is a first evidence that the data contain information regarding all parameters. And, indeed, the ECDF of the p-values are almost aligned on the 1:1 diagonal (Fig. 5).

5.4. Numerical results

To analyze the performances of the inference procedure we have simulated 500 data sets for each scenario, composed of 10 independent loci and 100 genes per locus.
Table 1

Accuracy of the MLE with SIS and SISR algorithms.

<table>
<thead>
<tr>
<th>$\theta$, $\theta_{anc}$</th>
<th>(0.4, 1.25, 40)</th>
<th>(0.4, 1.25, 400)</th>
<th>(0.4, 0.25, 40)</th>
<th>(0.4, 0.25, 400)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm with</td>
<td>SIS</td>
<td>SISR</td>
<td>SIS</td>
<td>SISR</td>
</tr>
<tr>
<td>$n_{H}$ = 100</td>
<td>$\theta$</td>
<td>0.17</td>
<td>0.12</td>
<td>0.56</td>
</tr>
<tr>
<td>$\theta_{anc}$</td>
<td>0.063</td>
<td>0.051</td>
<td>-0.02</td>
<td>-0.017</td>
</tr>
<tr>
<td>$D$</td>
<td>0.076</td>
<td>0.016</td>
<td>0.048</td>
<td>-0.042</td>
</tr>
<tr>
<td>rel. RMSE</td>
<td>$\theta$</td>
<td>0.47</td>
<td>0.46</td>
<td>0.71</td>
</tr>
<tr>
<td>$\theta_{anc}$</td>
<td>0.28</td>
<td>0.28</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>$D$</td>
<td>0.53</td>
<td>0.48</td>
<td>0.37</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Gain in accuracy of the MLE

The numerical results regarding the relative bias and relative RMSE are given in Table 1. We find that the resampling technique allows overall to reduce the relative RMSE and the relative bias, mostly for the parameter $\theta$.

(i) For the baseline scenario ($\theta = 0.4$, $D = 1.25$, $\theta_{anc} = 40$), the resampling procedure allows to reduce the relative bias on $\theta$ by 30%, the relative bias on $D$ by 20% and the relative bias on $\theta_{anc}$ by 80% and also the relative RMSE on $\theta_{anc}$ by 10%, other values being similar.

(ii) In the situation ($\theta = 0.4$, $D = 1.25$, $\theta_{anc} = 400$) of a stronger contraction but not too recent, the resampling procedure allows to reduce the relative bias on $\theta$ by 40%, and also the relative RMSE on $\theta$ by 35%, other values being similar.

(iii) Concerning the scenario ($\theta = 0.4$, $D = 0.25$, $\theta_{anc} = 40$) of a more recent contraction than the baseline situation but with the same strength, the resampling procedure allows to reduce the relative bias on $\theta$ by 79% and the relative RMSE on $\theta$ by 10%. However, the relative bias and relative RMSE on $D$ increase by a factor of 2.5 and 1.8 respectively.

(iv) We observe the same trend with the situation ($\theta = 0.4$, $D = 0.25$, $\theta_{anc} = 400$) of a more recent and stronger contraction than the baseline situation. Indeed, when using resampling, the relative bias and relative RMSE on $\theta$ decrease by a factor 4.5 and 3 respectively while the relative bias and relative RMSE on the parameter $D$ increase by a factor 2.5 approximately.

Reducing the size of the collection of histories thanks to the resampling

We compare here the inference method relying on estimates of the likelihood either based on SIS or on SISR with a smaller collection of histories per locus and per point of the parameter space.

We first compare both methods on the baseline situation ($\theta = 0.4$, $D = 1.25$, $\theta_{anc} = 40$) of a weakly and not too recent contraction. In this case, the SIS algorithm performs well due to the large amount of information in the genetic data (Leblois et al., 2014).

We find out that SISR with 50 ancestral histories produces comparable results to SIS with 100 ancestral histories. With the same number of ancestral histories, the relative bias and relative RMSE are lower with SISR than with SIS as explained above, and the ECDF of the $p$-values are closer to the diagonal as shown in Fig. 5. We conclude that resampling improves the parameter estimation in a situation where the SIS performs well, dividing by 2 the required number of ancestral histories.

We then compare both methods on a more difficult situation ($\theta = 0.4$, $D = 1.25$, $\theta_{anc} = 400$) of a stronger contraction. In this situation, the SIS procedure performs less well but we obtain satisfying results with $n_{H} = 2000$ sampled histories (Leblois et al., 2014), which leads to reasonable computation time. Indeed, for a single data set with one hundred gene copies and ten loci, analyses are carried out in 8 h in C++ process time on average. Here we decrease the number of ancestral histories in both SIS and SISR, in order to show that the SISR performs well when the SIS does not.

For both procedures we explored $n_{H} = 100, 200$ and 400 ancestral histories per parameter point. With 400 ancestral histories, the analyses are carried out in 1.7 h in C++ process time on average, for a single data set with one hundred gene copies and ten loci.

Fig. 6 shows that with the same number of explored ancestral histories, we obtain lower relative bias and relative RMSE and also better ECDF, which means better coverage properties of the CIs. By comparing Fig. 6(c) with 6(b), we also find that the SISR procedure provides comparable results (relative bias, relative RMSE and ECDF) to the SIS procedure with half of the number of explored ancestral histories, as in the baseline scenario. The comparison of Fig. 6(e) with Fig. 6(d) shows the same results.

Improvements in more difficult situations

We compare both SIS and SISR procedure on two more difficult situations ($\theta = 0.4$, $D = 0.25$, $\theta_{anc} = 40$) and ($\theta = 0.4$, $D = 0.25$, $\theta_{anc} = 400$) of very recent contraction, very strong for the second case. Thus in these two cases, the proposal distribution is inefficient. In the extreme case of ($\theta = 0.4$, $D = 0.25$, $\theta_{anc} = 400$), the SIS algorithm does not provide satisfactory CI coverage properties, even with 200,000 sampled histories.

We first consider the situation ($\theta = 0.4$, $D = 0.25$, $\theta_{anc} = 40$) in which the magnitude of the contraction is the same as the baseline case but here it occurred much more recently. In this situation, the SIS procedure performs less well and we do not obtain satisfactory results with 2000 sampled histories (Leblois et al., 2014). With the same number of explored ancestral histories, applying the resampling technique, we obtain lower relative bias and relative RMSE for $\theta$ but higher for $D$. We also obtain less good CI coverage properties for $\theta$ and $D$ (see Fig. 7). We conclude that the method detects the contraction, estimates the date and the ancestral size of the population but does not find enough information in the data about the current size of the population. Indeed, the likelihood surface is quite flat.

We then consider the extreme situation with a very recent and stronger past contraction ($\theta = 0.4$, $D = 0.25$, $\theta_{anc} = 400$). Increasing greatly the number of ancestral histories sampled per parameter point up to 200,000 and consequently the computation times by a factor 100, decreases relative bias and relative RMSE on $\theta$ but does not provide satisfactory CI coverage (Leblois et al., 2014). The resampling technique allows us to decrease relative bias and relative RMSE on $\theta$ for a fixed number of ancestral histories sampled and also to provide better CI coverage (see Fig. 8). Comparing Figs. 8(c) and 8(d), we obtain approximately the same relative bias and relative RMSE on $\theta$ and close CI coverage performance with SIS with 100 times less histories sampled than with SIS. In this scenario as in the previous one there is not enough information in the data and hence the likelihood surface is flat. We also face an issue due to the strength of the
contraction and the inefficiency of the proposal distribution in this situation, in addition to the previous difficulty. The resampling technique allows a global gain, since it corrects the inefficiency of the proposal distribution in disequilibrium situations. However, in this situation as in the previous one, the resampling technique does not bring an improvement to the lack of information in the data about the parameter \( \theta \) when the contraction is too recent.

**About flat likelihood surfaces**

The inference method of Section 5.1 suffers from two major defects when the likelihood surface is flat. First any error regarding the likelihood estimate at some point of the parameter space can lead to artificial local maxima or minima in the smoothed surface, depending on how the Kriging method performs. Second even if we recover the true flat likelihood surface, the distribution of the
likelihood-ratio statistic is not correctly approximated by the $\chi^2$ distribution because of the lack of information on the parameter in the data. Both defects can be seen in our numerical studies. First the median of the SIS likelihood estimates is much lower than the reference value in Figs. 2 to 3(d). It indicates that the SIS likelihood estimates are very often below the true value of the likelihood which can lead to artificial local minimum in the smoothed likelihood surface obtained with the Kriging algorithm. When comparing Fig. 9(a) and (b), the SIS likelihood estimates introduce a local minimum at $\theta = 2N\mu \approx 0.001$ while the
(a) $SIS_n = 2000$. (b) $SIS_n = 20,000$. (c) $SIS_n = 200,000$. (d) $SISR_n = 2000$. Fig. 8. ECDF of $p$-values of likelihood ratio tests for the scenario $\theta = 0.4$, $D = 0.25$ and $\theta_{\text{anc}} = 400$. (a) SIS with $n_H = 2000$ sampled histories (b) SIS with $n_H = 20,000$ sampled histories (c) SIS with $n_H = 200,000$ sampled histories (d) SISR with $n_H = 2000$ sampled histories, on 200 simulated data sets. See Fig. 5 for details.

more reliable SISR likelihood estimates manage to recover the flat likelihood surface. Likewise on $D = T/2N$, the SIS likelihood estimates introduce a local minimum around $D \approx 0.4$, see Fig. 5(c), while the SISR likelihood estimates recover the flat likelihood surface, see Fig. 5(d). The second defect can be seen on the $p$-value's ECDF in Fig. 8. Even with the SISR likelihood estimates which manage to recover the flat likelihood surface, the Figure exhibits a departure from the uniform distribution, meaning that the $\chi^2$ approximation is not accurate. Flat likelihood surface as in Fig. 9 and very large (even if untrustworthy) CIs, for instance the 95% CI with SISR likelihood estimates is approximately $[0.00016; 2.5]$ on $\theta = 2N\mu$ and $[0.14; 0.72]$ on $D = T/2N$, should act as warnings that the data do not carry much information about the parameters of interest.
6. Cynopterus sphinx data set

We applied the inference method of Section 5.1 on the fruit bat *Cynopterus sphinx* data set presented in Storz and Beaumont (2002), consisting of allelic frequencies computed from a sample of 246 individuals, hence 492 genes, genotyped at 8 microsatellite loci. Using a coalescent-based Monte Carlo Markov chain (MCMC) algorithm, Storz and Beaumont (2002) found a strong evidence for a pronounced population contraction with (i) a flat posterior distribution for the strength $N_{\text{ratio}}$ of the contraction, characterized by large 50% and 90% highest probability density (HPD) intervals of $[2 \cdot 10^{-5} - 0.05]$ and $[< 10^{-6} - 0.13]$, respectively; and (ii) a more peaked posterior distribution for the scaled time of occurrence, $D = T/2N$, with a mode around 0.3 and 90% HPD interval of $[0.10 - 0.65]$. (Note that the parameter $t_f$ of Storz and Beaumont (2002) is defined as $T/N$ rather than $T/2N$ as in this work.)

Here, we compare their results with those obtained with our methods, with an emphasis on the gain in precision and in computation cost using the SISR algorithm with different resampling probabilities vs. the SIS algorithm. To assess the accuracy of the different algorithms on this real data for which we do not have concrete expectations, we rely on the following steps.

We know from the results presented in the previous sections that both SIS and SISR algorithms may give inaccurate point estimates and CIs for some parameters if the number of simulated histories is too low to infer the likelihood at each parameter point with enough precision. For this reason, we first compared parameter estimates obtained by analyzing the data with the SIS algorithm using $n_H = 1000; 10,000; 100,000$ and $1,000,000$ sampled histories per point. Moreover, to evaluate the IS variance of the log-likelihood estimate by the importance sampling algorithm, we also considered the likelihood RMSE estimate from kriged duplicate points. It is computed from independent pairs of likelihood estimates at the same parameter point for the points retained at the Kriging step.

Our results, presented in Table 2, show a strong decrease in the MLEs of $\theta$ and $N_{\text{ratio}}$, from 1.1 to 0.42 and 0.0034 to 0.00111 respectively, with an associated shift and narrowing of CIs. An opposite and weaker trend of increase is also observed for $D$, from 0.44 to 0.49, and for $\theta_{\text{anc}}$, from 320 to 380, with a slight shift of CIs for both parameters. As expected, increasing the number of simulated histories leads to a decrease in the variance of the likelihood estimate at each parameter point. But more interestingly, it also leads to an important increase of the log-likelihood value at the MLE $\ln(L(\phi))$, from $-664$ to $-607$. This first step implied a very high computational burden but allowed us to see a clear trend towards an improvement of the SIS inference for
A first global conclusion is that all SISR analyses with different resampling strategies yield similar results both in terms of parameter estimates and maximum log-likelihood value over the surface. Moreover, results from SISR analyses differ from those obtained with the SIS algorithm with the same number of sampled histories, but are closer to the best results we get with the SIS algorithm when considering 1,000,000 histories. Our results thus show that SISR analyses on this data set lead to more accurate inferences than SIS with similar computation costs.

According to Table 3, the most significant improvement due to resampling concerns the parameter $\theta$, which is the most poorly estimated by the SIS algorithm with $n_H = 1000$. To a lesser extent, resampling also improves inference of the parameter $D$ compared to the SIS algorithm, but conclusions about the parameter $\theta_{anc}$ are less obvious, even if almost all analyses lead to values closer to the best estimate we have than the one obtained without resampling.

Tackling precisely the effects of the different resampling strategies on this real data set analysis is more difficult as we do not have in hand the true values of the different parameters and also because our results show subtle differences among parameters for different resampling strategies. Thus, we can only globally conclude that resampling, especially when using PCL, always lowers the likelihood estimation variance and increases the maximum likelihood value on this data set compared to the analysis without resampling. It should also be noted that resampling according to a distribution which does not depend on the IS weights but only on the PCL also improves the inference precision compared to SIS in terms of likelihood estimation variance, maximum likelihood value and parameter estimates (results not shown).

Overall, our results on this fruit bat data set qualitatively agree with those of Storz and Beaumont (2002) showing a strong evidence for a past population contraction. However, a comparison between their results and our own inferences shows slight quantitative differences in point estimates and associated CIs. Our analyses globally show greater precision for all parameter estimates with CIs that are much narrower than their credibility intervals.

### Table 2

Evolution of the MLE with SIS when increasing the number of sampled histories per point, on the bat data set.

<table>
<thead>
<tr>
<th>$n_H$</th>
<th>$\hat{\theta}$</th>
<th>$\hat{D}$</th>
<th>$\theta_{anc}$</th>
<th>$N_{anc}$</th>
<th>ln(lik)</th>
<th>lik-RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>1.1 (0.76 – 1.5)</td>
<td>0.44 (0.35 – 0.55)</td>
<td>320 (180 – 560)</td>
<td>0.0034</td>
<td>−663.74</td>
<td>9.2</td>
</tr>
<tr>
<td>10,000</td>
<td>0.80 (0.53 – 1.16)</td>
<td>0.45 (0.37 – 0.55)</td>
<td>340 (205 – 553)</td>
<td>0.0024</td>
<td>−639.08</td>
<td>8.3</td>
</tr>
<tr>
<td>100,000</td>
<td>0.62 (0.38 – 0.94)</td>
<td>0.48 (0.40 – 0.57)</td>
<td>460 (254 – 492)</td>
<td>0.0014</td>
<td>−620.68</td>
<td>5.7</td>
</tr>
<tr>
<td>1,000,000</td>
<td>0.42 (0.24 – 0.68)</td>
<td>0.49 (0.41 – 0.59)</td>
<td>380 (240 – 589)</td>
<td>0.0011</td>
<td>−606.59</td>
<td>4.4</td>
</tr>
</tbody>
</table>

### Table 3

Accuracy of the MLE with SIS and SISR algorithms on the bat data set.

<table>
<thead>
<tr>
<th>$(\alpha, \beta, k)$</th>
<th>$\hat{\theta}$</th>
<th>$\hat{D}$</th>
<th>$\theta_{anc}$</th>
<th>$N_{anc}$</th>
<th>ln(lik)</th>
<th>lik-RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIS (1, 0.0005, 50)</td>
<td>0.20 (0.085 – 0.4)</td>
<td>0.54 (0.46 – 0.62)</td>
<td>490 (270 – 490)</td>
<td>0.00041</td>
<td>−586.81</td>
<td>6.7</td>
</tr>
<tr>
<td>(1, 0.001, 50)</td>
<td>0.21 (0.091 – 0.39)</td>
<td>0.53 (0.45 – 0.63)</td>
<td>390 (250 – 490)</td>
<td>0.00053</td>
<td>−586.78</td>
<td>5.3</td>
</tr>
<tr>
<td>(0.7, 0.0005, 50)</td>
<td>0.19 (0.075 – 0.4)</td>
<td>0.53 (0.45 – 0.63)</td>
<td>410 (230 – 510)</td>
<td>0.00046</td>
<td>−588.25</td>
<td>6.5</td>
</tr>
<tr>
<td>(0.5, 0.0005, 50)</td>
<td>0.21 (0.088 – 0.4)</td>
<td>0.53 (0.44 – 0.61)</td>
<td>400 (240 – NA)</td>
<td>0.00053</td>
<td>−588.43</td>
<td>6.4</td>
</tr>
<tr>
<td>(0.5, 0.001, 50)</td>
<td>0.21 (0.091 – 0.4)</td>
<td>0.54 (0.46 – 0.61)</td>
<td>490 (330 – 490)</td>
<td>0.00042</td>
<td>−588.99</td>
<td>7.4</td>
</tr>
</tbody>
</table>
intervals (see above). This is especially true for the strength \( N_{\text{aio}} \) of the contraction, for which we found good inference precision using our method compared to the flat posterior distributions obtained by Storz and Beaumont (2002). Note that all results from Storz and Beaumont (2002) reported here so far were obtained using a non-hierarchical Bayesian model in which the model parameters cannot differ between loci, similarly to what is done in our method. Considering a hierarchical model with variability among loci, they found slightly different results. They obtained a more peaked posterior distribution for \( N_{\text{aio}} \) and conversely, a flatter posterior distribution for the timing. Compared to our analyses, this latter analysis of Storz and Beaumont (2002) principally suggests a much weaker past contraction (e.g. \( N_{\text{aio}} \) around 0.03 for their hierarchical analysis vs. 0.0005 for our method) that occurred slightly further in the past with our \( D \) being about 0.55 vs. 0.18. Moreover, despite Storz and Beaumont (2002) do not present their results in terms of \( \theta \) and \( \theta_{\text{inc}} \), we found that our scaled estimates of current population size (i.e. \( \theta \)) are congruent with their unscaled estimates of current population sizes and mutation rates. The large difference observed under the hierarchical model in the amplitude of the past contraction inferred by Storz and Beaumont (2002) thus comes mostly from differences in past population size estimates. Our SIS and SISR analyses inferred ancestral population sizes that are hundred times larger than their estimates. This latter result does not seem to be due (1) to a potential bias in our methods as ancestral population size estimates do not decrease with increasing precision (low RMSE) of estimates of likelihood (the opposite may actually occur); nor (2) to a great variability among loci that would have been best modeled by the hierarchical model of Storz and Beaumont (2002) because a leave one out inference procedure with our method did not find any loci strongly differing from the others (results not shown).

7. Limits and perspectives

A first limitation of our method concerns the size of the data sets to be analyzed. Here, we considered a sample size of 100 haploid individuals genotyped at 10 independent loci for all our simulations, and about 250 diploid individuals genotyped at 8 loci for the fruit bat data set. Under similar demographic models of past changes in population size, Leblois et al. (2014) considered up to 500 haploid individuals genotyped at 50 loci. Data sets of such size can be analyzed on a desktop computer within a day or two depending on the number of histories explored. Larger sample sizes are feasible but then the analyses must be parallelized over large number of cores, e.g. 10 or more, on a computer grid for processing a single data set in a single day. Note however that simulation tests considering hundreds of simulated data sets for a single scenario, as done in this paper, needs much more computation time and may thus be quickly intractable for very large data sets.

A second important limitation concerns the type of genetic markers that can be analyzed. All tests presented in this study considered allelic data analyzed under a strict stepwise mutation model (SMM, typical for microsatellite makers) but different mutation models have been implemented and tested in our Migraine software (e.g. the parent independent mutation model, PIM, and the generalized stepwise mutation model, GSM, for allelic data; and the infinite site model, ISM, for DNA sequences). The resampling technique developed here can thus easily be extended to all those models, notably allowing the analysis of short non-recombining DNA sequences in addition to microsatellite data. Interestingly, under the ISM without recombination, the topology of trees that are compatible with the observed data is almost fixed, which strongly reduces the space of histories to explore. Thus, under simple demographic models such as a single population of constant size or with a single continuous past change in population size, and for sufficiently small data sets, dynamic programming can be used to sum over all ancestral histories and thus the exact likelihood can be computed as in Faisal et al. (2015). Under those conditions, the exact likelihood calculation will lead to more accurate (and sometimes faster) estimates than approximate likelihood approaches such as importance sampling algorithms. However, the work of Faisal et al. (2015) also showed that, when increasing the scaled mutation rate and sample size, the number of histories to be considered severely increases and the algorithm rapidly becomes too slow for practical inferences. For example, for \( \theta = 6 \) and 10 sampled haploid individuals sequenced at 10 independent loci, the maximum likelihood estimates of \( \theta \) is calculated in 254 s but the same calculation takes 6967 seconds when considering only 20 sampled haploid individuals, showing a non linear increase with sample size. Moreover, when the mutation model allows for reversible mutations on a finite number of sites or alleles, as for the SMM considered in this study, the number of possible histories explodes to the extent that the exact likelihood computation is no longer possible and only approximate algorithms such as importance sampling allow the inference.

Moreover, and more importantly, our method can only consider non-recombining independent loci. Given the current exploitation of new sequencing techniques that provide longer and longer DNA sequences, analyzing DNA sequences without a model for recombination is an important limit. However, there is no method based on the class of importance sampling algorithms used here that can so far handle realistic recombination models (e.g. multiple recombination points within DNA sequences). Existing IS algorithms dealing with recombination only consider a single recombination point, i.e. recombination between independent pairs of non-recombining loci (e.g. Jenkins and Griffiths, 2011), and considering more recombination points seems computationally infeasible for the moment. Finally, given the burst of single nucleotide polymorphism (SNP) data, it would be interesting to develop IS algorithms specifically adapted to SNPs, but we are not aware of any development, application or test of IS algorithms for SNP data. A third limitation concerns the maximum number of model parameters that our method can deal with, i.e. the level of complexity of the demographic and mutational models used. This limit is mostly due to (1) the way we extrapolate the likelihood surface on the explored parameter space using a kriging technique; and (2) the computation of profile likelihoods used for confidence interval calculations and plots of likelihood profile surfaces. Here, we considered models with three parameters, but more complex models can be analyzed. Simulated and real data sets have been routinely analyzed under the same demographic model as the one used here but with an additional parameter for the GSM mutation model, thus with four parameters (e.g. Leblois et al., 2014; Vignaud et al., 2014a,b; Lalis et al., 2016; Zenboudji et al., 2016). More recently, models considering two past population size changes and a generalized stepwise mutation model, with up to five parameters, have been treated successfully (e.g. Rousset et al., 0000). However, models with more than eight parameters would clearly be difficult to handle, due in particular to limitations in some libraries on which our R code depends, as well as to a general increase in time of all other explorations of the likelihood surface for profile likelihood computations.

8. Conclusions

In this study, we proposed to improve sequential importance sampling algorithms for likelihood inference of demographic parameters by adding a resampling procedure based on Liu et al. (2001). The new resampling probability distribution we considered depends on the SIS weights, as proposed in Liu et al. (2001);
Liu (2008), but also on the pairwise composite likelihood of the sample, providing additional information about the future trend of each sampled history. To evaluate the gain in efficiency due to this new strategy, we focused on varying population size models for which no efficient proposal distribution is available.

We first showed using simulations that resampling allows to reduce the variance and the bias in the likelihood estimate at a parameter point. This step allowed us to show that resampling is more efficient when the checkpoints correspond to a number of coalescent events undergone by each history rather than considering coalescent and mutations events together, and that checkpoints must be frequent, more precisely after each coalescence event. This step also showed that the information provided by the composite likelihood allows a stronger decrease in variance of the likelihood estimates than the information provided only by the SIS weights.

Then we showed that the increased precision in estimation of likelihood also improved the likelihood-based inferences. According to numerical results based on the analysis of simulated data sets under various scenarios of past population size contraction with different strength and timing, we conclude that the resampling procedure helps to correct the inefficiency of the SIS proposal distribution. Under those time-inhomogeneous models, the stronger the population size contraction is, the less efficient the proposal distribution is because it is computed under equilibrium hypothesis. For a similar precision of inference, SISR provides at least a two-fold, and sometimes a much higher gain in computational efficiency for these models. However, when the contraction is very recent, the genetic data does not contain much information about the parameter θ, resulting in flat likelihood surfaces. Consequently, our simulation results show that (i) when the contraction of the population size is not very recent, the resampling procedure divides the computation cost by at least a factor 2; (ii) when the contraction is not very strong but very recent, it is not obvious how the resampling improves the inference because we face an important lack of information for the θ parameter; and finally (iii) in extreme cases of both very recent and very strong contractions, the resampling procedure divides the computation cost by a factor 100, partially correcting the inefficiency of the SIS proposal distribution but not the lack of information in the data, and thus leads to potential bias in θ estimates and associated incorrect CI coverage properties.

Finally, we analyzed a fruit bat microsatellite data set previously analyzed by Storz and Beaumont (2002) using coalescent-based MCMC algorithms and found that the SISR method allows to infer the past demographic history of this population with computation times reduced by at least a factor one hundred compared to SIS. Our results also shows a slightly more pronounced past contraction of population size, with larger ancestral size, than previously found.

Acknowledgments

The first author was financially supported by the Labex Numéris (Solutions Numériques, Matériel et Modélisation pour l’Environnement et le Vivant, ANR-10-LABX-20) and by the Labex CeMEB (Centre Méditerranéen de l’Environnement et de la Biodiversité), Projet Investissement d’Avenir 2011–2019. Part of this work was carried out by using the resources of the INRA MIGALE (http://migale.jouy.inra.fr) and GENOTOUL (Toulouse Midi-Pyrénées) bioinformatics platforms, the computing grids of the CBGP lab and the Montpellier Bioinformatics Biodiversity platform.

This study was supported by the Agence Nationale de la Recherche (project IM-Model CORALFISH 2010-BLAN-1726-01) and by the Institut National de Recherche en Agronomie (Project INRA Starting Group “IGCIPop”).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.tpb.2016.09.002.

References


Faisal, Muhammad, Futschik, Andreas, Vogl, Claus, 2015. Exact likelihood calculation under the infinite sites model. Computation 3 (4), 701–713.


C. Merle et al. / Theoretical Population Biology 114 (2017) 70–87